الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

جامعة وهران للعلوم و التكنولوجيا محمد بوضياف

THÈSE

En vue de l'obtention du Diplôme de Doctorat

Présentée par : ZEBOUDJ Meriem

Soutenue le : 12 Décembre 2022

Intitulé

Conception et développement d'un environnement d'accessibilité des sites Web pour les malvoyants

Faculté	***: Mathématique et Informatique***
Département	***: Informatique***
Domaine	***: Informatique***
Filière	***: Informatique***
Intitulé de la Formation	***: Génie Informatique***

Devant le Jury Composé de :

Membres de Jury	*Grade*	*Qualité*	*Domiciliation*
FIZAZI Hadria	*Pr*	*Président*	*USTO-MB*
BELKADI Khaled	*Pr*	*Encadreur*	*USTO-MB*
BENDELLA Fatima	*Pr*	*Examinateur*	*USTO-MB*
BELALEM Ghalem	*Pr*	*Examinateur*	*U-Oran1*
BOUAMRANE Karim	*Pr*	*Examinateur*	*U-Oran1*

Année Universitaire : 2022-2023

Résumé

La facilité et la rapidité de propager une information sur le web permettent la production d'une grande masse d'informations. En conséquence, les internautes malvoyants (personnes déficientes visuelles) examinent beaucoup de données afin de trier les résultats trouvés, évaluer l'accessibilité du contenu et finalement accéder à l'information recherchée. Pour ces raisons une procédure d'amélioration de l'accessibilité au web, d'une part et l'optimisation de la recherche d'information d'autre part est toujours sollicitée pour ces utilisateurs. L'approche proposée comprend principalement deux étapes: la première consiste à l'optimisation de la recherche d'information par la reformulation de la requête web en se basant sur les métaheuristiques (FireFly, Bat, PSO et Pigeon algorithms), tandis que la deuxième étape consiste à l'adaptation des pages web selon les préférences de ces utilisateurs.

Mot clés : Malvoyants, Personnes Déficientes Visuelles, Recherche d'Information, Reformulation des Requêtes Web, Métaheuristiques, Accessibilité au Web.

Abstract

However, the ease and growth of disseminating information on the Web allow the production of giant information mass. As a result, visually impaired users examine a lot of data to sort the results found, evaluate the content accessibility and finally access the information sought. For these reasons, a procedure for improving web accessibility, on the one hand and information retrieval optimization on the other hand is always requested for these users. The proposed approach consists mainly of two steps: the first is to optimize the information retrieval by query web reformulation based on metaheuristics (FireFly, Bat, PSO, and Pigeon algorithms), while the second step is to adapt the web pages according to the preferences of these users.

Keywords: Visually Impaired, Information Retrieval, Query Web Reformulation, Metaheuristics, Web Accessibility.

ملخص

تسمح سهولة وسرعة نشر المعلومات على الويب بإنتاج كتلة كبيرة من المعلومات. ونتيجة لذلك يفحص مستخدمو الإنترنت الذين يعانون من ضعف البصر الكثير من البيانات من أجل فرز النتائج التي تم العثور عليها وتقييم إمكانية الوصول إلى المحتوى والبلوغ أخيرًا إلى المعلومات المطلوبة. ولهذه الأسباب، يطلب دائما لهؤلاء المستخدمين إجراء تحسين إمكانية الوصول إلى الشبكة، من ناحية، والاستفادة المثلى من البحث عن المعلومات من ناحية أخرى. يتكون النهج المقترح من خطوتين أساسيتين : الأولى تتمثل في تحسين البحث عن المعلومات من خلال إعادة صياغة طلب الويب بناءً على خوارزميات الاستدلال (خوارزميات FireFly و Bat و PSO و Pigeon)، بينما تتمثل الخطوة الثانية في تكييف صفحات الويب وفقًا لتفضيلات هؤلاء المستخدمين..

الكلمات الرئيسية: ضعف البصر، البحث عن المعلومات، إعادة صياغة طلبات الويب، إمكانية الوصول إلى الويب.

DÉDICACE

À ma famille.

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui ont contribué de manière directe ou indirecte à l'aboutissement de ce travail et précisément Mme DEKHICI Latifa et MAKHLOUFI Ouardia.

J'exprime tout particulièrement mes remerciements et respects à mon directeur de thèse Pr. BELKADI Khaled, pour sa disponibilité, son écoute et ses conseils qui m'ont été toujours précieux. Sa confiance, son investissement scientifique et humain ont été essentiels à la réalisation de ce travail.

J'adresse tous mes remerciements au Pr. FIZAZI Hadria, qui m'a fait l'honneur de présider le jury.

J'exprime également ma gratitude aux Pr. BENDELLA Fatima, Pr. BELALEM Ghalem et Pr. BOUAMRANE Karim qui ont accepté d'examiner ce travail et de me faire l'honneur de participer au jury.

J'exprime ma reconnaissance à ma famille pour sa patience infinie et son soutien inestimable.

Enfin, je remercie tous mes amis pour leur soutien et pour leurs encouragements.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

AG	Algorithme Génétique
BA	Bat Algorithm
Bat-QR	Bat Algorithm Query Reformulation
BITV	Germany Barrierefreie Informations Technik Verordnung
CISI	Centre for Inventions and Scientific Information
CNR	Content Navigation Rules
CSS	Cascading Style Sheets
CVD	Color Vision Deficiency
DOM	Document Model Object
FA	FireFly Algorithm
FAE	Functional Accessibility Evaluator
FA-QR	FireFly Algorithm Query Reformulation
FIRE	Forum for Information Retrieval Evaluation
FP_Growth	Frequent Pattern Growth
GA-QR	Genetic Algorithm Query Reformulation
HTML	HyperText Markup Language
IMA	Improved Memetic Algorithm
JAWS	Job Access With Speech
Kwaresmi	Knowledge-based Web Automated Evaluation with REconfigurable Guidelines OptiMIzation
LWGD	Language for Web Guideline Definition
MAP	Mean Average Precision
MAUVE	Multi-guideline Accessibility and Usability Validation Environment
NSGA	Non dominated Sorting Genetic Algorithm
NVDA	Non Visual Desktop Access
OCAWA	Operational Control and Analysis for Web Accessibility
PA-QR	Pigeon Algorithm Query Reformulation
PIO	Pigeon Inspired Optimization
PSO	Particle Swarm Optimization
PSO-QR	Particle Swarm Optimization Query Reformulation
RGAA	Référentiel Général d'Amélioration de l'Accessibilité

RI	Recherche d'Informations
RP	Réinjection de Pertinence
SGSL	Simple Guideline Specification Language
SI	Swarm Intelligent
SIRSD	Semantic Information Retrieval in Sports Domain
SRI	Système de Recherche d'Information
STReSS	Simulation of Tactile Receptors by Skin Strech
TAW	Test Accessibilidad Web
TF_IDF	Term Frequency-Inverse Document Frequency
TIC	Technologies de l'Information et de la Communication
TREC	Text REtrieval Conference
TTS	Text To Speech
UAAG	User Agent Accessibility GuideLines
W3C	World Wide Web Consortium
WaaT	Web Analytics Automation Testing
WAI	Web Accessibility Initiative
WAI-ARIA	WAI-Accessible Rich Internet Applications
WAVE	Web Accessibility Evaluation Tool
WCAG	Web Content Accessibility Guidelines
WebbIE	Web Browser Internet Explorer

Introduction Générale

Contexte

Au cours des dernières décennies, la recherche d'informations apparait comme une branche de recherche dominante ainsi que l'exploration du web en raison de la croissance et l'évolution des données dans l'Internet. La généralisation d'Internet permet à des personnes déficientes visuelles (Malvoyants) ou aveugles d'atteindre une richesse d'informations. En dépit des progrès significatifs dans les technologies de l'information et de la communication, ces utilisateurs rencontrent toujours des obstacles lors de l'accès au contenu web (Murphy and al., 2008) face à la quantité élevée des données qui peuvent s'afficher sur l'écran et des liens qu'il faut survoler, et aussi pour mettre en ligne leurs informations. De plus, de nombreuses études ont montré que l'accessibilité du web est inférieure à l'optimale (Marincu and McMillin, 2004).

Problématique

La recherche d'informations sur un site web ou d'une page se fait selon une stratégie appelée « essai-erreur ». L'utilisateur malvoyant ouvre une page web pour s'apercevoir que l'information recherchée ne correspond pas à son besoin et/ou ne respecte pas les critères d'accessibilité. Il ferme ainsi la page puis décide d'ouvrir une autre, d'où beaucoup de facteurs influent sur la qualité de ces résultats (Deveaud and Bellot, 2012). En outre, si la requête n'exprime pas ces besoins et le contenu web est inaccessible, ceci implique que, certaines tâches ne sont pas modifiées.

Contributions

Pour tenter de répondre à problématiques exposées, nous nous intéressons dans cette thèse à la reformulation des requêtes dans le contexte web et l'accessibilité de ce dernier pour les utilisateurs ayant une déficience visuelle. Notre objectif est de trouver des moyens informatiques adéquats et faciles pour ces utilisateurs afin de faciliter leurs tâches de navigation tout en les optimisant.

Notre approche de reformulation proposée est basée essentiellement sur les métaheuristiques (FireFly, Bat, PSO et Pigeon algorithms) dans le but de trouver les documents pertinents (pages web) dans un délai raisonnable. La procédure globale de la reformulation consiste, dans un premier temps, à récupérer les pseudo-documents de la requête initiale, à extraire leurs mots-clés pour constituer des solutions potentielles par

FP_Growth et enfin à lancer les métaheuristiques pour déterminer la meilleure requête. Nous évaluons en profondeur notre approche de reformulation proposée sur TREC (Text REtrieval Conference), FIRE (Forum for Information Retrieval Evaluation) et CISI (Centre for Inventions and Scientific Information). Nous menons d'abord une expérience préliminaire pour fixer les paramètres de FireFly, Bat, PSO et Pigeon. Ensuite, nous comparons les résultats aux méthodes de l'état de l'art.

De plus, pour défier les difficultés des pages web récupérées, nous proposons une méthode de personnalisation du contenu des pages web. Cette personnalisation consiste à transformer les éléments d'une page web telles que la couleur du fond, la couleur du texte et la taille de police selon les préférences des utilisateurs ayant une déficience visuelle. L'étape de personnalisation proposée est répartie en trois phases. La phase « création du profil utilisateur » qui permet à l'utilisateur de modifier les paramètres qui vont être appliqués sur toutes les pages web qu'il va consulter. La phase « traitement de données » qui est responsable de l'analyse et de la transformation du contenu des pages web selon les préférences de chaque utilisateur (choisis précédemment). Et la phase « Concrétisation des résultats » qui représente l'interface utilisateur du navigateur et qui lui sert principalement à consulter les pages web, et notamment voire le résultat de la personnalisation (selon leurs préférences).

Les outils utilisés lors de la phase de conception et qui permettent à un utilisateur ayant une déficience visuelle de naviguer facilement sont : la fonction du Zoom (loupe Windows), une loupe et la synthèse vocale SI_VOX.

Organisation de la thèse

La thèse est structurée en six chapitres comme suit :

- Dans le premier chapitre nous déterminons la notion des handicapes visuels et nous présentons les différents outils, techniques et travaux de recherche existant autour de l'accessibilité du web pour aider et faciliter l'accès à l'information à ces personnes;
- Dans le deuxième chapitre nous montrons les diverses problématiques de visualisation des interfaces web pour les déficients visuels ;
- Dans le troisième chapitre nous présentons les concepts, les techniques et les modèles de représentation utilisés dans la recherche d'information ainsi que les diverses méthodes conçues pour la reformulation des requêtes;

- Dans le quatrième nous détaillons notre proposition d'adaptation de métaheuristiques pour la reformulation des requêtes web ;
- Dans le cinquième chapitre nous décrivons les aspects de la conception et de la modélisation de notre proposition ;
- Dans le dernier chapitre nous présentons les résultats de simulation de différentes approches de reformulation des requêtes et leurs comparaisons, aussi les différentes interfaces utilisées pour l'adaptation des pages web selon les préférences des utilisateurs déficients visuels ainsi que les résultats obtenus.

Nous terminons par une conclusion générale et quelques perspectives.

Chapitre I.

Généralités sur l’accessibilité du web

I.1 Introduction

L'une des principales causes de souffrance humaine est la déficience visuelle. Elle représente un grave problème de santé publique. En Algérie, il y a environs 2 millions d'handicapés et ce, selon l'Office National des Statistiques (ONS), mais selon d'autres sources, le véritable chiffre serait proche des 6 millions, dont 32% est le pourcentage des personnes atteintes d'un trouble de la vision. Pour cela, il est essentiel que le web soit accessible afin de fournir l'égalité d'accès à l'information et les fonctionnalités.

Dans ce chapitre nous présentons dans la première section des notions générales des handicapés visuels et nous présentons les différentes techniques pour aider et faciliter l'accès à l'information à ces personnes. L'accessibilité est devenue ces dernières années un enjeu majeur. On met en évidence son importance dans la deuxième section, ainsi que les normes d'accessibilité que doivent vérifier les pages web.

I.2 Notions générales concernant les malvoyants

I.2.1 L'œil humain et son fonctionnement

La vision est l'un de nos cinq sens les plus sollicités, lorsqu'on parle sur cette dernière on réfère toujours à l'œil (ou globe oculaire).

L'œil humain représente l'organe le plus important pour la communication avec l'extérieur, de même, il est complexe dans son fonctionnement (Figure I.1.).

La cornée est transparente et constitue la première lentille de l'objectif de l'œil. Le cristallin est le zoom qui permet d'ajuster la focale à toutes les distances. La pupille est contrôlée par l'iris, un muscle circulaire coloré qui a la faculté de se contracter pour s'adapter aux différentes conditions d'éclairage comme un diaphragme photo. La rétine, enfin, est comme le film photographique sur lequel s'imprègne la lumière, transformant les photons en énergie électrique. Cette énergie est ensuite transmise par le nerf optique au cerveau qui la transforme en images.

La rétine qui reçoit ainsi l'image adaptée par le système cornée-cristallin est composée de plusieurs capteurs. Ces récepteurs ne sont pas tous identiques. Chaque type permettant l'acquisition d'informations lumineuses différentes. Les cônes sont responsables de la vision colorée, tandis que les bâtonnets permettent de capter l'intensité lumineuse. Les cônes qui permettent de percevoir les couleurs se décomposent en trois type de cellules différentes : celles qui vont être sensibles au rouge, celles qui seront sensibles au bleu et celles qui seront

sensibles au vert. La combinaison de ces trois couleurs primaires permet de créer un très grand nombre de couleurs perçues (Jonquet, 2013).

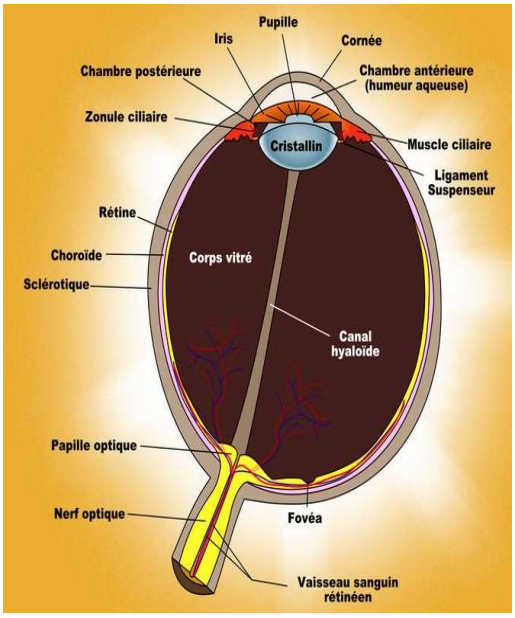


Figure I.1. Coupe transversale de l'œil.

L'œil humain est capable de distinguer des couleurs ayant une longueur d'onde comprise entre 400 et 800 nanomètres (environ), cette plage constitue une infime partie du spectre d'ondes électromagnétiques. En dessous des 400 nanomètres on trouve par exemple les ultraviolets et les rayons X et gamma, au dessus des 800 nanomètres il y a les infrarouges puis les ondes radios (Bonavero, 2015).

Les bâtonnets se situent en périphérie de la rétine et permettent la vision dans des conditions d'éclairage faible (claire ou sombre). Également dans tout mécanisme, chacune de ses fonctions peut être mal effectuée et causer, dans ce cas-ci, une déficience visuelle différente.

I.2.2 La déficience visuelle

L'explication du mot malvoyant est la personne qui ne voit pas bien ou pas du tout (Dictionnaire-Malvoyant, 2016). La déficience visuelle n'a pas de définition unique. Elle se fonde avant tout sur les mesures de la diminution de l'acuité visuelle (une métrique permettant d'évaluer la capacité de l'œil à discriminer deux points) et/ou celle du champ visuel (la capacité à percevoir l'espace visuel lorsque l'œil est immobile). D'après l'OMS (Organisation mondiale de la Santé), on compte 5 catégories de déficiences visuelles (Kammoun, 2013) :

- Catégorie I : La déficience visuelle moyenne (Réalisation des activités presque normalement avec une aide simple : lunettes, loupe). Avec une acuité visuelle corrigée comprise en 3/10 et 1/10 et un champ visuel supérieur à 20° ;
- Catégorie II : La déficience visuelle sévère (Possibilité de s'appuyer sur la vision avec des aides spécifiques mais réalisation des activités plus lente et fatigabilité importante). Avec une acuité visuelle comprise entre 1/20 et 1/10 ;
- Catégorie III : La déficience visuelle profonde (La vision seule empêche une activité classique, même avec des aides techniques. Appui sur l'ouïe et le toucher). Avec une acuité visuelle comprise entre 1/50 et 1/20, ou champ visuel compris entre 5° et 10° ;
- Catégorie IV : La cécité presque totale (Appui sur les autres sens). Avec une acuité visuelle inférieure à 1/50 mais perception de la lumière ;
- Catégorie V : La cécité absolue (Absence totale de perception de lumière. Appui total sur les autres sens).

Les catégories I et II correspondent à ce qu'il est convenu d'appeler la malvoyance. On parle aussi de basse vision ou d'amblyopie ou encore de vision réduite. Diverses affections de l'œil et du système visuel peuvent causer la basse vision. Des malformations ou des anomalies congénitales, des blessures, certaines maladies de l'organisme et le vieillissement peuvent tous conduire à une baisse de l'acuité visuelle.

I.2.3 Différents déficits visuels

L'œil est un organe à la fois complexe et fragile. Comme tout organe, il peut être sujet à des malformations ou des maladies, les raisons sont multiples et les impacts différents. Un certain nombre d'atteintes sont considérées comme bénignes et peuvent être soit corrigées, soit soignées. À l'inverse, il existe des atteintes de la vision dues à des pathologies aux caractéristiques très différentes, qui ne possèdent actuellement aucun traitement. Une partie de ces pathologies sont qualifiées d'évolutives car elles endommagent progressivement le champ visuel. D'autres sont héréditaires, c'est-à-dire qu'elles se transmettent de génération en génération, ou même congénitales (présentes à la naissance mais ne sont pas nécessairement héréditaires).

Cinq grands types d'atteintes de la vision sont répertoriés. Leur développement ainsi que la gêne visuelle qu'ils induisent varient en fonction de la gravité de la maladie (Jonquet, 2013):

- **La vision centrale** : certaines pathologies touchent la vision périphérique. Petit à petit celle-ci réduit et la personne ne voit plus que ce qu'il y a au centre de sa vision. Cette

vision est mesurée par l'acuité visuelle en dixième en lisant des lettres de taille différentes sur un tableau par les ophtalmologistes.

Figure I.2. Exemple de vision centrale.

- **La vision périphérique** : d'autres maladies brouillent la vision centrale pour ne laisser que la vision périphérique. Elle est l'opposée de la catégorie précédente dont laquelle le champ périphérique est atteint. Cette vision est déterminée par la mesure du champ visuel qui peut être réalisée par différentes techniques chez les ophtalmologistes.

Figure I.3. Exemple de vision périphérique.

- **La vision floue** : dans cette catégorie, le cristallin s'épaissit et trouble la vision, mais rend également les patients sensibles à la source lumineuse (personnes photophobes), bien qu'ils aient besoin que l'activité qu'ils réalisent soit éclairée.
- **La vision tachetée** : dans cette famille, les taches peuvent apparaitre à divers niveau.

Figure I.4. Exemple de vision tachetée.

- **La cécité totale** : ces personnes sont complètement aveugles. Elles ne peuvent même pas percevoir la luminosité.

I.2.4 Les pathologies visuelles

La déficience visuelle est un terme vaste qui regroupe nombreux pathologies différentes. Le schéma de la Figure I.5 montre la classification de l'ensemble de ces pathologies (Bonavero, 2015).

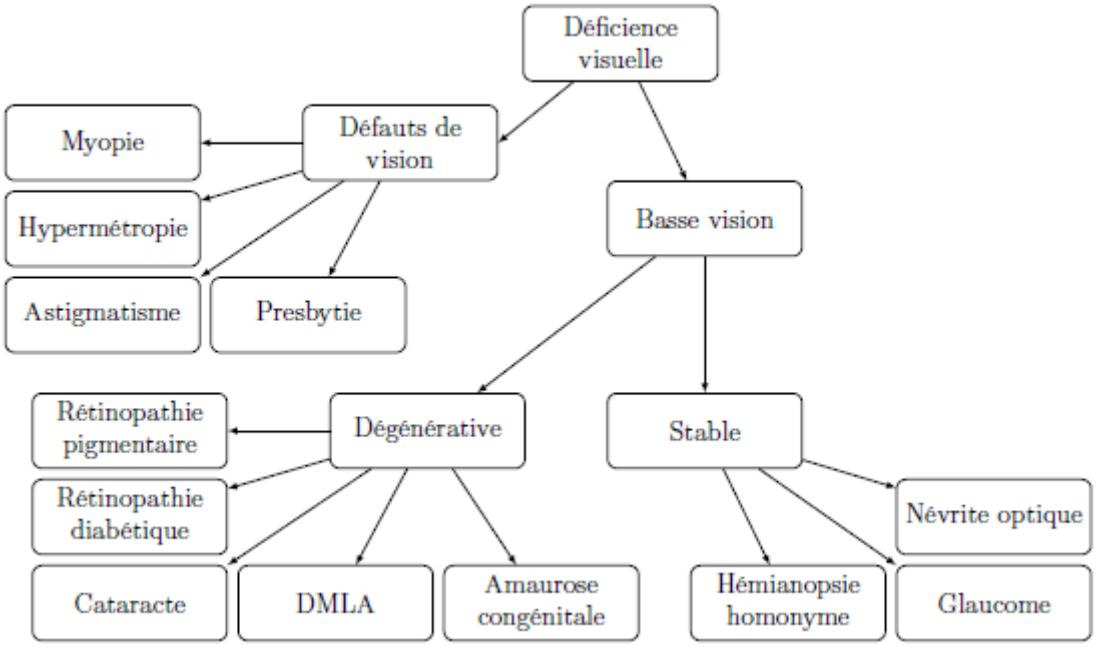


Figure I.5. Classification de l'ensemble des pathologies support (Bonavero, 2015).

I.3 Les handicapés visuels et l'informatique

I.3.1 Les handicapes visuels

L'informatique est l'une des disciplines où les personnes aveugles et/ou malvoyantes peuvent se retrouver le plus égales, à l'aide des outils adaptés. Contrairement à ce qui est souvent prétendu, les malvoyants peuvent accéder à l'informatique, et donc à Internet en utilisant des techniques spécifiques telles que la synthèse vocale ou l'afficheur Braille.

A titre d'exemples, l'informatique permet aux malvoyants de : relire et corriger ses textes sans l'aide d'un voyant ; correspondre avec des personnes qui ne connaissent pas le Braille ; rédiger des courriers officiels et privés, et grâce au scanner, lire soi-même son courrier ; accéder aux informations; faire des achats en ligne ; gérer ses réservations ; etc.

I.3.2 Les aides techniques

Plusieurs outils sont conçus aujourd'hui pour aider et faciliter l'accès à l'information des personnes malvoyantes. Ces outils regroupent des appareils entièrement dédiés à leur

usage et des systèmes constitués de périphériques matériels (hardware) et/ou logiciels (software) destinés à être utilisés sur ou avec un ordinateur standard.

Les principales aides techniques pour les personnes aveugles et /ou malvoyantes sont les suivantes :

- Un logiciel « ***lecteur d'écran*** » permet de restituer un contenu numérique. Il transmet les informations apparaissant à l'écran de l'ordinateur en format compatible avec l'afficheur braille (au toucher) et/ou le synthétiseur vocal (à l'oral). Les logiciels les plus utilisés sont :

- JAWS « Job Access With Speech » (Freedom Scientific-JAWS, 2016) est fabriqué par Freedom Scientific. Un utilisateur aveugle ou malvoyant peut obtenir facilement des données à partir de la page HTML. La raison est que l'information doit être réorganisée avant qu'elle ne soit présentée. JAWS ouvre un observateur virtuel tandis que la page web est affichée à l'écran par le navigateur. Puis il décompose la page entière en blocs, et présente ces blocs à l'utilisateur, en essayant d'établir le bon ordre, qui est généralement de gauche à droite et de haut en bas. Ainsi, l'utilisateur arrive à lire un bloc à la fois. Par exemple, la barre de menu en haut, puis la colonne de gauche où l'on trouve les principaux titres d'article, puis la colonne de droite où nous trouvons le texte de l'article sélectionné, et le pied de page à la fin. Deux modes d'activation sont possibles : GLI « Gestion de Licence par Internet » et DONGLE « activation sur clef USB ».

- NVDA « Non Visual Desktop Access » (PDF association, 2012), est développé par Michael Curran et James Teh. C'est un lecteur d'écran gratuit et open-source du système d'exploitation Microsoft Windows. Contrairement à ses homologues commerciaux, qui doivent être achetés pour être légalement utilisés pour tester les sites web, NVDA ne coûte pas d'argent. Il prend le code HTML de la page et le convertit en un document contenant des informations sémantiques : les liens, les champs de formulaire, des images, etc.

- VoiceOver (Accessibilité – OS X – VoiceOVer – Apple (FR), 2016) est le lecteur d'écran d'Apple du système d'exploitation Mac OS X. Il permet de lire vocalement le texte affiché à l'écran grâce à une synthèse vocale intégrée. Outre ces dernières, VoiceOver prend aussi en charge les afficheurs braille connectés à l'ordinateur (le système intègre des pilotes logiciels pour plus d'une cinquantaine d'afficheurs braille USB et sans fil). Il permet de contrôler un ordinateur même si l'utilisateur ne voie pas l'écran, d'accéder de manière autonome à la gestion des données, d'utiliser un traitement de texte, la messagerie, internet, le multimédia ainsi que de nombreuses applications dans divers domaines d'activité.

- ***La synthèse vocale*** (Morel and Dujour, 2001) est un système qui lit le contenu de l'écran de l'ordinateur de telle sorte qu'il soit émis via un haut-parleur. Elle englobe deux parties : matérielle et logicielle. On trouve dans la première partie la carte son et le haut-parleur, et dans la deuxième les programmes de lecteur d'écran et les voix des différentes langues. Elle est appréciée par certains utilisateurs malvoyants, souvent parce qu'à la lecture d'un document long, les yeux se fatiguent et l'attention baisse, malgré un système d'agrandissement.
- ***L'afficheur braille*** (Berthet, 2013) ou bien plage braille, ou encore barrette braille, est un dispositif électromécanique qui affiche en temps réel des caractères brailles reçues du lecteur d'écran, ces caractères apparaissent sous forme de points saillants sur une surface plane.

Il peut être utilisé simultanément ou en complément avec un système de synthèse vocale. Il existe plusieurs modèles selon le nombre de cellules allant de 20 à 80, ce qui détermine le volume de l'appareil. L'afficheur braille est un périphérique qui remplace le moniteur. Il est souvent placé à l'avant du clavier dactylographique.

- ***Les logiciels d'agrandissements*** (Burger, 2006) permettent de lire plus confortablement l'écran en agrandissant toutes les informations qui apparaissent, mais aussi en modifiant les couleurs, les polices, les contrastes suivant les besoins de la personne malvoyante, ainsi qu'en suivant automatiquement les menus, déplacements de souris, apparition d'objets, etc. Le logiciel **ZoomText** (Squared, 2009) est une solution d'accès à un ordinateur puissant conçu pour les déficients visuels, offrant un accès complet aux applications, documents, e-mail et Internet. Il est composé de deux technologies d'adaptation : grossissement de l'écran et la lecture de l'écran. Il est disponible en deux versions : « Magnifier » - agrandisseur d'écran - et « Magnifier/Reader » - agrandisseur et lecteur d'écran intégré -. **SteeringWheel** (Billah and al., 2018) est une interface de grossissement préservant la localité, exploite un ensemble simple de mouvements de rotation et avec un retour audio-haptique fourni par un cadran physique pour la navigation web pour les personnes à basse vision.

Il existe aussi **des imprimantes braille**, dites embosseuses Braille, qui permettent d'imprimer sur un support papier du texte en braille, elles impriment en relief en perforant la feuille et non pas avec de l'encre. Il s'agit de créer des bosses dans le papier qui représentent les lettres en Braille.

I.4 Les handicapés visuels et l'Internet

Le développement du web et la généralisation d'internet permettent aujourd'hui à des personnes aveugles ou malvoyantes d'atteindre une richesse d'informations et de services qui leurs étaient avant inaccessibles. Contrairement à ce qui est souvent supposé, les handicapés visuels peuvent accéder à l'informatique, et donc à Internet, en utilisant des aides techniques comme les lecteurs d'écran, les logiciels d'agrandissement ou encore les tablettes Braille. Bien que ces aides techniques soient nécessaires et régulièrement utilisés, les malvoyants évoquent souvent leur aspect frustrant. Plusieurs raisons sont citées, comme le manque d'organisation spatiale du contenu lu avec les lecteurs d'écran ou le fait de ne solliciter qu'un seul sens.

Le système d'enseignement à distance développé dans cette étude (Yrtay and al., 2015) a utilisé JAWS comme un logiciel de lecture d'écran permettant d'accéder facilement aux ressources pour les personnes ayant une déficience visuelle où la structure des pages HTML était conçue pour être compatibles avec JAWS. Une autre recherche consiste à adapter pour les malvoyants un logiciel de lecture d'écran multimodal « TactoColor » développant le TactoWeb de Petit (Jonquet, 2013) qui permet une exploration audio-tactile du web. Outre, le lecteur d'écran TactoWeb est une combinaison de deux outils : TactoGraph et MaskGen. Le premier utilise un appareil appelé STReSS (Simulation of Tactile Receptors by Skin Strech) qui est intégré sous forme d'une grosse souris, qui bouge sur une surface de 24×17 cm sur lequel est projetée tactilement la page internet ouverte, et les sensations sont transmises en fonction de ce que la personne malvoyante survole avec la souris. Puis, le MaskGen transforme l'image en informations audio-tactiles accessibles par le Tactograph.

TactoColor a traduit les composantes de la page en ajoutant la notion des couleurs comme les liens en Bleu, le menu en Rose, les images sont représentées par un rectangle orange, pour rendre l'exploration et la saisie de données plus facile.

La synthèse vocale est une technologie utilisée dans de nombreux domaines des TIC. Cette technologie peut apporter une solution à l'activité de lecture des personnes malvoyants en raison de la conversion d'un texte écrit en texte lu. Sur la base de ce problème, dans l'étude de (Yurtay and al., 2011), un système est conçu aux étudiants ayant une déficience visuelle pour avoir droit des bénéfices de la bibliothèque de l'université. A ce stade, le logiciel développé guide l'étudiant vocalement dont la gestion est fournie par les commandes de clavier. Un scanner est utilisé lors de la numérisation des documents et des livres imprimés

dans la bibliothèque, qui sont ensuite convertis au format d'images et le système d'OCR (Optical Character Recognition) traduit en format texte.

D'autres travaux ont été menés pour créer un logiciel de messagerie vocal (Shrivas and al., 2016), utilisant TTS (Text to Speech) comme une assistance vocale adaptée aux besoins des personnes malvoyantes.

I.4.1 Normes et règles d'accessibilité aux pages web

Le web est une ressource de plus en plus importante dans de nombreux aspects de la vie: l'éducation, l'emploi, le commerce, les loisirs …etc., mais l'accès à l'information est le plus important. Pour cela, il est essentiel que ce dernier soit accessible afin de fournir l'égalité d'accès à l'information et les fonctionnalités pour les personnes malvoyantes. Les personnes ayant une déficience visuelle peuvent percevoir, comprendre, naviguer et interagir à travers l'accessibilité du web.

Il est essentiel de rendre le texte, les images et les parties audio-vidéo des sites web accessibles à ces personnes plus précisément les rendre compréhensibles par les outils qu'ils utilisent pour la navigation sur internet.

Certains organismes ont déterminé leurs propres critères d'accessibilité dont le Web Accessibility Initiative (WAI) rattaché au World Wide Web Consortium (W3C) qui est à l'origine de la standardisation des langages HTML et CSS. Le WAI a été créé en avril 1997, il rassemble des gens de l'industrie, des organisations des personnes handicapées, les gouvernements et les laboratoires de recherche du monde entier afin de rendre le web accessible aux personnes handicapées, y compris auditive, cognitive, neurologique, physique, mutité, et la déficience visuelle.

Sa fonctionnalité permet de contrôler le contenu du web (sites et les applications), les outils de création (tels que les systèmes de gestion de contenu - Content Management Systems-), les navigateurs, et des techniques spécifique du W3C, y compris WAI-ARIA (WAI-Accessible Rich Internet Applications). Dans ce sens, le document WCAG (Web Content Accessibility Guidelines) (Caldwell and al, 2016) définit les règles pour l'accessibilité des contenus web aux personnes handicapées. Ces règles peuvent s'appliquer à n'importe quel site ou outil et ne se limitent pas aux administrations. Leur dernière version de Juin 2018 (WCAG 2.1) contient quatre principes avec un total de 12 guidelines. Chaque critère fournit des conseils pour la mise en œuvre des guidelines, un total de 78 guidelines. La norme contient les directives les plus récentes pour améliorer l'accessibilité pour 3 groupes

d'utilisateurs : les personnes malvoyantes, les personnes ayant des troubles cognitifs ou éducatifs, et les utilisateurs handicapés. Les quatre principes sont les suivants :

- **Perceptible** : l'information et les composants de l'interface utilisateur doivent être présentés aux utilisateurs de manière à ce qu'il puisse les percevoir. Elle ne peut être invisible à tous leurs sens ;
- **Utilisable** : la navigation et les composants de l'interface doivent être utilisables. Il ne peut pas exiger une interaction qu'un utilisateur ne peut effectuer ;
- **Compréhensible** : les informations et l'utilisation de l'interface doivent être compréhensibles. Aider l'utilisateur à éviter et corriger les erreurs, rendre le contenu lisible et compréhensible ;
- **Robuste** : le contenu doit être suffisamment robuste pour être interprété de façon fiable par une large variété d'agents utilisateurs.

Chacun de ces points se distribuent en plusieurs critères de succès correspondant à un niveau A, AA ou AAA. Ces critères constituent la base de l'évaluation de l'accessibilité d'un contenu web.

Les critères WCAG 2.1 sont répartis en 3 grandes catégories : développement, éditorial et graphisme. Ils s'articulent autour des 13 thèmes suivants :

- Images ;
- Cadres ;
- Tableaux ;
- Couleurs ;
- Liens ;
- Consultation ;
- Multimédia ;
- Scripts ;
- Éléments obligatoires ;
- Présentation ;
- Navigation ;
- Formulaires ;
- Structuration de l'information.

La nouvelle version WCAG 2.1 s'appuie sur la version précédente (WCAG 2.0). Entre les deux versions, il existe une compatibilité descendante, le contenu qui répond aux exigences de la version 2.1, répond aux exigences de la version 2.0. La version 2.1 (dont laquelle 17 critères sont ajoutés) ne contredit pas et n'annule pas la version 2.0. Il est conseillé de suivre les exigences de la nouvelle version : elles contribuent à une meilleure accessibilité des utilisateurs (Georgieva-Tsaneva and Sabev, 2018).

Pour examiner les outils (applications) de production de code source HTML (éditeur de pages) ou les services de publication en ligne, la recommandation est les « Authoring Tools Accessibility Guidelines -ATAG- ». Les ATAG 2.0 (Authoring Tool Accessibility Guidelines

(ATAG) 2.0, 2021) permettent d'une part à avoir un outil de génération de code, de publication ou de gestion de page accessible à tous, et d'autre part de prévoir l'accessibilité du contenu créé par le biais de ces outils. Les critères suivants font donc partie des normes :

- L'accessibilité de l'interface utilisateur par :
  - La compatibilité avec les différentes technologies d'assistance.
  - La capacité à être perceptible, compréhensible et utilisable.
- La capacité des outils à inciter à la génération de contenus accessibles :
  - Guider les utilisateurs dans l'édition de contenus accessibles par le biais de l'outil.
  - Inciter les utilisateurs à mettre en œuvre l'accessibilité des contenus.

En ce qui concerne les outils de consultation du contenu web, les « User Agent Accessibility GuideLines -UAAG- » permettent de règlementer cela en proposant des directives. Les UAAG 2.0 (User Agent Accessibility Guidelines (UAAG) 2.0, 2021) sont décomposées en 5 principes généraux qui eux-mêmes sont composés de directives.

- Garantir que l'interface utilisateur est perceptible ;
- Garantir que l'interface utilisateur soit utilisable ;
- Garantir que l'interface utilisateur soit compréhensible.
- Respecter les normes et conventions applicables ;
- Favoriser l'accès par les technologies d'aide ;

I.4.2 Les limites des normes

Les normes ont été déterminées selon des études des besoins des personnes en situation de déficience, particulièrement visuelle. Ces recommandations sont réparties en catégories et des niveaux (A, AA ou AAA) ce qui donne une validation des interfaces web plus ou moins rigoureuse, sachant que le niveau intermédiaire (AA) est celui le plus recherché par les producteurs de contenu.

Ces normes permettent de rendre les interfaces web accessibles, soit directement par un affichage règlementé, soit indirectement en adaptant le contenu par les technologies d'assistance. Malgré que ces normes et recommandations contrôlent le contenu de web, les outils de production, de conception et de développement, leurs applications ce n'est pas obligatoire et spécialement par des particuliers. Pour cela, la majorité de ressources web restent encore inaccessibles.

I.5 Le web et l'accessibilité aux personnes malvoyantes

I.5.1 Comment atteindre l'accessibilité du web ?

Afin d'atteindre l'accessibilité des sites web par les personnes handicapées et spécialement ayant une déficience visuelle, deux manières existent :

- La première manière dite active, qui se base sur l'utilisation de normes et de recommandations.
- La deuxième manière dite passive, qui s'appuie sur la transformation (transcodage) a posteriori de contenu web (images, tableaux, … etc.).

I.5.2 Adaptation des pages web

Adapter les interfaces en fonction du contexte, des caractéristiques et des préférences de l'utilisateur est un problème difficile rapporté par plusieurs études. Nombreux travaux de recherche sont focalisés sur le transcodage de pages web pour les personnes ayant une déficience visuelle. Le transcodage pour l'accessibilité du web est une catégorie de technologies permettant de transformer un contenu web inaccessible en contenu accessible. Il a été inventé pour aider les personnes handicapées à accéder à des pages web inaccessibles sans demander aux auteurs de modifier le contenu de leurs pages. Pour ce faire, il convertit le contenu à naviguer dans un serveur intermédiaire entre le serveur web et le navigateur web (Asakawa and al., 2019). La technologie a évolué avec la technologie de navigation vocale à partir de 1992 et a été activement utilisée dans les années 2000. Les applications de transcodeur peuvent fonctionner en tant que proxy, en tant que serveur web ou en tant qu'application cliente. Certains transcodeurs nécessitent des annotations pour effectuer des adaptations tandis que d'autres sont basé sur des systèmes de recherche d'informations.

Des enquêtes récentes (Araújo and al., 2020) (Chen and Wang, 2020) (Oguego and al., 2018) (Quinde and al., 2020), présentent des revues de littérature sur des études qui prennent en compte le contexte, les préférences et les besoins des utilisateurs pour traiter les aspects d'accessibilité et d'accès universel. La majorité de travaux de recherche traitent les caractéristiques de base telles que la taille, la couleur d'un texte et son arrière plan (Tibbitts and al., 2002) (Santucci, 2009) (Benzouak, 2019). D'autres se centralisent sur le rôle des éléments comme les menus, les listes, et les sections ou bien sur des formulaires des structures logiques de plus haut niveau (Lunn and al., 2008) (Lunn and al., 2009) (Islam and al, 2010).

Afin de rendre possible la manipulation d'abstraction de haute niveau, (Macias and al., 2002) et (Lunn and al., 2008) ont proposé leur propre langage ou méta-modèle. Ces approches nécessitent une étape de reconnaissance de ces éléments dans la page, qui est une étape peut être effectuée par reconnaissance de motifs mais elle peut également être effectuée à l'aide d'une analyse du code source de la page (Macias and al., 2002).

Au niveau de l'étape de transcodage de la page, la modification est réalisé soit sur des éléments HTML (Tibbitts and al., 2002) (Lunn and al., 2008), sur des règles CSS (Bighan and Ladner, 2007) (Mirri and al., 2012) ou bien sur quelques conditions des scripts comme JavaScript (Mirri and al., 2012) (Santucci, 2009). Cette modification peut être apportée soit sur la totalité de la page (dites approches globales) (Troiano and al., 2008) (Mereuta and al., 2014), soit sur une partie précise de la page (dites approches locales) (Mirri and al., 2013) (Ferretti and al., 2014a) (Ferretti and al., 2014b) (Ferretti and al., 2017).

L'approche de (Lunn and al., 2009) permet d'ajouter des règles de navigation CNR (Content Navigation Rules) à le contenu des pages HTML en introduisant des attributs ARIA dans les balises des pages. Cet ajout de balises améliore la qualité de la navigation au clavier et consiste une bonne accessibilité des interfaces web par les lecteurs d'écran. Dans d'autres approche, le contenue des pages est enrichi par l'ajout de texte alternatif sur les images (Bighan and al., 2006), qui est effectuée soit par reconnaissance optique des caractères présents sur l'image, soit en décodant l'URL de cette image.

Ce qui concerne l'aspect de conservation du contexte original de la page, les travaux de Bonavero (Bonavero, 2015) (Bonavero and al., 2015) (Bonavero and al., 2016) s'appuient sur les préférences de l'utilisateur en plus sur les préférences du designer (contexte original de la page). Dans cette approche les utilisateurs peuvent donner des préférences à la fois sur des éléments de bas niveau et sur des éléments de plus haut niveau d'abstraction. Ceci est effectué en utilisant un l'algorithme évolutionnaire NSGA (deux de ses versions NSGA-II et NSGA-III) permettant d'obtenir une bonne approximation. L'article (Bonacin and al., 2021) se base sur le Framework FAIBOUD pour l'adaptation automatique des interfaces pour les personnes souffrant d'un déficit en vision des couleurs (Color Vision Deficiency CVD) en tenant compte leurs caractéristiques individuelles, leur contexte et de leurs préférences. FAIBOUD s'appuie sur des ontologies faisant référence à des structures qui expriment formellement la sémantique de la connaissance du domaine. Il explore les faits d'ontologie pour effectuer automatiquement des modifications (par exemple, une recoloration) sur des éléments d'interface tels que des images, du texte et de l'arrière-plan en fonction des algorithmes

modélisés et des paramètres récupérés. Ferati et son équipe (Ferati and al., 2016) ont construit un prototype intitulé middleware qui rend les sites web utilisables en les adaptant en fonction du niveau de déficience visuelle de l'utilisateur. Dans ce travail trois Workshop ont été organisés avec divers participants et huit différentes techniques d'adaptation sont utilisées : transformation de l'image (redimensionnement), transformation du texte, transformation de couleur (entre texte et son arrière plan), filtrage d'images (permettre à l'utilisateur de différencier les images pertinentes et non pertinentes dans le contexte du site web), Filtrage de contenu (permettre à l'utilisateur d'identifier le contenu principal du site web), Changement de contexte (permettre à l'utilisateur de basculer entre les modes d'adaptation en fonction de ses besoins), Lentille amplificatrice et un narrateur vocal.

La table I.1 présente une comparaison des travaux de recherche existants selon les critères des éléments étudiés et transcodés (transformés) dans une page.

Travaux de recherche	Éléments traités	Champ d'application	Reconnaissance automatique	Conservation du contexte original
(Macias and al., 2002)	Méta-modèle BML	Global	x	
(Tibbitts and al., 2002)	Basique, tables	Global/local		Couleur derrière les images
(Bighan and al, 2006)	Basique, image	Local/ ajout de texte alt.		x
(Bighan and al, 2007)	DOM, CSS	Global		
(Lunn and al., 2008)	Annotation Sém., CSS	Global		Sém. des éléments
(Troiano and al., 2008)	Couleur	Global	palette	x
(Lunn and al., 2009)	Annotation Sém., CSS	Global		Sém. des éléments
(Santucci, 2009)	Basique	Global		
(Islam and al., 2010)	Labels et zone saisies	Ajout de labels	x	
(Mirri and al., 2012)	DOM, CSS, Scripts	Global		
(Ferretti and al., 2014a)	Basique	Local		
(Mereuta and al., 2014)	Couleur	Global	x	x
(Bonavero and al., 2015)	DOM, Méta-modèle	Global		x
(Ferati and al., 2016)	HTML, CSS,	Global		

	DOM			
(Benzouak, 2019)	HTML, CSS	Global		x
(Bonacin and al., 2021)	HTML, Images, CSS	Global/local	x	

Tableau I.1. Comparaison des éléments étudies et modifies dans les pages.

I.5.3 Les barres d'outils

Pour faciliter la navigation et/ou modifier l'affichage des pages, des barres d'outils peuvent être associées aux principaux navigateurs. Accessibar (mozdev.org – accessibar : index –, 2016) est une extension de barre d'outils, pour Firefox et Mozilla qui vise à fournir diverses fonctions d'accessibilité pour les utilisateurs permettant de modifier l'affichage comme la couleur du texte et du fond (70 couleurs sont définies sur la session en cours), la taille des caractères, l'interligne, ou encore de désactivation des images. De plus cette barre d'outils intègre un synthétiseur vocal capable de lire à haute voix l'interface utilisateur ainsi que le contenu de la page web, aussi l'option Zoom (pour Firefox nightly builds seulement).

MozBraille ((mozdev.org – mozbraille : index –, 2016) est une autre barre d'outils sous forme d'une extension pour transformer le navigateur Mozilla ou Firefox afin d'obtenir un navigateur internet autonome conçu pour les non-voyants et les malvoyants. De plus avec cette barre l'utilisateur n'a pas besoin d'un troisième outil comme un lecteur d'écran. Elle possède son propre lecteur d'écran et offre à l'utilisateur trois types d'affichage : premièrement une sortie braille sur un terminal braille, deuxièmement une sortie vers un synthétiseur vocal, et troisièmement une vue avec des caractères agrandis. **Select and Speak** (Select and Speak - text to Speech - Chrome Web Store, 2016) est une autre extension du navigateur « Google Chrome ». Elle utilise le TTS iSpeech (dans seize différentes langues) de qualité humaine pour lire tout texte sélectionné dans le navigateur ; Elle offre la possibilité de configurer les options de voix et de vitesse en changeant la configuration dans la page des options.

En plus des différents navigateurs classiques auxquels on peut ajouter des améliorations, il est possible d'utiliser un navigateur spécialisé, permettant une lecture simplifiée des informations disponibles sur la page web, comme Voxiweb et WebbIE.

Voxiweb (VoxiWeb, 2016) existe sous formes d'un site Internet qui se contrôle avec seulement cinq touches du clavier facilement localisables (exemple : les *flèches haut* et *bas* pour défiler entre les choix, la touche *entrée* pour valider, la touche *Échap* pour retourner à la page d'accueil et la touche *Retour arrière* pour revenir en arrière/remonter d'un niveau) dont tout le contenu résultant de différents sites est restitué sur une seule interface. Ce contenu

récupéré en temps réel sur les sites Internet, est nettoyé d'éléments inutiles (menus, publicités, liens…), présenté de manière structurée et claire (pas plus de 5 niveaux), accessible et vocalisée à l'aide d'une synthèse vocal intégrée. Aussi, il existe sous forme d'une application disponible sur smartphone et tablette (iOS et Android). La navigation se fait alors avec des gestuels de la main ou à la voix.

WebbIE (WebbIE- le navigateur web gratuit pour les personnes aveugles ou malvoyantes, 2016) est un navigateur web gratuit basé sur Internet Explorer de Microsoft. Il fonctionne avec le lecteur d'écran utilisé par la personne malvoyante/non voyante, affiche le contenu des sites web en mode texte et donne des informations sur la page consultée, et donne la possibilité de modifier l'affichage des pages (les couleurs, la taille du texte, l'espace entre les lignes, etc.) selon les besoins d'utilisateur.

I.5.4 Les validateurs

Un grand nombre d'outils a été développé pour l'évaluation automatique ou semi-automatique de l'accessibilité des sites web pour les personnes handicapées, mais la plupart d'entre eux n'ont pas été mis à jour au fil du temps pour suivre l'évolution des normes et des guides en matière d'accessibilité.

Une liste complète des outils d'évaluation de l'accessibilité du web disponibles est disponible sur le site web du W3C-Web Accessibility Initiative (WAI). Sa fonctionnalité permet de rechercher des outils d'évaluation selon différents critères : référentiels, langue, type d'outil, technologie, assistance apportée, périmètre et type de licence. Les outils d'évaluation de l'accessibilité peuvent être classés selon divers critères (Abascal and al., 2019):

- Gratuit contre commercial : alors que certains d'entre eux sont des outils commerciaux, la majorité d'entre eux sont disponibles gratuitement.
- Plateforme : selon l'endroit où ils sont exécutés, ils peuvent être classés dans des applications locales ou autonomes, des services en ligne ou des extensions de navigateur.
- Périmètre d'évaluation : distingue l'évaluation de pages individuelles, d'ensembles de pages et de sites web complets.
- Évaluation uniquement par rapport à l'évaluation et à la réparation : alors que la plupart des outils n'offrent que la possibilité d'évaluer, certains outils sont également capables de fournir des conseils sur le processus de réparation.

- Moyens de signaler les problèmes d'accessibilité : les rapports d'évaluation générés peuvent contenir des conseils d'évaluation étape par étape, ou peuvent afficher des informations sur les obstacles dans les pages web, ou même des présentations modifiées de pages web.
- Guidelines utilisées : les normes ou les ensembles de guidelines qu'ils utilisent pour évaluer l'accessibilité.

Un des plus anciens validateurs existants est **AChecker** (Gay and Li, 2010) (AChecker _ IDI Accessibility Checker_, 2016), un projet open source lancé en 2004 et développé par l'Université de Toronto, qui permet la validation des documents HTML à l'égard de cinq différents guides (WCAG 1.0, WCAG 2.0, BITV 1.0 (Germany Barrierefreie Informations technik Verordnung), Section 508 (U.S.) et Stanca Act (Italy)).

Test Accessibilidad Web (TAW) (TAW - Servicios de accesibilidad y movilidad web, 2016), est un validateur de la Fondation Española CTIC30, collaboré avec le Gouvernement de la Principauté des Asturies. Cet outil fournit la validation pour WCAG 2.0 et mobileOK, une norme promue par le W3C pour assurer la convivialité et l'interopérabilité des pages web sur les appareils mobiles. **Accessibility Valet** (Accessibility Valet, 2016) est un autre outil qui permet de vérifier les pages web en analysant le balisage pour la conformité aux directives d'accessibilité web: en particulier le WCAG et la Section 508.Une URL à la fois peut être vérifiée avec cet outil en ligne en mode gratuit ou une utilisation illimitée avec abonnement payant. Toutes les options de rapport HTML affichent le balisage sous une forme normalisée, mettant en surbrillance un marquage valide, obsolète et faux, ainsi que des éléments qui sont mal placés. Tous les avertissements d'accessibilité sont affichés dans un rapport généré.

Functional Accessibility Evaluator (FAE) (Run FAE: Functional accessibility Evaluator 2.0, 2016) évalue un site web ou une page web unique basée sur les exigences du niveau A et AA des lignes directrices sur l'accessibilité du contenu web WCAG 2.0 du W3C. Les résultats de l'évaluation sont décomposés de cinq catégories: Navigation et orientation, Equivalences de texte, Scripting, Styling et standards HTML. Le jugement de la performance globale dans chaque catégorie est un pourcentage, divisé entre Passer, avertir et Échouer. **OCAWA** (Operational Control and Analysis for web Accessibility) (OCAWA Validateur d'accessibilité, 2017) est développé par Urbilog et Orange. Il permet de vérifier l'accessibilité d'une page web ou bien un site complet selon l'ensemble des référentiels de règles RGAA 2.2, WCAG (1.0 et 2.0). Les utilisateurs de ce validateur peuvent soumettre l'URL du site web

ou télécharger un fichier HTML et l'outil affiche un rapport d'audit d'accessibilité avec des liens vers les violations découvertes.

Une autre approche différente **WaaT (Web A**nalytics **A**utomation **t**esting**)** (Fernandes and al., 2014) est fondée sur une base de données des ontologies selon les directives WCAG 2.0 et WAI-ARIA. Un moteur d'inférence de règles est capable d'extraire la définition de la ligne directrice à partir de cette base de données afin d'effectuer la validation de l'accessibilité: la description des règles et les requêtes de base de données sont effectuées à travers deux langages bien connus (SWRL et SPARQL respectivement).

Afin de faciliter le travail des webmasters, Colas (Colas, 2009) a créé des outils (guide, validateur et correcteur) favorisant la démarche vers l'accessibilité des sites web. D'autre part, elle a conçu un système d'aide pour accélérer la vitesse de lecture en braille et celle de la saisie sur un clavier virtuel, l'algorithme des fourmis artificielles est utilisé pour organiser ces touches. Cette organisation permet à des personnes handicapées de réduire leurs mouvements pour saisir un type de texte donné.

Certains types de validateur d'accessibilité sont basés sur le langage XML pour la définition des lignes directrices. **Kwaresmi** (Knowledge-based Web Automated Evaluation with REconfigurable Guidelines OptiMIzation (Beirekdar and al., 2002) est le premier outil qui a utilisé une représentation basée sur ce langage appelée SGSL (Simple Guideline Specification Language). Un autre environnement d'analyse et d'évaluation de sites web fondé sur le langage XML : **MAUVE** (Multi-guideline Accessibility and Usability Validation Environment) (Schiavone and Paterno, 2015), il est composé d'un langage abstrait nommé LWGD (Language for Web Guideline Definition) pour définir les lignes directrices et se caractérise par la possibilité de spécifier et de mettre à jour ces lignes qui doivent être validées sans nécessiter de modifications dans la mise en œuvre de l'outil, permet de vérifier à la fois HTML et CSS pour détecter les problèmes d'accessibilité et capable de valider les sites dynamiques via l'ajout de plugins au niveau des navigateurs. **Total Validator** (TotalValidator, 2016) est un logiciel de validation « 5 en 1 » développé par homonymous British firm, comprenant un validateur HTML, un validateur d'accessibilité selon les normes WCAG (1.0 et 2.0) et US 508, un validateur CSS, un vérificateur orthographique, et un vérificateur de liens brisés. Cet outil est disponible en tant que logiciel de bureau ou en extensions d'un navigateur, il est compatible avec HTML5 et CSS3.

D'autre validateurs utilisent une approche «graphique» pour montrer les résultats de la validation, visualiser la page web en cours d'analyse et localiser les erreurs sur la page de

l'interface utilisateur à l'aide de marqueurs, d'étiquettes ou d'autres éléments graphiques, **WAVE** (WAVE, 2016) est un exemple de ces validateurs.

La table I.2 présente une comparaison entre ces validateurs.

	Méthodes de soumission du site web			Directives d'accessibilité référencées			
Outils	**URL**	**Fichier**	**Coller**	**WCAG1**	**WCAG2**	**Sec 508**	**Autre**
AChecker	√	√	√	√	√	√	B, W, S
TAW	√	×	×	×	√	×	M
Accessibility Valet	√	×	×	√	×	√	×
FAE	√	×	×	×	√	×	×
OCAWA	√	√	×	√	√	×	×
WaaT	√	×	×	×	√	×	×
MAUVE	√	√	√	×	√	×	V, S
Total Validator	√	×	×	√	√	√	×
WAVE	√	×	×	×	√	×	W, S

Tableau I.2. Comparaison entre les validateurs.

Avec: B: BITV, M: Mobile OK, S: Stanca Act, V: Visualy Impaired et W: W3C HTML/CSS Validator.

I.6 Conclusion

Actuellement face aux nouveaux services administratifs qui sont inaccessibles pour les personnes handicapées, ces personnes se trouvent encore en marge de la société. Les sites administratifs non accessibles ne font qu'affermir la fracture sociale entre les personnes valides et les personnes handicapées. Afin d'expliquer la grande quantité de sites web non accessibles, les Webmestres évoquent les raisons suivantes : l'absence de formation en accessibilité, des coûts de développement a priori trop importants, le manque d'intérêt pour ces profils d'usagers, la crainte d'appauvrir le site et le temps de recherche. Or, comme nous l'avons présenté, ces idées sont fausses mais subsistent. Pour que les webmestres révisent leur jugement, il faut mettre les moyens nécessaires à la diffusion d'informations réelles sur l'accessibilité et l'optimisation de la recherche.

Chapitre II.

Problématiques de visualisation des interfaces web

II.1 Introduction

L'accès aux informations présentées sur le web peut poser des grandes difficultés aux personnes déficientes visuelles. En effet, le problème majeur est la liberté offerte aux concepteurs et développeurs de pages web. Dans ce chapitre résume les diverses problématiques de visualisation des interfaces web pour ces personnes.

II.2 Problèmes sur la page d'origine

La majorité de pages web ne respectent pas les normes et les règles d'accessibilités. Cependant même si elles sont conformes à ces normes et règles, ces pages ne peuvent être accessibles pour toute personne. Cela est dû à plusieurs raisons, y compris diverses pathologies.

II.2.1 Éblouissements (brillance importante)

La sensation d'éblouissement apparait lorsque les yeux ne réagissent pas assez vite quand la luminosité augmente soudainement. Une personne reçoit trop de lumière (à partir d'un écran d'ordinateur) et elle ne voie alors qu'une image blanche qui s'estompe progressivement.

Un nombre élevé de pages web sont sur fond (arrière-plan) blanc (ou très clair). En effet, malgré que ces pages peuvent parfois respecter les normes et les règles d'accessibilités mais ceci ne convient pas du tout aux personnes qui sont très sensibles à la lumière et donc facilement éblouies. En effet, lorsque le fond est trop lumineux, le texte s'affiche plus fin et devient entièrement illisible. Sa couleur plan déborde sur la lettre ce qui donne des contours flous.

Figure II.1. Effet d'éblouissement par l'arrière-plan (lettre gauche : pas d'éblouissement, lettre droite : éblouissement).

II.2.2 Contraste faible

Le contraste est la capacité de deux éléments à ressortir l'une par rapport à l'autre. Plus le contraste est élevé et plus la discrimination sera facile. Sur une page web, Un faible

contraste entre le texte et l'arrière-plan peut la rendre illisible et donc peut poser des difficultés de lecture à beaucoup d'utilisateurs sans qu'ils aient nécessairement une déficience visuelle avérée. Pour cela, il est important d'avoir un bon contraste entre Texte et son fond correspondant aux normes afin d'aboutir à une lecture confortable pour ces personnes.

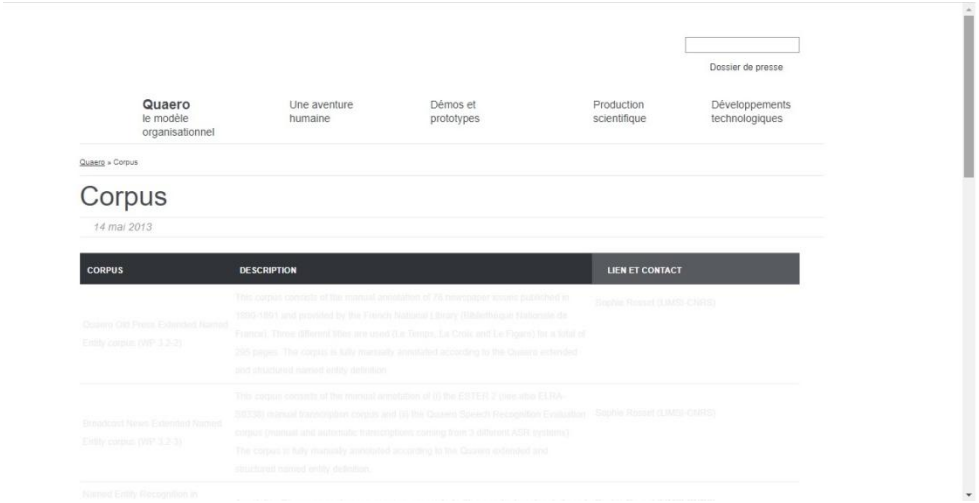


Figure II.2. Exemples de contrastes faibles (Site: http://www.quaero.org/).

II.2.3 Texte au format image

Certains sites web conçus avec la technologie Flash (aussi en HTML) ou des technologies similaires utilisent des images pour dispatcher les informations textuelles. Bien que ce procédé soit strictement déconseillé par les normes et les règles d'accessibilités, mais il reste néanmoins toujours existe.

Plusieurs difficultés et obstacles peuvent survenir, notamment lors de la consultation de ces textes par les utilisateurs malvoyants. Ces derniers, ne peuvent ni sélectionnés, ni copiés, et ni lus que visuellement puisqu'ils sont affichés sur des images. Un lecteur d'écran ne sera donc pas en mesure d'y accéder. De plus, l'agrandissement de ces textes peut lui aussi poser des problèmes. Le fait d'agrandir une image qu'elle que soit le niveau de grossissement, sa qualité sera diminue ainsi que la netteté du texte, et ce qui réduit le confort de lecture.

Figure II.3. Exemples d'un texte au format image (Site : http://techno-flash.com/).

II.2.4 Images d'arrière-plan

L'arrière-plan d'un site web est une image qui occupe tout l'écran lorsque les utilisateurs survolent ce site. Il est donc important de bien choisir cette image parce qu'elle influence à la clarté et la lisibilité de contenu, par exemple si elles sont insérées au-dessous de contenus textuels, et ce qui réduit le confort de lecture.

Figure II.4. Exemples d'une image insérée au-dessous de contenus textuels (Site : https://www.univ-oran2.dz/).

II.2.5 Le daltonisme

Le daltonisme est une anomalie de la vision affectant la perception de certaines couleurs. Sur une page web, cette anomalie peut entraîner des problèmes de lecture à cause de la déformation dans la perception d'une des deux couleurs : soit de l'arrière-plan ou bien soit du texte.

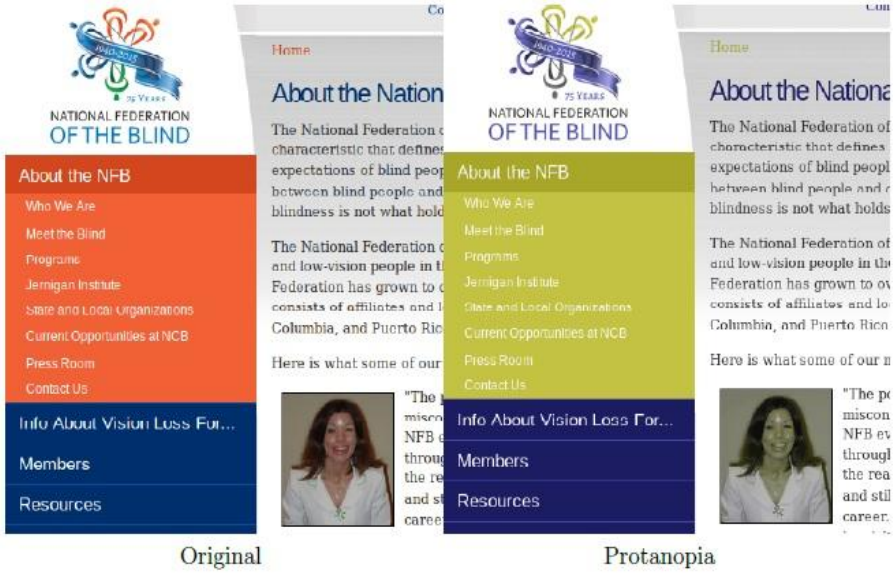


Figure II.5. Exemple d'un type de daltonisme (la protanopie).

II.3 Souhaits et préférences de l'utilisateur

Du fait de sa culture, de son vécu ou de sa santé, chaque utilisateur a ses propres préférences et habitudes dépend de leur confort, pour certain il peut devenir des besoins lorsqu'il n'est pas possible de s'en passer. La compensation d'une déficience est un facteur d'important pour exprimer ces besoins. En effet, chaque individu a un potentiel d'adaptation différent. On peut trouver deux utilisateurs ayant la même pathologie mais vont développer des méthodes différentes afin de compenser leur handicap.

Afin d'accéder à l'information il faut entre autres différencier la structure des divers éléments de la page web, mais aussi comprendre ce qui s'y existe et comment les rendre accessible. Pour cela l'utilisateur peut par exemple augmenter le contraste entre le texte et son arrière plan s'il est insuffisant pour lui. Il peut également modifier la taille du texte et la famille de police afin de rende le contenu lisible. Aussi, il peut changer la couleur si la couleur utilisée lui pose un problème comme ce pourrait être le cas si la personne est daltonienne. La figure II.6 montre une façon de décomposer les besoins et une évolution possible vers une série de souhaits correspondant aux préférences.

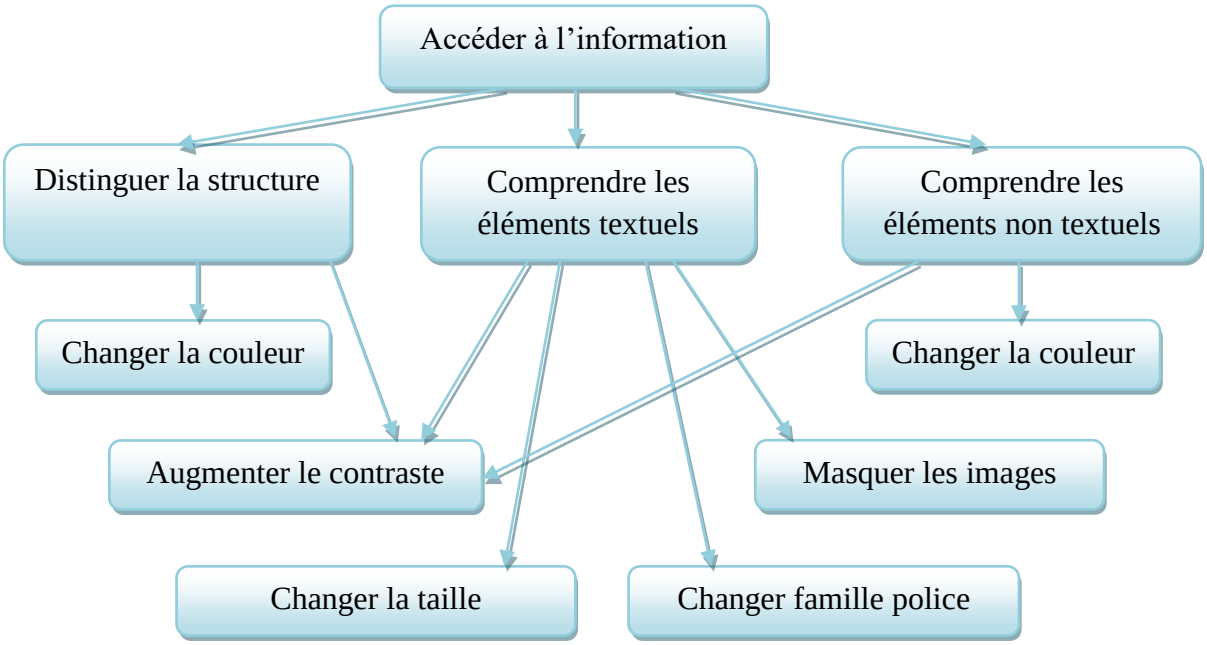


Figure II.6. Du besoin d'accès à l'information aux souhaits (Bonavero, 2015).

II.4 Conclusion

Malheureusement, les problèmes d'accessibilité sont nombreux sur les pages web. Malgré que les normes et recommandations contrôlent le contenu de web, la majorité de ces pages web restent encore inaccessibles. Nous avons vu dans ce chapitre les différents problèmes de visualisation des interfaces web. Une approche basée sur la satisfaction de préférences d'utilisateur malvoyant sera développée durant nos travaux de recherche.

Chapitre III.

La recherche d’information et la reformulation des requêtes web

III.1 Introduction

Selon Salton, la recherche d'information (RI) est un domaine de l'informatique qui consiste à acquérir, organiser, stocker; rechercher et sélectionner l'information selon les besoins des utilisateurs (Salton, 1970) (Salton and McGill, 1983). Les principaux concepts dans ce domaine sont les documents, le besoin en information, la requête, la pertinence et les modèles de la RI.

Dans ce chapitre nous nous s'intéressons au système de recherche de l'information qui est peut être défini comme un ensemble de programme informatique permet de retourner à partir d'une collection de documents, ceux dont le contenu correspond à un besoin en informations d'un utilisateur, exprimé à l'aide d'une requête.

Aussi nous traitons les concepts, les techniques et les modèles de représentation utilisés dans la RI. Nous commençons par la présentation du processus de la RI, puis nous définissons les différentes étapes du processus de la RI, ensuite nous détaillons les principaux modèles de la RI. Etant donné que la reformulation des requêtes représente le cœur du notre sujet, nous présentons les différentes techniques développées pour cette dernière.

III.2 Le Processus de la Recherche d'Information

L'interface du système de la recherche d'information (SRI) permet de saisir les requêtes, ces dernières sont traduises (description structurée) par la même méthode utilisée lors de l'indexation de documents, afin de pouvoir mettre en relation l'ensemble des informations contenues dans le corpus documentaire d'une part, et les besoins d'information d'un utilisateur exprimés par les requêtes d'autre part. Le SRI doit retourner à l'utilisateur le maximum de documents pertinents. Le système est composé de trois modules principaux :

- Le module d'indexation, qui permet une représentation des documents et des requêtes ;
- Le module d'appariement (mise en correspondance) requête-document, qui permet de répondre à l'interrogation ;
- Le module de reformulation de la requête.

Les différentes étapes du processus de RI, sont représentées dans la Figure III.1.

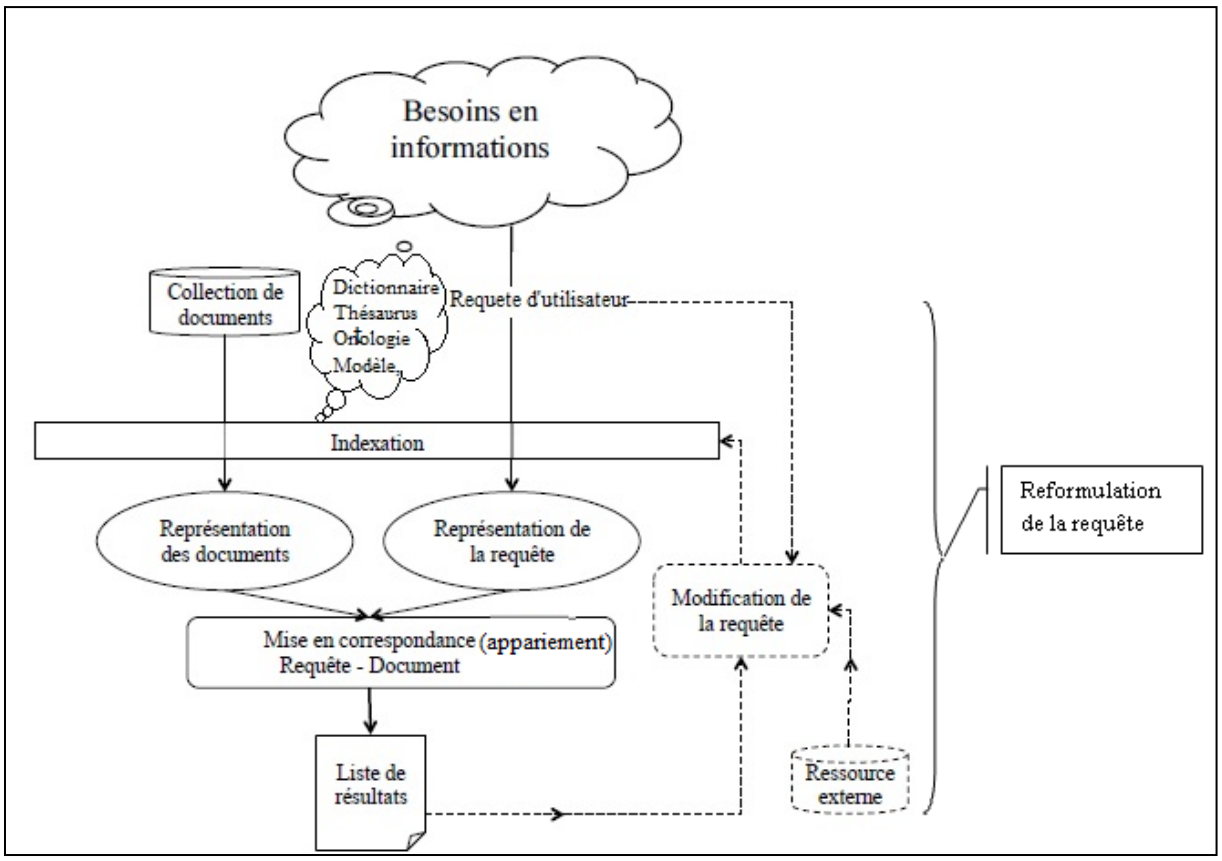


Figure III.1. Le Processus de la Recherche d'Information (Audeh, 2014).

III.2.1 Collection de documents (Corpus)

La collection de documents est un ensemble documentaire qui peut être des documents complets ou bien des parties des documents. Le document est l'unité d'information recherchée pour répondre aux besoins en information de l'utilisateur. Ce dernier peut être un texte, une page web, une image, une bande vidéo,…etc.

III.2.2 Besoin en information

Dans la recherche d'information, un besoin d'information est exprimé par une requête qui permet l'interrogation d'un SRI. Différents types de langages d'interrogation sont présentés dans la littérature : interrogation en langage naturel, interrogation en langage booléen ou bien interrogation graphique. En général, une requête est un ensemble de mots clés qui peuvent probablement être reliés entre eux par des opérateurs définis par un langage adapté au modèle de recherche.

III.2.3 L'indexation

L'indexation (ou l'analyse pour la requête) est l'une des importantes phases dans le processus du RI. Elle permet de transformer un document textuel (ou une requête) à une représentation exploitable par un modèle de RI afin d'extraire un ensemble des mots clés

servant comme descripteurs. Ces descripteurs sont ordonnés dans une structure appelée dictionnaire (thesaurus) constituant le langage d'indexation. Il existe trois modes d'indexation :

- **Indexation manuelle** :

  Le choix et l'extraction des documents sont réalisés par un spécialiste du domaine correspondant ou par un documentaliste;

- **Indexation automatique** :

  Le choix et l'extraction des documents sont réalisés de façon entièrement automatisé;

- **Indexation semi-automatique** :

  Le choix final des documents est laissé au spécialiste du domaine correspondant ou documentaliste, mais l'extraction est réalisée par le système.

III.2.4 Mise en correspondance (appariement) requête-document

La fonction d'appariement (matching en anglais) permet de juger la pertinence d'un document. A ce stade, une probabilité ou mesure de similitude (correspondance) entre les descripteurs des documents de la collection et la requête indexée est calculée. Cette fonction de correspondance permet de classer les documents renvoyés à l'utilisateur par ordre de pertinence.

III.2.5 Modification de la requête (la reformulation)

La modification des requêtes est parfois obligatoire parce que ces requêtes expriment le besoin d'information des utilisateurs. Le choix des termes et celui des paramètres sont les deux principaux facteurs causant la mauvaise qualité des requêtes. La fonction de modification de la requête peut être manuelle, automatique ou bien semi-automatique. Cette dernière sera détaillée dans la section III.4.

III.3 Taxonomie des modèles de RI

Un modèle de RI est une abstraction d'un processus de recherche d'information. Selon l'équipe de Salton (Salton and al., 1983) le premier objectif de ce modèle est de fournir un cadre théorique pour modéliser la mesure de pertinence. Il existe trois principaux modèles :

- Les modèles ensemblistes ;
- Les modèles algébriques ;
- Les modèles probabilistes ;

La figure III.2 montre une taxonomie des principaux modèles de la RI.

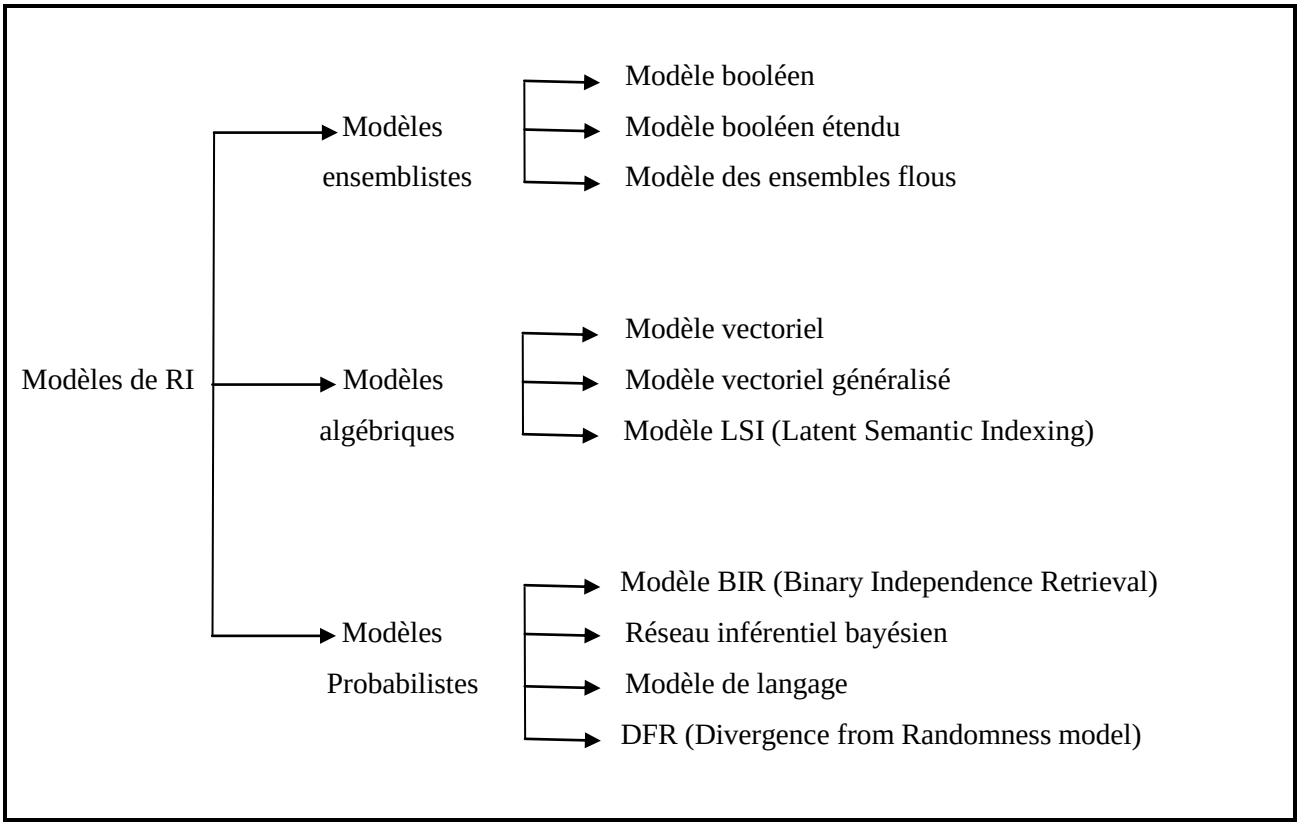


Figure III.2. Taxonomie des modèles de RI.

III.3.1 Les modèles ensemblistes

Les modèles ensemblistes s'appuient sur la théorie des ensembles, où les opérateurs logique (OU, ET et NON) relient les termes de la requête. Ces opérateurs effectuent des opérations d'union, d'intersection et de différence (OU, ET et NON respectivement) entre les ensembles de résultats de chaque terme. Parmi les modèles ensemblistes nous citons :

- **Modèle booléen**

Le modèle booléen (Salton and al., 1993) est le premier modèle de la RI. Dans ce modèle, un document est représenté par un ensemble de termes d= $t^1, t^2, t^3, \dots t^n$, et la requête est représentée sous forme une expression logique dont les termes sont reliés par des opérateurs booléens : AND (∧), OR (∨) et NOT (¬). Il considère qu'un document est soit pertinent ou soit non pertinent.

La correspondance entre un document d_i et une requête q_h est inclue dans l'ensemble {0, 1} et repose sur un appariement exact, c'est-à-dire qu'elle restitue seulement des documents appartenant à l'ensemble décrit par la requête. Cette fonction est exprimée par :

$$RSV(q_h, d_i) = \begin{cases} 1 & Si\ d_i\ appartient\ \text{à}\ l'ensemble\ décrit\ par\ la\ requête \\ 0 & Sinon \end{cases} \qquad \text{Équation 1}$$

L'avantage majeur de ce modèle est sa simplicité de mise en œuvre (Frakes and Baeza-Yates, 1992) et sa transparence à l'utilisateur dans la mesure. Par contre, ce modèle a reçu plusieurs critiques : la premier est que la sélection des documents qui répondent à la requête est basé sur une décision binaire, ainsi que ces derniers sont retournés dans un ordre quelconque, et la formulation des requêtes pour les utilisateurs non expert est difficile (Belkin and Croft, 1992). Ces problèmes peuvent être corrigées par l'utilisation non-binaire des poids des termes, présentée dans le modèle booléen étendu (Salton and McGill, 1983) ou le modèle des ensembles flous (Ogawa and al., 1991).

- **Modèle booléen étendu**

Afin de résoudre les problèmes du modèle booléen, Salton et son équipe (Salton and al., 1983) ont introduit en 1983 le modèle booléen étendu. Ce modèle peut être vu comme une hybridation qui inclut les propriétés des modèles booléen et vectoriel. Le modèle booléen étendu repose sur l'appariement approché et utilise une distance algébrique pour mesurer la pertinence d'un document et une requête.

Considérons w_{tj} le poids du terme d'un document D_j compris entre 0 et 1, avec D_j= ($w_{1j}, w_{2j}, \ldots\ w_{tj}$) ; La similarité entre la requête q (à deux termes t_1 et t_2) et le document D_j décrite selon l'opérateur booléen utilisé (conjonction ou disjonction), et se calcule comme suit :

$$RSV\ (\ D_j, t_1 \vee t_2\ ) = \frac{\sqrt{(w_{1j}^2 + w_{2j}^2)}}{\sqrt{2}} \qquad \textit{Équation 2}$$

$$RSV\ (\ D_j, t_1 \wedge t_2\ ) = \frac{\sqrt{((1-w_{1j})^2 + (1-w_{2j})^2)}}{\sqrt{2}} \qquad \textit{Équation 3}$$

- **Modèle des ensembles flous**

La représentation des documents et des requêtes reflète partiellement les contenus sémantiques des documents et des requêtes (Boughanem and al., 2009). Chaque terme d'un document a un poids qui mesure à quel point le terme caractérise le document. La fonction d'approximation est un sous-ensemble d'un grand ensemble U caractérisée par la fonction μ_A : U → [0, 1] qui affecte à chaque élément u de U la valeur μ_A (u) ∈ [0, 1]. La similarité est calculée par l'Équation 4:

$$RSV\ (d, q) = \frac{\sum_{k=1}^{n} r^{k-1} T(a_k)}{\sum_{k=1}^{n} r^{k-1}} \qquad \textit{Équation 4}$$

Où les $T(a_k)$ sont l'ensemble flou relatif à a_k ; ils sont considérés dans l'ordre décroissant pour les requêtes OU, et croissant pour les requêtes ET, la valeur de r est déterminée expérimentalement.

III.3.2 Les modèles algébriques

Les modèles algébriques se basent sur une représentation vectorielle pour le document et la requête. La mise en correspondance entre la requête et le document est définie par des mesures de similarité entre les vecteurs représentant les documents et les requêtes (à deux termes t_1 et t_2), comme c'est illustré dans la Figure III.3 qui considère un index de deux termes.

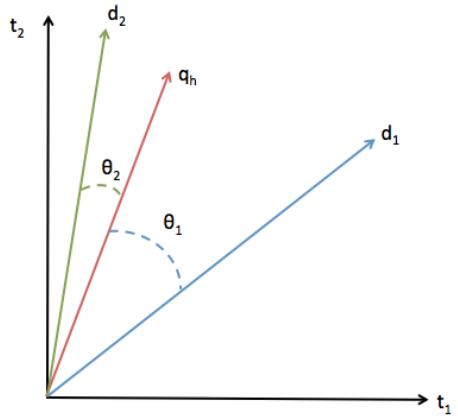


Figure III.3. Représentation algébrique des documents et des requêtes dans l'espace à deux dimensions.

- **Le modèle vectoriel**

Le modèle vectoriel basic (nommé aussi VSM pour Vector Space Model), a été proposé par Gérard Salton et son équipe dans le système de recherche documentaire SMART (Salton's Magical Automatic Retriever of Text) (Salton , 1991). Ce modèle consiste à représenter les requêtes et les documents par des vecteurs d'indexation dans l'espace vectoriel engendré par les termes que le système a rencontré durant d'indexation.

La mesure de similarité entre le document représenté par un vecteur $\vec{d} = (d_1, d_2 \ldots d_n)$, avec d_i au poids d'un terme i dans le document, et la requête définie par $\vec{q} = (q_1, q_2 \ldots q_n)$, avec q_i correspond au poids du terme i dans la requête (prend généralement 0 ou 1 selon que le terme appartient ou pas à la requête), est calculée suivant les propriétés et les mesures issues de la théorie des espaces vectoriels. Plusieurs mesures sont utilisées pour déterminer la similarité entre le document et la requête (Soulier, 2014), nous citons :

	Notation vectoriel	Notation ensembliste	
Produit scalaire	$RSV(d, q) = \sum_i d_i \times q_i$	$\lvert d \cap q \rvert$	*Équation 5*
Cosinus	$RSV(d, q) = \frac{\sum_i d_i \times q_i}{\sqrt{\sum_i d_i^2 \times q_i^2}}$	$\frac{\lvert d \cap q \rvert}{\sqrt{\lvert d \rvert} \times \sqrt{\lvert q \rvert}}$	*Équation 6*
Coefficient de Dice	$RSV(d, q) = \frac{\sum_i d_i \times q_i}{\frac{1}{2}((\sum_i d_i^2) + (\sum_i q_i^2))}$	$\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	*Équation 7*
Mesure de Jaccard	$RSV(d, q) = \frac{\sum_i d_i \times q_i}{\sum_i d_i^2 + \sum_i q_i^2 - \sum_i d_i \times q_i}$	$\frac{\lvert d \cap q \rvert}{\lvert d \cup q \rvert}$	*Équation 8*
Mesure de recouvrement	$RSV(d, q) = \frac{\sum_i d_i \times q_i}{min\ (\sum_i d_i, \sum_i q_i)}$	$\frac{\lvert d \cap q \rvert}{min\ (\lvert d \rvert, \lvert q \rvert)}$	*Équation 9*

Les documents possédant les plus hauts degrés de correspondance sont considérés comme la réponse de la requête. D'une façon générale, les résultats de recherche tendent à démontrer que les systèmes de recherche vectoriels sont plus performants en termes de précision que les systèmes de recherche booléens (Turtle and Croft, 1991).

En plus de son simplicité de mise en œuvre, l'avantage du modèle vectoriel est qu'il permet de renvoyer des listes ordonnées de documents selon une fonction d'appariement approximatif. Le principal inconvénient de ce modèle est le fait qu'il suppose que les termes d'indexation forment une base (l'hypothèse d'indépendance des termes −bag of words−). Alors qu'ils existent plusieurs relations sémantiques (comme les ontologies) qui s'expriment un terme en fonction des autres. Song et Croft (Song and Croft, 1999) ont proposé une solution pour ce problème qui permet de regrouper les termes successifs qui peuvent avoir le même sens (la considération des N-grammes). Le modèle LSI (Furnas and al., 1988) est une autre solution qui a été proposée pour réduire les dimensions de représentation des document et d'exprimer un terme en fonction des autres.

- **Le modèle vectoriel généralisé**

En 1985, Wong et al. (Wong and al., 1985) ont introduit le modèle vectoriel généralisé pour établir un cadre formel dans lequel les dépendances entre les termes de l'index peuvent être représentés d'une manière facile. Ce modèle est caractérisé par sa complexité et sa lenteur par rapport au modèle vectoriel basique.

- **Le modèle LSI (Latent Semantic Indexing)**

Afin de résoudre le problème de la représentation des requêtes et des documents dans le modèle vectoriel, Furnas et son équipe (Furnas and al., 1988) ont proposé le modèle d'analyse sémantique latente. Ce modèle permet de diminuer l'espace des termes, en

rassemblant les termes similaires dans les mêmes dimensions. Les vecteurs de termes de l'index sont alors représentés dans un nouvel espace plus réduit, composé de concepts de haut niveau, en utilisant la technique SVD (Singular Value Decomposition) :

$$A = U \times \sum \times V^T \qquad \textit{Équation 10}$$

Où A (m × n) est la matrice documents-termes originale, U (m × r) est une matrice dont ces colonnes sont les vecteurs propres de AA^T, Σ (r × r) est une matrice diagonale (contient les valeurs singulières de A, $\Sigma = \text{diag}(\sigma_1,...\sigma_r)$) et V^T(r × n) est une matrice orthogonale. La valeur r est définie par r = inf (n, m) où m est le nombre de termes et n est le nombre de documents.

Les valeurs de Σ sont des nombres réels et non négatifs, et sont triées dans l'ordre décroissant.

$$\Sigma = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{pmatrix}$$

La représentation par des mots-clés possède beaucoup de bruit. Précisément, ce bruit s'interprète par les valeurs les moins élevées de Σ. En plus, la technique de LSI tend à supprimer ce bruit en écartant ces valeurs, ce qui réduit la dimension de Σ à k, cette matrice réduite est notée Σ_k (Figure III.4).

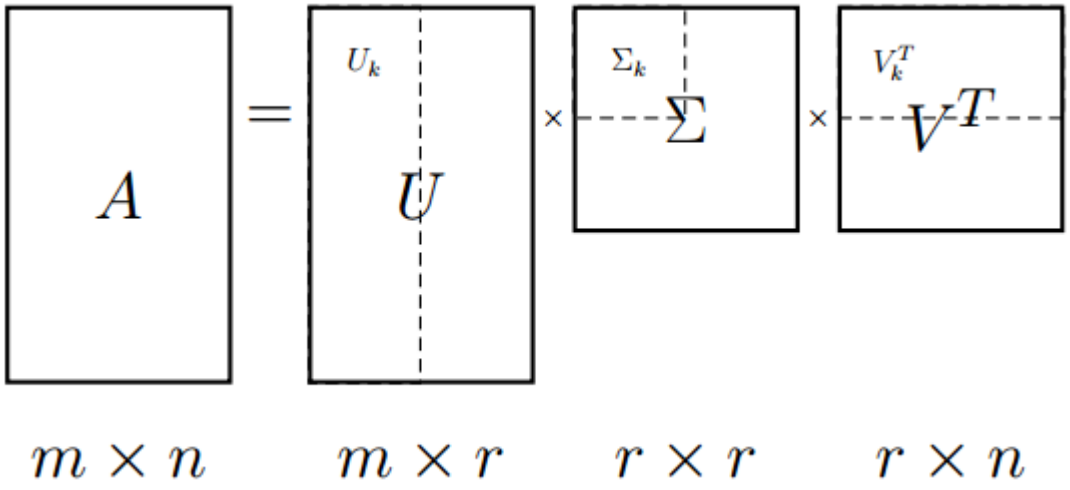


Figure III.4. Le modèle LSI.

Quand une requête est soumise, elle est aussi traduite dans ce nouvel espace en utilisant la formule suivante :

$$q = q^T \times U_k \times \sum_k^{-1} \qquad \textit{Équation 11}$$

La diminution de la dimension de l'espace vectoriel se fait par une compression sans perte considérable. La quantité de l'information préservée par la décomposition de A en $U_k \times \Sigma_k \times V_k^T$ peut être évaluée comme suit :

$$Contraste = \frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{r} \sigma_i} \qquad \text{Équation 12}$$

- **Le modèle connexionniste**

Le modèle connexionniste imite les phénomènes mentaux ou comportementaux du l'être humain à l'aide des réseaux de neurones. Chaque neurone présente une unité de traitement de base qui, lorsqu'elle est stimulée par des signaux d'entrée, peut émettre des signaux de sortie comme un réactif. Formellement, le réseau de neurones R est composé d'un ensemble de couches L_i et un ensemble de liens C_i :

$$R= ((L_1, L_2 \ldots L_n), (C_1, C_2 \ldots C_{n-1}))$$

Chaque couche L_i est constituée des neurones n_{ij} :

$$\forall\ 1 \leq I \leq n,\ L_i = \{n_{i1}, n_{i2} \ldots n_{i|L_i|}\}$$

Où chaque neurone n_{ij} est caractérisé par un poids w_{ij} et une fonction d'activation g_{ij} , la relation entre les différents neurones est définie comme suit :

$$\forall\ 1 \leq I \leq n,\ C_i : L_i \times L_{i+1} \mapsto \mathbb{R}$$
$$(n_{ij}, n_{i+1k}) \mapsto C_i\,(n_{ij}, n_{i+1k})$$

Dans le cadre de la recherche d'information il existe plusieurs représentations du réseau de neurones (Wilkinson and Hingston, 1991). Boughanem et son équipe (Boughanem and al., 2000) ont proposé une architecture d'un réseau de neurones pour la recherche d'information . Cette architecture est composée de trois couches (Figure III.5), dont la première couche modélise les termes d'une requête, la deuxième modélise les termes d'indexation et la dernière couche modélise les documents de la collection.

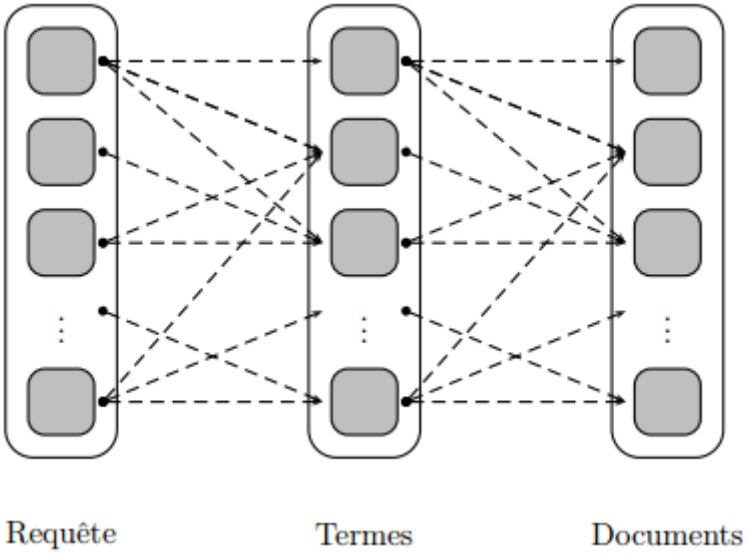


Figure III.5. Un modèle de réseau de neurones pour la recherche d'information.

III.3.3 Les modèles probabilistes

Les modèles probabilistes s'appuient sur la théorie de la décision. L'objectif est de calculer la probabilité de la pertinence d'un document vis-à-vis d'une requête (Robertson and Walker, 1994) (Robertson and al., 1995).

- **Le modèle probabiliste**

Au début des années 1960, Maron et Kuhns (Maron and Kuhns, 1960) ont proposé le premier modèle probabiliste pour modéliser le processus de sélection des documents dans un système de recherche d'information en s'appuyant sur la théorie de la décision. Dans ce modèle une requête et un document sont représentés par un vecteur comme dans le modèle vectoriel, mais les poids des termes sont binaires (soient $D_i = \{ D_1, D_2 \ldots D_n \}, D_i$ prend 1 si le terme t_i est présent dans le document et 0 sinon). Le principe de base du ce modèle consiste à déterminer les probabilités $p\ (L = 1|d)$ (la probabilité que le document d soit pertinent pour la requête q) et $p\ (L = 0|d)$ (la probabilité que le document d ne soit pas pertinent pour la requête q).

Si le document d contient les termes $D_1, D_2 \ldots D_n$, alors $p\ (L = 1|d)$ qui mesure la probabilité que d soit pertinent pourra se ramener à la probabilité qu'un document soit pertinent sachant qu'il contient les termes ($D_1, D_2 \ldots D_n$), cette probabilité sera notée $P\ (L = 1|\ D_1, D_2 \ldots D_n)$ (Maron and Kuhns, 1960) (Robertson and Jones, 1976). La fonction de similarité entre un document d et une requête q est défini par :

$$RSV\ (d, q) = log \frac{P\ (L = 1|\ D_1, D_2 \ldots D_n)}{P\ (L = 0|\ D_1, D_2 \ldots D_n)} \qquad \text{Équation 13}$$

De nombreuses approches du modèle probabiliste ont été présentées dans la littérature (Maron and al., 1960) (Bookstein,1983) (Fuhr, 1989), le modèle le plus utilisé est le modèle Okapi BM25 (Robertson and Walker, 1994). Les atouts majeurs de ce modèle consistent en la considération de la longueur des documents dans le calcul de la pertinence et de la fréquence des termes dans la collection.

- **Le réseau inférentiel bayésien**

Les réseaux bayésiens (RB), aussi appelés réseau de croyance. Ils sont des graphes orientés acycliques (GAO) ou réseaux causals définis par un triplet (N, A, P) (Jensen, 2007) dans lesquels les nœuds (N) représentent des variables aléatoires, les liens (A) sont des dépendances entre ces variables, et P est la distribution des probabilités dans le RB (Figure III.6). C'est l'ensemble des probabilités conditionnelles [(1)] des nœuds sachant leurs parents dans le RB. Différentes formes de modèles Bayésiens pour la RI ont été proposées dans la littérature selon différents points de vue. On distingue deux principaux modèles de réseaux, les réseaux d'inférence et les réseaux de croyance.

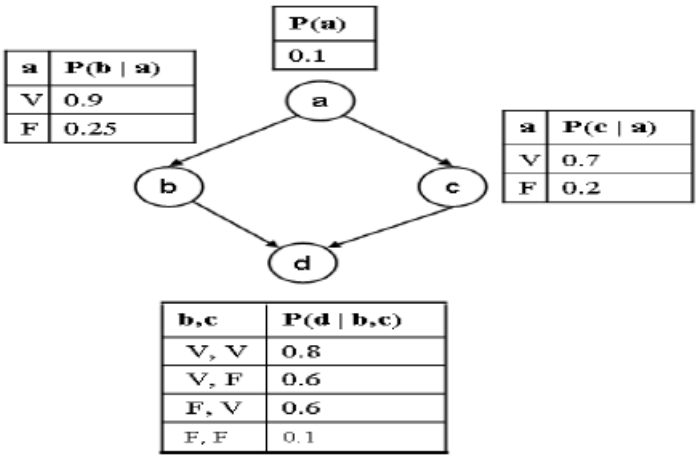


Figure III.6. Exemple d'un réseau bayésien.

Les réseaux bayésiens d'inférence : les probabilités initiales sont attachées aux racines du graphe en calculant de proche en proche le degré de croyance associé aux autres nœuds du graphe.

Les réseaux inférentiels bayésiens examinent le problème de la RI d'un point de vue épistémologique [(2)]. Ils associent des variables aléatoires aux termes d'indexation, aux requêtes de l'utilisateur et aux documents. Les termes des documents et de l'indexation sont représentés comme des nœuds.

[1] Une probabilité conditionnelle P(B|A) exprime la force (strength) de l'influence de la relation de *A* à *B*.
[2] L'approche épistémologique interprète les probabilités comme un degré de croyance dont les spécifications viennent de statistiques expérimentales.

Une variable aléatoire associée à un document correspond à l'évènement où ce document est observé, de même pur la requête. Les arcs sont orientés du nœud document vers les nœuds termes, l'observation d'un document dans la collection est la condition d'une augmentation de la valeur des variables associées aux termes d'indexation (Boughanem and al., 2007). Aussi, les arcs sont dirigés des nœuds associés aux termes d'indexation vers le nœud de la requête. La figure III.7, issue de (Turtle and Croft, 1989), décrit la forme graphique d'un réseau inférentiel bayésien simple de pertinence d'un document à une requête de trois termes.

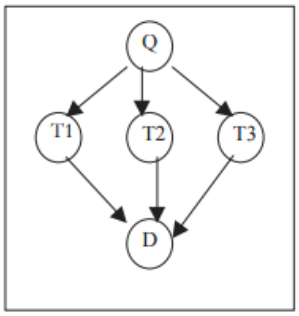


Figure III.7. Modèle de réseau inférentiel pour la RI.

L'évènement « la requête est accomplie » (Q = 1) est réalisé si le sujet lié à un terme est vrai (T1= 1, T2= 1 ou T3= 1), ou une combinaison de ces évènements.

Les trois sujets sont inférés par l'évènement « le document est pertinent » (D = 1).

Par l'enchaînement de règles de probabilités, la probabilité jointe des autres nœuds du graphe est :

$$P(D, T1, T2, T3, Q) = P(D) \times P(T1|D) \times P(T2|D, T1) \times P(T3|D, T1, T2) \times P(Q|D, T1, T2, T3)$$

Équation 14

Le modèle inférentiel bayésien a été mis en œuvre dans le système Inquery (Allan and al., 1998) .Il est utilisé pour formuler des requêtes simples basées sur des mots-clés, des requêtes booléennes, des requêtes basées sur des phrases[3] (Figure III.8) ou bien une combinaison des trois types (Croft and al., 1995). Il propose des opérateurs de moyenne et de moyenne pondérée, des opérateurs booléens probabilistes ou stricts (conservant les probabilités), des opérateurs de proximité et de synonymie. Pour ce faire, une procédure de prétraitement de la requête consiste à générer une forme inférentielle prête à être évaluée.

[3] Cai et son équipe ont proposé dans (Cai and al., 2009) un modèle de recherche documentaire bayésien basé sur les phrases.

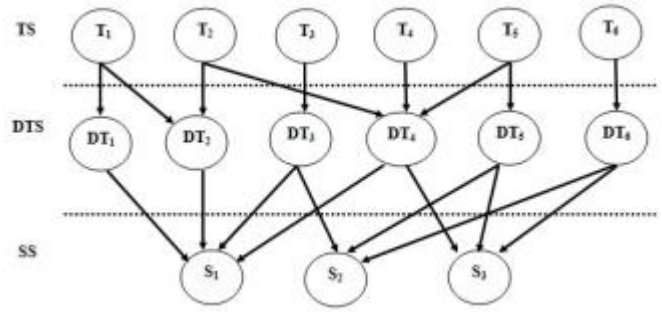


Figure III.8. Structure du modèle bayésien basé sur les phrases.

D'autres travaux ont été basés sur les réseaux bayésiens. Citons par exemple les "Belief Networks" introduits par Ribeiro-Neto et Muntz (Ribeiro-Neto and Muntz, 1996), les travaux de Vogt (Vogt, 1999), et ceux de Garrouch (Garrouch, 2017).

Les réseaux bayésiens de croyance : Les réseaux bayésiens de croyance (belief network), sont introduits en 1996 par Ribeiro-Neto et Muntz (Ribeiro-Neto and Muntz, 1996). Ils sont aussi basés sur une étude épistémologique pour l'interprétation des probabilités, mais travaillent dans un espace différent. Cependant, ils s'écartent de l'inférence réseau par l'adoption d'une topologie de réseau différente bien définie, les réseaux de croyance sont une généralisation des réseaux d'inférence.

Le problème majeur des réseaux bayésiens reste toujours le calcul des probabilités, qui nécessite un temps exponentiel au nombre de termes dans la requête, même si l'introduction des quatre formes canoniques présenté par Turtle et Croft (Turtle and Croft, 1990) résout partiellement le problème.

- **Le modèle de langage**

L'hypothèse de base dans le modèle de langage est : « *un utilisateur en interaction avec un système de recherche fournit une requête en pensant à un ou plusieurs documents qu'il souhaite retrouver* ». Dans ce modèle un document n'est considéré pertinent que si la requête d'utilisateur ressemble à celle inférée par le document. On cherche alors à estimer la probabilité que la requête soit inférée par le document (Boughanem and al., 2004). Les modèles de langages permettent d'estimer cette probabilité et l'utilisent pour ordonner les documents.

On considère une fonction de probabilité P qui assigne une probabilité P(s) à un mot ou à une séquence de mots s = $m_1 m_2 \ldots m_n$ en une langue :

$$P(s) = \prod_{i=1}^{n} P(m_i \mid m_i \ldots . m_{i-1}) \qquad \textit{Équation 15}$$

Etant donné qu'un document d incarne un sous-langage, pour lequel on tente de construire un modèle de langue MD. Le score du document d vise à vis d'une requête q = $t_1 t_2 \ldots t_n$ est déterminé par la probabilité que son modèle génère la requête :

$$RSV(d, q) = P(Q|MD) = P(t_1 t_2 \ldots t_n|MD) \qquad \text{Équation 16}$$

III.4 Reformulation des requêtes

Rechercher des informations, se déplacer ou se repérer sur le web peuvent parfois poser de grands problèmes face aux différentes tâches qu'on doit réaliser, mais aussi face à la quantité élevée des données qui peuvent s'afficher sur l'écran et des liens qu'il faut survoler. Les internautes doivent examiner beaucoup de données afin de trier les résultats trouvés pour au final accéder à l'information recherchée. Pour cela, mettre en correspondance un besoin d'information et un document (ou bien lien) pertinent par la reformulation des requêtes est souvent la meilleure solution.

On peut définir le concept de « reformulation de la requête » comme une procédure qui permet de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur, en utilisant des fonctions, comme la proximité et la pondération, fournis par le modèle de recherche.

Les dénominations des diverses techniques visant à modifier ou réajuster d'une façon ou d'une autre une requête ne sont pas toujours claires et se contredisent même au gré des publications. Par exemple, dans les travaux de Bhogal et son (Bhogal and al. ; 2007, 2013) ainsi que dans (Carpineto and al., 2012), (Audeh, 2014) et plus récemment dans (Kumar and al., 2022) et (Azad and al., 2022) , ces techniques sont désignées par "méthodes d'expansion de requêtes" alors que dans le travail de Hlaoua (Hlaoua, 2007), dans (Klampanos and al., 2009) et dans (Breja and Jain, 2021) l'expansion de requêtes ne désigne qu'un cas particulier de modification de requête parmi d'autres. Dans les paragraphes suivants, nous désignons par "techniques de reformulation de requêtes" toute méthode de modification ou de réajustement de requêtes.

Le concept de reformulation est à la fois un processus itératif et interactif entre l'utilisateur et les moteurs de recherche pour obtenir des résultats satisfaisants (Lu and al., 2018). Les résultats récupérés jouent un rôle important dans la stratégie de reformulation. Bien que ce concept soit étudié par de nombreux chercheurs, il est devenu parmi l'un des principales notions dans le domaine de la recherche d'information et il a fait l'objet de beaucoup de travaux. Ces travaux apportent des solutions aux deux importantes questions :

1. Comment peut-on retrouver plus de documents pertinents par rapport à une requête donnée ?
2. Comment peut-on mieux exprimer la requête de l'utilisateur de façon à mieux répondre à son besoin ?

Les principales techniques d'amélioration des SRI par reformulation de la requête initiale en y ajoutant de nouveaux termes sont illustrées dans la Figure III.9. Ces techniques sont distribuées en trois principales voies : Combinaison de diverses présentations de la requête, expansion de la requête et réinjection de pertinence. Nous présentons dans les paragraphes suivants ces trois voies.

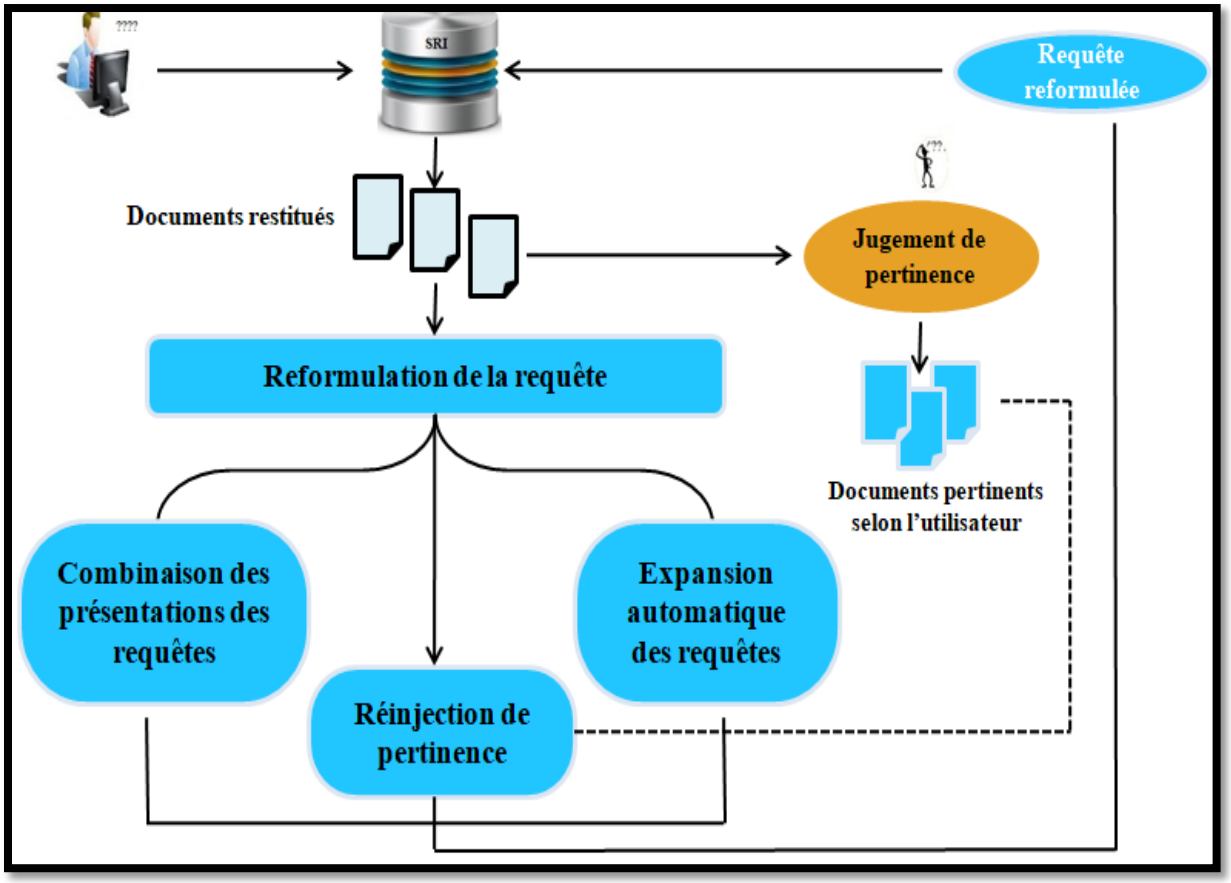


Figure III.9. Les techniques d'amélioration des SRI par reformulation de la requête.

III.4.1 Expansion et combinaison de requêtes

L'expansion directe de la requête est désignée comme un traitement pour agrandir le champ de recherche pour cette requête. Elle consiste principalement à rajouter à la requête initiale de l'utilisateur des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections.

Plus précisément, dans la catégorie des ressources linguistiques, l'objectif est d'utiliser un vocabulaire contrôlé issu de ressources externes. Ces ressources utilisées sont sous la forme d'un thésaurus.

Le thésaurus (Wei and al, 2000) est une structure de données où sont enregistrées les associations entre les termes. Il regroupe diverses informations de type statistique (pondération des termes) et linguistique (équivalence, association, hiérarchie), et sa construction se fait généralement pendant le processus d'indexation (Hlaoua 2007), et peut être manuelle, comme c'est le cas pour « WordNet [(4)] » (Azad and Deepak, 2019), ou automatique, par exemple comme un thésaurus basé sur les similarités, ou un thésaurus statistique.

Wordnet (Voorhees, 1994) était le premier projet qui est utilisé pour l'expansion de la requête (Miller, 1995). L'idée de Voorhees est d'ajouter des synonymes choisis manuellement dans Wordnet aux termes de la requête, il a conclu que le succès de cette technique dépend de deux éléments : la qualité de la requête initiale et celle des synonymes trouvés. Gonzalo et son équipe (Gonzalo and al., 1998) ont utilisé le même principe mais, pour l'indexation de documents, où le choix de synsets de l'index a été fait manuellement. Une étude (Navigli and Velardi, 2003) est basée sur le choix des meilleurs termes d'expansion une fois la désambiguïsation (le choix du synset[(5)] -Synonym Set- dans WordNet) faite. Alors que cette étude indique que l'utilisation de la définition (le glossaire de WordNet) est plus adéquate pour l'expansion de la requête que les synonymes et les hyperonymes, les résultats ne sont pas concluants à cause des limites des expériences. Shah et Croft (Shah and Croft, 2004) utilisent la technique de (Cronen-Townsend and al., 2002) pour pondérer les termes de la requête par les scores de clarté. Dans cette dernière technique, les termes qui ont un score élevé sont considérés comme précis et ne sont donc pas étendus, ceux avec un score faible sont considérés comme ambigus et ne sont pas étendus non plus. Seulement les termes de clarté « moyenne » sont étendus par des synonymes de WordNet. Song et son équipe (Song and al., 2007) ont proposé une autre approche d'expansion sémantique de requêtes (dite approche SemanQE+Ontologie) en combinant les règles d'association avec les ontologies et les techniques de traitement automatique des langues. La méthode proposée utilise la sémantique explicite aussi les propriétés linguistiques des corpus non structurés de texte, les propriétés

[4] WordNet est une base de données lexicale (Baziz, 2005), conçue manuellement par des experts dans le but de classifier, répertorier, et mettre en relation de plusieurs manières le contenu lexical et sémantique de la langue anglaise.

[5] Un groupe de mots interchangeables, dénotant un sens ou un usage particulier.

contextuelles des termes importants découverts par les règles d'association et ajoute les entrées d'ontologie à la requête en s'appuyant sur une technique de désambiguïsation de sens des mots. Les auteurs ont choisi pour la génération des règles d'association, l'algorithme Apriori. Les synsets de WordNet et les collections TREC sont utilisés comme des entrées d'ontologie. Un intéressant travail sur l'utilisation de WordNet pour l'expansion de la requête est celle de Fang (Fang, 2008) qui s'appuie sur la pondération des termes d'expansion provenant de WordNet. Avec des expériences sur six collections de TREC (de taille inférieure à 600000 documents), une amélioration significative est trouvée lors de l'expansion en utilisant les définitions de synsets. Ce travail a été effectué à l'aide d'un modèle de recherche axiomatique [6] (Fang and Zhai, 2005), et la pondération des termes d'expansion a été adaptée au modèle axiomatique, avec l'amélioration des résultats. Une autre approche d'expansion de la requête en utilisant WordNet a été faite dans (Zhang and al., 2009). Zhang et son équipe ont basé sur l'ajout simple de tous les synonymes d'un synset (désambiguïsé) par leur méthode de désambiguïsation.

D'autres méthodes s'appuient sur l'ajout des variantes morphologiques des termes employés par l'utilisateur à la requête, dont le but d'assurer la restitution des documents indexés par des variantes des termes composant la requête.

Les associations conçues manuellement interprètent des relations de synonymie et de hiérarchie. Les thésaurus construits de façon manuelle sont une bonne solution pour l'expansion de requête. Cependant, leur construction et la maintenance des informations sémantiques qu'ils possèdent sont coûteuses en temps et sollicitent le recours des experts des domaines considérés. Pour cela, ils restent peu utilisés par les SRI.

En ce qui concerne la deuxième catégorie, les ressources sont construites en se basant sur une analyse statistique des collections. Il s'agit de chercher des associations de différents termes pour ajouter des termes voisins à la requête.

Les premières techniques employées dans cette catégorie (Minker and al., 1972) s'appuient sur la classification automatique des mots (term clustering), où pour chaque terme de la requête, ajoutent tous les termes du même groupe que celui-ci. Ces groupes sont créés en utilisant les statistiques de cooccurrence de termes dans toute la collection. L'effet de cette cooccurrence entre les termes sur la performance d'un modèle de RI (en précision et en rappel) n'a pas été étudié. D'autres techniques sont proposées par Qiu et Frei (Qiu and Frei,

[6] L'idée d'un modèle de recherche axiomatique est de choisir la fonction de correspondance selon des propriétés exprimables mathématiquement qu'elle doit vérifier (axiome), par exemple convexité, croissance etc.

1993), ils ont introduit le thésaurus de similarité[7] pour l'expansion des requêtes. Dans cette approche, le thésaurus de similarité est fondé sur la manière dont laquelle les termes de la collection sont indexés par les documents. Les auteurs ont utilisé une méthode probabiliste pour calculer la probabilité qu'un terme soit similaire à une requête dans l'espace vectoriel de document (DVS : Document Vector Space). Ils ont constaté que leur méthode a une bonne performance avec des collections de taille élevée (dans leurs expériences, la plus grande collection possède moins de 12000 documents) et avec un grand nombre de termes d'expansion (800 termes d'expansion pour la plus grande collection). Jing et Croft (Jing and Croft, 1994) ont proposé l'utilisation de la technique « phrase finder ». Dans cette dernière, un thésaurus d'association qu'ils produisent en examinant le texte de tous les paragraphes dans les documents, pour calculer la fréquence d'occurrence des termes et des expressions. L'avantage de cette approche est au niveau de la considération des expressions et la complexité des calculs.

Peu d'articles sur ce genre d'expansion de la requête sont proposés plus tard, car l'inconvénient majeur de ces techniques est la taille énorme de la collection de documents, ce qui induit à des méthodes coûteuses sur les collections de nos jours en matière du temps de traitement. Aussi, sur une collection de plusieurs documents tels que le web, la similarité sémantique entre les termes calculés sur tous les documents est moins précise que celle qu'on peut calculer sur un sous-ensemble de documents lié au sujet de la requête qu'on désire étendre. La segmentation du corpus en plusieurs corpus de domaines est une proposition qui permet de diminuer la taille du corpus de recherche pour une requête en lui affectant un corpus local (El Ghali, 2016). Malgré cela, divers problèmes se posent, parmi lesquels, la limitation des corrélations de termes pour les termes de la requête qui sont présents dans le corpus et l'absence de quelques termes de la requête dans le corpus local du domaine. Afin de résoudre ces problèmes et pour enrichir les connaissances extraites du corpus de domaine, Gan et Hong (Gan and Hong, 2015) ont présenté une approche d'expansion qui utilise le corpus local du domaine de la requête, et les pages Wikipédia liées au domaine du corpus local. L'approche présentée s'appuie sur la construction d'un réseau de Markov en utilisant les corrélations entre termes. Cette construction se fait en trois étapes : la première étape consiste à construire le réseau de Markov à l'aide de corpus local (appelé B-Markov network); la deuxième étape consiste à construire le réseau de Markov à l'aide de corpus

[7] Un thésaurus de similarité est une matrice, construite en prenant en considération les similarités terme à terme plutôt que les simples données de cooccurrence.

Wikipédia (appelé W-Markov network); et la dernière étape le W-Markov network est superposée à B-Markov network pour former un nouveau réseau plus vaste et plus riche en corrélation pour chaque terme, ce réseau est appelé C-Markov network. Dans (Keikha and al., 2018), les auteurs ont choisi Wikipédia comme source pour extraire les articles pertinents et ont utilisé des méthodes supervisées et non supervisées pour sélectionner les termes d'expansion candidats. Les chercheurs de l'article (Azad et Deepak, 2019) ont également utilisé Wikipédia combiné à WordNet comme sources de données pour les termes d'expansion. Les résultats montrent que cette combinaison des deux sources de données a donné de bons résultats par rapport aux deux méthodes individuellement.

D'autres travaux se sont focalisés sur les ontologies (Aminu and al., 2018; 2019) (Misbah and Maruf, 2016) (Oyebisi Oyefolahan and al., 2018). D'après Sharma et son équipe (Sharma and al., 2021) l'expansion des requêtes par les ontologies se base sur la relation sémantique entre les mots, ce qui aide à récupérer des résultats plus pertinents, et donc à fournir des résultats plus précis en donnant une valeur élevée à la fois de rappel et de précision, mais créer une ontologie est en soi une tâche difficile.

(Aminu and al., 2018) présente une approche d'expansion des requêtes comme cadre de développement d'un système de recherche d'informations basé sur les ontologies. Le système proposé comprend quatre couches : La première est la couche de recherche de requête comprenant l'interface utilisateur, où les termes de la requête et les termes candidats sont extraits en fonction de la conception de l'ontologie à l'aide de SPARQL. La Deuxième, est décrite comme la couche centrale du système proposé, car elle gère l'extension de la recherche de requête initiale pour récupérer les résultats pertinents avec l'utilisation du WordNet. La troisième, est la couche de recherche sémantique permet d'interroger les termes de recherche, qui sont développées (argumentées) et mises en correspondance avec les concepts d'ontologie du domaine proposé. La dernière couche permet de récupérer les résultats. Dans (Devi and Gandhi, 2015), les auteurs ont présenté une approche nommée Semantic Information Retrieval in Sports Domain (SIRSD), qui est basée sur les ontologies et WordNet pour améliorer les résultats de la requête renvoyée. La recherche a décrit l'importance de l'expansion des requêtes à l'aide de l'ontologie par rapport à l'expansion des requêtes par réinjection de pertinence (recherche traditionnelle). Le travail de recherche (Tulasi and al., 2017) s'appuie aussi sur le même domaine d'ontologie (Sport) de l'approche précédente. Un autre système utilisant les ontologies et l'indexation pour la reformulation des requêtes est proposé dans (Misbah and Maruf, 2016). Les auteurs présentent une approche en

couches pour l'expansion des requêtes en associant des annotations sémantiques et en attribuant des poids à ces concepts pour une récupération de requête avec précision.

III.4.2 Combinaison de requêtes

Diverses techniques de la recherche d'information (Simonnot, 1996) utilisent une seule représentation de requête en comparant aux diverses représentations de document. Lee (Lee, 1998) a prouvé dans son approche que lorsqu'en exploitant des représentations multiples de requêtes ou en utilisant différents algorithmes de recherche ou bien différentes techniques de réinjection, on atteint une recherche plus efficace.

L'augmentation du rappel d'une requête peut être faite par la combinaison des représentations de cette dernière, et l'augmentation de la précision peut être faite par la combinaison des algorithmes de recherche. La base théorique de la combinaison des évidences a été proposée par Ingwersen dans ces travaux (Ingwersen, 1994; 1996). Il a montré que des représentations multiples du même objet (pour une requête) permettent une meilleure perception de l'objet qu'une seule bonne représentation. Néanmoins, il faut que chaque source d'évidence employée offre non seulement un point de vue différent sur l'objet, mais que ces points de vue aient différentes bases cognitives. Les représentations multiples d'une requête peuvent fournir différentes interprétations du besoin en information.

Belkin et al. ont proposé dans (Belkin and al., 1995) une méthode de combinaison de multiples représentations de requêtes. Elle permet de calculer les scores des documents directement à partir de la fonction de mise en correspondance document vis-à-vis une requête en utilisant le même système de recherche mais avec différentes versions de la requête, et les résultats acquis par chaque version sont combinés pour choisir une seule liste finale. Ces versions sont obtenues soit par des présentations d'une même requête dans plusieurs langages (qui sont différents), soit par des expressions d'une même requête par des chercheurs différents.

III.4.3 Réinjection de pertinence

L'intervention de l'utilisateur au niveau de jugement de la pertinence des documents restitués par le SRI joue un rôle important pour raffiner le processus de reformulation de requêtes (Robertson and al., 1995). Les informations de pertinence sont utilisées par un ensemble de méthodes appelées « la Réinjection de Pertinence RP » (Relevance Feedback). Ces mêmes méthodes, peuvent être utilisées en ne prenant pas en considération le jugement

des utilisateurs, dans ce cas elles sont appelées « Techniques de Réinjection de Pertinence Implicite » ou bien « Pseudo-RP ».

La RP est une approche évolutive et interactive. Elle se base sur l'utilisation de la requête initiale pour amorcer la recherche, ensuite modifier cette dernière en s'appuyant sur les jugements de pertinence (respectivement non-pertinence) de l'utilisateur afin de répondre les termes de la requête initiale, ou y ajouter ou supprimer d'autres termes contenus dans les documents pertinents et/ou non pertinents. La nouvelle requête obtenue à chaque itération de feedback, permet de cibler les documents pertinents.

- **Le processus de la réinjection de pertinence**

Le processus de la RP, comme élaboré sur la Figure III.10, comporte essentiellement trois phases : l'échantillonnage, l'extraction des évidences et la réécriture de la requête.

- L'échantillonnage : cette phase permet de construire un échantillon de documents à partir des éléments évalués par l'utilisateur. Cet échantillon est caractérisé par le nombre d'éléments jugés et le nombre d'éléments jugés pertinents.

- L'extraction des évidences est la phase la plus importante dans le processus de RP, et elle permet (White, 2004) :

  - L'extraction des termes pertinents qui serviront à l'enrichissement de la requête initiale. Ces termes choisis par le système de la RP sont précisément ceux qui différencient le plus entre les documents marqués pertinents et ceux qui ne le sont pas.
  - La repondération des termes de la requête, en augmentant le poids des termes contenus dans les documents pertinents et en diminuant le poids des termes qui apparaissent dans les documents non-pertinents.

- La réécriture de la requête permet de formuler une nouvelle requête en combinant la requête initiale de l'utilisateur avec les informations extraites dans la phase précédente.

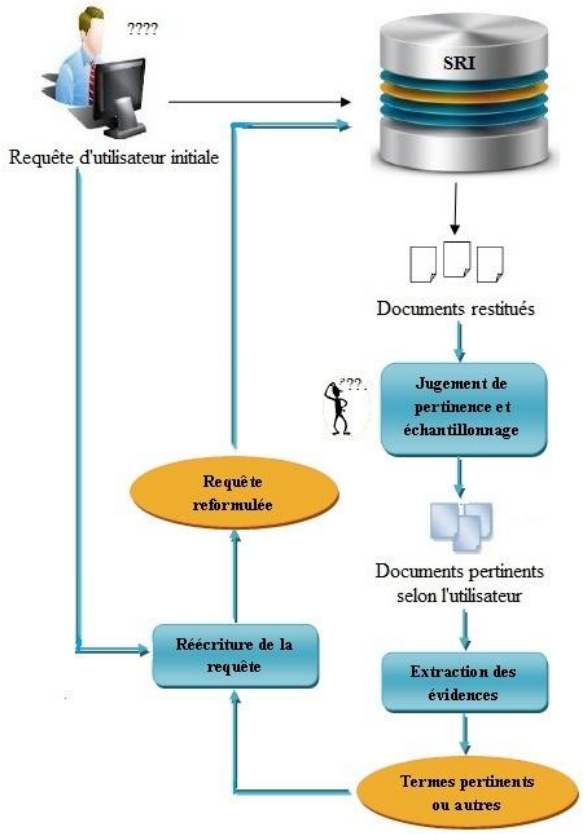


Figure III.10. Le Processus de la réinjection de pertinence.

- **Principales approches de réinjection de pertinence en RI**

Diverses techniques ont été développées dans ce contexte, la plus reconnue est celle de (Rocchio, 1971) adaptée au modèle vectoriel. Rocchio a introduit la première technique de reformulation automatique des requêtes. Il a considéré que la restitution des documents pertinents est liée à l'idée de « requête optimale ». Cette idée est supposée pour maximiser la discrimination entre les vecteurs des documents pertinents et celui des documents non-pertinents. Néanmoins, comme l'utilisateur n'est pas qualifié pour soumettre une requête optimale au système, la RP consiste à rapprocher le vecteur de la requête initiale du vecteur moyen des documents pertinents et de l'éloigner du vecteur moyen des documents non-pertinents.

La formule originale de Rocchio est déterminée par l'équation suivante :

$$Q_{nlle} = Q_{init} + \frac{1}{n_r}\sum_{\forall j \in [1,n_r]} D_j - \frac{1}{n_n}\sum_{\forall i \in [1,n_n]} D_i \qquad \textit{Équation 17}$$

Avec Q_{nlle} est le vecteur de la nouvelle requête, Q_{init} est le vecteur de la requête initiale, n_r et $n_n$sont le nombre de documents pertinents et non pertinents respectivement, D_j est le vecteur du $j^{ème}$ document pertinent et D_i est le vecteur du $i^{ème}$ document non pertinent.

L'ensemble des termes de la requête sera éliminé si le poids d'un terme de cette dernière décroît vers zéro ou au-dessous de zéro.

Une variante de cette équation, permet de pondérer la contribution relative de la requête initiale, des documents pertinents et des documents non-pertinents dans le processus de RP. La formule la plus utilisée (Formule standard), est indiquée par l'équation suivante :

$$Q_{nlle} = \alpha Q_{init} + \frac{\beta}{n_r}\sum_{\forall j \in [1,n_r]} D_j - \frac{\gamma}{n_n}\sum_{\forall i \in [1,n_n]} D_i \qquad \textit{Équation 18}$$

Où α, β et γ représentant le degré d'effet de chaque composant sur le processus de RP.

Dans le modèle probabiliste la RP est mise en place dans le modèle de mesure de pertinence. Robertson et Sparck-Jones (Robertson and Jones, 1976) ont utilisé une formule de pondération des termes fondée sur la distribution des termes de la requête dans les documents pertinents (respectivement non pertinents) qui sont jugés par l'utilisateur. Cette formule est la suivante :

$$W_i = log\,\frac{r_i/R - r_i}{n_i - r_i/(N - n_i) - (R - r_i)} \qquad \textit{Équation 19}$$

Avec :

W_i : Poids du terme dans la requête ;

r_i : Nombre de documents pertinents contenant le terme t_i ;

R : Nombre de documents pertinents pour la requête ;

n_i : Nombre de documents contenant le terme t_i ;

Et N : Nombre de documents dans la collection.

Une méthodologie de repondération a été déterminée par Croft (Croft, 1983) en utilisant une version révisée de la formule de pondération de Sparck-Jones :

- *Recherche initiale* : $W_{ijk} = (C + idf_i).f_{ik}$ *Équation 20*
- *Feedback (repondération)* : $W_{ijk} = \left[C + log\,\frac{p_{ij}(1-q_{ij})}{q_{ij}(1-p_{ij})}\right].f_{ik}$ *Équation 21*

Avec :

W_{ijk} : Le poids du terme t_i dans la requête Q_j et le document k ;

idf_i : Fréquence absolue du terme t_i dans la collection ;

p_{ij} : Probabilité que le terme t_i soit assigné à un ensemble de documents pertinents pour une requête Q_j :

$$\begin{cases} p_{ij} = \frac{r_i + 0.5}{R + 1.0} & Si \quad r_i > 0 \\ p_{ij} = 0.01 & Si \quad r_i = 0 \end{cases} \qquad \textit{Équation 22}$$

q_{ij} : Probabilité que le terme t_i apparaisse dans un ensemble de documents non pertinents pour une requête Q_j :

$$q_{ij} = \frac{n - r_i + 0.5}{N - R + 1.0} \qquad \textit{Équation 23}$$

$$f_{ik} = K + (1 - K).\frac{f_{req_{ik}}}{max\,(f_{req_k})} \qquad \textit{Équation 24}$$

Avec :

$f_{req_{ik}}$: Fréquence du terme t_i dans le document *k* ;

$max\,(f_{req_k})$: Fréquence maximale du terme t_i dans le document *k* ;

C, K : Constantes.

Dans le modèle inférentiel la réinjection de pertinence a été présentée dans (Haines and Croft, 1993). Haines et Croft ont évalué la probabilité de pertinence d'un terme en fonction de son occurrence dans les documents pertinents. Les résultats obtenus sont comparables à ceux observés dans le modèle vectoriel.

Une autre technique de recherche d'information intégrant la RP est présentée dans (Tamine, 2000). Elle s'appuie sur une combinaison d'un processus d'optimisation génétique (technique de nichage et d'adaptation des opérateurs génétiques), à un modèle de recherche de base [(8)]. La technique de nichage permet de rappeler le maximum de documents pertinents pour une même requête mais qui sont dissemblant, et l'adaptation des opérateurs génétiques permet de diminuer la complexité d'exploration en régulant les transformations génétiques opérées sur les individus requêtes. La RP est en effet exploité pour effectuer des croisements et mutations traduisant un procédé de reformulation de requête.

- **La pseudo-réinjection de pertinence**

Afin de résoudre les difficultés de la RP, cette dernière a été remplacée par la Pseudo-RP (PRP) (Cui and al. 2002). La PRP est une technique alternative à la RP (Hlaoua, 2007), nommée aussi la RP implicite ou Aveugle « Blind Relevance Feedback » car elle s'effectue de manière automatique à l'aveugle pour construire une nouvelle requête.

La pseudo-RP se base sur l'hypothèse que les documents mieux classés (les k premiers documents) sont considérés comme étant pertinents. Le SRI utilise alors les k premiers documents pour reformuler la requête (Vaidyanathan and al., 2016) (Xu and al.,2018)

[8] Une composante permet de modéliser le fond documentaire et supporter le mécanisme inhérent de sélection des documents

(Reuben and al., 2022). Le principe derrière la réinjection implicite est qu'une itération de réinjection appuyée sur les documents les plus similaires à la requête initiale de l'utilisateur pourrait donner une meilleure restitution des documents.

La pseudo-RP a été développée la première fois en 1979 dans (Croft and Harper, 1979), en tant qu'une solution afin d'estimer des probabilités dans le modèle probabiliste pour une première recherche. Depuis, plusieurs travaux ont été proposés pour améliorer les classements des documents par rapport à une première recherche (nommé les pseudo-documents pertinents « Pseudo-DPs ») (Arampatzis and al., 2021) (Khennak and Drias, 2017a) (Valcarce and al., 2019).

Parmi les études intéressantes dans ce domaine, nous trouvons le travail de Billerbeck et Zobel (Billerbeck and Zobel, 2006). Ils ont présenté une technique d'expansion de requêtes par les représentants de documents, qui permet de choisir les termes candidat d'expansion à partir des résumés des pseudo-DPs, et qui sont maintenus et créés en mémoire pour chaque document. Ces termes sont sélectionnés après une pondération des termes du document par la mesure Tf-Idf. Cette technique améliore la précision moyenne, et cela est due au fait qu'elle utilise des résumés qui sont des représentants appropriés du contenu des documents malgré leur petite taille par rapport à l'ensemble local original. Donc toutes les phases d'expansion sont minimisées en fonction du temps. Cependant, l'enregistrement des ces représentants du corpus en mémoire s'avère coûteux en matière d'espace.

Une autre technique d'expansion de requêtes par l'approche théorétique d'information est présentée dans (Imran and Sharan, 2010). Elle consiste à extraire les termes candidats d'expansion en s'appuyant sur l'information de cooccurrence et de les classer en utilisant les mesures théorétiques d'information, qui se présentent en tant que : la divergence de Kullback-Leibler (KLD) et une variante de la KLD qu'il propose sur leur travail. Le principal objectif de cet article, est de trouver une meilleure méthode pour sélectionner les termes des pseudo-DPs qui co-occurrent avec les termes de la requête. Pour cela, les auteurs proposent une technique, qui se présente comme le passage des termes des n documents les mieux ordonnés par plusieurs filtres. Chaque filtre sélectionne un nombre de termes co-occurrent à faire passer à l'étape suivante et élimine ce qui reste. Après les résultats d'expérimentations effectuées, Imran et Sharan ont déduit que les mesures théorétiques d'information appliquées sur les termes co-occurrents peuvent aider à améliorer l'efficacité de la recherche.

Dans ce travail (Ganguly and al., 2011), les auteurs ont proposé une technique d'expansion de requêtes appelée SBQE (Sentence Based Query Expansion), qui s'appuie sur

le calcul des similarités de la requête avec des phrases dans les documents les mieux classés. Cette technique est fondée sur l'extraction des phrases à partir des pseudo-documents pertinents, ainsi que leurs classements par rapport à leurs similarités aux phrases de la requête et l'ajout des plus similaires à la requête originale. Dans le but d'obtenir la similarité entre une phrase de requête et une phrase de document, les auteurs ont utilisé la mesure de similarité Cosinus. Ganguly et al., raisonnent que les documents les mieux ordonnés utilisés en tant que pseudo-DPs ont été favorisés et classés par modèles de langue, mais également les requêtes auxquelles s'applique l'approche d'expansion sont préparées pour un processus de recherche d'information par modèles de langue. Les résultats d'expérimentations effectuées sur les collections TREC montrent que, cette technique surpasse les techniques existantes de la réinjection par des modèles de langue pour 300 sujets de TREC. Mais, elle n'a pas permis de dépasser la performance de la vraie RP.

Une autre approche d'expansion de requêtes est présentée dans (Cui and al., 2003). L'idée principale des auteurs est de profiter du lien entre les termes des requêtes et les documents choisis dans les sessions de recherche des utilisateurs. Ils supposent que le document qui est tout le temps visité par les utilisateurs pour la même requête dans plusieurs sessions de recherche, les termes de ce document sont pertinents pour la reformulation d'une nouvelle requête similaire. Ils conservent une liste de tous les documents consultés pour une requête particulière. La probabilité que le document soit visité lorsqu'un mot d'interrogation est présent dans une requête est calculée pour déterminer la pertinence du document et le plus grand registre des requêtes améliore la précision de récupération du système. Aussi, sur les historiques des utilisateurs, les auteurs (Wang and Zhai, 2008) présentent une méthode d'expansion de requêtes qui permet l'extraction de motifs (patterns) à partir de ces historiques. L'utilisation de ces motifs est proposée pour décider si un terme associé à la requête doit être ou non ajouté.

Rasheed et son équipe (Rasheed and al., 2021) ont testé cinq différentes techniques hybrides afin de récupérer les documents pertinents en utilisant la méthode d'expansion de requête traditionnelle. Après, la méthode du renforcement du terme de la requête (Boosting Query Term) a été proposée pour reformuler la requête d'origine. Ces techniques sont nommées comme suit: Kullback–Leibler DivergenceBoosting Query Term (KLDBQT), Bose-Einstein1 Boosting Query Term (Bo1BQT), Information Gain Boosting Query Term (IGBQT), Condorcet Boosting Query Term (CBQT), et Borda Boosting Query Term (BBQT).

Les résultats montrent que la méthode BBQT est la meilleure par rapport aux autres méthodes en termes de la précision moyenne.

Ces approches sont largement utilisées dans les systèmes de reformulation de requêtes et apporte plusieurs avantages aux Système de recherche d'information (Saint Réquier and al, 2010). Malgré cela, cette technique a un inconvénient évident (Croft and Harper, 1979) (Xu and Croft 1996). Si une grande quantité des documents les mieux ordonnés par le SRI possèdent peu d'informations pertinentes ou aucune, alors les termes ajoutés par la réinjection pour l'élargissement de la requête sont susceptibles de causer une dégradation des performances. Ainsi, les effets de la Pseudo-RP dépendent beaucoup plus de la qualité de la récupération initiale.

III.4.4 Autres approches

D'autres approches de reformulation des requêtes sont présentées dans la littérature, et elles ne sont pas figurées dans la classification citée précédemment. Les journaux de requêtes ont été largement utilisés dans ces différentes approches de Fouille de données web «Web Mining», afin d'améliorer l'efficacité des moteurs de recherche et leurs utilisabilités (Kunpeng and al., 2009)(Baeza-Yates and al., 2005). De telles méthodes extraient des informations d'interactions à partir des fichiers logs pour améliorer de nombreuses fonctionnalités des moteurs de recherche, tels que la classification de requêtes, la reformulation de requêtes, la publicité ciblée, le classement des documents, … etc. Avec un journal contenant un nombre élevé de requêtes, il est facile de créer un représentant pour chaque document dans une collection, composé des requêtes qui ont été assez proches de ce document.

Le graphe de clic « click graph » utilisé dans (Zahera and al. 2010), est un graphe bipartite entre les requêtes et les documents cliqués (URLs). Ces articles utilisent une technique importante pour décrire les informations contenues dans les journaux de requêtes, dont les arêtes connectent une requête avec les URLs qui ont été sélectionnés par les utilisateurs comme des documents pertinents à cette requête. Les requêtes sont considérées similaires, si elles ont un plus grand nombre de documents cliqués en commun. Comme les arêtes du graphe de clique peuvent capturer quelques relations sémantiques entre les requêtes et les URLs, donc il est nécessaire de représenter les requêtes sous forme de vecteurs de documents quand on tient compte seulement des informations du graphe.

Les fichiers logs sont également exploités par Gao et Nie (Gao and Nie, 2012) pour construire des modèles de traduction statistiques entre concepts. Cette technique traite les termes en considérant également leur contexte dans les phrases pour éviter le problème

d'ambiguïté de terme. Pour cet objectif, les auteurs proposent deux manières de placer les termes dans leur contexte. D'une part, les termes contigus peuvent former une phrase, les termes à proximité peuvent fournir des contraintes contextuelles mutuelles moins strictes mais utiles également. Ils ont estimé que les relations extraites entre de tels groupes restreints de termes, considérés comme des concepts, sont censées être moins bruyantes que les relations entre les termes. Les modèles statistiques de traduction entre concepts, permettent l'expansion de requêtes en interprétant les concepts de la requête à étendre. La méthode de (Cui and al., 2002) permet d'extraire des corrélations probabilistes entre les termes de la requête et les termes des documents à l'aide des journaux de requêtes, dans le but de diminuer l'écart entre le vocabulaire de l'utilisateur et celui utilisé dans les documents du corpus. Pour chaque terme de requête tous les termes de documents qui sont corrélés sont choisis en s'appuyant sur une probabilité conditionnelle. Après, le poids de corrélation de chaque terme de document est calculé par rapport à la nouvelle requête en combinant les probabilités de corrélation avec tous les termes de la requête. Par conséquent, une liste de termes candidat d'expansion pondérés est donnée en sortie de la technique, dont les mieux classés sont rajoutés à la nouvelle requête de l'utilisateur.

III.5 Les métaheuristiques pour la reformulation des requêtes web

Cependant, lorsque l'espace de recherche devient très large, l'énumération de toutes les solutions ne serait pas possible à cause de la très longue durée que le processus de recherche prendra. Les métaheuristques sont l'une des méthodes ayant la possibilité de résoudre ce type de problème. Elles sont devenues de plus en plus intéressantes ; Ce qui fait, de nouvelles métaheuristiques sont apparues presque chaque année (Jain and al., 2019) (Molina and al., 2020).

Parmi les études intéressantes utilisant les métaheuristques, l'approche de (Singh and al., 2016). Elle permet d'utiliser le graphe des termes afin de fournir la suggestion de requête en évaluant la similarité. Le graphique est construit en fonction des mots de documents pertinents de la requête de l'utilisateur. L'association entre le graphe des termes est basée sur la similarité et considérée comme la valeur de fitness pour l'algorithme génétique (AG). Veningston et Shanmugalakshmi (Veningston and Shanmugalakshmi, 2014) ont utilisé les mêmes étapes de l'approche précédente sauf au niveau de la métaheuristique, ils ont choisi l'algorithme de colonies de fourmis (Ant Colony en englais). Une autre approche d'optimisation inspirée de la nature (Fahsi and al., 2010) se base sur les techniques de

relevance feedback par les AG, elle consiste à utiliser le "Wordnet" pour ajouter tous les termes possibles qui n'apparaissent pas dans les réactions de pertinence (feedback), et l'utilisation d'un AG afin de déterminer la meilleure requête qui correspond le mieux au besoins à l'aide d'une fonction de fitness floue. Bhatnagar et Pareek (Bhatnagar and Pareek, 2015) ont utilisé une approche nommée expansion de la requête basée sur l'algorithme génétique (Genetic Algorithm-Based Query Expansion, GABQE) pour améliorer l'efficacité de la réinjection de pseudo pertinence où les termes pertinents sont générés à partir d'une liste de documents initialement récupérées, et classés sur la base de la mesure de cooccurrence, tandis que l'AG est appliqué pour générer les combinaisons des termes. Une autre approche (Al-Khateeb and al., 2017) est basée sur l'algorithme génétique pour renvoyer des résultats plus pertinents à l'utilisateur via la reformulation de la requête. L'approche s'appuie sur WordNet pour sélectionner la bonne signification des mots-clés, le moteur de recherche Google et l'AG pour sélectionner le meilleur résultat de chaque génération. Dans (Sharma and al., 2019), les auteurs se sont basés sur Cuckoo Search algorithm et l'optimisation par les essaims de particules accélérées (PSO accélérée) pour la reformulation des requêtes. Ils ont également utilisé la technique de logique floue pour améliorer les performances de la PSO accélérée en contrôlant divers paramètres. Une autre approche dans ce domaine est proposée par (Deulkar and Narvekar, 2015) dans laquelle Deulkar et Narvekar ont prouvé la capacité de l'Algorithme Mimétique Amélioré (Improved Memetic Algorithm, IMA) pour récupérer les documents les plus pertinents pour les requêtes complexes. L'IMA est basé sur la méthode de sélection hybride où la population initiale est générée aléatoirement, ainsi que les chromosomes. Khennak et Drias ont appliqué les algorithmes d'intelligence en essaim (Swarm intelligent, SI) pour l'expansion de requête web (Khennak and Drias, 2018). Les SI utilisés sont l'algorithme FireFly (Khennak and Drias, 2016), l'algorithme inspiré des chauves-souris (Khennak and Drias, 2017b) et l'optimisation accélérée de l'essaim de particules (Khennak and Drias, 2017a). Ces algorithmes ont été utilisés pour extraire les meilleurs termes d'expansion et sélectionner les meilleurs documents pertinents à l'aide de la fonction de notation des documents (document-scoring function) Okapi BM25 et qui sont réalisés sur la base de données MEDLINE. Pour améliorer les résultats de l'expansion de la requête, la fonction de dominance de Pareto (strength Pareto dominance) a été testée comme fitness (Khennak and Drias, 2015) et qui a été exprimés par la cooccurrence et la proximité.

D'autres travaux (Bhopale and Tiwari, 2020) (Djenouri and al., 2018) (Thirugnanasambandam and al., 2021) se sont basés sur les algorithmes d'intelligence en

essaim pour résoudre le problème de recherche d'informations sur les documents (Documents Information Retrieval, DIR).

III.6 Évaluation d'un système de recherche d'information

L'objectif de la RI est de trouver des documents pertinents à une requête, et donc utiles pour l'utilisateur. L'évaluation d'un système est une phase très importante dans le domaine de la RI dans la mesure où elle permet de mesurer la performance du système en comparant les résultats obtenus avec ceux espérés (les résultats pertinents). Plus les réponses du système correspond à celles que l'utilisateur souhaite, mieux est le système.

Afin d'évaluer la performance d'un SRI à renvoyer des documents pertinents, il est nécessaire de fixer des collections de documents, des requêtes sur ces collections, ainsi que l'ensemble des documents pertinents correspondant aux requêtes. C'est ce que l'on appelle en RI des collections de tests (ou encore collections de références). On associe à ces collections un ensemble de mesures, qui vont permettre d'évaluer les résultats d'un système.

III.6.1 Collections de test

Une collection est composée d'un ensemble de documents (chaque document possédant un identifiant unique), de requêtes formulant des besoins en information et une liste de documents pertinents pour chaque requête (liste de jugements de pertinence associés). Les listes de jugements de pertinence (ou vérité terrain) sont réalisées par des utilisateurs, afin de déterminer l'ensemble des documents pertinents pour chacune des requêtes. Plusieurs collections de test ont été développées afin d'évaluer les SRI par exemple : TREC (Text REtrieval Conference) and CLEF (Cross-Language Evaluation Forum), INEX (INitiative for the Evaluation of XML Retrieval), Cranfield et MEDLINE.

III.6.2 Mesures d'évaluation

Les mesures d'évaluation permettent d'estimer quantitativement l'efficacité d'un SRI. Deux notions fondamentales sont utilisées pour qualifier cette efficacité : le silence et le bruit documentaires. Le silence documentaire fait référence aux documents pertinents non retournés par le système tandis que le bruit documentaire fait référence aux documents non-pertinents retournés par le système. Donc un bon SRI est celui qui retourne le maximum de documents pertinents (minimiser le silence) sans augmenter le nombre de documents non pertinents retournés (minimiser le bruit). La Figure III.11 montre les différentes notions utilisées lors de l'évaluation d'un SRI.

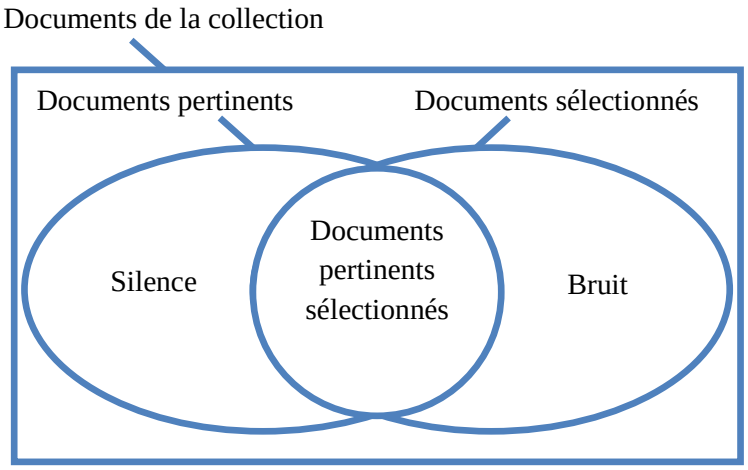


Figure III.11. Ensembles de documents utilisés pour l'évaluation d'un SRI.

Les critères de mesure des performances les plus connus sont le rappel (R) et la précision (P). Le rappel mesure la proportion des documents pertinents renvoyés par le système parmi tous ceux qui sont pertinents pour la requête, et la précision mesure la proportion des documents pertinents parmi l'ensemble de ceux renvoyés par le système.

$$\text{Rappel} = \frac{1}{Q}\sum_{q_h \in Q} \frac{|Sel_{q_h} \cap Pert_{q_h}|}{|Pert_{q_h}|} \qquad \textit{Équation 25}$$

$$\text{Précision} = \frac{1}{Q}\sum_{q_h \in Q} \frac{|Sel_{q_h} \cap Pert_{q_h}|}{|Sel_{q_h}|} \qquad \textit{Équation 26}$$

Avec :

Q : l'ensemble des requêtes $q_h \in Q$.

Sel_{q_h}: L'ensemble des documents sélectionnés par le SRI pour la requête q_h.

$Pert_{q_h}$: L'ensemble des documents pertinents pour la requête q_h.

Il existe aussi la F-mesure, qui est une mesure qui combine la précision, le rappel et leur pondération. Cette mesure permet d'évaluer la performance globale du SRI, et elle se calcule comme suit :

$$\text{F-mesure} = \frac{1}{Q}\sum_{q_h \in Q} 2.\frac{Rappel * Précision}{Rappel + Précision} \qquad \textit{Équation 27}$$

La moyenne des précisions moyennes (Mean Average Precision, MAP) correspond à la précision moyenne obtenue à chaque rang considéré par rapport aux r premiers résultats de la liste l_h retournée (qui contient R documents) pour la requête q_h :

$$\text{MAP} = \frac{1}{Q}\sum_{q_h \in Q} \frac{1}{r} \sum_{R=1}^{r} Précision(q_h)@\text{R} \qquad \textit{Équation 28}$$

III.7 Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information. La qualité d'un système de recherche d'information dépend de sa capacité de retrouver des documents pertinents dont le contenu répond aux besoins de l'utilisateur. Par conséquence, ce système fournit parfois des résultats qui ne lui conviennent pas et qui ne satisfont pas l'utilisateur. Donc, afin de s'approcher de ce dernier le plus possible à la pertinence, une étape de reformulation de la requête est souvent utilisée. Nous nous intéressons dans le cadre de cette thèse à cette étape et nous avons présenté les différentes approches existantes dans la littérature.

Nous avons remarqué que l'inconvénient des approches d'expansion et de combinaison de requêtes se présente dans le fait qu'elles sont coûteuses en matière de temps de traitement. Celui des approches de RP est le fait que les documents retournés sont fournis par les jugements de pertinence des utilisateurs. Les approches de Pseudo-RP ont des meilleure solutions pour ce problème, mais leur dépendance aux résultats de la première recherche fait leur faiblesse en cas de présence de documents non-pertinents parmi les documents les mieux classés par le système.

Nous entamons dans le chapitre suivant l'adaptation de métaheuristiques pour la reformulation des requêtes web.

Chapitre IV.

Les métaheuristiques pour la reformulation des requêtes web

IV.1 Introduction

L'objectif de la recherche d'information est donc trouver parmi les documents existants, ceux qui répondent vraiment au besoin de l'utilisateur.

En effet, quand l'espace de recherche devient très vaste, l'énumération de toutes les solutions ne serait pas faisable à cause de la très longue durée que le processus de recherche prendra. Les métaheuristiques bio-inspirées représentent l'une des méthodes ayant la possibilité de résoudre ce type de problèmes. Elles sont devenues de plus en plus intéressantes. Pour cela, les chercheurs ont commencé à s'intéresser de plus en plus à la manière dont la nature résout les problèmes, ce qui fait, de nouvelles métaheuristiques apparaissent presque chaque année. Dans ce chapitre, nous présentons d'abord les métaheuristiques utilisées dans notre approche, puis nous détaillons notre proposition d'adaptation de ces dernières pour la reformulation des requêtes web.

IV.2 Les métaheuristiques utilisées

Afin de récupérer les documents pertinents dans un délai raisonnable, nous utilisons les métaheuristiques évolutionnaires bio-inspirées suivantes : l'algorithme des lucioles (FireFly Algorithm), l'optimisation par essaims particulaires (PSO), l'algorithme de chauve-souris (Bat Algorithm) et l'algorithme de Pigeon (PIO).

IV.2.1. Algorithme de lucioles (FireFly Algorithm)

- **Inspiration**

Les lucioles sont de petits coléoptères ailés capables de produire une lumière clignotante froide pour une attraction mutuelle. Les femelles peuvent imiter les signaux lumineux d'autres espèces afin d'attirer des mâles qu'elles capturent et dévorent. Ces insectes sont capables de produire de la lumière à l'intérieur de leur corps grâce à des organes spéciaux situés très près de la surface de la peau. Les lucioles ont un mécanisme de type condensateur, qui se décharge lentement jusqu'à ce qu'un certain seuil est atteint, ils libèrent l'énergie sous forme de lumière. Le phénomène se répète de façon cyclique (Yang, 2010).

- **Algorithme**

L'Algorithme de lucioles (FireFly Algorithm, FA) introduit par Dr Xin-She Yan à l'université Cambridge en 2007 (Yang, 2008), il est inspiré par l'atténuation de la lumière sur la distance et l'attraction mutuelle mais il considère toutes les lucioles comme unisexes.

Cet algorithme s'est avéré efficace pour résoudre des multitudes problèmes d'optimisation (Yang, 2020). Il y a eu des évolutions importantes depuis son émergence il y a

une dizaine d'années (Yang and He, 2018). De nombreux travaux de recherche se sont basés sur cet algorithme pour résoudre leurs problèmes. FireFly a été utilisé pour résoudre le problème de flux hybride monoprocesseur (Dekhici and Belkadi, 2017), la planification flexible des opérations (Fuyu and al., 2018), la reconnaissance des expressions faciales (Mistry and al., 2017), la planification des soins à domicile (Dekhici and al., 2019), les soins de santé (Nayak and al., 2020), l'analyse des images (Dey and al., 2020) ainsi que pour la reformulation des requêtes web (Zeboudj and Belkadi, 2022). Certains chercheurs ont déclaré que FA est un algorithme puissant peut résoudre même certains problèmes NP-complets (Kumar and Kumar, 2020). Le pseudo-code de l'algorithme de lucioles est donné dans l'algorithme IV.1.

Algorithme IV.1. L'algorithme de lucioles

Début
γ: coefficient d'absorption de lumière
Définir la fonction objectif f(**x**)
Générer une population initiale de positions de lucioles $\boldsymbol{x_i}$ (i=1..., n)
Déterminer les intensités de lumière I_i à $\boldsymbol{x_i}$ via f(**x**i)
Tant que (t<Nbr_iter) **faire**
    Pour i = 1**à** n **faire** //toutes les lucioles
    Pour j=1 **à** n **faire** //toutes les lucioles
        si (I_j> I_i) **alors**
            Attractivité $\beta_{i,j}$ varie selon la distance $r_{i,j}$
            Déplacer luciole I vers j avec l'attractivité $\beta_{i,j}$
        sinon
            Déplacer luciole i aléatoirement
        fin si
        Evaluer la nouvelle solution
        Mettre à jour l'intensité I_i
        Vérifier si luciole i est la meilleure.
    Fin j
    Fin i
Trouver la meilleure luciole
t++
Fait
Fin

- **Paramètres**

➢ Règles de déplacement : Xin-She Yang (Yang, 2008) a pris en considération les trois règles suivantes :

– Toutes les lucioles sont unisexes, elles vont se déplacer vers d'autres plus attrayantes et plus lumineuses quelque soit leur sexe;

- L'attraction d'une luciole est proportionnelle à l'intensité des lucioles adjacentes et diminue dès que la distance entre deux lucioles augmente. Si aucune luciole n'est attrayante qu'une luciole particulière, cette dernière se déplacera de manière aléatoire ;
- L'intensité lumineuse d'une luciole représente la fonction objectif qui détermine le meilleur chemin.

➢ Attractivité : L'attraction d'une luciole est proportionnelle à l'intensité des lucioles adjacentes. La forme de cette fonction prend n'importe quelle fonction monotone décroissante, dont la forme générale est la suivante:

$$\beta_{i,j} = \beta_0^* e^{-\gamma r_{i,j}^m} \qquad \textit{Équation 29}$$

Avec β_0 qui représente l'attractivité à $r = 0$, r représente la distance entre deux lucioles et γ représente le coefficient constant d'absorption de lumière.

➢ Distance : La distance entre 2 lucioles i et j au x_i et x_j est déterminée par la distance cartésienne de la manière suivante:

$$r_{i,j} = \sqrt{\sum_{k=1}^{q}(x_{i,k} - x_{j,k})^2} \qquad \textit{Équation 30}$$

Avec $x_{i,k}$ est la $k^{ème}$ composante de la $i^{ème}$ luciole.

➢ Mouvement : Le déplacement d'une luciole par une autre luciole j plus lumineuse, est défini par :

$$x_i = (1 - \beta_{i,j})x_i + \beta_{i,j}x_j + \alpha(rand - {}^1/_2) \qquad \textit{Équation 31}$$

Avec le premier terme et le second dans l'équation sont dus à l'attraction. Tant dis que le troisième terme est la randomisation. α est le paramètre aléatoire (peut être constant) et *rand* est un générateur de nombre aléatoire uniforme distribué dans l'intervalle [0, 1].

IV.2.2. Optimisation par essaim de particules (PSO)

- **Inspiration**

Les algorithmes d'optimisation par essaim particulaire trouvent leur origine dans les modèles de simulation des mouvements de volée d'oiseaux « boid-bird-oid object-» de l'infographie Craig Reynolds à la fin des années 1980; en analysant les nuées d'oiseaux, Reynolds a proposé de les simuler à partir de trois règles simples (Reynolds, 1987) :

- Chaque individu doit éviter de heurter ses voisins ;
- Chaque individu tend à s'approcher des vitesses et des directions générales du groupe local, c.-à-d. des voisins immédiats ;
- Chaque individu cherche à s'approcher du centre de gravité du groupe local.

L'application de ces règles simples permet de construire des simulations graphiques d'un réalisme étonnant.

Les rassemblements d'animaux présentent divers avantages adaptatifs relevant au moins en partie de l'échange d'informations. De la même manière que les algorithmes évolutionnaires qui s'inspirent des mécanismes évolutifs pour mettre en œuvre des procédures d'optimisation, la PSO s'inspire des phénomènes de rassemblement et de nuée, considérant qu'en tant que processus adaptatifs, ils sont potentiellement porteurs de capacités d'optimisation. Les premiers articles concernant la PSO datent de 1995, l'appellation d'origine étant, en anglais, « Particle Swarm Optimization » (Kennedy and Eberhart, 1995).

- **Algorithme**

L'algorithme de la PSO conserve plusieurs solutions candidates dans l'espace de recherche et pendant chaque itération, chaque solution candidate est évaluée par la fonction objectif. Au départ, les particules sont positionnées dans l'espace de recherche. Puis, chaque particule est modélisée par sa position (la solution candidate) dans l'espace de recherche, par sa fitness et par sa vitesse. A chaque instant, toutes les particules ajustent leurs positions et vitesses, par rapport à leurs meilleures positions. Finalement, la PSO mémorise la meilleure fitness trouvée parmi tous les particules. Le pseudo-code de l'algorithme est donné dans l'algorithme IV.2.

Algorithme IV.2. L'algorithme d'optimisation par essaim particulaire (PSO)

Début
Initialiser la population x_i *(i=1..., n)* et les vitesses des particules
Définir la fonction objectif ***f(x)***
Répéter jusqu'à la fin d'itération.
    Répéter pour chaque particule.
        Générer la nouvelle valeur de la vitesse en utilisant l'équation 32
        Calculer la nouvelle position en utilisant l'équation 33
        Evaluation de la valeur de fitness.
        Trouver la meilleure position pour chaque particule.
    Fin
Trouver la meilleure position globale
Fin
Fin

La PSO a été utilisée avec succès dans plusieurs domaines tels que : planification (Ding and Gu, 2020), datamining (Moradi and Gholampour, 2016), Classification (Li and al., 2021), les réseaux de neurone (Liu and al., 2021), et traitement d'images (Lei and al., 2021).

- **Paramètres**

L'algorithme de la PSO consiste en trois étapes suivantes qui sont répétées jusqu'à ce que la condition d'arrêt soit vérifiée :

- Evaluer la fonction de fitness pour chaque particule ;
- Mise à jour de la meilleur fitness;
- Mise à jour de la vitesse et la position de chaque particule par l'équation 32 et 33.

$$v_i^{t+1} = \mathrm{w}\ v_i^t + c_1 \mathrm{r}_1 (best_i - x_i^t) + c_2 \mathrm{r}_2 (g_{best} - x_i^t) \qquad \textit{Équation 32}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \qquad \textit{Équation 33}$$

Avec : v_i la vitesse de particule *i*;

x_i la position de la particule *i*;

w, c_1 et c_2 des coefficients constants (*w* appelé facteur d'inertie, c_1 et c_2 appelés coefficients d'accélération);

r_1 et r_2 sont des nombres aléatoires triés à chaque itération ;

g_{best} est la meilleure solution trouvée ;

$best_i$ est la meilleure solution trouvée par le particule *i*.

Le facteur d'inertie *w* contrôle l'influence de l'ancienne vitesse sur la vitesse courante, afin de permettre aux particules d'éviter les minima locaux. Le paramètre c_1 contrôle le comportement de la particule dans sa recherche autour de sa meilleure position et c_2 contrôle l'influence de l'essaim sur le comportement de la particule.

IV.2.3. L'algorithme de chauve-souris (Bat Algorithm)

- **Inspiration**

Les chauves-souris sont des animaux fascinants qui peuvent trouver leurs proies et discriminer différents types d'insectes, même dans l'obscurité complète, et elles ont la capacité de l'écholocation. L'écholocalisation a été définie par un sonar biologique qui permet de détecter la distance à travers les impulsions.

- **Algorithme**

L'algorithme de chauve-souris (Bat Algorithm, BA) a été proposé la première fois par Yang en 2010 (Yang, 2010). BA est inspiré par une idéalisation de l'écholocation en prenant les règles suivantes:

- Chaque chauve-souris vole de manière aléatoire à la position x_i avec une vitesse v_i, une fréquence variante ou une longueur d'onde f_i et une intensité A_i. Comme elle cherche et trouve sa proie, elle change de fréquence, taux d'émission d'impulsions r et l'intensité.
- La recherche est amplifiée par un vol aléatoire local. La sélection des meilleures continue jusqu'à ce que des critères d'arrêt soient satisfaits.

L'algorithme de chauve-souris a été appliqué avec succès dans divers domaine de recherche tel que : la planification (Osaba and al., 2019), la segmentation d'image (Yue and Zhang, 2020), et les réseaux de neurones (Gajic and al., 2021). Le pseudo-code de l'algorithme est donné dans l'algorithme IV.3.

Algorithme IV.3. L'algorithme de chauve-souris

Début
Définir la fonction objectif $f(x_i)$
Initialiser la population de chacune des chauves-souris x_i $(i=1..., n)$ et leur vitesse v_i
Initialiser les fréquences d'impulsion f_i, l'intensité A_i et l'impulsions r_i
Calculer la fréquence de la position
Tant que (t < Nbr_iter) **faire**
  Générer de nouvelles solutions par l'ajustement des fréquences selon Equation 34,
  Mettre à jour la vitesse selon Equation 35
  Générer une nouvelle chauve souris x_i selon Equation 36
  Pour i=1 à Nbr_max **faire**
    si ($rand > r_i$)
      Générer une solution locale autour de la meilleure solution selon Equation 37
    fin si
    Générer une nouvelle solution en volant aléatoirement selon Equation 38
    si ($rand < A_i$ and $f(x_i) < f(g_{best})$)
      Accepter les nouvelles solutions
      Mettre à jour les intensités A_i et les fréquences de pulsation r_i selon Equation 39 et 40
    fin si
  Classer les chauves-souris et trouver la meilleure solution courante
  Fait
  t++
Fait
Fin

- **Paramètres**

➢ Fréquence : La fréquence f_i de l'écholocalisation émise par la chauve-souris *i* est un nombre entier ou un nombre réel, générée de manière uniforme en fonction des fréquences sélectionnées minimale f_{min} et maximale f_{max}.

$$f_i=f_{min}+(f_{max}-f_{min})\beta \qquad \textit{Équation 34}$$

Avec β un nombre aléatoire de l'intervalle [0,1].

➢ La vélocité : La vélocité suggère combien de chauve-souris devraient être changées à un certain moment. Les chauves-souris communiquent entre elles grâce à la meilleure location (solution) g_{best} globale.

$$v_i^{t+1}= v_i^t +(x_i^t-g_{best})f_i \qquad \textit{Équation 35}$$

➢ Position : la nouvelle position est générée selon l'équation suivante :

$$x_i^{t+1}= x_i^t + v_i^{t+1} \qquad \textit{Équation 36}$$

Elle peut être autour de la meilleure solution globale en utilisant l'intensité :

$$x_i=g_{best} + \varepsilon.A_{moy} \qquad \textit{Équation 37}$$

Ou de manière aléatoire :

$$x_i=best_i + \varepsilon.A_{moy} \qquad \textit{Équation 38}$$

Avec : $best_i$ la meilleure solution globale actuelle;
ε un nombre aléatoire de l'intervalle [0,1] ;
A_{moy} l'intensité moyenne des chauves-souris.

➢ L'intensité : définit le volume sonore moyen de toutes les chauves-souris. Elle se calcule par l'équation suivante :

$$A_i^{t+1}=\alpha\, A_i^t \qquad \textit{Équation 39}$$

Avec α un nombre aléatoire de l'intervalle [0,1].

➢ Fréquence des pulsations : r_i la valeur du taux de pulsation permet de choisir la bonne décision dans la partie de recherche locale autour de la meilleure solution globale trouvée, selon l'équation suivante :

$$r_i^{t+1}= r_i^0(1-e^{(-\gamma t)}) \qquad \textit{Équation 40}$$

Avec r_i^0 le taux initial da la pulsation et $\gamma>0$.

IV.2.4. Algorithme de Pigeon (PIO)

- **Inspiration**

Le pigeon est un oiseau populaire dans le monde. Il était considéré comme un outil de communication important, en temps de paix et de guerre. Pendant la première et la seconde Guerre mondiale, les pigeons ont joué un rôle crucial dans l'ensemencement des renseignements militaires en raison de leur dissimulation et de leur véracité.

Des recherches approfondies (Guilford and al., 2004) ont été menées sur des pigeons où ils ont découvert que les pigeons peuvent facilement trouver leur destination, en utilisant trois méthodes: le champ magnétique, le soleil et les points de repère.

- **Algorithme**

Inspirés par ces pigeons voyageurs, Duan et Qiao (Duan and Qiao, 2014) ont présenté l'algorithme de pigeon (Pigeon-inspired optimization, PIO), qui a été appliqué à la planification de trajectoire de robot aérien et la détection de cible.

Le PIO a été utilisé dans différents domaines de recherche comme : l'optimisation (Cui and al., 2019) (Asaju and al., 2021), la sélection des caractéristiques dans un système de détection d'intrusion (Alazzam and al., 2020) et la classification des images satellitaire (Benyettou and Fizazi, 2020). Le pseudo-code de l'algorithme de pigeon est donné dans l'algorithme IV.4.

Algorithme IV.4. L'algorithme de pigeon (PIO)

Début

1. Initialisation des paramètres

Initialiser les paramètres :
- Np : nombre des pigeons
- D : dimension de l'espace de recherche
- R : le facteur de carte et de boussole
- Espace de recherche : les limites de l'espace de recherche
- Nc_{1max} : le nombre maximum de générations quand l'opération carte et boussole est effectuée
- Nc_{2max} : le nombre maximum de générations quand l'opération de repère est effectuée.

Initialiser les positions x_i et les vitesses v_i pour chaque pigeon
Définir $x_p = x_i$, $N_c = 1$
Calculer les valeurs de fitness de différents pigeons
$x_g = min\ [f(x_p)]$

2. Map et opérateur de boussole

pour N_c=1 à Nc_{1max} **faire**
  pour i=1 à N_p **faire**
    Tant que x_i appartient à l'espace de recherche **faire**
    Calculer x_i et v_i selon Equations 41 et 42

Fait
fin pour
Evaluer x_i, et mettre à jours x_p et g_{best}
fin pour

3. Opérateur de repère
pour $N_c = Nc_{1max} + 1$ à Nc_{2max} **faire**
Tant que x_p appartient à l'espace de recherche **faire**
Classer tous les pigeons selon leurs valeurs de fitness
$N_P = N_P/2$
Garder la moitié des individus avec de meilleures valeurs de fitness
x_c= valeur moyenne des pigeons restants
Calculer x_i selon Equation 45
Fait
Evaluer x_i, et mettre à jours x_p et g_{best}
fin pour
4. Sortie
Trouver le meilleur pigeon
Fin

- **Paramètres**

Le PIO de base comprend deux opérateurs: Map et opérateur de boussole (Map and compass operator), et opérateur de repère (Landmark operator). Le premier opérateur est basé sur le champ magnétique et le soleil, tandis que le deuxième opérateur est basé sur les repères.

- Map et opérateur de boussole :

Dans un espace de recherche de dimension D, chaque pigeon *i* a une position x_i et une vitesse v_i, qui sont mises à jour à chaque itération. La nouvelle position x_i et vitesse v_i du pigeon seront calculées itérativement en utilisant les équations suivantes :

$$v_i^{t+1} = v_i^t . e^{-R\,t} + rand.(g_{best} - x_i^t) \qquad \textit{Équation 41}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \qquad \textit{Équation 42}$$

Avec : *R* le facteur de carte et de boussole;

rand est un nombre aléatoire généré à partir d'une distribution uniforme dans [0,1] ;

g_{best} est la meilleure position globale actuelle, et qui peut être calculée en comparant toutes les positions.

- opérateur de repère :

Dans cet opérateur, la qualité de l'individu pigeon est exprimée par la fonction objectif. N_p désigne la moitié du nombre de pigeons à chaque génération, et x_c est le centre des pigeons à la $t^{ième}$ itération, et qui sont calculés par les équations suivantes :

$$N_p^{t+1} = \frac{N_p^t}{2} \qquad \textit{Équation 43}$$

$$x_c^{t+1} = \frac{\sum x_i^{t+1}.\ fitness\,(x_i^{t+1})}{N_p(\sum fitness\,(x_i^{t+1}.)} \qquad \textit{Équation 44}$$

$$x_i^{t+1} = x_i^t + rand.(x_c^{t+1} - x_i^t) \qquad \textit{Équation 45}$$

IV.3 Les métaheuristiques pour la reformulation des requêtes web

Dans cette section, nous présentons notre propre modélisation de reformulation des requêtes web en se basant sur les métaheuristiques. En particulier, nous décrivons la représentation de la solution, l'espace de recherche, la fonction objectif (fitness) et les différents paramètres.

IV.3.1. Représentation de la solution

Au cours de la reformulation de la requête, la décision centrale est la détermination du schéma de représentation approprié pour les termes de reformulation candidats (solution). Le choix de l'encodage pour ces solutions candidates est l'un des enjeux les plus importants dans la conception d'un algorithme basé sur les métaheuristiques. La solution pour notre problème est une requête représentée par un vecteur de termes (mots-clés), qui est définie comme :

$$Q_F = < t_1, t_2, \dots t_n > \qquad \textit{Équation 46}$$

avec $1 \leq n \leq 5$;

IV.3.2. Initialisation de la population

Les métaheuristiques sont des algorithmes à base de population. La taille de cette population détermine le nombre de solutions ou la taille de l'espace de recherche (*R*) dont le but de diriger la recherche vers le meilleur emplacement (meilleure requête). Dans notre cas, l'espace de recherche contient des termes liés à l'information (items fréquents générés par FP_Growth) ainsi que leurs fréquences. Pour *p* items, on a 2^p itemsets possibles (cette section sera développée dans le chapitre V).

IV.3.3. Fonction objectif

Chaque solution dans l'espace de recherche est associée à une valeur numérique. La qualité d'une solution est proportionnelle à la valeur de la fonction objectif. Dans notre cas, la solution est une combinaison de termes (mots clés) ; la meilleure façon d'évaluer ses performances est de prendre en considération la fréquence de ses termes appartenant à l'espace de recherche, et la meilleure fréquence est alors considérée comme la valeur de fitness de la requête reformulée.

Pour cela, OkapiBM25 a été choisi comme fitness qui est une méthode de pondération des termes dans les documents et les requêtes selon le modèle de pertinence probabiliste développé par Robertson et Sparck Jones dans le système d'information « Okapi » de l'Université de Londres (Robertson and Walker, 1994). Par conséquent, pour l'ensemble des requêtes Q'_l,, l'équation 47 est utilisée pour mesurer sa fitness.

$$w_i^{BM25}\,(Q'_l,\, t_n,\, D_k) = log\,\frac{K-n_l+0{,}5}{n_l+0{,}5} \times \left[\sum_{t\in Q'_l} \frac{(k_1+1)tf}{k_1\left(1+\frac{b(dl-avgl)}{avgl}\right)+tf}\right] \qquad \textit{Équation 47}$$

Avec : *K* le nombre de documents dans la collection;

n_l le nombre de documents contenant t_n ;

tf la fréquence du terme t_n dans chaque document D_k;

$k_1\ et\ b$ des constantes;

dl la longueur du document;

avgl la longueur moyenne du document;

La meilleure requête reformulée Q_b est présentée comme une particule avec une fréquence maximale:

$$f(\,Q_F) = max\,(w_i^{BM25}\,(Q_i^{'},\, t_n,\, D_k)) \qquad \textit{Équation 48}$$

IV.3.4. Adaptation de l'algorithme de lucioles (FireFly Algorithm)

Chaque luciole *i* dans la population est un vecteur de mots-clés ($Q_i^{'} = < t'_1, t'_2, \ldots t'_n >$) où I est le nombre des items fréquents contenant au plus 5 termes choisis au hasard dans l'ensemble des chemins obtenus par FP_Growth. Par conséquent, la position du luciole est représentée par $x_{i,n} = \{w_{i,1},\ w_{i,2,}\ \ldots\ w_{i,n}\}$, où $w_{i,n}$ est le poids de la requête *i* (ces mots-clés) dans l'espace de recherche.

Après la comparaison de la luminosité de chaque luciole Q'_i, avec toutes les autres lucioles, les positions des lucioles sont mises à jour en fonction des règles de déplacement et de leurs voisines. Ces règles sont généralement la position initiale, la distance entre deux lucioles et un mouvement aléatoire.

Pour adapter l'algorithme des lucioles au problème de reformulation des requêtes, quelques modifications ont été apportées à ces règles. Au niveau de l'attractivité (Equation 29), la distance de Manhattan a été utilisée pour trouver la distance entre deux lucioles (requêtes) Q'_1 et Q'_2. Le choix de cette distance est très important car une fois que sa valeur décroit, l'attractivité augmente.

$$ManhattanDist = \sum_{n=1}^{I} |x_{i,n} - x_{j,n}| \qquad \text{Équation 49}$$

Avec $x_{i,n}$ et $x_{j,n}$ sont les pondérations des termes *n* dans Q'_i et Q'_j respectivement.

Un exemple qui calcule cette distance entre $Q'_1 = < t'_1, t'_2, t'_4, t'_5 >$ et $Q'_2 = < t'_3, t'_4, t'_5 >$ est le suivant :

	t'_1,	t'_2,	t'_3,	t'_4	t'_5
Q'_1	1	1	0	1	1
Q'_2	0	0	1	1	1

Tableau IV. 1. Les fréquences des termes de Q'_1et Q'_2.

$ManhattanDist(Q'_1, Q'_2)$= |1-0|+|1-0|+|0-1|+|1-1|+|1-1|=3.

Lors de la boucle de comparaison deux à deux, la meilleure solution Q_b est mise à jour itérativement. Le processus de comparaison se répète jusqu'à ce que la condition d'arrêt soit satisfaite. Cette condition correspond dans notre cas au nombre d'itérations fixé au démarrage ou lorsque les résultats de l'algorithme deviennent stables, c'est-à-dire que la requête ne change pas pour éviter la perte du temps.

Les principales étapes de l'algorithme FireFly proposé pour la reformulation des requêtes sont présentées dans l'algorithme IV.5.

Algorithme IV.5. L'algorithme de lucioles proposé

Début
Définir le coefficient d'absorption de lumière γ et le paramètre α
Générer une population initiale des lucioles Q_i'
Déterminer les intensités de lumière selon l'*Equation 48*
Tant que (t < au nombre maximum d'itérations et la requête reformulée change) **faire**
  Pour i = 1 **à** N **faire** //toutes les lucioles
  Pour j=1 **à** N **faire** //toutes les lucioles
    si ($f(Q_i') < f(Q_j')$)**alors**
      Attractivité $\beta_{i,j}$ varie selon la distance $r_{i,j}$ (*Equation 49*)
      Déplacer requête Q_i'vers Q_j'selon *Equation 29*
    sinon
      Déplacer luciole i aléatoirement
    fin si
    Evaluer la nouvelle solution
    Mettre à jour l'intensité I_i
    Vérifier si la requête Q_i' est la meilleure requête reformulée Q_b
  Fin j
  Fin i
Trouver la meilleure requête
t++
Fait
Fin

IV.3.5. Adaptation de l'algorithme d'optimisation par essaim de particules (PSO)

Le principe de la PSO est de commencer par une population de solutions aléatoires représentées par des requêtes Q_i' (Équation 50) et faire la recherche pour un optimum (la meilleure requête) par la mise à jour des générations et fouiller l'espace des solutions en suivant les optimums actuels jusqu'à la satisfaction des critères d'arrêt qui correspondent dans notre cas au nombre d'itérations fixé au démarrage ou lorsque la requête ne change pas.

Afin d'adapter la PSO au problème de reformulation des requêtes, quelques modifications sont apportées. Pour modéliser l'essaim, à l'itération *t*, chaque particule est associée à une position $Q_i^{'t}$ et une vitesse v_i^t. La position de chaque solution (requête) est évaluée en calculant sa valeur de fitness en utilisant l'équation 48 et la particule avec la valeur de fitness la plus élevée est considérée comme la meilleure solution globale Q_b. En utilisant sa nouvelle vitesse v_i^{t+1}, la position $Q_i^{'t}$ de chaque particule *i* est mise à jour selon l'équation suivante:

$$Q_i'^{t+1} = Q_i'^{t} + v_i^{t+1} \qquad \textit{Équation 50}$$

L'équation se formule par l'addition de deux vecteurs contenant des mots clé ; La valeur du k^{ème} élément du vecteur de position $Q_i^{'t}$ est un terme (mot-clé) tandis que la valeur du k^{ème} élément du vecteur vitesse peut être 0 ou bien un terme. Le résultat est déterminé par un vecteur dans lequel chaque k^{ème} élément obtient la valeur de $Q_{i,k}^{'t}$ si $v_{i,k}^{'t+1}$ de $v_i^{'t+1}$ est égale à 0, sinon le k^{ème} élément de la nouvelle position $Q_i^{'t+1}$ est mis à $v_{i,k}^{'t+1}$. La formule d'ajustement de position est définie par l'équation suivante :

$$.Q_{i,k}^{'t+1} = \begin{cases} Q_{i,k}^{'t}, & si\ \ v_{i,k}^{'t+1} = 0 \\ v_{i,k}^{'t+1}, & sinon \end{cases} \qquad \textit{Équation 51}$$

Chaque particule se déplace avec une vitesse variable vers sa propre meilleure position *best$_i$* dans le passé et la meilleure position globale Q_b selon l'équation 32. Pour calculer cette vitesse, nous devons d'abord effectuer deux différentes opérations de soustraction. La première opération permet de calculer la différence entre la meilleure position *best$_i$* dans le passé et la position actuelle Q_i' . L'équation suivante illustre la manière dont cette opération est appliquée :

$$best_{i,k} - Q'^{t}_{i,k} = \begin{cases} 0, & si\ best_{i,k} = Q'^{t}_{i,k} \\ best_{i,k}, & sinon \end{cases} \qquad \text{Équation 52}$$

Et $$c_1 r_1 (best_{i,k} - Q'^{t}_{i,k}) = \begin{cases} 0, & si\ c_1 > r_1 \\ best_{i,k} - Q'^{t}_{i,k}, & sinon \end{cases} \qquad \text{Équation 53}$$

La deuxième opération permet de calculer la différence entre la meilleure position globale Q_b et la position actuelle Q'_i. L'équation suivante illustre la manière dont cette opération est appliquée :

$$Q_{b,k} - Q'^{t}_{i,k} = \begin{cases} 0, & si\ Q_{b,k} = Q'^{t}_{i,k} \\ Q_{b,k}, & sinon \end{cases} \qquad \text{Équation 54}$$

Et $$c_2 r_2 (Q_{b,k} - Q'^{t}_{i,k}) = \begin{cases} 0, & si\ c_2 > r_2 \\ Q_{b,k} - Q'^{t}_{i,k}, & sinon \end{cases} \qquad \text{Équation 55}$$

La nouvelle vitesse est calculée par l'équation suivante :

$$v_I^{t+1} = w\, v_I^t + c_1 r_1 (best_i - Q'^{t}_i) + c_2 r_2 (Q_b - Q'^{t}_i) \qquad \text{Équation 56}$$

Avec $$w\, v^t_{i,k} = \begin{cases} 0, & si\ w = 0 \\ v'^{t}_{i,k}, & sinon \end{cases} \qquad \text{Équation 57}$$

Les principales étapes de l'algorithme d'essaim particulaire proposé pour la reformulation de la requête sont résumées dans l'algorithme IV.6.

Algorithme IV.6. L'algorithme d'optimisation par essaim particulaire

Début
Définir la fonction objectif *f(* Q_F)
Initialiser la population Q'_i et les vitesses v_i des particules
Tant que (t < au nombre maximum d'itérations et la requête reformulée change) **faire**
  Pour i=1 à Nbr_max **faire**
    Générer la nouvelle valeur de la vitesse en utilisant l'équation 56
    Calculer la nouvelle position en utilisant l'équation 50
    Evaluation de la valeur de fitness.
    Trouver la meilleure position pour chaque requête.
  Fin
Trouver la meilleure requête reformulée Q_b
t++
Fait
Fin

IV.3.6. Adaptation de l'algorithme des chauves-souris (Bat)

Le mécanisme d'optimisation par cet algorithme démarre par une population des chauves souris représentée par des requêtes Q_i' distribuée aléatoirement dans l'espace de recherche R.

Pour trouver l'optimum d'une fonction objectif, chaque chauve souris représente une solution (Équation 46), on évalue la fitness des requêtes à chaque itération en se basant sur l'équation 48, et on change leur position en utilisant leurs fréquences f_i, vitesses v_i (initialement v_i=0), la meilleure position courante de chaque chauve souris, et la meilleure solution courante trouvée $best_i$ par toute la population. La meilleure solution Q_b est maintenue après la vérification des critères d'arrêt qui correspondent dans notre cas au nombre d'itérations fixé au démarrage ou lorsque la requête ne change pas.

A l'itération t, chaque solution Q_i' dans la population a sa position $Q_i'^t$ et sa vitesse v_i. La valeur du k^{ème} élément du vecteur de position $Q_i'^t$ est un terme et celle du vecteur vitesse peut être 0 ou bien un terme. En se basant sur l'Equation 36, la solution Q_i' se déplace de la position courante $Q_i'^t$ vers la position suivante $Q_i'^{t+1}$ en utilisant sa nouvelle vitesse $v_i'^{t+1}$.

Pour adapter l'algorithme des chauves souris au problème de reformulation des requêtes, le mouvement vers la position suivante s'exprime par l'addition de deux vecteurs contenant des mots clé (Equation 36). Le résultat est déterminé par le vecteur suivant :

$$.Q_{i,k}'^{t+1} = \begin{cases} Q_{i,k}'^{t}, & si\ \ v_{i,k}'^{t+1} = 0 \\ v_{i,k}'^{t+1}, & sinon \end{cases} \qquad \textit{Équation 58}$$

Pour calculer la vitesse (Equation 35), nous devons d'abord calculer la différence entre la position actuelle Q_i' et la meilleure solution globale Q_b. Cette différence est effectuée par la soustraction de deux positions, et son résultat est représenté par un vecteur Q_F des éléments (termes) dans lesquels chaque élément k prend la valeur de $Q_{b,k}$ si $Q_{i,k}'$ est différent de $Q_{b,k}$; Sinon la valeur du k^{ième} élément est fixée à 0. L'équation suivante illustre la manière dont l'opération de soustraction entre Q_I' et Q_b est appliquée :

$$.Q_{i,k}'^{t} - Q_{b,k} = \begin{cases} 0, & si\ \ Q_{i,k}'^{t} = Q_{b,k} \\ Q_{b,k}, & sinon \end{cases} \qquad \textit{Équation 59}$$

La nouvelle vitesse est calculée par l'équation suivante :

$$v_i'^{t+1} = v_i'^{t} + (Q_i'^{t} - Q_b)f_i \qquad \textit{Équation 60}$$

Avec

$$(Q_i'^{t} - Q_b)f_i = \begin{cases} 0, & si\ rand > f_i \\ Q_i'^{t} - Q_b, & sinon \end{cases} \qquad \textit{Équation 61}$$

Les principales étapes de l'algorithme des chauves-souris proposé pour la reformulation de la requête sont résumées dans l'algorithme IV.7.

Algorithme IV.7. L'algorithme des chauves-souris proposé

Début
Définir la fonction objectif *f(* Q_F)
Initialiser la population de chacune des chauves-souris Q_i' et leur vitesse v_i
Initialiser les fréquences d'impulsion f_i, l'intensité A_i et l'impulsion r_i
Calculer la fréquence de la position
Tant que (t< au nombre maximum d'itérations et la requête reformulée change) **faire**
Générer de nouvelles solutions par l'ajustement des fréquences selon Equation 34,
Mettre à jour la vitesse selon Equation 61
Générer la nouvelle requête selon Equation 58
Pour i=1 à Nbr_max **faire**
si (*rand* > r_i)
Générer une solution locale autour de la meilleure solution selon Equation 37
fin si
Générer une nouvelle solution aléatoirement selon Equation 38
si (*rand* < A_i et f(Q_i') < f(Q_b))
Accepter les nouvelles solutions
fin si
Classer les requêtes et trouver la meilleure requête reformulée Q_b
Fait
t++
Fait
Fin

IV.3.7. Adaptation de l'algorithme des Pigeons (PIO)

L'algorithme du pigeon se compose de deux opérateurs: Map et opérateur de boussole (Map and compass operator), et opérateur de repère (Landmark operator).

Afin d'adapter le PIO au problème de reformulation des requêtes, quelques modifications sont apportées à ces deux opérateurs. Dans le premier opérateur, chaque pigeon *i* est associé à une position $Q_i'^t$ et une vitesse v_i^t. La position de chaque solution est mise à jour selon l'équation suivante:

$$Q_i'^{t+1} = Q_i'^t + v_i^{t+1} \qquad \textit{Équation 62}$$

Les valeurs des vecteurs Q_i' et v_i sont des mots clé (ou bien 0 pour v_i). Le résultat de l'équation 62 est un vecteur déterminé par la formule suivante :

$$.Q_{i,k}^{'t+1} = \begin{cases} Q_{i,k}^{'t}, & si\ \ v_{i,k}^{'t+1} = 0 \\ v_{i,k}^{'t+1}, & sinon \end{cases} \qquad \textit{Équation 63}$$

Chaque pigeon se déplace avec une vitesse variable vers la meilleure position globale Q_b selon l'équation suivante :

$$v_i^{t+1} = v_i^t . e^{-R\,t} + rand.(Q_b - Q_{i,k}^{'t}) \qquad \textit{Équation 64}$$

Où *R* est le facteur de carte et de boussole;

Pour calculer cette vitesse, nous devons d'abord calculer la différence entre la meilleure position globale Q_b et la position actuelle Q_i' . L'équation suivante montre la façon dont cette opération est effectuée :

$$Q_{b,k} - Q_{i,k}^{'t} = \begin{cases} 0, & si\ \ Q_{b,k} = Q_{i,k}^{'t} \\ Q_{b,k}, & sinon \end{cases} \qquad \textit{Équation 65}$$

Et $$rand.(Q_{b,k} - Q_{i,k}^{'t}) = \begin{cases} 0, & si\ \ rand = 0 \\ Q_{b,k} - Q_{i,k}^{'t}, & sinon \end{cases} \qquad \textit{Équation 66}$$

La première partie de l'équation 64 se base sur le facteur de carte et de boussole dont R ∈ [0, 1]. Il est important de noter que ce facteur est une valeur constante qui joue un rôle important dans la convergence de l'algorithme. Plus R devient petit, plus la valeur de $e^{-R\,t}$ devient grande. Par conséquent, les pigeons héritent d'une plus grande vitesse, ce qui favorise une convergence rapide et une meilleure capacité de recherche globale ; à l'inverse, un grand R ne conduirait qu'à une recherche locale, pour cela :

$$v_i^t . e^{-R\,t} = \begin{cases} 0, & si\ \ R < rand \\ v_{i,k}^{'t}, & sinon \end{cases} \qquad \textit{Équation 67}$$

Dans le deuxième opérateur (opérateur de repère), la moitié du nombre de pigeons diminue de Np à chaque génération et leur qualité est exprimée par la fonction objectif (Equation 48). Cependant, les pigeons sont encore loin de leur destination et ne connaissent pas les repères. La règle de mise à jour de la position du pigeon *i* à la t^{ième} itération peut être donnée par :

$$C_i^{'t} = \frac{\sum w_i^t .\ fitness\,(Q_i^{'t})}{N_p(\sum fitness\,(Q_i^{'t})} \qquad \textit{Équation 68}$$

Avec w_i est le poids de la requête *i* dans l'espace de recherche et N_p est la moitié du nombre de pigeons à la t^{ième} itération.

Le centre des pigeons est représenté sous forme d'une requête (un vecteur de terme) Q_c, dont lequel leur poids est égale à la valeur de $C_i^{'t}$ ou bien proche de cette valeur, et donc :

$$Q_{i,k}^{'t+1} = Q_{i,k}^{'t} + rand.(Q_{c,k}^{t+1} - Q_{i,k}^{'t}) \quad \textit{Équation 69}$$

Avec $$Q_{c,k}^{t+1} - Q_{i,k}^{'t} = \begin{cases} 0, & si \ \ Q_{c,k}^{t+1} = Q_{i,k}^{'t} \\ Q_{c,k}^{t+1}, & sinon \end{cases} \quad \textit{Équation 70}$$

Et $$rand.(Q_{c,k}^{t+1} - Q_{i,k}^{'t}) = \begin{cases} 0, & si \ rand = 0 \\ Q_{c,k}^{t+1} - Q_{i,k}^{'t}, & sinon \end{cases} \quad \textit{Équation 71}$$

Les principales étapes de l'algorithme des pigeons proposé pour la reformulation de la requête sont résumées dans l'algorithme IV.8.

Algorithme IV.8. L'algorithme de pigeon proposé

Début

1. Initialisation des paramètres

Initialiser les paramètres :
- Np : nombre de pigeons
- D : dimension de l'espace de recherche (D=1)
- R : facteur de carte et de boussole
- Nc_{1max} : le nombre maximum de générations pour que l'opération carte et boussole soit effectuée
- Nc_{2max} : le nombre maximum de générations pour que l'opération de repère soit effectuée.

Initialiser les positions Q_i' et les vitesses v_i pour chaque pigeon
Définir $best_i = Q_i'$, $N_c = 1$
Calculer les valeurs de fitness des différents pigeons
$Q_b = max\ [f(best_i)]$

2. Map et opérateur de boussole

Tant que (t < au nombre maximum d'itérations et la requête reformulée change)
faire
pour N_c=1 à Nc_{1max} **faire**
pour i=1 à N_p **faire**
Tant que Q_i' appartient à l'espace de recherche **faire**
Calculer Q_i' et v_i selon Equations 62 et 64
Fait
fin pour
Evaluer Q_i' , et mettre à jour $best_i$ et Q_b
fin pour

3. Opérateur de repère

pour $N_c = Nc_{1max} + 1$ à Nc_{2max} **faire**
Tant que x_p appartient à l'espace de recherche **faire**
Classer tous les pigeons selon leurs valeurs de fitness
$N_P = N_P/2$
Calculer Q_i' selon Equation 69
Fait

Evaluer Q_i' , et mettre à jour $best_i$ et Q_b
fin pour
t++
Fait

4. Sortie
Trouver la meilleure requête reformulée Q_b
Fin

IV.4 Conclusion

Dans ce chapitre, nous avons proposé de considérer le problème de reformulation des requêtes comme un problème d'optimisation et de l'aborder avec les métaheuristiques. La procédure globale consiste, dans un premier temps, à sélectionner les documents pseudo-pertinents à la requête initiale, à extraire leurs mots-clés pour construire des solutions potentielles et enfin à lancer les métaheuristiques pour déterminer la meilleure requête reformulée. Les étapes de l'approche proposée seront détaillées dans le chapitre suivant.

Chapitre V.

Modélisation et conception

V.1 Introduction

Comme nous l'avons déjà mentionnai, l'objectif de notre travail est l'amélioration de l'accessibilité du web aux personnes ayant une déficience visuelle, d'une part en contribuant à compenser les problèmes liés directement à la non-conformité des pages web et d'autre part en optimisant la recherche pour ces personnes. Ce chapitre décrit les aspects de conception et de modélisation de notre proposition afin de mieux répondre à l'objectif.

V.2 Approche proposée

Après avoir traité la partie théorique, nous voulons fournir une approche pour les utilisateurs ayant une déficience visuelle afin de faciliter leurs tâches de navigation tout en les optimisant. L'approche proposée comprend principalement deux étapes: la première consiste à l'optimisation de la recherche d'information par la reformulation de la requête web en se basant sur les métaheuristiques, tandis que la deuxième étape consiste à l'adaptation des pages web selon les préférences de ces utilisateurs. La Figure V.1 illustre les deux principales étapes du l'approche proposée, et la Figure V.2 montre les détails de ces étapes.

Figure V.1. Les principales étapes de l'approche proposée.

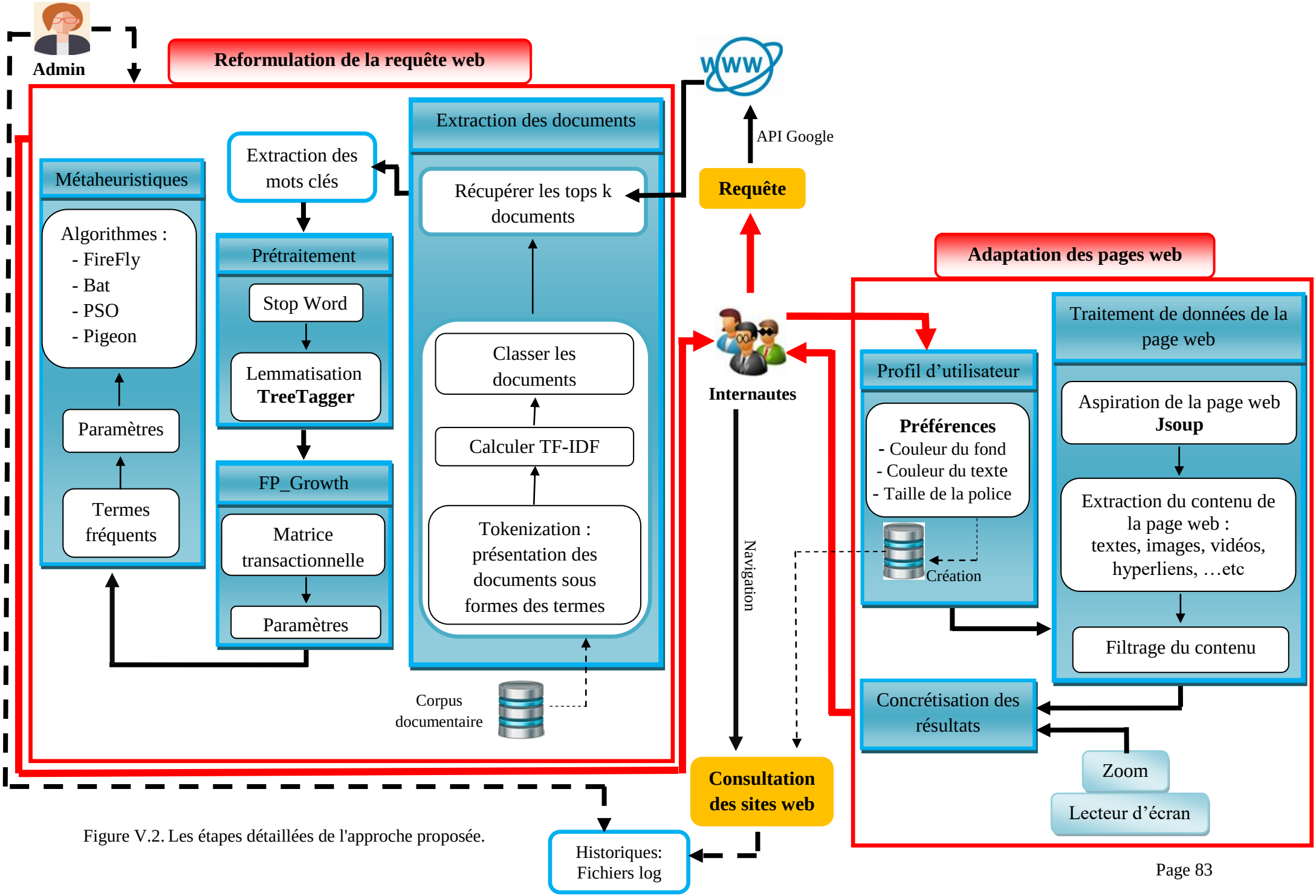


Figure V.2. Les étapes détaillées de l'approche proposée.

V.3 Reformulation de requête web

Dans un système de recherche d'information, mettre en correspondance le besoin d'information de l'utilisateur est l'un des problèmes majeurs. Une approche de reformulation de la requête est donc souvent nécessaire, dans laquelle nous nous intéressons aux métaheuristques pour mieux exprimer ce besoin dans le contexte web.

Notre objectif est de trouver une requête efficace dans un laps de temps considérablement plus court. Une bonne requête est celle qui contient des mots pertinents relativement au besoin d'information, alors qu'une requête efficace est celle qui obtient de bons résultats (documents ou liens qui sont pertinents selon l'utilisateur) selon les métriques d'évaluation. Le diagramme schématique de la reformulation de requête proposée est représenté dans la Figure V.3.

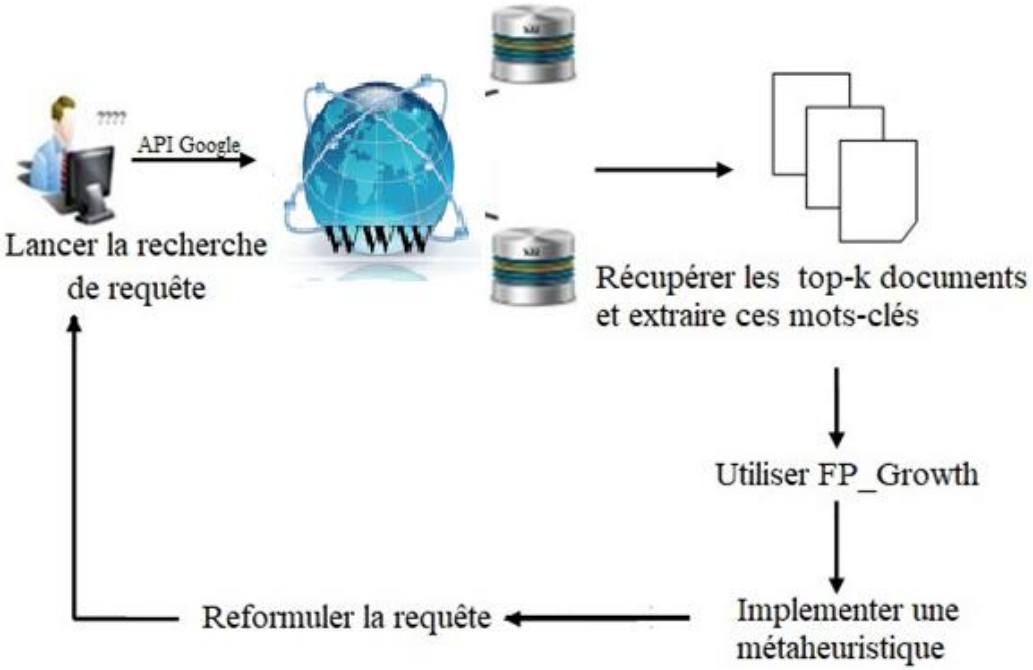


Figure V.3. L'organigramme de la reformulation de requête proposée.

À l'arrivée de la requête de l'utilisateur au moteur de recherche web, la première étape consiste à récupérer les documents pertinents (document, URL, etc.). Dans la deuxième étape, les termes extraits (mots clés) de ces documents sont l'entrée de FP_Growth (frequent-pattern Growth) afin de générer des itemsets fréquents. Dans la dernière étape, chacun de ces items est considéré comme un chemin d'entrée d'une métaheuristique choisi pour converger vers les top-k chemins optimaux et les mots-clés présents dans le chemin sont extraits.

V.3.1. L'API de Google web Service

Notre approche de reformulation de requête est sous forme d'une interface, qui s'appuie sur l'API de Google (Custom Search JSON API | Custom Search | Google Developers, 2017) et qui nécessite à s'inscrire auprès de Google et obtenir une clé. Elle prend en entrée la requête initiale formulée par l'utilisateur ainsi que le nombre de pages à retourner (Dix résultats dans chaque page). Ensuite, on extrait les termes (Metakeywords) de chaque résultat récupéré à cette requête (de façon explicite).

Dans le but d'enrichir les termes de reformulation, des corpus documentaires sont aussi utilisés dont lesquels les documents sont présentés sous formes de termes, puis la mesure TF_IDF (term frequency-inverse document frequency) est calculée pour évaluer l'importance d'un terme contenu dans un document puis le classer.

Une phase de prétraitement est utilisée par la suite, dont laquelle nous supprimons tous les StopWord (mots inutiles). Ensuite, les termes utiles seront annotés avec des informations sur le type de mots (noms, verbes, infinitifs et participes) et des informations de lemmatisation à l'aide de treeTagger (TreeTagger - a part-of-speech tagger for many languages, 2017) afin de prendre la forme racine.

V.3.2. FP_Growth

Sur la base des termes extraits de top_k documents, les termes sont associés à FP_Growth afin de construire les itemsets fréquents en se basant sur les règles d'association. Le FP_Growth utilise la stratégie "diviser et dominer" (divide-and-conquer), il consiste d'abord à compresser la base de données (BD) représentant les itemsets fréquents en une structure compacte nommée FP-tree (Frequent Pattern tree) dont ces branches possèdent toutes les associations possibles des items. Chacune de ces associations peut être décomposée en fragments (pattern fragment) qui constituent les itemsets fréquents.

L'avantage de cet algorithme est qu'il n'effectue que deux balayages de la BD transactionnelle et donc souvent valide, même sur les grands ensembles de données (Afuan and al., 2019). FP_Growth calcule d'abord les listes d'éléments fréquents en fonction du support minimum fourni, puis trie le résultat par fréquence dans l'ordre décroissant lors de son premier balayage. Dans son deuxième balayage, la BD est compressée dans FP-tree. Les différentes étapes de l'algorithme FP_Growth peuvent être résumées dans l'algorithme V.1 (Han and al., 2000).

Algorithme V.1. FP_Growth

Début

1. Balayer la base de transactions T une première fois
   Créer L, la liste des items fréquents avec leur support
   Trier L en ordre décroissant du support
2. Créer l'arbre N contenant une racine étiquetée « Null »
3. Procédure FP_Growth
   Si FP-tree contient un seul chemin P **alors**
   Pour chaque combinaison β de P **faire**
   Générer l'itemsets β ∪ α de support = minimum des supports des nœuds de β
   Fin Pour
   Sinon pour chaque α_i dans l'indexe de FP-tree **faire**
   Générer l'itemset β = α_i ∪ α de support α_i.support
   Construire l'itemset condition de base de β
   Construire $\mathrm{FP-tree}_{\beta}$
   Si $\mathrm{FP-tree}_{\beta} \neq 0$ **alors**
   FP_Growth($\mathrm{FP-tree}_{\beta}$, β)
   Fin Si
   Fin Pour
   Fin Si

Fin

Exemple explicatif de FP_Growth : Afin d'expliquer le principe de FP_Growth, nous choisissons la BD décrite dans le Tableau V.1 qui contient des items descriptifs (correspond dans notre cas aux termes extrais utiles).

Id Transaction (Documents)	**Liste d'item TID**
T1	I1, I2, I5, I6
T2	I2, I4, I7
T3	I2, I1, I3, I6
T4	I1, I2, I4, I7

Tableau V.1. Exemple.

Items	Support
I2	4
I1	3
I4	2
I6	2
I7	2
I3	1
I5	1

Tableau V. 2. Items associés à leur support.

Premièrement, FP_Growth parcourt la BD et lit tous les items existants en calculant leur support et les compare avec le seuil de support préfixé, et par conséquent tous les items ayant un support supérieur ou égal au seuil appelé item fréquent seront retenus (Dans notre exemple le seuil de support est égal à 1). FP_Growth enregistre les items fréquents dans une liste L où ils figurent par ordre décroissant du support. Dans notre cas le Tableau V.2 représente les items retenus ainsi que leurs nombre d'occurrences respectif, et la liste correspondante est : L = I2 : 4, I1 : 3, I4 : 2, I6 : 2, I7 : 2, I3 : 1, I5 : 1.

Deuxièmement, un FP-Tree est construit par la création d'une racine vide et un second parcours de la BD où chaque transaction est décrite dans l'ordre des items donnés par la liste L. Par exemple, la première transaction T1 : I1, I2, I5, I6 sera sauvegardée en T1 : I2, I1, I6, I5. Cette première transaction, une fois ordonnée, va construire la première branche du FP-Tree avec 4 nœuds (I2 : 1), (I1 : 1), (I6 : 1) et (I5 : 1). La seconde transaction T2, contenant T2 dans l'ordre de L : I2, I4, I7, construit une branche où le nœud I2 est lié à la racine, le nœud I4 lié à I2 et le nœud I7 lié à I4. Cette branche partage un préfixe commun I2, avec T1. Ainsi, lors de la construction du nœud I4 :1, on doit incrémenter de 1 le compteur du nœud I2 (I2 :2).

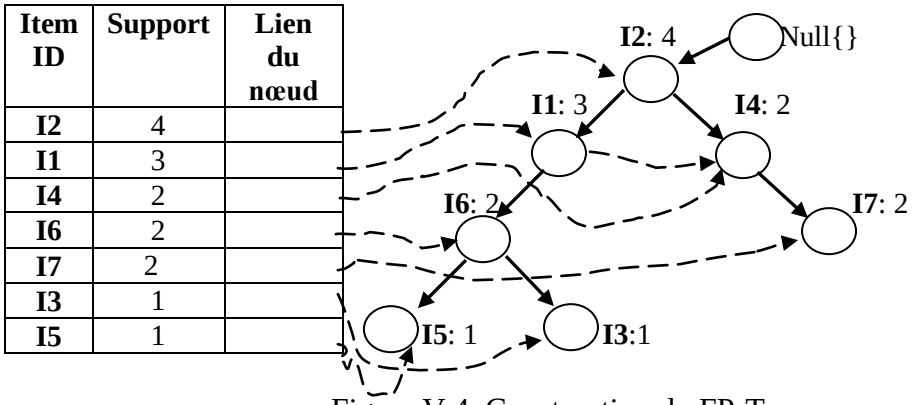


Figure V.4. Construction du FP-Tree.

En dernière étape, le FP-Tree est exploré par la création des sub-fragments conditionnels de base. En fait, pour trouver ces fragments, on extrait pour chaque fragment de

longueur 1 (suffix pattern) l'ensemble des préfixes existants dans le chemin du FP-Tree (Conditional Pattern Base, CPB). L'itemset fréquent est obtenu par la concaténation du suffixe avec les fragments fréquents extraits des FP-Tree conditionnels (Tableau V.3).

Item	Motifs conditionnels	FP-Tree conditionnels	Motifs fréquents
I5	I1, I2, I6, I2, I1, I3	(I2 : 2, I1 :1)	I2, I5 :2, I1, I5 :2 :I2, I1, I5 :2
I4	I2 :1, I2, I1 :1	(I2 :2)	I2, I4 :2

Tableau V.3. Fouille du FP-Tree.

L'utilisation de l'algorithme FP_Growth permet de générer tous les chemins contenant des itemsets fréquents. Afin de minimiser le temps de calcul et donner une reformulation d'une requête plus adéquate, on sélectionne seulement les chemins admettant au maximum cinq mots (motifs fréquents). Ces chemins sont considérés comme une entrée pour une métaheuristique choisie.

V.3.3. Les métaheuristiques

Notre approche de reformulation de requêtes proposée est basée essentiellement sur les métaheuristiques évolutionnaires bio-inspirées, qui sont : FireFly, PSO, Bat et PIO algorithms. Ces derniers sont utilisés afin d'extraire les meilleurs termes de reformulation dans un intervalle de temps rapide et raisonnable. L'adaptation de ces algorithmes était détaillée dans le chapitre précédent (Section IV.3).

V.4 Adaptation des pages web

Les internautes peuvent consulter les pages web en les adaptant selon leurs préférences. L'étape d'adaptation des pages web proposée est répartie en trois phases, qui sont illustrée dans la figure ci-dessous.

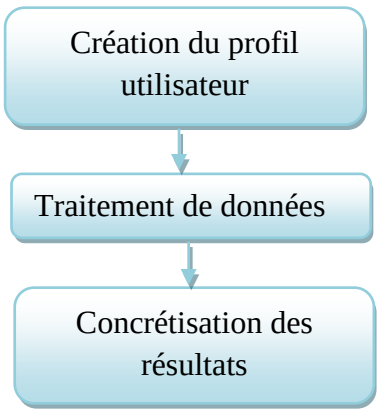


Figure V.5. L'organigramme de l'adaptation de pages web.

V.3.1. Création du profil utilisateur

La phase « création du profil utilisateur » permet à l'utilisateur de modifier les paramètres qui vont être appliqués sur toutes les pages web qu'il va consulter.

L'acquisition des préférences n'est pas quelque chose d'évident, il n'est pas éventuel de demander à quelqu'un d'exprimer l'ensemble de ses préférences. Notre approche s'appuie sur des préférences portant sur des éléments de bas niveau d'une page web: la couleur du fond, la couleur du texte et la taille de police ; et cela dépend du type de déficience visuelle de chaque utilisateur, qui peut créer et enregistrer son profil dés qu'il fixe ces préférences.

V.3.2. Traitement des données

La partie cœur du mécanisme de l'adaptation des pages web est la phase traitement de données, car c'est elle qui est responsable de l'analyse et de la transformation du contenu des pages web selon les préférences de chaque utilisateur (choisis précédemment). Le diagramme schématique de cette partie est illustré dans la Figure V.6.

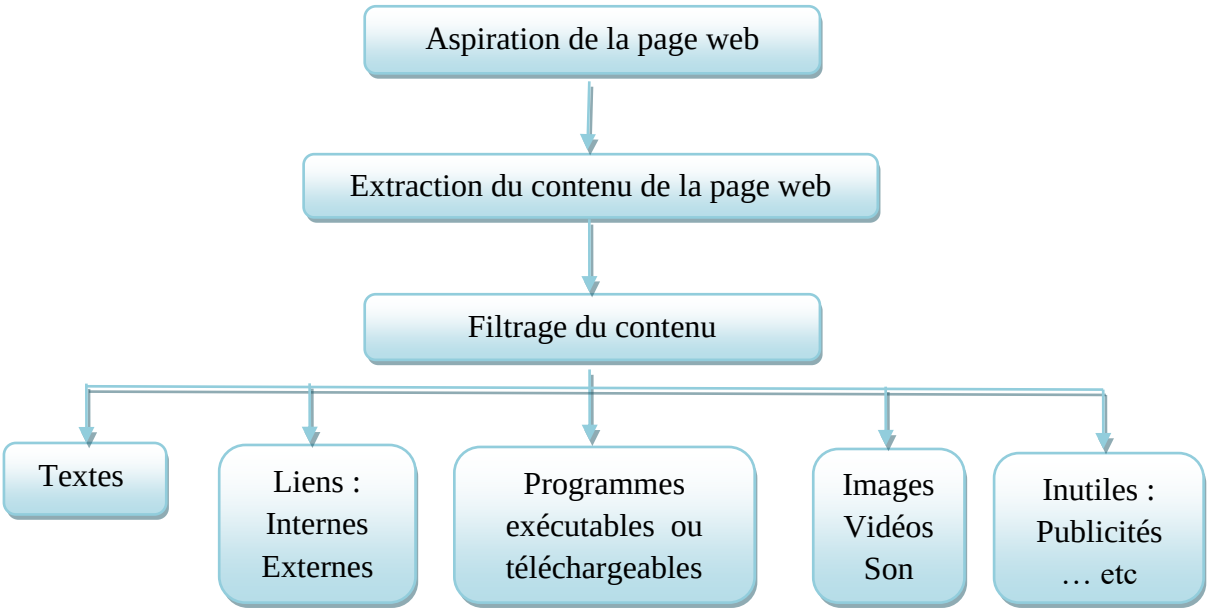


Figure V.6. L'organigramme du processus de traitement de données.

La phase du traitement de données se compose de trois processus majeurs :

- **Aspiration de la page web :** C'est le processus clé de la phase traitement de données. Il est responsable de l'analyse de tous les éléments visibles (les textes, les images, les fichiers son, les vidéos, les liens hypertextes, et les programmes exécutables ou téléchargeables) et invisibles (codes HTML) du document HTML. La structure

DOM[9] (Document Model Object) est le résultat de l'analyse du document. Ce processus est implémenté à l'aide de la bibliothèque Jsoup [10] .

- **Extraction du contenu de la page web :** Ce processus permet d'identifier et de télécharger toutes les ressources d'un document HTML (page Web consultée).
- **Filtrage du contenu :** Ce processus sert à rassembler et organiser les données visibles extraites de la page web selon leurs différents types, et à supprimer les données inutiles (spam, publicité…etc.).

V.4.1. Concrétisation des résultats

Le dernier processus est l'environnement de navigation qui représente l'interface utilisateur du navigateur et qui lui sert principalement à consulter les pages web, et notamment voire le résultat de la personnalisation (selon leurs préférences).

Les outils utilisés lors de la phase de conception et qui permettent à un utilisateur ayant une déficience visuelle de naviguer facilement sont : une fonction Zoom (loupe Windows), une loupe et la synthèse vocale SI_VOX.

SI _VOX est développée par des étudiants de Polytech'Nice de l'école polytechnique universitaire de Nice Sophia Antipolis en 2004 dans le cadre d'un projet. Cette bibliothèque est appuyée sur le synthétiseur vocal MBROLA (MBROLA Voices, 2022) et est placée sous la protection de la licence MBROLA qui permet de transformer un texte "écrit" en texte "parlé" dans divers langue.

V.5 Conclusion

Dans ce chapitre, nous avons détaillé la partie modélisation et conception de notre travail. En effet, l'approche proposée comprend deux importantes phases: la première phase est la reformulation de la requête web, tandis que la deuxième phase est l'adaptation des pages web selon les préférences des utilisateurs. Le chapitre d'expérimentations suivant montre la réalisation de la solution proposée pour les utilisateurs ayant une déficience visuelle afin de faciliter leurs tâches de navigation.

[9] DOM est une interface de programmation normalisée par le W3C, qui permet à des scripts d'examiner et de modifier le contenu du navigateur web, sa structure et ses styles.

[10] Jsoup est Java HTML Parser. Jsoup est une libraire de Java utilisée pour parsemer des données du document HTML. Jsoup fournit des API qui sont utilisés pour extraire et manipuler des données venant de l'URL ou du fichier HTML.

Chapitre VI.

Implémentation et expérimentation des résultats

VI.1 Introduction

La mise en œuvre de divers algorithmes a permis la réalisation de nombreuses expérimentations. Ces dernières sont nécessaires dans la phase de la validation de l'approche globale.

Ce chapitre est consacré à l'évaluation des performances de l'approche proposée. Dans la première partie nous décrivons les ensembles de données (collections), les critères d'évaluation, les algorithmes utilisés pour la comparaison et les résultats expérimentaux pour la reformulation des requêtes web. Dans la deuxième partie nous présentons les différentes interfaces utilisées pour l'adaptation des pages web selon les préférences des utilisateurs ayant une déficience visuelle, puis les résultats obtenus.

PARTIE 01 : EVALUATION DE L'ETAPE DE REFORMULATION DES REQUETES WEB

VI.2 Les ensembles de données

Des expériences approfondies sont réalisées sur trois différentes collections de tests en anglais: Text Retrieval Conference TREC-3 (disques 1 et 2) (Text Retrieval Conference (TREC),2020), Forum for Information Retrieval Evaluation FIRE 2011 Ad-hoc (Forum for Information Retrieval Evaluation, 2020) et Centre for Inventions and Scientific Information CISI de l'ensemble de collections standards SMART (CISI (a dataset for Information Retrieval) _ Kaggle, 2021). Ces collections contiennent un ensemble de documents, un ensemble de requêtes et des jugements de pertinence (une liste des documents pertinents pour chaque requête). Elles sont utilisées pour avoir suffisamment de données d'apprentissage car elles contiennent un très grand nombre de sujets et l'ensemble de documents sur lesquels la recherche d'information est effectuée. Les descriptions détaillées de ces ensembles de données sont présentées dans le Tableau VI.1.

Les ensembles de données	Taille	Nombre de requêtes	Numéros de requêtes	Documents
TREC-3	6 Gb	50	151–200	7,41,856
FIRE 2011	1.76 Gb	50	126 -175	3,92,577
CISI	2.23 MB	112	1-112	1,460

Tableau VI. 1. Détails de tous les ensembles de données utilisés.

VI.3 Les mesures d'évaluation

Les mesures d'évaluation: rappel (R), précision (P), Moyenne des précisions moyennes (MAP) et F-mesure sont utilisées afin d'évaluer la performance globale de notre approche de reformulation des requêtes (Section III.6.2).

VI.4 Choix des valeurs des paramètres

Avant de discuter sur les performances de l'approche proposée, nous identifions d'abord les meilleures valeurs à utiliser pour les paramètres de l'algorithme FP_Growth et des métaheuristiques utilisés (FirFly, Bat, PSO et Pigeon).

Afin de choisir les valeurs des paramètres, nous avons utilisé dans les expériences préliminaires seulement les dix premières requêtes de chaque ensemble de données (TREC, FIRE et SICI), puis en calculant la précision de dix premiers documents retournés (P@10).

VI.4.1. FP_Growth

L'algorithme d'extraction de règles d'association FP_Growth extrait touts les itemsets fréquents dont le support dépasse des seuils préfixés (min_Sup). Sa première phase est la plus coûteuse en terme de temps d'exécution ce qui pourra causer un inconvénient pour les utilisateurs, car le nombre de ces items fréquents dépend exponentiellement du nombre d'items manipulés. Pour que cette étape soit exécutable, on doit augmenter le seuil de support. Dans notre cas, nous avons choisi min_Sup=1.

VI.4.2. Les métaheuristiques

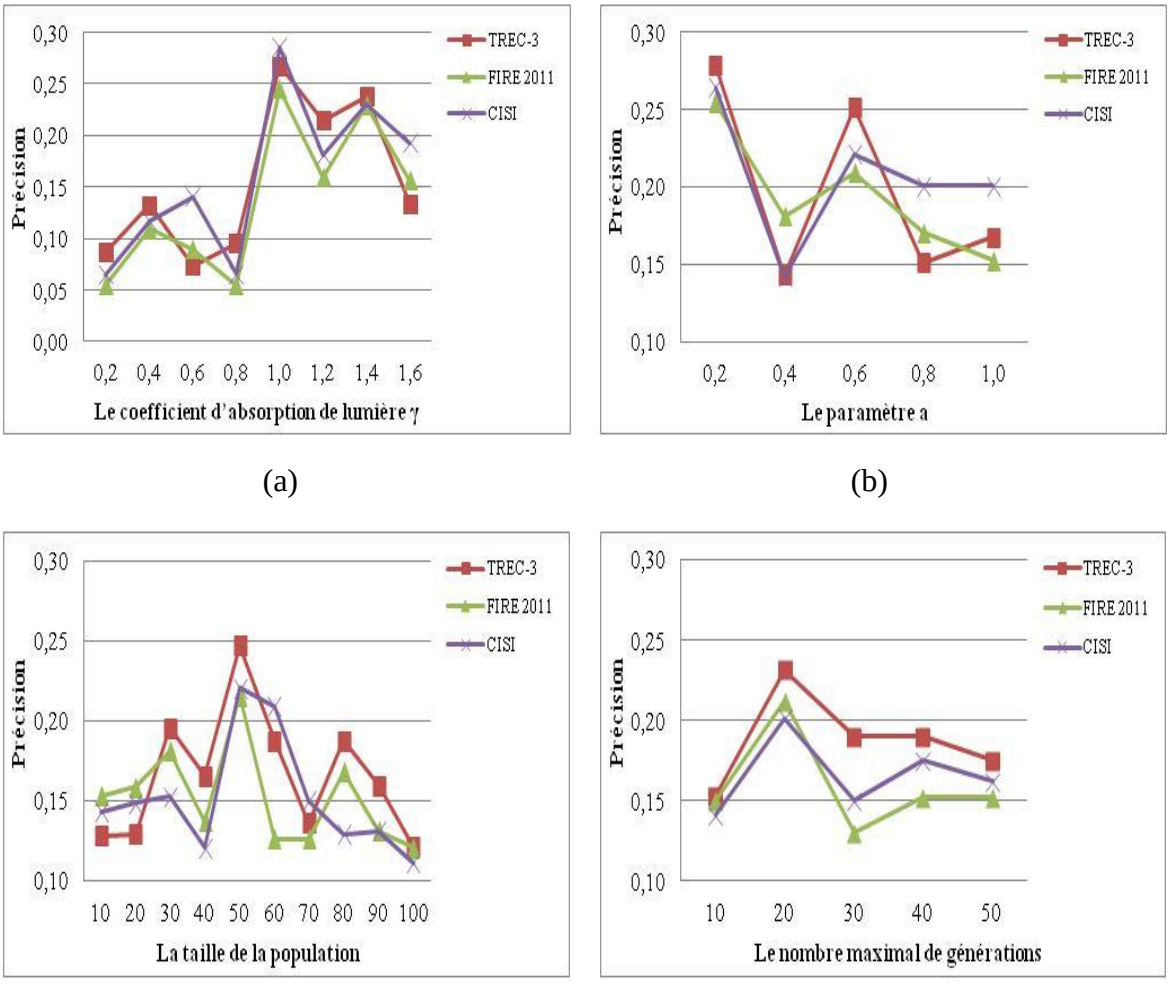
- **Algorithme des lucioles (FireFly Algorithm)**

Nous ajustons les valeurs des paramètres de l'algorithme des lucioles (FA) suivantes : le coefficient d'absorption de lumière γ, le paramètre α, la taille de la population N et le nombre maximal de générations T.

Les figures VI.1(a), VI.1(b), VI.1(c) et VI.1(d) montrent les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de FA pour les trois ensembles de données. Le Tableau VI.2 présente les valeurs utilisées pour la fixation des valeurs de ces paramètres.

Paramètres / Figures	γ	A	*N*	*T*
Figure VI.1(a)	[0.2 – 1.6]	1	10	10
Figure VI.1(b)	1	[0.2 – 1]	10	10
Figure VI.1(c)	1	0.2	[10 – 100]	10
Figure VI.1(d)	1	0.2	50	[10-50]

Tableau VI.2. Les valeurs utilisées pour la fixation des paramètres de FA.



(a) (b)

(c) (d)

Figure VI.1. Les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de FA.

Sur la Figure VI.1(a), nous avons varié le coefficient d'absorption de lumière γ entre les valeurs 0.2 et 1.6. La valeur de γ qui donne la précision la plus élevée est égale à 1. Sur la Figure VI.1(b), nous avons varié le paramètre *a* entre les valeurs 0.2 et 1 en l'incrémentant par 0.2. Nous remarquons qu'il y a deux valeurs qui donnent de bonnes performances et qui sont :

0.2 et 0.6. D'après la Figure VI.1(c), la meilleure valeur de la taille de la population qui apporte une meilleure efficacité de récupération est de 50. Finalement, selon la Figure VI.1(d) nous apercevons que la valeur appropriée du nombre maximal de générations est égale à 20.

Sur la base de ces expériences préliminaires réalisées sur les trois ensembles de données, les valeurs des paramètres de l'algorithme des lucioles sont présentées dans le Tableau VI.3 ci dessous.

Paramètres	**Valeurs**
Le coefficient d'absorption de lumière γ	1
Le paramètre α	0.2
La taille de la population *N*	50
Le nombre maximal de générations *T*	20

Tableau VI.3. Les valeurs des paramètres de FA choisies.

- **Algorithme des chauves-souris (Bat)**

Nous ajustons les valeurs des paramètres de l'algorithme de chauve-souris suivantes : la fréquence d'impulsion f_{max}, l'impulsion *r*, l'intensité *A*, la taille de la population N et le nombre maximal de générations T.

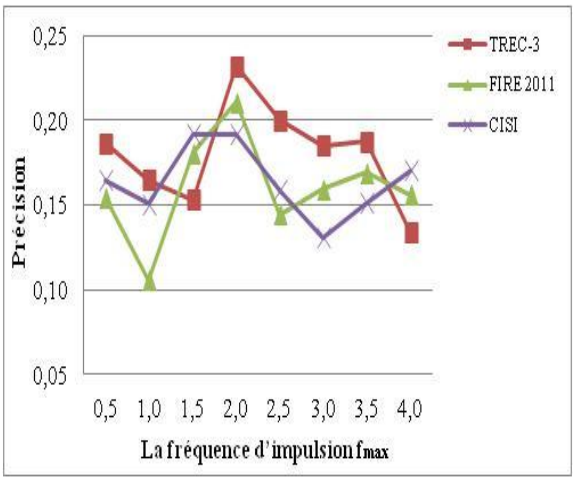
Les figures VI.2(a), VI.2(b), VI.2(c), VI.2(d) et VI.2(e) montrent les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de Bat algorithm pour les trois ensembles de données. Le Tableau VI.4 présente les valeurs utilisées pour la fixation des valeurs de ces paramètres.

Paramètres / **Figures**	f_{max}	***r***	***A***	***N***	***T***
Figure VI.2(a)	[0 – 4]	0.5	0.5	10	10
Figure VI.2(b)	2	[0.1 – 0.8]	0.5	10	10
Figure VI.2(c)	2	0.4	[0.1 – 0.8]	10	10
Figure VI.2(d)	2	0.4	0.2	[10 – 100]	10
Figure VI.2(e)	2	0.4	0.2	30	[10-50]

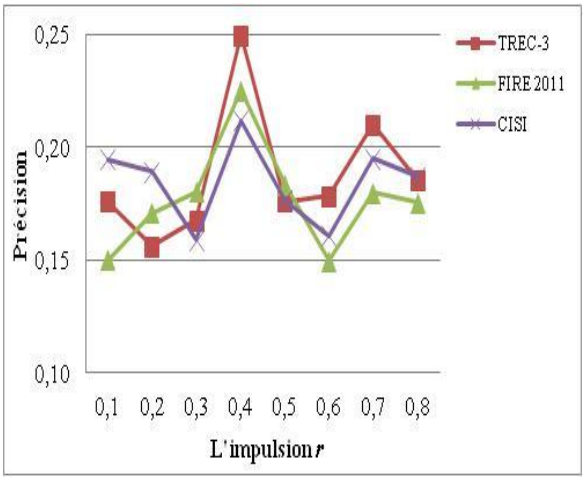
Tableau VI.4. Les valeurs utilisées pour la fixation des paramètres de Bat algorithm.

Sur la Figure VI.2(a), nous avons fait varier la fréquence d'impulsion f_{max} entre les valeurs 0 et 4 en l'incrémentant par 0.5. D'après les résultats obtenus la meilleure valeur de f_{max} qui donne une bonne précision est égale à 2. Sur la Figure VI.2(b), nous avons fait varier l'impulsion *r* entre les valeurs 0.1 et 0.8. Nous observons que la valeur 0.4 donne la meilleure précision par rapport aux autres valeurs. Sur la Figure VI.2(c), nous avons fait varier

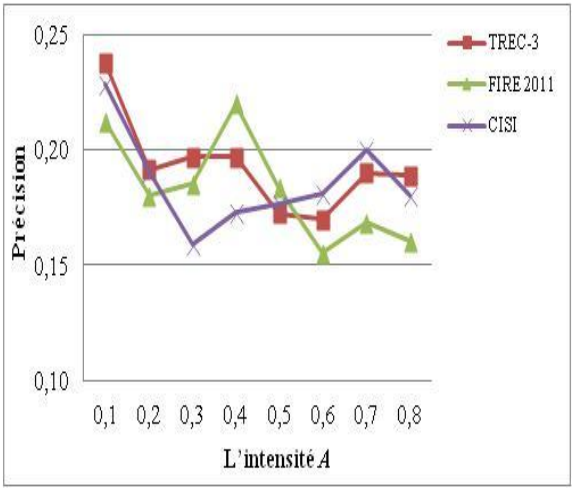
l'intensité *A* entre les valeurs 0.1 et 0.8. La meilleure valeur de *A* qui apporte une meilleure efficacité de récupération est 0.1. Sur la Figure VI.2(d), Nous remarquons qu'il y a deux valeurs de la taille de la population qui donnent de bonnes performances et qui sont: 30 et 40. Finalement, selon la Figure VI.2(e) nous apercevons qu'il y a aussi deux valeurs de nombre maximal de générations qui donnent de bonnes précisions et qui sont 20 et 30.



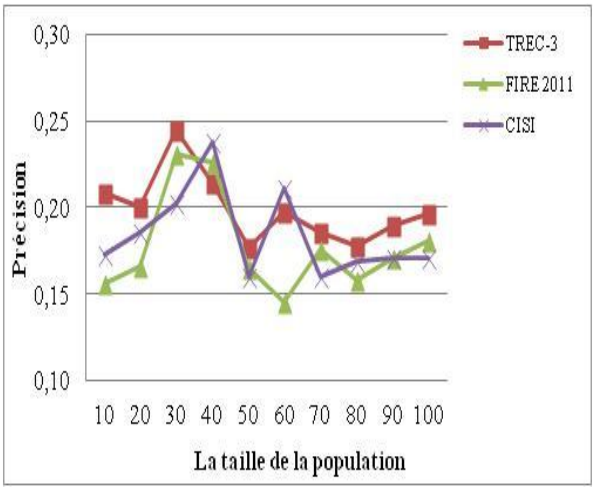
(a)



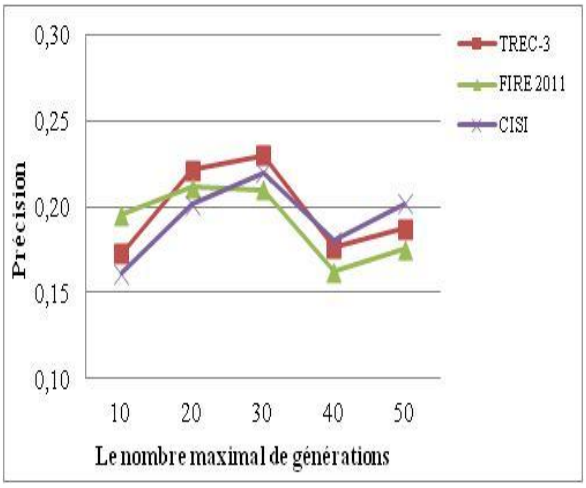
(b)



(c)



(d)



(e)

Figure VI.2. Les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de Bat algorithm.

Le Tableau VI.5 présente les valeurs des paramètres de l'algorithme des chauves-souris selon les expériences préliminaires réalisées sur les trois ensembles de données.

Paramètres	**Valeurs**
La fréquence d'impulsion f_{max}	2.0
L'impulsion r	0.4
L'intensité A	0.1
La taille de la population N	30
Le nombre maximal de générations T	30

Tableau VI.5. Les valeurs des paramètres de Bat algorithm choisies.

- **Algorithme d'optimisation par essaim de particules (PSO)**

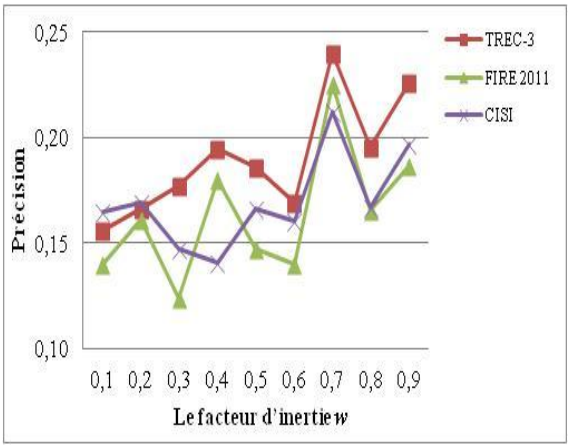
Nous ajustons les valeurs des paramètres de l'algorithme d'optimisation par essaim de particules suivants: le facteur d'inertie w, les coefficients d'accélération $c_1$et c_2, la taille de la population N et le nombre maximal de générations T.

Les figures VI.3(a), VI.3(b), VI.3(c), VI.3(d) et VI.3(e) donnent les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de PSO pour les trois ensembles de données. Le Tableau VI.6 présente les valeurs utilisées pour la fixation des valeurs de ces paramètres.

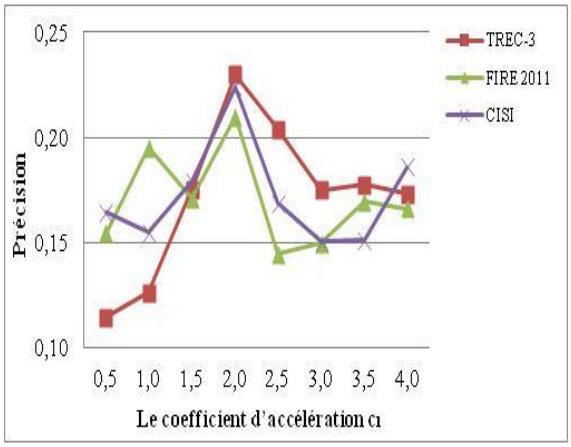
Paramètres / Figures	W	c_1	c_2	N	T
Figure VI.3(a)	[0.1 – 0.9]	1	1	10	10
Figure VI.3(b)	0.7	[0.5 – 4]	1	10	10
Figure VI.3(c)	0.7	2	[0.5 – 4]	10	10
Figure VI.3(d)	0.7	2	2	[10 – 100]	10
Figure VI.3(e)	0.7	2	2	20	[10-50]

Tableau 1. Les valeurs utilisées pour la fixation des paramètres de PSO algorithm.

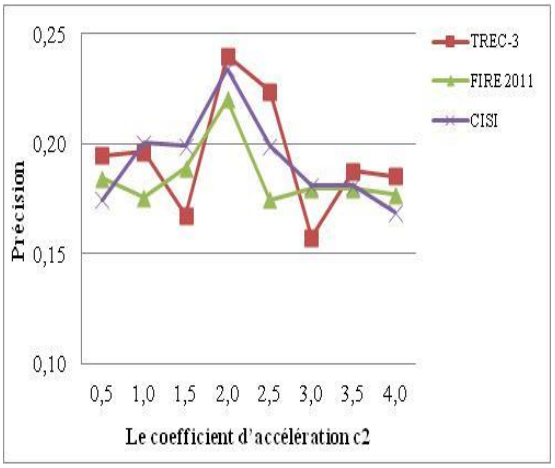
Sur la Figure VI.3(a), nous avons fait varier le facteur d'inertie *w* entre les valeurs 0.1 et 0.9. La valeur de *w* qui donne la meilleure précision est égale à 0.7. Sur les Figures VI.3(b) et VI.3(c), nous avons varié les coefficients d'accélération c_1 et c_2 respectivement entre les valeurs 0.5 et 4 en les incrémentant par 0.5. Nous remarquons que la valeur 2 donne de bonnes performances pour ces deux coefficients.



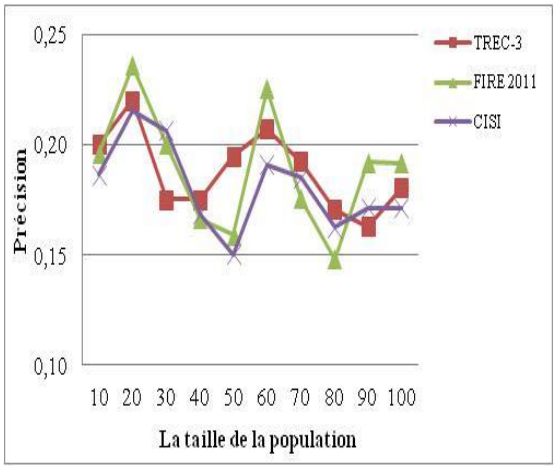
(a)



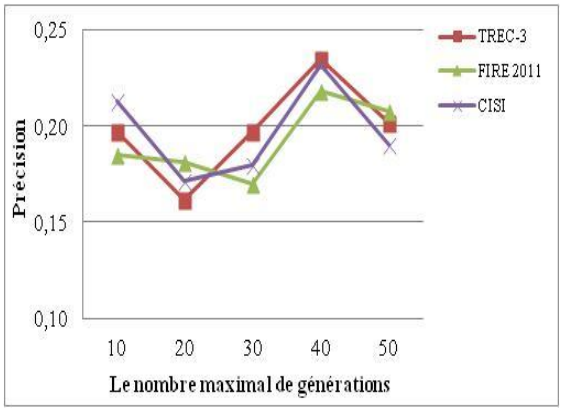
(b)



(c)



(d)



(e)

Figure VI.3. Les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de PSO algorithm.

Sur la Figure VI.3(d), Nous observons qu'il y a deux valeurs de la taille de la population qui donnent de meilleurs résultats et qui sont: 20 et 60. Finalement, selon la Figure VI.3(e) nous apercevons que la valeur appropriée du nombre maximal de générations est égale à 40.

Sur la base de ces expériences préliminaires réalisées sur les trois ensembles de données, les valeurs des paramètres de l'algorithme d'optimisation par essaim de particules sont présentées dans le Tableau VI.6.

Paramètres	**Valeurs**
Le facteur d'inertie w	0.7
Le coefficient d'accélération c_1	2
Le coefficient d'accélération c_2	2
La taille de la population N	20
Le nombre maximal de générations T	40

Tableau VI.6. Les valeurs des paramètres de la PSO algorithm choisies.

- **Algorithme des pigeons (PIO)**

Nous ajustons les valeurs des paramètres de l'algorithme des pigeons suivants: le facteur de carte et de boussole R, le nombre maximum de générations que l'opération carte et boussole soit effectuée N_{c1max}, le nombre maximum de générations que l'opération de repère est effectuée N_{c2max}, la taille de la population N et le nombre maximal de générations T.

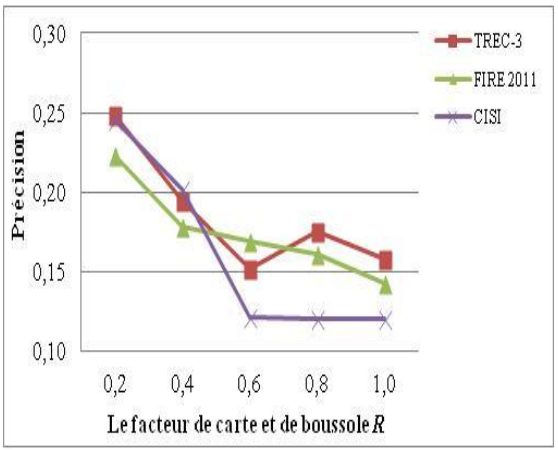
Les figures VI.4(a), VI.4(b), VI.4(c), VI.4(d) et VI.4(e) montrent les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de pigeon pour les trois

ensembles de données. Le Tableau VI.7 présente les valeurs utilisées pour la fixation des valeurs de ces paramètres.

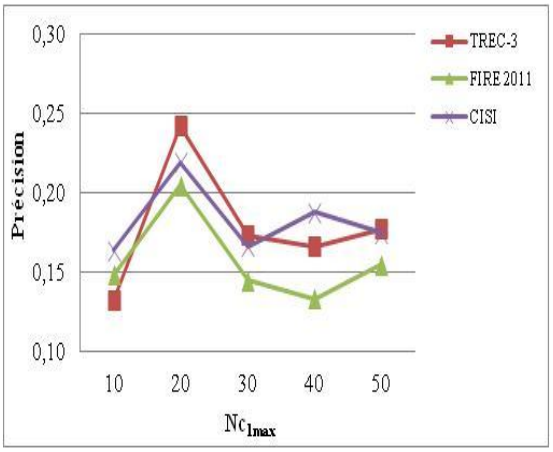
Paramètres / **Figures**	R	N_{c1max}	N_{c2max}	N	T
Figure VI.4(a)	[0.2 – 1]	10	20	10	10
Figure VI.4(b)	0.2	[10– 50]	20	10	10
Figure VI.4(c)	0.2	20	[10– 50]	10	10
Figure VI.4(d)	0.2	20	40	[10 – 100]	10
Figure VI.4(e)	0.2	20	40	60	[10-50]

Tableau 2. Les valeurs utilisées pour la fixation des paramètres de l'algorithme de pigeon.

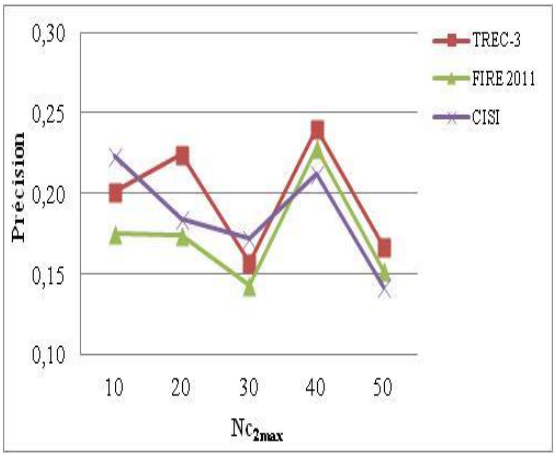
Sur la Figure VI.4(a), nous avons fait varier le facteur de carte et de boussole R entre les valeurs 0.2 et 1 en l'incrémentant par 0.2. Selon les résultats obtenus la valeur 0.2 donne la meilleure précision par rapport aux autres valeurs. Sur les Figures VI.4(b) et VI.4(c), nous avons fait varier N_{c1max} et N_{c2ma} entre 10 et 50. Les nombres maximum de générations que l'opération carte/boussole et l'opération de repère sont effectuées, et qui apportent une bonne efficacité de récupération sont égales à 20 et 40 respectivement. Sur la Figure VI.4(d), Nous remarquons que la valeur de la taille de la population qui donne une meilleure précision est égale à 60. D'après la Figure VI.4(e), la valeur de nombre maximal de générations qui donnent une précision plus élevé est égale à 30.



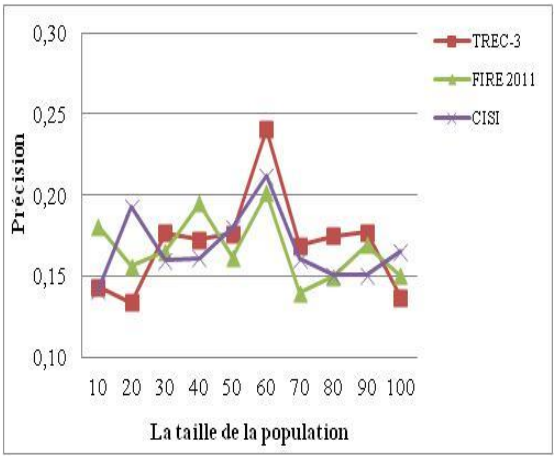
(a)



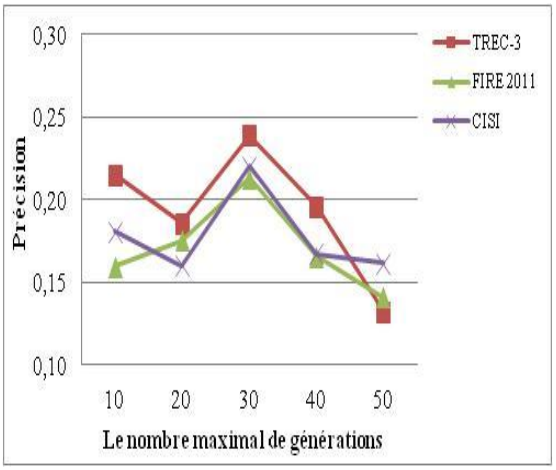
(b)



(c)



(d)



(e)

Figure VI.4. Les valeurs de Précision obtenues lors du réglage des valeurs des paramètres de l'algorithme des pigeons.

Le Tableau VI.7 montre les valeurs des paramètres de l'algorithme des pigeons selon les expériences préliminaires réalisées sur les trois ensembles de données.

Paramètres	**Valeurs**
Le facteur de carte et de boussole R	0.2
N_{c1max}	20
N_{c2max}	40
La taille de la population N	60
Le nombre maximal de générations T	30

Tableau VI.7. Les valeurs des paramètres de l'algorithme de pigeon choisis.

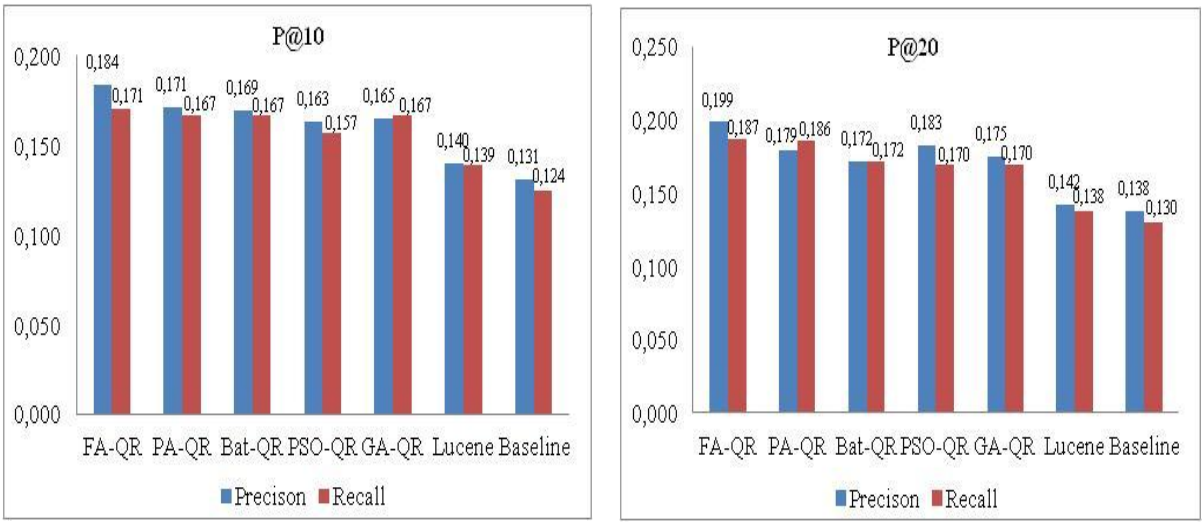
VI.5 Résultats et discussions

Afin d'évaluer les performances de l'approche de reformulation des requêtes proposée, il est recommandé de la comparer avec des approches récentes de reformulation de requêtes

basées sur les mêmes algorithmes ou bien méthodes. Parfois, malgré l'utilisation des mêmes ensembles de données et des mêmes approches, des contradictions dans les résultats seront détectées, empêchant une comparaison équitable en raison de l'utilisation d'une grande variété de paramètres de configuration tels que le filtrage des mots vides, les modèles de classement, … etc. Pour cela, l'approche proposée basée sur FireFly algorithm (FA-QR) a s'été comparée à l'approche de base (Baseline-sans reformulation) et au Lucene [(11)] ainsi qu'aux différentes métaheuristiques qui sont : l'optimisation des essaims de particules (PSO-QR), algorithme de pigeon (PA-QR), l'algorithme de chauve-souris (Bat-QR) et les algorithmes génétiques (GA-QR).

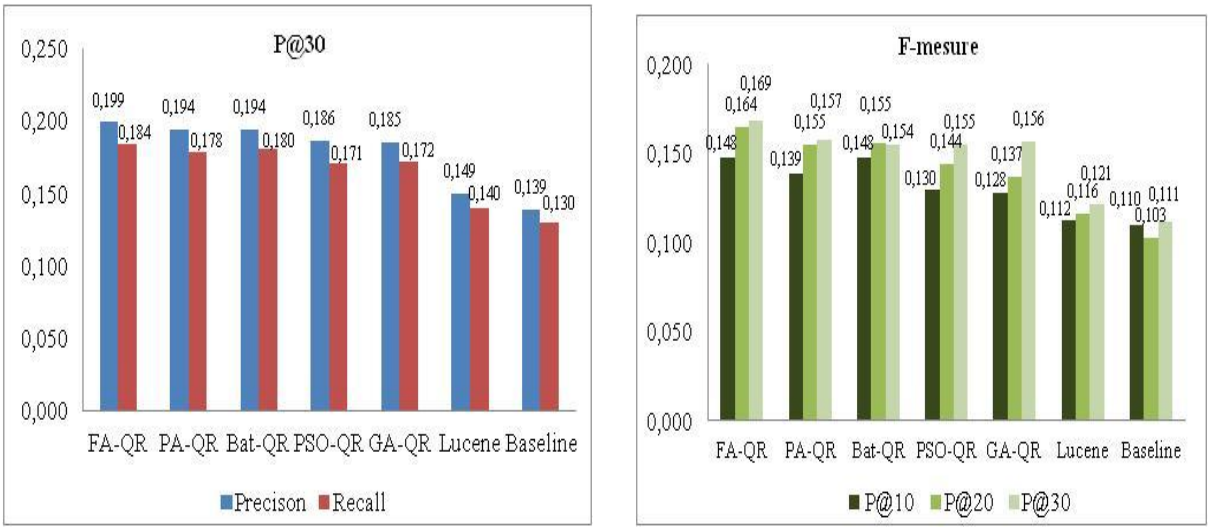
La Figure VI.5 affiche les performances de récupération de l'approche proposée basée sur les métaheuristiques pour tous les documents récupérés (P@X) en termes de précision, rappel et F-mesure sur l'ensemble de données TREC.

P@X indique la proportion de documents pertinents retrouvés parmi les X premiers documents de la liste renvoyée pour les requêtes données. Dans notre cas, X prend les valeurs 10, 20 et 30, respectivement. Par exemple, P@10 est la proportion des documents pertinents restitués parmi les 10 premiers documents retrouvés.



(a) (b)

[11] Lucene est une bibliothèque open source écrite en Java publiée par l'Apache Software Foundation, qui permet d'indexer et de chercher du texte dans une série de documents (un texte Word, un fichier PDF, un ensemble de fichiers, une page web sur un serveur distant, des informations stockées dans une base de données, etc.)



(c) (d)

Figure VI.5. Valeurs de précision, rappel et F-mesures de toutes les approches sur l'ensemble de données TREC.

Selon la Figure VI.5, nous pouvons voir évidemment que les approches proposées basées sur les métaheuristiques produisent les valeurs les plus élevées et réalisent une amélioration significative par rapport aux Lucene et l'approche de base, notamment pour P@30 (Figure VI.5(d)) la F-mesure donne de meilleurs résultats par rapport à P@10 et P@20 pour toutes les approches, et cela est dû probablement que lorsqu'il y a beaucoup de documents, il y a beaucoup de termes qui permettent l'enrichissement de la requête.

De même, pour les métaheuristiques, les résultats obtenus présentent quelques similarités sur le rappel. Cependant, la précision donne une certaine différence. Pour P@30, l'algorithme FireFly (FA-QR) est plus efficace pour récupérer des résultats pertinents. Il améliore les performances de la recherche en atteignant 0,199 par rapport à la méthode PA-QR avec 0,194, Bat-QR avec 0,194, PSO-QR avec 0,186 et GA-QR avec 0,185.

En termes de MAP (Tableau VI.8), nous observons que la FA-QR rapporte les meilleurs résultats par rapport aux différentes autres méthodes. FA-QR donne 42,56 % d'améliorations par rapport à PA-QR avec 33,46%, Bat-QR avec 31,32 %, PSO-QR avec 30,45 %, GA-QR avec 28,83 % et Lucene avec 5,75 %. De plus, comparé à Lucene (MAP-Gain-L) où 21,82 %, 23,34 %, 24,17 % et 26,19 ont été obtenus par GA-QR, PSO-QR, Bat-QR et PA-QR respectivement par rapport à FA-QR avec 34, 80%.

Approche	MAP	MAP-Gain	MAP-Gain-L
FA-QR	0,1937	42,565%	34,804%
PA-QR	0,1813	33,461%	26,196%
Bat-QR	0,1784	31,325%	24,176%
PSO-QR	0,1773	30,451%	23,349%
GA-QR	0,1751	28,837%	21,823%
Lucene	0,1437	5,757%	
Baseline	0,1359		

Tableau VI.8. Comparaison de toutes les approches en terme de MAP sur l'ensemble de données TREC.

La Figure VI.6 montre les valeurs de précision, rappel et F-mesure de chaque approche sur l'ensemble de données FIRE après avoir récupéré 10, 20 et 30 documents respectivement.

Nous pouvons clairement voir que l'approche proposée basée sur les métaheuristiques produit les valeurs les plus élevées et réalise une amélioration substantielle par rapport aux Lucene et l'approche de base, par exemples, pour PA-QR la précision s'améliore de 0,168 (P@10) et 0,173 (P@20) à 0,182 (P@30), pour Bat-QR la précision s'améliore de 0,169 et 0,177 à 0,178 ; par contre pour l'approche de base la précision s'améliore de 0,128 (P@10) à 0,133(P@20 et P@30) et Lucene s'améliore de 0,133 (P@10) à 0,138 (P@20 et P@30).

Nous observons que pour toutes les approches, les valeurs de la précision et le rappel sont élevés pour P@30 comparé aux P@10 et P@20 respectivement, par exemple, en terme de précision Bat-QR s'augmente de 0.169(P@10) et 0.177(P@20) à 0.178(P@30). De même en terme de rappel l'approche s'augmente de 0.159(P@10) et 0.166(P@20) à 0.167(P@30).

En terme de F-mesure (Figure VI.6(d)), nous remarquons aussi que FA-QR rapporte les meilleurs résultats par rapport aux autres approches proposées, par exemple pour P@30, FA-QR archive 0.169 par rapport à la méthode PA-QR et PSO-QR avec 0,155, Bat-QR avec 0,152, et GA-QR avec 0,142.

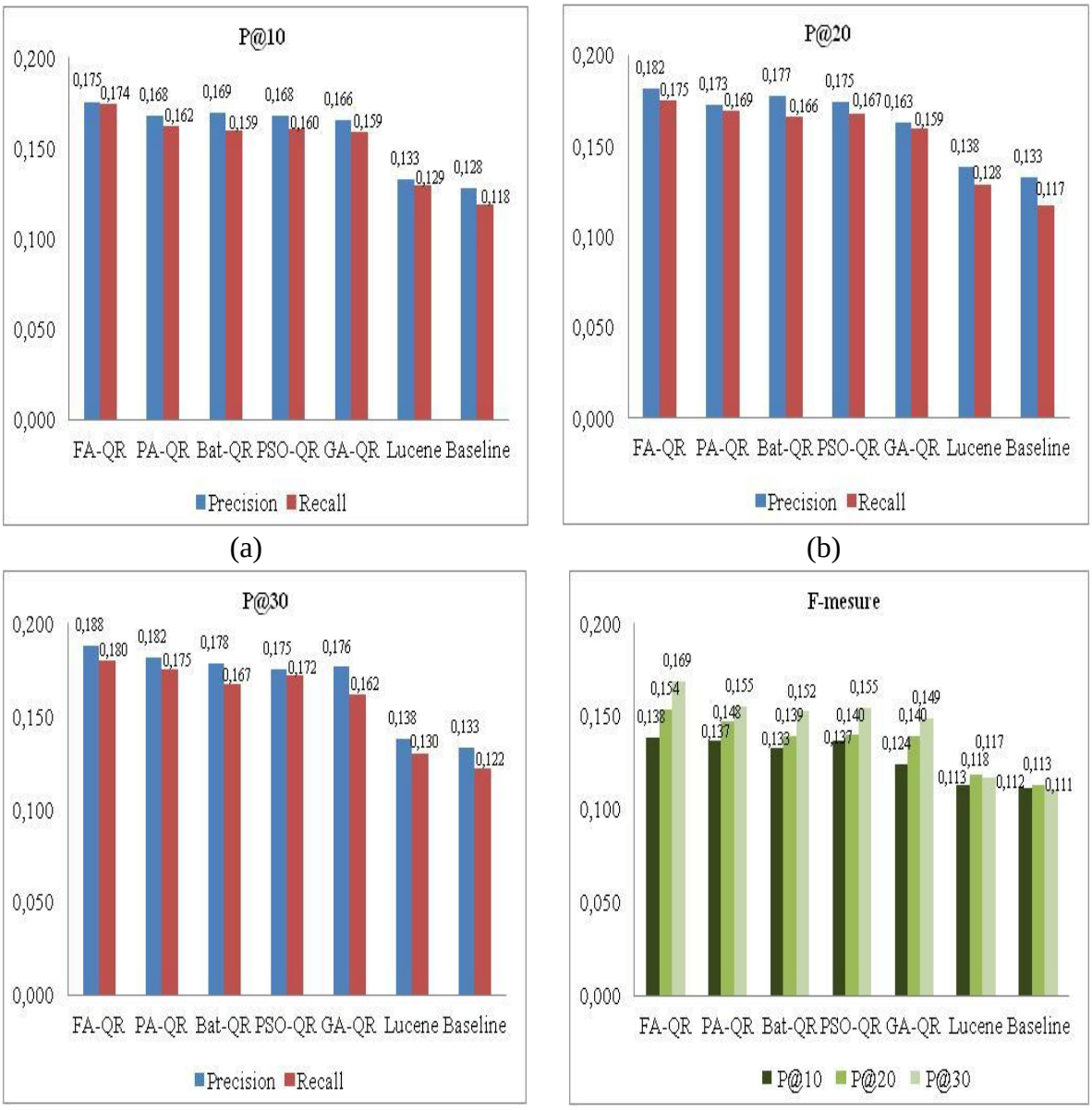


Figure VI. 6. Valeurs de précision, rappel et F-mesures de toutes les approches sur l'ensemble de données FIRE.

À partir du Tableau VI.9, nous pouvons voir une nette supériorité de FA-QR proposée par rapport à Lucene (+26,33%), et cette supériorité est plus significative par rapport à Baseline (+38,47 %). Nous pouvons également observer que PA-QR, Bat-QR, PSO-QR et GA-QR proposées surpassent Baseline autour de 32,75%, 33,48%, 31,69% et 28,42% ; et Lucene autour de 21,11%, 21,78%, 20,15% et 17,16% respectivement.

Approche	MAP	MAP-Gain	MAP-Gain-L
FA-QR	0,1815	38,470%	26,332%
PA-QR	0,1741	32,756%	21,118%
Bat-QR	0,1750	33,482%	21,781%
PSO-QR	0,1727	31,695%	20,150%
GA-QR	0,1684	28,424%	17,166%
Lucene	0,1362	3,854%	
Baseline	0,1311		

Tableau VI.9. Comparaison de toutes les approches en terme de MAP sur l'ensemble de données FIRE.

La Figure VI.7 affiche les performances de récupération de l'approche proposée basé sur les métaheuristiques pour 10, 20 et 30 documents récupérés en termes de précision, rappel et F-mesure sur l'ensemble de donnée CISI.

Nous pouvons voir que les approches proposées basés sur les métaheuristiques produisent de bons résultats par rapport aux Lucene et l'approche de base, précisément pour P@30 (Figure VI.7(d)) où la F-mesure donne les valeurs les plus élevés par rapport à P@10 et P@20, par exemple PA-QR produit 0,141 pour P@10, 0,149 pour P@20 et 0,157 pour P@30. De même GA-QR produit 0,122; 0,135 ; 0,149 pour P@10, P@20 et P@30 respectivement.

Nous remarquons aussi que FA-QR rapporte les meilleurs résultats par rapport aux autres approches proposées. En termes de la précision, l'approche améliore les performances de la recherche en atteignant 0,191 par rapport à la méthode PA-QR avec 0,185, Bat-QR avec 0,180, PSO-QR avec 0,177 et GA-QR avec 0,176 pour P@30. En terme de rappel, l'approche produit 0,182 par rapport aux autres méthodes où 0,177; 0,172 ; 0,169 et 0,163 ont été produits par GA-QR, PSO-QR, Bat-QR et PA-QR.

À partir du Tableau VI.10, nous pouvons voir une nette supériorité en terme de MAP de FA-QR proposée par rapport aux Lucene (+41,43%) et Baseline (+29,03%). Nous pouvons également observer que PA-QR, Bat-QR, PSO-QR et GA-QR proposées surpasse Baseline autour de 36,03%, 33,48%, 34,63% et 32,63% ; et Lucene autour de 24,10%, 22,83%, 21,01% et 17,71% respectivement.

Approche	MAP	MAP-Gain	MAP-Gain-L
FA-QR	0,1854	41,434%	29,035%
PA-QR	0,1783	36,030%	24,105%
Bat-QR	0,1765	34,636%	22,834%
PSO-QR	0,1739	32,639%	21,012%
GA-QR	0,1692	29,027%	17,716%
Lucene	0,1389	5,938%	
Baseline	0,1324		

Tableau VI.10. Comparaison de toutes les approches en terme de MAP sur l'ensemble de données CISI.

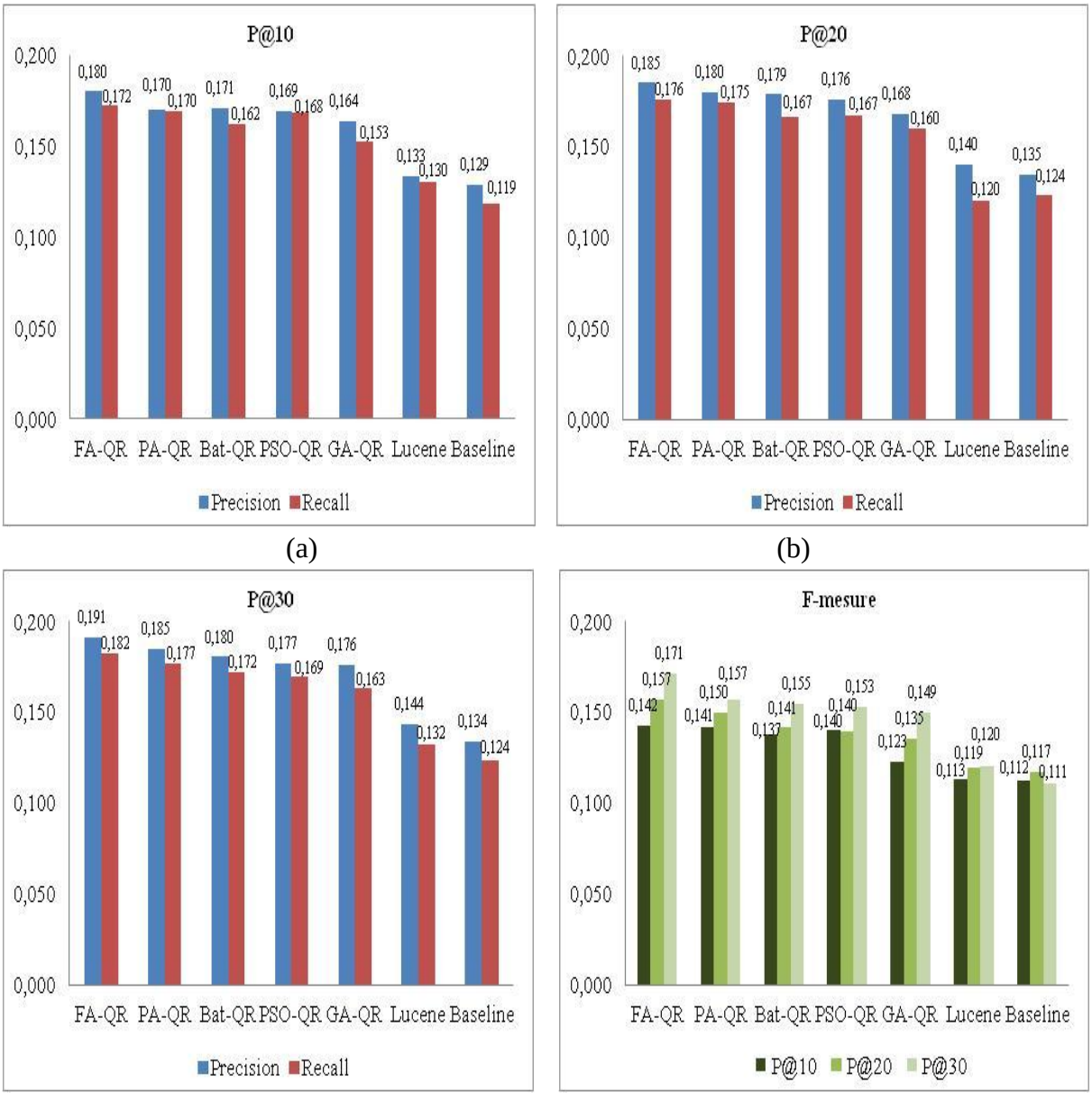


Figure VI.7. Valeurs de précision, rappel et F-mesures de toutes les approches sur l'ensemble de données CISI.

Selon les résultats expérimentaux, nous remarquons que les améliorations obtenues par les approches proposées sur l'ensemble de données TREC-3 sont un peu supérieures par rapport aux ensembles de données FIRE 2011 et CISI. Cela est probablement dû au fait que l'ensemble TREC-3 inclue des articles de presse généralement considérés comme des données textuelles de haute qualité avec moins de bruit. Au contraire, les autres ensembles de données incluent des collections de web et contiennent plusieurs sources de documents hétérogènes.

Nous observons aussi que la précision du système peut être améliorée en prenant en considération les 30 premiers résultats renvoyés pour toutes les approches et sur les trois ensembles de données.

De plus, l'utilisation de l'algorithme FireFly pour la reformulation des requêtes web améliore l'efficacité de la récupération des documents.

D'une manière globale, nos approches de reformulation des requêtes web en s'appuyant sur les métaheuristiques portent les propriétés suivantes :

- La Précision de la requête est l'une des principaux facteurs utilisés pour mesurer la qualité de la requête.
- L'algorithme d'extraction de règles d'association FP_Growth extrait touts les itemsets fréquents dont le support dépasse des seuils préfixés. Sa première phase est la plus coûteuse en termes de temps d'exécution ce qui pourra causer un inconvénient pour les utilisateurs, car le nombre de ces items fréquents dépend exponentiellement du nombre d'items manipulés (pour p items, on a 2^p itemsets possibles). Pour que cette étape soit exécutable, on doit parfois augmenter le seuil de support, ce qui donne un rôle central à la condition de support et rend plus difficile l'extraction d'items.
- Les APIs Google sont très utiles pour faire les tests, mais elles restent insuffisantes dans la phase d'évaluation, qui serait plus subjective avec les autres collections de tests. Néanmoins les évaluations prises en compte restent crédibles pour prouver l'efficacité des résultats dans le contexte de la Recherche d'Information sur le web.

PARTIE 02 : EVALUATION DE L'ETAPE DE L'ADAPTATION DES PAGES WEB

VI.6 Accès à l'application

Dans cette section nous présentons par des captures d'écran un exemple d'utilisation de notre application depuis la reformulation de la requête, jusqu'à l'adaptation de la page web.

L'accès à l'application nécessite une authentification soit en tant que 'administrateur' soit en tant que 'utilisateur'. La Figure VI.8 présente la première interface d'authentification.

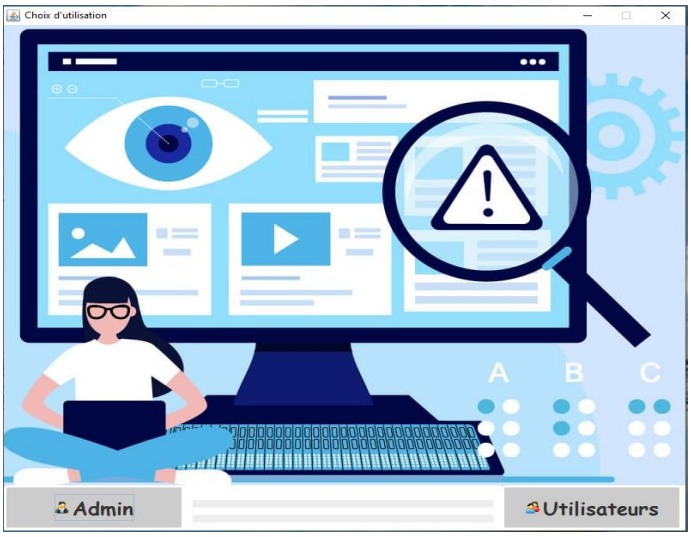


Figure VI.8. Accès à l'application.

La Figure VI.9 montre l'authentification en tant que 'administrateur'. Le compte administrateur est un utilisateur générique. Ainsi cet utilisateur a presque tous les autorisations, il peut accéder à tout ce qui se trouver sur l'application.

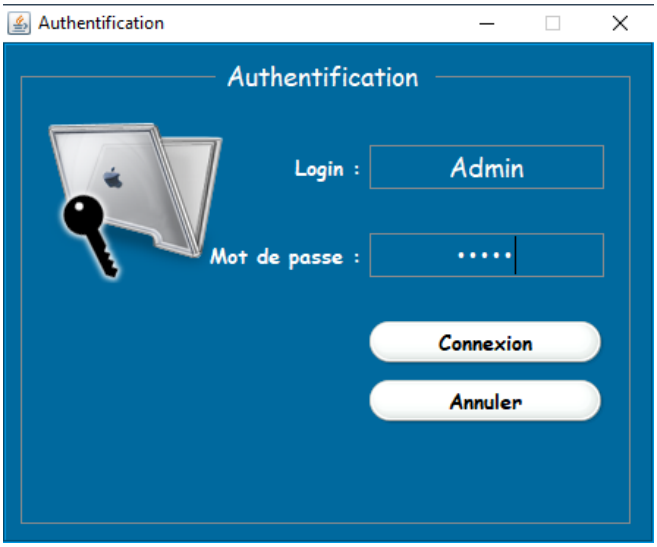


Figure VI.9. Accès à l'application en tant que administrateur.

Une fois l'administrateur est connecté, la Figure VI.10 sera affichée qui illustre l'interface des approches de reformulation proposées.

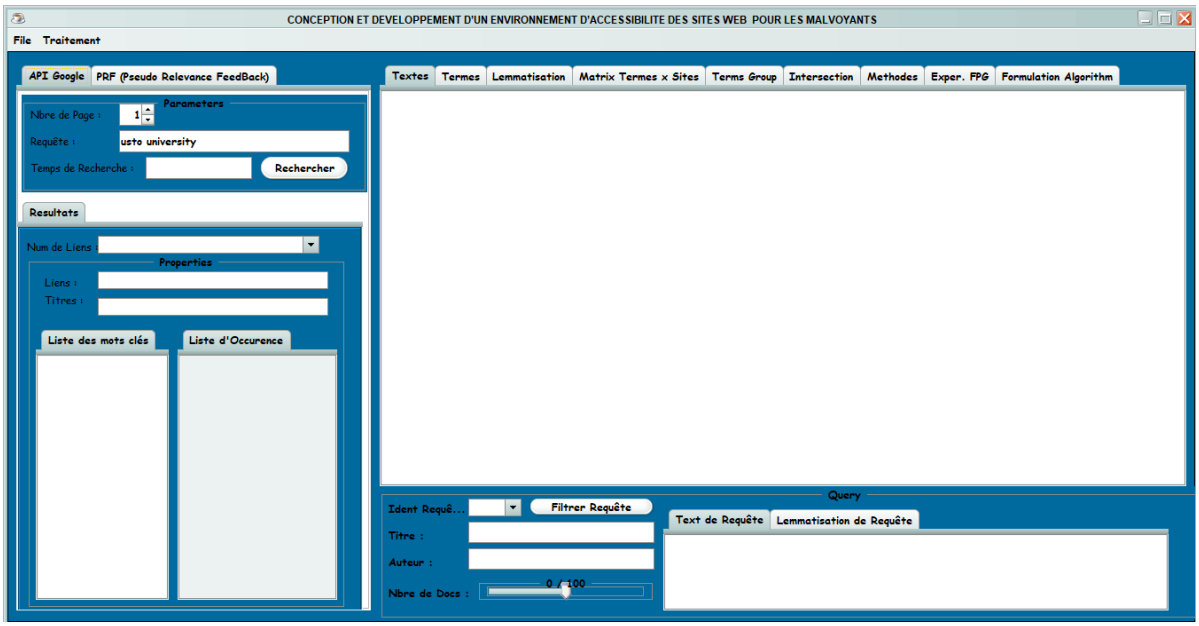


Figure VI.10. Interface des approches de reformulation.

La Figure VI.11 montre l'authentification en tant que 'utilisateur'. Cette authentification nécessite un login et un mot de passe pour y accéder si l'utilisateur possède un compte au préalable.

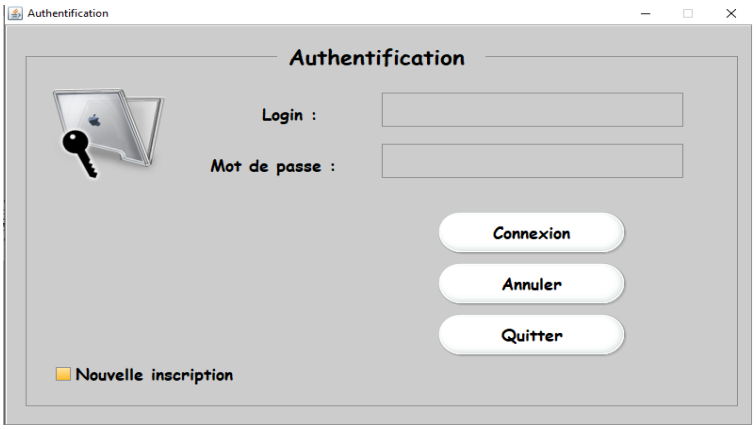


Figure VI.11. Accès à l'application en tant que utilisateur.

Sinon une création d'un compte utilisateur sera demandée, qui sollicite le remplissage du formulaire d'inscription illustré dans la Figure VI.12.

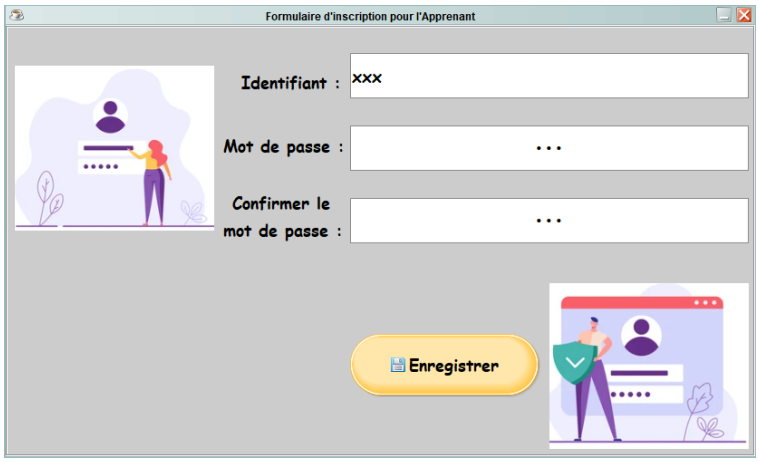
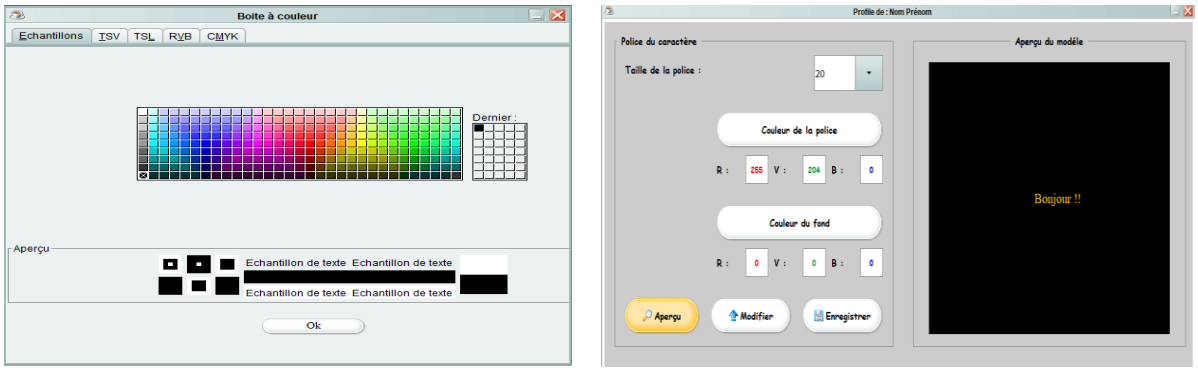


Figure VI.12. Création d'un nouveau profil utilisateur.

La Figure VI.13 montre la partie de la modification des paramètres d'affichage selon les préférences de l'utilisateur (la couleur du texte, la couleur de fond, la taille) qui vont être appliqués sur toutes les pages web qu'il va consulter.



(a) (b)

Figure VI.13. Modification des paramètres d'affichage.

La Figure VI.14 présente la navigation d'une personne malvoyantes selon les préférences choisies précédemment.

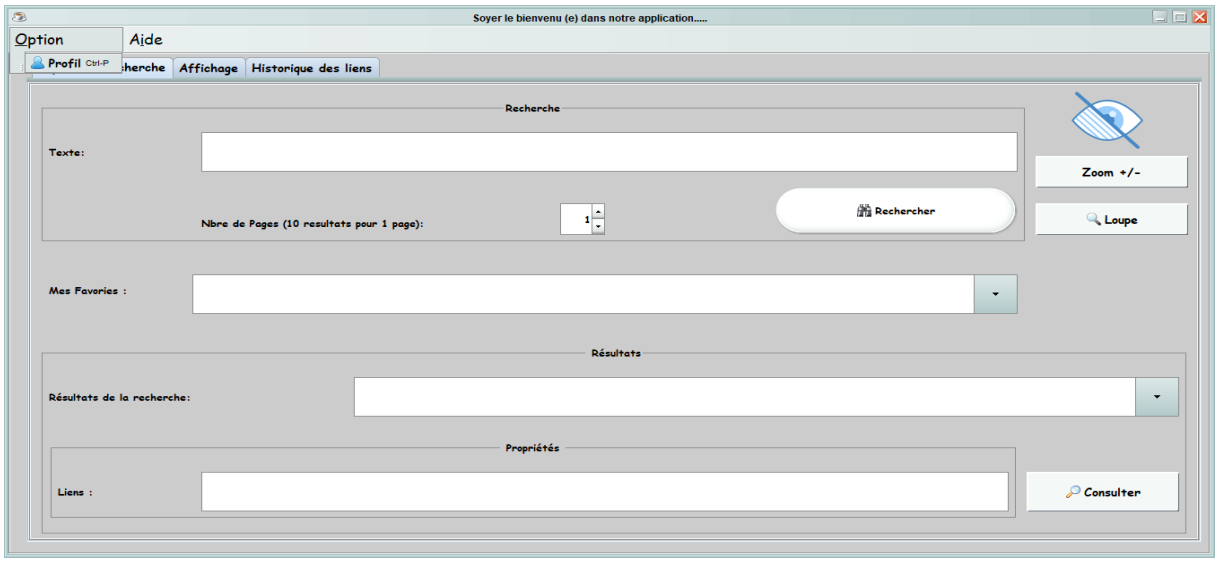


Figure VI.14. Navigation d'une personne malvoyante.

Afin de faciliter la navigation pour les personnes malvoyantes, un processus d'organisation de la page web est fait, qui est montré dans la figure ci-dessous.

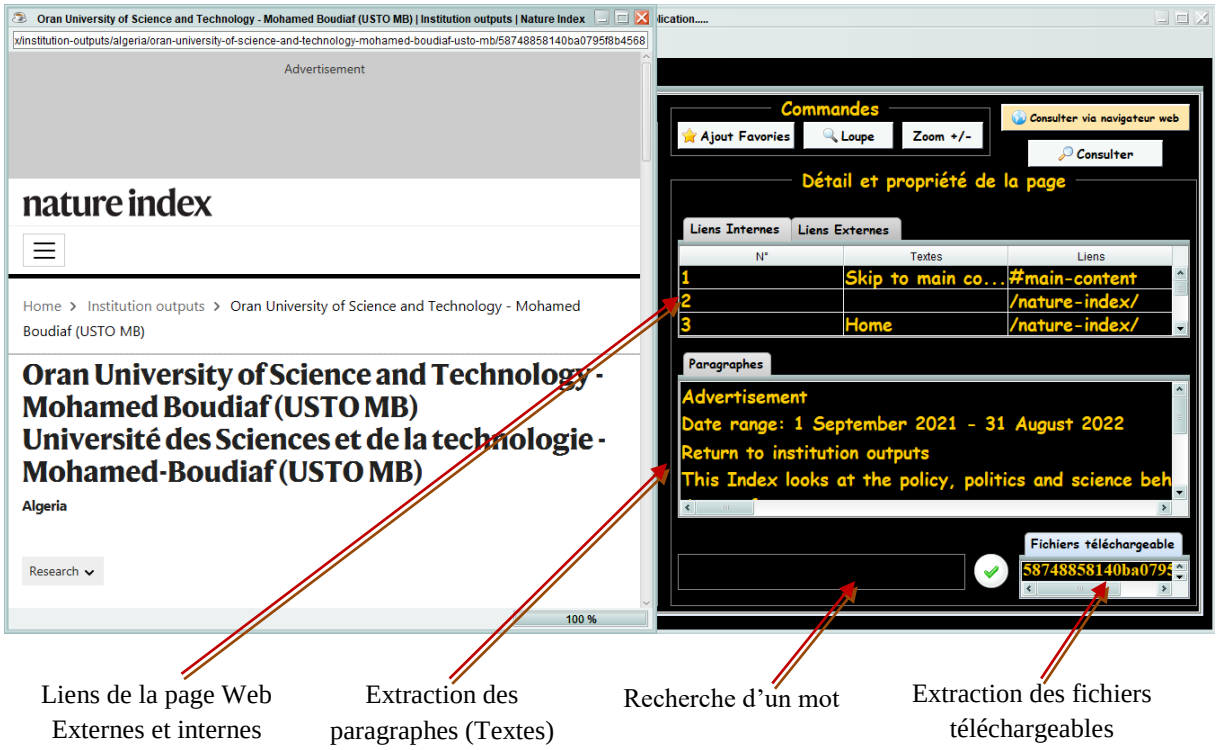


Figure VI.15. Organisation de la page web.

De plus, l'outil « loupe» peut être utilisé afin d'agrandir la page web.

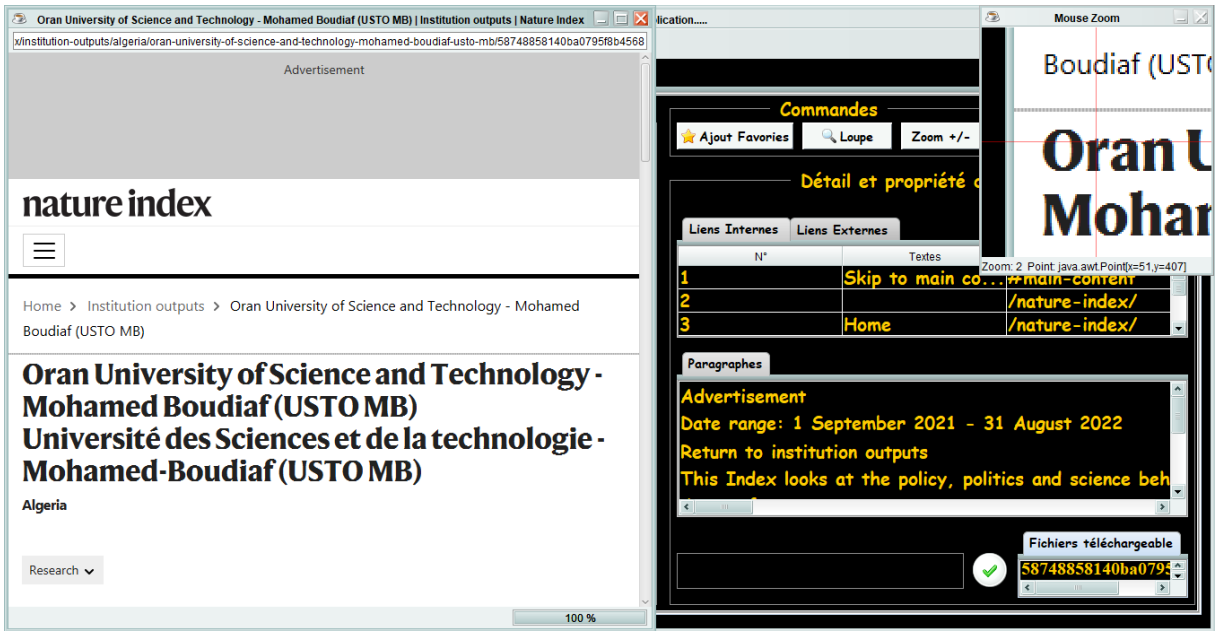


Figure VI.16. Utilisation de la loupe.

VI.7 Les raccourcis clavier

Le Tableau VI.11 présente les principaux raccourcis clavier utilisés dans notre application.

Touche de raccourci	**Description**
Alt+O	Accès à la barre de menu Option
Alt+A	Accès à la barre de menu Affichage
Alt+D	Accès à la barre de menu Aide
Alt+P	Accès au profil d'utilisateur
Ctrl + H	Aller à l'historique des liens
Alt+L	Lire un paragraphe
Alt+N	consulter un lien
Alt+F	Ajouter un lien aux favoris
Alt+Z	Lancer le Zoom
Ctrl+I	Aller aux liens internes
Ctrl+E	Aller aux liens externes
Ctrl + F	Recherche d'un mot
Ctrl+J	Aller aux fichiers téléchargeables
F1	Lire le texte
Alt+F4	Quitter

Tableau VI.11. Les raccourcis clavier utilisés.

VI.8 Etude comparative

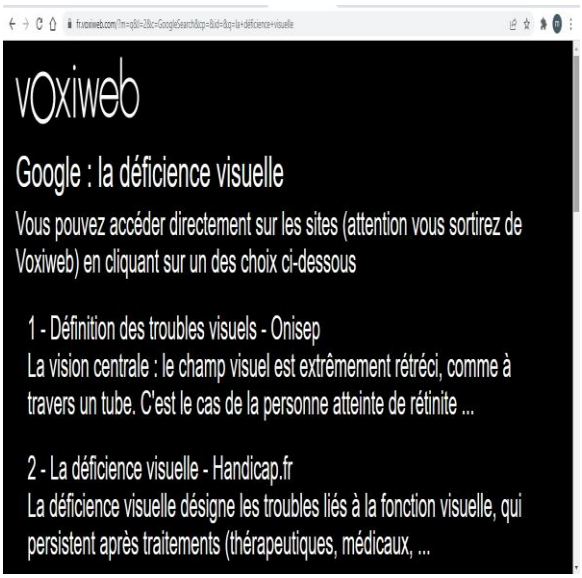
Afin d'évaluer les performances de l'approche d'adaptation des pages web pour les personnes malvoyantes, il est recommandé de la comparer avec des approches utilisant le même contexte. Néanmoins, divers critères influent sur les résultats, nous citons par exemple : le débit d'Internet, la performance de l'ordinateur, le langage de programmation, …etc. Pour cela, l'approche proposée a été comparée au WebbIE et au Voxiweb (présentés dans le chapitre II) en se basant sur le temps d'adaptation de la page web (depuis la recherche d'une requête jusqu'à l'affichage de la page web). Les résultats expérimentaux sont montrés sur la Figure ci-dessous.

Figure VI.17. Résultats expérimentaux.

Selon la Figure VI.17, nous pouvons voir évidemment que le temps d'affichage de l'approche proposée est plus rapide que WebbIE et cela à cause de l'extraction des informations utiles pour les personnes malvoyantes.

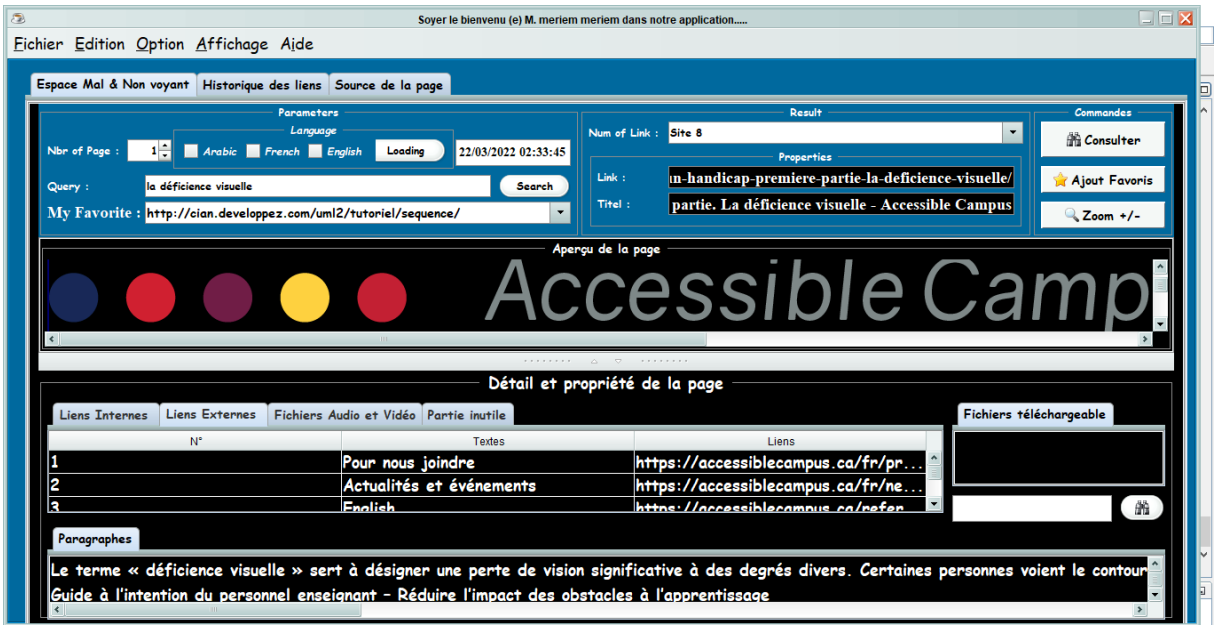
Nous observons aussi qu'il y a une amélioration significative par rapport au site d'Internet Voxiweb (accédé par Google chrome), et cela dû probablement que ce dernier est organisé sous formes de rubriques (catégories), donc pour lancer la recherche d'une requête il faut survoler les rubriques suivantes : 'Vie pratique' puis 'Au quotidien' et à la fin 'Recherche Google'.

La figure VI.18 montre un exemple d'affichage d'une page web sur les trois différentes applications : (a) Voxiweb, (b) WebbIE et (c) notre application.



Voxiweb

WebbIE



Approche proposée

Figure VI.18. Un exemple d'affichage d'une page web dans les trois applications.

De plus, afin de mieux examiner l'efficacité de notre approche par rapport aux autres applications (WebbIE et Voxiweb), un tableau comparatif entre ces dernières est présenté ci-dessous en prenant en considération plusieurs fonctionnalités.

Fonctionnalités	**Approche proposée**	**WebbIE**	**Voxiweb**
Accès gratuit	Oui	Oui	Non
Personnalisation d'affichage	Oui	Oui (prédéfinie)	Oui (prédéfinie)
Vocalisation	Oui	Non	Oui
Sauvegarde favoris	Oui	Oui	Oui
Téléchargement des fichiers	Oui	Oui	Oui
Catégorisation (rubriques)	Non	Oui	Oui
Traduction	Oui	Oui	Oui
Consultation des liens (interne et externe)	Oui	Oui	Oui
Extraction des textes à partir des images	Non	Non	Oui

Tableau VI.12. Comparaison entre les différentes applications.

Selon les résultats obtenus, nous remarquons que notre approche a porté quelques améliorations par rapport à WebbIE et Voxiweb, au niveau de la personnalisation de l'affichage et la rapidité de navigation sur le web, et ce qui répond à l'objectif de faciliter l'accès pour les personnes ayant une déficience visuelle au web en diminuant leurs temps de recherche d'information.

VI.9 Conclusion

Au cours de ce chapitre nous avons montré l'évaluation de l'approche proposée. Deux parties sont présentées ; La première partie a été destinée à la reformulation des requêtes web tandis que la deuxième partie a été destinée à l'adaptation des pages web. Les résultats obtenus montrent l'efficacité de notre approche pour les deux parties. Nous avons pu concevoir un environnement d'accessibilités des sites web pour les personnes malvoyantes selon leurs préférences.

Conclusion Générale et Perspectives

L'Internet apparaît comme un outil qui offre aux personnes déficientes visuelles (Malvoyants) un accès à l'information. Il constitue également un moyen de réaliser plus facilement et rapidement des tâches administratives.

Cependant, la facilité de diffuser une information sur le web permet la production d'une grande masse d'informations. En conséquence, les internautes déficients visuels examinent beaucoup de données afin de trier les résultats trouvés, et finalement accéder à l'information recherchée.

Néanmoins, l'accès à l'information recherchée pour ces personnes est inférieure à l'optimale, à cause de plusieurs raisons : l'absence de formation en accessibilité, le manque d'intérêt pour ces profils d'usagers, le temps de recherche, etc. …. Or, comme nous l'avons présenté, ces idées sont fausses mais subsistent.

Pour cela, une approche de reformulation est donc souvent nécessaire pour enrichir la requête d'utilisateur permettant de mieux exprimer son besoin d'une part, et d'autre part rendre l'accès aux informations web accessible et plus facile.

Notre approche de reformulation proposée est basée essentiellement sur les métaheuristiques (FireFly, Bat, PSO et Pigeon algorithms) dans le but de trouvez les documents pertinents (pages web) dans un délai raisonnable. Afin d'obtenir de meilleurs résultats, nous avons évalué notre approche sur trois différents ensembles de données TREC (Text REtrieval Conference), FIRE (Forum for Information Retrieval Evaluation) et CISI (Centre for Inventions and Scientific Information).

De plus, pour défier les difficultés de l'accessibilité des pages web récupérées, nous avons proposé une approche de personnalisation du contenu des pages web selon les préférences de l'utilisateur.

Les résultats expérimentaux montrent que les métaheuristiques proposées pour la reformulation des requêtes web réussissent à récupérer les documents pertinents, et l'approche de personnalisation du contenu des pages web a porté des améliorations au niveau de la personnalisation de l'affichage et la rapidité de navigation sur le web.

En perspectives pour notre travail on peut proposer de:

- Tester d'autres métaheuristiques récentes.
- Traiter la reformulation des requêtes web en Arabe ;
- Utiliser l'historique de navigation et le profil de l'utilisateur ;
- Ajouter la traduction ;
- Reconnaissance des objets et des textes à partir des images ;
- Intégrer l'approche proposée dans un navigateur (sous forme de Script) ;
- Créer un validateur web dédié spécialement pour les malvoyants.

Bibliographies

Abascal, J., Arrue, M., & Valencia, X. (2019). Tools for Web Accessibility Evaluation. Dans Y. Yesilada, & S. Harper, *Web Accessibility - Foundation for Research, Second Edition* (pp. 479-503). London: Springer.

Accessibilité – OS X – VoiceOVer – Apple (FR). (s.d.). Consulté le 5 11, 2016, sur http://www.apple.com/fr/accessibility/osx/voiceover/

Accessibility Valet. (s.d.). Consulté le 11 17, 2016, sur http://valet.webthing.com/access/url.html

*AChecker _ IDI Accessibility Checker_*. (s.d.). Consulté le 11 17, 2016, sur http://www.atutor.ca/achecker/

Afuan, L., Ashari, A., & Suyanto, Y. (2019). Query Expansion in Information Retrieval using Frequent Pattern (FP) Growth Algorithm for Frequent Itemset Search and Association Rules Mining. *International Journal of Advanced Computer Scien , 10* (2), 263-264.

Alazzam, H., Sharieh, A., & Sabri, K. (2020). A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer. *Expert Systems with Applications , 148*, 1-14.

Al-Khateeb, B., Al-Kubaisi, A. J., & Al-Janabi, S. T. (2017). Query reformulation using WordNet and genetic algorithm. *New Trends in Information & Communications Technology Applications (NTICT)* (pp. 91-96). Baghdad, Iraq: IEEE.

Allan, J., Callan, J., Sanderson, M., Xu, J., & Wegmann, S. (1998). INQERY at TREC-7. *The Seventh Text REtrieval (TREC-7),* (pp. 201–216).

Aminu, E. F., Oyefolahan, I. O., Abdullahi, M. B., & Salaudeen, M. T. (2018). Enhanced query expansion algorithm: framework for effective ontology based information retrieval system . *International Conference on Information and Communication Technology and Its Applications (ICTA),* (pp. 496-502). Minna, Nigeria.

Aminu, E. F., Oyefolahan, I. O., Abdullahi, M. B., & Salaudeen, M. T. (2019). Enhanced query expansion algorithm: framework for effective ontology based information retrieval system. *i-Manager's Journal on Computer Science , 6* (4), 1-11.

Arampatzis, A., Peikos, G., & Symeonidis, S. (2021). Pseudo relevance feedback optimization. *Information Retrieval Journal* .

Araújo, R. J., Reis, J. C., & Bonacin, R. (2020). Understanding interface recoloring aspects by colorblind people: a user study. *Universal Access in The Information Society , 19* (1), 81-98.

Asaju, B., Okwonu, F., Shola, P., Balogun, B. S., & Adubisi, O. (2021). A Modified Binary Pigeon-Inspired Algorithm for Solving the Multi-dimensional Knapsack Problem. *Journal of Intelligent Systems , 30* (1), 90-103.

Asakawa, C., Takagi, H., & Fukuda, K. (2019). Transcoding. Dans Y. Yesilada, & S. Harper, *Web Accessibility - Foundation for Research* (pp. 569-602). London: Springer.

association, P. (2012). *NVDA goes PDF/UA - Popular, free screen reader to provide full access to PDF for blind and vision-impaired users.* Conference de l' association PDF.

Audeh, B. (2014). Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web. Thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint-Etienne, France.

Authoring Tool Accessibility Guidelines (ATAG) 2.0. (s.d.). Consulté le 7 14, 2021, sur https://www.w3.org/TR/ATAG20/

Azad, H. K., & Deepak, K. (2019). A New Approach for Query Expansion using Wikipedia and WordNet. *Information Sciences. , 492*, 147-163.

Azad, H. K., Deepak, A., Chakraborty, C., & Abhishek, K. (2022). Improving query expansion using pseudo-relevant web knowledge for information retrieval. *Pattern Recognition Letters , 158*, 148-156.

B.McMullin, C. (5 July 2004). *« A comparative assessment of Web accessibility and technical standards conformance in four EU states »* (Vol. 9).

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). Query Recommendation Using Query Logs in Search Engines. *Current Trends in Database Technology - EDBT 2004 Workshops. 3268*, p. 588-596. Lecture Notes in Computer Science.

Baziz, M. (2005). Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information. thèse de doctorat de l'université de Paul Sabatier, Département informatique, France.

Beirekdar, A., Vanderdonckt, J., & Noirhomme-Fraiture, M. (2002). Kwaresmi—knowledge-based Web automated evaluation with REconfigurable guidelineS optimization. Dans P. Forbrig, Q. Limbourg, B. Urban, & J. Vanderdonckt, *DSV-IS 2002* (Vol. 2545, pp. 362–376). Springer, Heidelberg.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management , 31* (3), 431- 448.

Belkin, N., & Croft, W. B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin. *Communication of the ACM , 35* (12), 29-38.

Benyettou, Y., & Fizazi, H. (2020). Application of the Pigeon Method to the Classification of Captured Data. *Acta Scientific COMPUTER SCIENCES , 2* (12), 27-35.

Benzouak, A. (2019). Développement d'un environnement de travail pour les personnes mal voyantes sur le web. mémoire de magister, Université des sciences et de la technologie d'Oran Mohamed Boudiaf (USTO-MB), département informatique, Option : Modélisation, Optimisation et Evaluation des Performances des Systèmes (MOEPS), Algérie.

Berthet, L. (2013). L'avenir du braille face aux nouvelles technologies informatiques, Dans quelle mesure subit-il leurs influences ?. Diplôme universitaire de basse vision. Université Claude Bernard-Lyon 1, Institut des sciences et techniques de la réadaptation, France.

Bhatnagar, P., & Pareek, N. (2015). Genetic algorithm-based query expansion for improved information retrieval. *Intelligent Computing, Communication and Devices , 308*, 47-55.

Bhogal, J., & A., M. (2013). Ontology Based Query Expansion with a Probabilistic Retrieval Mode. Dans M. Lupu, E. Kanoulas, & F. Loizides, *Multidisciplinary Information Retrieval* (Vol. 8201, pp. 5-16). Berlin: Springer-Verlag.

Bhogal, J., Macfarlane, A., & Peter, S. (2007). A review of ontology based query expansion. *Information Processing & Management , 43* (4), 866-886.

Bhopale, A. P., & Tiwari, A. (2020). Swarm optimized cluster based framework for information retrieval. *Expert Systems with Applications , 154*, 1-16.

Bigham, J. P., & Ladner, R. E. (2007). Accessmonkey : a collaborative scripting framework for web users and developers. *the 2007 international cross-disciplinary conference on Web accessibility (W4A)* (pp. 25-34). Banff, Canada: ACM.

Bigham, J. P., Kaminsky, R. S., Ladner, R. E., Danielsson, O. M., & Hempton, G. L. (2006). Webinsight : Making web images accessible. *the 8th International ACM SIGACCESS Conference on Computers and Accessibility, ser. Assets '06.* (pp. 181–188). New York, NY, USA: ACM.

Billah, S. M., Ashok, V., Porter, D. E., & Ramakrishnan, I. V. (2018). A Locality-Preserving Magnification Interface for Low Vision Web Browsing. *the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). New York, NY, USA: Association for Computing Machinery.

Billerbeck, B., & Zobel, J. (2006). Efficient query expansion with auxiliary data structures . *Information Systems , 31* (7), 573-584.

Bonacin, R., Reis, J. C., & Araujo, R. J. (2021). An ontology-based framework for improving color vision deficiency accessibility. *Universal Access in The Information Society* , 1-26.

Bonavero, Y. (2015). Une approche basée sur les préférences et les méta-heuristiques pour améliorer l'accessibilité des pages Web pour les personnes déficientes visuelles. thèse de doctorat, Université de Montpellier, Préparée au sein de l'école doctorale I2S* et de l'unité de recherche UMR 5506, Spécialité: Informatique, France.

Bonavero, Y., Huchard, M., & Meynard, M. (2016). Algorithmes évolutionnaires pour l'adaptation personnalisée de pages Web. *IFRATH 9ème Conférence sur les Aides Techniques pour les Personnes en Situation de Handicap* (pp. 175-180). France: Handicap 2016 : La recherche au service de la qualité de vie et de l'autonomie.

Bonavero, Y., Huchard, M., & Meynard, M. (2015). Reconciling user and designer preferences in adapting web pages for people with low vision. *the 12th Web for All Conference, W4A '15* (pp. 1-10). Florence, Italy: ACM.

Bookstein, A. (1983). Outline of a general probabilistic retrieval model. *Journal of Documentation , 39* (2), 63-72.

Boughanem, M., Brini, A., & Dubois, D. (2009). Possibilistic networks for information retrieval. *International Journal of Approximate Reasoning , 50*, 957–968.

Boughanem, M., Chrisment, C., Mothe, J., Soule-Dupuy, C., & Tamine, L. (2000). Connectionist and Genetic Approaches for Information Retrieval. Dans F. Crestani, & G. Pasi, *Soft Computing in Information Retrieval, Studies in Fuzziness and Soft Computing* (pp. 173–198). Physica, Heidelberg.

Boughanem, M., Loiseau, Y., & Prade, H. (2007). Refining aggregation functions for improving document ranking in information retrieval. *International Conference on Scalable Uncertainty Management (SUM 2007)*, (pp. 255–267). Washington, DC, USA.

Breja, M., & Jain, S. (2021). Why-Type Question to Query Reformulation for Efficient Document Retrieval. *International Journal of Information Retrieval Research , 12* (1), 1-18.

Burger, D., & BrailleNet, A. (2006). L'accès au web et à la lecture numérique des publics diversement empêchés. *Bulletin des Bibliothèques de France , 51* (3), 58-63.

Cai, K., Chen, C., & Bu, J. (2009). Exploration of term relationship for bayesian network based sentence retrieval. *Pattern Recogn. Lett. , 30* (9), 805–811.

Caldwell, B., Cooper, M., Reid, L., & Vanderheiden, G. (s.d.). *Web Content Accessibility Guidelines (WCAG) 2.0*. Consulté le 5 22, 2016, sur W3C Recommendation 11 December 2008: https://www.w3.org/TR/2008/REC-WCAG20-20081211/

Carpineto, C., & R., G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys , 44* (1), 1-50.

Chen, S., & Wang, J. (2021). Individual differences and personalized learning: a review and appraisal. *Universal Access in the Information Society , 20*, 833–849.

CISI (a dataset for Information Retrieval) _ Kaggle. (s.d.). Consulté le 11 24, 2021, sur https://www.kaggle.com/dmaso01dsta/cisi-a-dataset-for-information-retrieval/metadata,

Colas, S. (2009). Individual differences and personalized learning: a review and appraisal. thèse de doctorat, Université François-Rabelais, Discipline: Informatique, France.

Croft, W. B., Cook, R., & Wilder, D. (1995). *Providing government information on the internet: Experiences with THOMAS.* University of Massachusetts.

Croft, W., & Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *of Documentation , 35* (4), 285-295.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 299–306). New York, NY, United States: Association for Computing Machinery.

Cui, H., Wen, J., Nie, J., & Ma, W. (2002). Probabilistic query expansion using query logs. *WWW '02: Proceedings of the 11th international conference on World Wide Web* (pp. 325–332). New York, NY, USA: Association for Computing Machinery.

Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2003). Query expansion by mining user logs. *IEEE Trans. Knowledge and Data Engineering , 15* (4), 829–839.

Cui, Z., Zhang, J., Wang, Y., Cao, Y., Cai, X., Zhang, W., et al. (2019). A pigeon-inspired optimization algorithm for many-objective optimization problems. *Science China Informations Sciences , 62* (70212).

Custom Search JSON API | Custom Search | Google Developers. (s.d.). Consulté le 8 5, 2017, sur https://developers.google.com/custom-search/json-api/v1/overview

Dekhici, L., & Belkadi, K. (2017). A Firefly Algorithm for the Mono-Processors Hybrid Flow Shop Problem. *International Journal of Advanced Computer Science and Applications , 8* (12), 424-433.

Dekhici, L., Redjem, R., Belkadi, K., & Mhamedi, A. E. (2019). Discretization of the Firefly Algorithm for Home Car. *Canadian Journal of Electrical and Computer Engineering , 42* (1), 20-26.

Deulkar, K., & Narvekar, M. (2015). An Improved Memetic Algorithm for Web Search. *Procedia Computer Science , 45*, 52-59.

Deveaud, R., & Bellot, P. (2012). Combinaison de ressources générales pour une contextualisation implicite de requêtes. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN* (pp. 479–486). Grenoble, France: ATALA/AFCP.

Devi, M. U., & Gandhi, G. M. (2015). WordNet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports Domain. *Journal of Computer Science , 11* (2), 361-371.

Dey, N., Chaki, J., Moraru, L., Fong, S., & Yang, X. (2020). Firefly Algorithm and Its Variants in Digital Image Processing: A Comprehensive Review. Dans N. Dey, *Applications of Firefly Algorithm and its Variants* (pp. 1-28). Singapore: Springer.

Dictionnaire, Malvoyant. (s.d.). Consulté le 8 5, 2016, sur http://www.linternaute.com/dictionnaire/fr/definition/malvoyant/

Ding, H., & Gu, X. (2020). Improved particle swarm optimization algorithm based novel encoding and decoding schemes for flexible job shop scheduling problem. *Computers & Operations Research , 121*.

Djenouri, Y., Belhadi, A., & Belkebir, R. (2018). Bees swarm optimization guided by data mining techniques for document information retrieval. *Expert Systems with Applications , 94*, 126–136.

Duan, H., & Qiao, P. (2014). Pigeon-inspired optimization: A new swarm intelligence optimizer for air robot path planning. *International Journal of Intelligent Computing and Cybernetics , 7* (1), 24-37.

El Ghali, B. (2016). Modèles Contextuels pour la Recommandation et l'Expansion de Requêtes en Recherche d'Information. thèse de doctorat, Université Mohammed V, Faculté des Science , Rabat, Maroc.

Fahsi, M., Lehirech, A., & Mimoune, M. (2010). Utilisation Du Service Web Google Dans La Reformulation Requête Par Les Algorithmes Génétiques. *RIST , 18* (1).

Fang, H. (2008). A Re-examination of Query Expansion Using Lexical Resources. *the 46th Annual Meeting of the Association for Computational Linguistics* (pp. 139–147). Columbus, Ohio, USA: The Association for Computer Linguistics.

Fang, H., & Zhai, C. (2005). An Exploration of Axiomatic Approaches to Information Retrieval. *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 480–487). New York, NY, United States: Association for Computing Machinery.

Ferati, M., Vogel, B., Kurti, A., Raufi, B., & Astals, D. (2016). Web Accessibility for Visually Impaired People: Requirements and Design Issues. Dans A. Ebert, S. Humayoun, N. Seyff, A. Perini, & S. Barbosa, *Usability- and Accessibility-Focused Requirements EngineeringUsARE UsARE 2012 2014, Lecture Notes in Computer Science* (Vol. 9312). Cham: Springer.

Fernandes, N., Kaklanis, N., Votis, K., Tzovaras, D., & Carric, L. (2014). An Analysis of Personalized Web Accessibility. *W4A '14: the 11th Web for All Conference* (pp. 1-10). New York, NY, USA: Association for Computing Machinery.

Ferretti, S., Mirri, S., Prandi, C., & Salomoni, P. (2014). Exploiting reinforcement learning to profile users and personalize web pages. *IEEE 38th Annual Computer Software and Applications Conference, COMPSAC Workshops 2014*, (pp. 252–257). Vasteras, Sweden: IEEE.

Ferretti, S., Mirri, S., Prandi, C., & Salomoni, P. (2017). On personalizing Web content through reinforcement learning. *Universal Access in the Information Society , 16*, 395–410.

Ferretti, S., Mirri, S., Prandi, C., & Salomoni, P. (2014). User centered and context dependent personalization through experiential transcoding. *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)* (pp. 486-491). Las Vegas, NV, USA: IEEE.

Forum for Information Retrieval Evaluation. (s.d.). Consulté le 3 26, 2020, sur http://fire.irsi.res.in/fire/static/data

Frakes, W. B., & Baeza-Yates, R. A. (1992). Information Retrieval Data Structures and Algorithms». *Prentice Hall PTR* , 131-160.

Freedom Scientific, JAWS,. (s.d.). Consulté le 5 10, 2016, sur http://www.freedomscientific.com/Products/Blindness/JAWS

Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information processing and management , 25* (1), 55–72.

Fuyu, W., Weining, L., & Yan, L. (2018). Variable neighborhood improved firefly algorithm for flexible operation scheduling problem. *U.P.B. Sci. Bull., Series C , 80* (2), 41-56.

Gajic, L., Cvetnic, D., Zivkovic, M., Bezdan, T., Bacanin, N., & Milosevic, S. (2021). Multi-layer Perceptron Training Using Hybridized Bat Algorithm. Dans S. Smys, J. Tavares, R. Bestak, & F. Shi, *Computational Vision and Bio-Inspired Computing.* Singapore: Springer.

Gan, L., & Hong, H. (2015). Improving Query Expansion for Information Retrieval Using Wikipedia. *International Journal of Database Theory and Application , 8* (3), 27-40.

Ganguly, D., Leveling, J., & Jones, G. (2011). Query Expansion for Language Modeling Using Sentence Similarities. Dans A. Hanbury, A. Rauber, & A. de Vries, *Multidisciplinary Information Retrieval, IRFC 2011, Lecture Notes in Computer Science* (Vol. 6653). Berlin, Heidelberg: Springer.

Garrouch, M. K. (2017). Modèles de Recherche d'information basés sur les Réseaux Bayésiens et les Réseaux Possibilistes. thèse de doctorat, Université de SFAX, Discipline: Sciences de l'informatique, Tunisie.

Gay, G. R., & Li, C. Q. (2010). AChecker: Open, Interactive, Customizable, Web Accessibility Checking. *International Cross-Disciplinary Conference on Web Accessibility—W4A 2010* (pp. 1-2). Raleigh, NC, USA: ACM.

Georgieva-Tsaneva, .., & Sabev, N. (2018). Technologies, Standards, and Approaches to Ensure Web Accessibility for Visually Impaired People. *the Digital Presentation and Preservation of Cultural and Scientific Heritage. 8*, pp. 143-150. Sofia, Bulgaria: Institute of Mathematics and Informatics.

Gonzalo, J., Verdejo, F., Chugur, I., & Cigarran, J. (1998). Indexing with WordNet synsets can improve Text Retrieval. *the COLING/ACL 98 Workshop on Usage of WordNet for NLP, cmp-lg/9808002*. Montreal.

Guilford, T., Roberts, S., & Biro, D. (2004). Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models. *Journal of Theoretical Biology , 227* (1), 25-38.

Haines, D., & Croft, W. (1993). Relevance feedback and inference network. *16th Annual International ACM SIGIR Conference on Research and development in Information Retrieval* (pp. 2-11). New York, NY, United States: ACM.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record , 29* (2), 1-12.

Hlaoua, L. (2007). Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurés. thèse de doctorat, Université Paul Sabatier de Toulouse, Spécialité : Informatique, France.

Imran, H., & Sharan, A. (2010). A Framework for Automatic Query Expansion. *Web Information Systems and Mining (WISM 2010). 6318*, pp. 386–393. Berlin: Springer.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation , 25* (1), 3-50.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. *the Seventeenth Annual International ACM SI-GIR Conference on Research and Development in Information Retrieval* (pp. 101-110). ACM Press.

Islam, M., Borodin, Y., & Ramakrishnan, I. (2010). Mixture model based label association techniques for web accessibility. *UIST '10: Proceedings of the 23nd annual ACM symposium on User interface software and technology* (pp. 67–76). New York, NY, USA: ACM.

J. Lazar, P. B. (2003). « Web accessibility in the mid-Atlantic United States: a study of 50 home pages ». *In Universal Access in the Information Society , 2* (4), pp 331-341.

Jain, M., Singh, V., & Rani, A. (2019). A novel nature-inspired algorithm for optimization: Squirrel search algorithm. *Swarm and Evolutionary Computation , 44*, 148-175.

Jensen, F. (2001). Belief Updating in Bayesian Networks. In: Bayesian Networks and Decision Graphs. Dans *Statistics for Engineering and Information Science.* New York, NY: Springer.

Jonquet, A. (2013). TactoColor: Conception et évaluation d'une interface d'exploration spatiale du web pour malvoyant. Mémoire présenté à la Faculté des Arts et des Sciences en vue de l'obtention de la Maîtrise en Communication Médiatique, Université de Montréal, Département de Communication, Canada.

K. Williamson, S. D. (2001). « The Internet for the blind and visually impaired »,. *In Journal of Computer-Mediated Communication , 7* (1), pp 0.

Kammoun, S. (2013). Assistance à la navigation pour les non-voyants : vers un positionnement, un SIG et un suivi adapté. thèse de doctorat, Université de Toulouse, Spécialité : Informatique, France.

Keikha, A., Ensan, F., & Bagheri, E. (2018). Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems , 50*, 455–478.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *IEEE International Conference on Neural Networks* (p. 1942_1948). Piscataway, NJ: IEEE.

Khennak, I., & Drias, H. (2016). A Firefly Algorithm-based Approach for Pseudo-Relevance Feedback: Application to Medical Database. *Journal of Medical Systems , 40* (240).

Khennak, I., & Drias, H. (2017a). An accelerated PSO for query expansion in web information retrieval: application to medical dataset. *Appl Intell , 47* (3), 793–808.

Khennak, I., & Drias, H. (2017b). Bat-Inspired Algorithm Based Query Expansion for Medical Web Information Retrieval. *Journal of medical systems , 41* (2).

Khennak, I., & Drias, H. (2018). Data mining techniques and nature-inspired algorithms for query expansion. *International Conference on Learning and Optimization Algorithms: Theory and Applications (LOPAL '18). 28*, pp. 1-6. New York, USA: Association for Computing Machinery.

Khennak, I., & Drias, H. (2015). Strength Pareto Fitness Assignment for Generating Expansion Features. *New Contributions in Information Systems and Technologies , 353*, 133-142.

Klampanos, I., Manning, C., Prabhakar, R., & Hinrich, S. (2009). Introduction to information retrieval. *Information Retrieval , 12*, 609–612.

Kumar, R., Tripathi, K., & Sharma, S. (2022). Optimal Query Expansion Based on Hybrid Group Mean Enhanced Chimp Optimization Using Iterative Deep Learning. *Electronics , 11* (10), 1-24.

Kumar, V., & Kumar, D. (2020). A Systematic Review on Firefly Algorithm: Past, Present, and Future. *Arch Computat Methods Eng* , 1-34.

Kunpeng, Z., Yoke, S., & Geok, S. (2009). Wavelet analysis of sensor signals for tool condition monitoring: A review and some new results. *International Journal of Machine Tools and Manufacture , 49* (7-8), 537-553.

Lee, J. (1998). Combining the evidence of different relevance feedback methods for information retrieval. *Information Processing and Management , 36* (4), 681-691.

Lei, H., Lei, T., & Yuenian, T. (2021). Sports image detection based on particle swarm optimization algorithm. *Microprocessors and Microsystems , 80*.

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing , 113* (B).

Liu, X., Zhang, D., Zhang, J., Zhang, T., & Zhu, H. (2021). path planning method based on the particle swarm optimization trained fuzzy neural network algorithm. *Cluster Computing , 24*, 1901–1915.

Lu, J., Wei, Y., Sun, X., Li, B., Wen, W., & Zhou, C. (2018). Interactive Query Reformulation for Source-Code Search With Word Relations. *IEEE Access , 6*, 75660-75668.

Lunn, D., Bechhofer, S., & Harper, S. (2008). A user evaluation of the sadie transcoder. *the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 137–144). Halifax, Nova Scotia, Canada: ACM.

Lunn, D., Harper, S., & Bechhofer, S. (2009). Combining sadie and axsjax to improve the accessibility of web content. *International Cross-Disciplinary Conference on Web Accessibility, W4A 2009* (pp. 75-78). Madrid, Spain: ACM.

Macias, M., Gonzalez, J., & Sanchez, F. (2002). On adaptability of Web sites for visually handicapped people. Dans P. B. De Bra, *Adaptive Hypermedia and Adaptive Web-Based Systems* (Vol. 2347, pp. 264–273). Berlin, Heidelberg: Springer.

Maron, M., & Kuhns, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery , 7*, 216-244.

MBROLA Voices. (s.d.). Consulté le 2 16, 2022, sur https://chromium.googlesource.com/chromiumos/third_party/espeak-ng/+/HEAD/docs/mbrola.md

Mereuta, A., Aupetit, S., Monmarché, N., & Slimane, M. (2014). Web page textual color contrast compensation for CVD users using optimization methods. *Journal of Mathematical Modelling and Algorithms in Operations Research , 13* (4), 447–470.

Minker, J., Wilson, G., & Zimmerman, B. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval , 8*, 329–348.

Mirri, S., Prandi, C., & Salomoni, P. (2013). Experiential adaptation to provide user-centered web content personalization. *IARIA The Sixth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies*, (pp. 31-36). Venice, Italy.

Mirri, S., Salomoni, P., Prandi, C., & Muratori, L. (2012). Gapforape : an augmented browsing system to improve web 2.0 accessibility. *New Review of Hypermedia and Multimedia , 18* (3), 205–229.

Misbah, F., & Maruf, P. (2016). A Non Trivial Approach for Ontology based Query Expansion Using Scoring. *Australian Journal of Basic and Applied Sciences , 10* (13), 94-100.

Mistry, K., Zhang, L., Sexton, G., Zeng, Y., & He, M. (2017). Facial expression recognition using firefly-based feature optimization. *IEEE congress on evolutionary*. Donostia-San Sebastián, Spain: IEEE.

Molina, D., Poyatos, J., Ser, J., García, S., Hussain, A., & Herrera, F. (2020). Comprehensive Taxonomies of Nature- and Bio-inspired Optimization: Inspiration Versus Algorithmic Behavior, Critical Analysis Recommendations. *Cognitive Computation , 12*, 897–939.

Moradi, P., & Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing , 43*, 117–130.

Morel, M., & Dujour, A. L. (2001). Kali », synthèse vocale à partir du texte : de la conception à la mise en oeuvre. *Traitement Automatique des Langues , 42*, 193-221.

mozdev.org – accessibar : index –. (s.d.). Consulté le 6 4, 2016, sur http://accessibar.mozdev.org/

mozdev.org – mozbraille : index –. (s.d.). Consulté le 6 4, 2016, sur http://mozbraille.mozdev.org/

Murphy, E., Kuber, R., McAllister, G., Strain, P., & Yu, W. (2008). An empirical investigation into the difficulties experienced by visually impaired Internet users. *Universal Access in the Information Society , 7* (1), 79-91.

Navigli, R., & Velardi, P. (2003). An analysis of ontology-based query expansion strategies. *14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining*, (pp. 42-49). 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining.

Nayak, D. J., Naik, B., Dinesh, P., Vakula, K., & Byomakesha, P. (2020). Firefly Algorithm in Biomedical and Health Care: Advances, Issues. *SN Computer Science , 1* (6), 1-36.

OCAWA Validateur d'accessibilité. (s.d.). Consulté le 10 13, 2017, sur http://www.ocawa.com/fr/Accueil.htm

Ogawa, Y., Morita, T., & Kobayashi, K. (1991). A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. *Fuzzy Sets and Systems , 39* (2), 163–179.

Oguego, C. L., Augusto, J., Muñoz, A., & Springett, M. (2018). A survey on managing users' preferences in ambient intelligence. *Universal Access in the Information Society , 7* (2), 97–114.

Osaba, E., Yang, X., Fister, I., Ser, J., Garcia, P., & Pardavila, A. (2019). Discrete and Improved Bat Algorithm for solving a medical goods distribution problem with pharmacological waste collection. *Swarm and Evolutionary Computation , 44*, 273-286.

Qiu, Y., & Frei, H. (1993). Concept Based Query Expansion. *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 160–169). New York, NY, USA: ACM.

Quinde, M., Khan, N., Augusto, J., Wyk, V. A., & Stewart, J. (2020). Context-aware solutions for asthma condition management: a survey. *Universal Access in the Information Society , 19* (3), 571–593.

Rasheed, I., Banka, H., & Khan, H. M. (2021). Pseudo-relevance feedback based query expansion using boosting algorithm. *Artificial Intelligence Review , 54*, 6101–6124.

Reynolds, C. (1987). Flocks, Herds, and Schools: A Distributed Behavioral Model. *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques* (pp. 25-34). New York, NY, USA: ACM.

Ribeiro-Neto, B., & Muntz, R. (1996). A belief network model for IR. *the 19th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 253–260). Zurich, Suisse: ACM.

Robertson, S., & Jones, J. (1976). Relevance weighting of search terms. *Journal of the American Siciety for Information Science , 27* (3), 129-146.

Robertson, S., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval . *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-241). Berlin: Ireland: Springer-Verlag.

Robertson, S., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-241). Ireland: Springer-Verlag; Berlin; Heidelberg.

Robertson, S., Walker, S., & Hnacock-Beaulieu, M. (1995). Large Test Collection Experiments on an Operational Interactive System: Okapi at TREC. *Information Processing and Management , 31* (3), 260-345.

Rocchio, J. (1971). Relevance feedback in information retrieval. *the SMART retrieval system-experiments in automatic document processing* (pp. 313-323). USA: ACM.

Run FAE: Functional accessibility Evaluator 2.0. (s.d.). Consulté le 11 17, 2016, sur https://fae.disability.illinois.edu/anonymous/?Anonymous%20Report=/

Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York : McGraw-Hill, Book Company.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 49–58). New York, NY, USA: ACM.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended Boolean information retrieval. *Extended Boolean information retrieval , 26* (11), 1022-1036.

Santucci, G. (2009). Vis-a-wis: Improving visual accessibility through automatic web content adaptation. Dans C. Stephanidis, *Universal Access in Human-Computer Interaction. Applications and Services* (éd. UAHCI 2009. Lecture Notes in Computer Science, Vol. 5616, pp. 787–796). Berlin, Heidelberg: Springer.

Schiavone, A. G., & Paterno, F. (2015). An extensible environment for guideline-based accessibility evaluation of dynamic Web applications. *Universal Access in the Information Society , 14* (1), 111–132.

Select and Speak - text to Speech - Chrome Web Store. (s.d.). Consulté le 6 4, 2016, sur https://chrome.google.com/webstore/detail/select-and-speak-text-to/gfjopfpjmkcfgjpogepmdjmcnihfpokn

Shah, C., & Croft, W. (2004). Evaluating High Accuracy Retrieval Techniques Chirag Shah. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 2-9). New York, NY, USA: ACM.

Sharma, D. K., Pamula, R., & Chauhan, D. S. (2019). A hybrid evolutionary algorithm based automatic query expansion for enhancing document retrieval system. *Journal of Ambient Intelligence and Humanized Computing* .

Sharma, D. K., Pamula, R., & Chauhan, D. S. (2021). Semantic approaches for query expansion. *Evolutionary Intelligence , 14*, 1101–1116.

Shrivas, N., Patil, K., Singh, S., & Shanwad, S. (2016). Email system based on voice response for visually impaired people. *International Journal of Engineering Science and Computing , 6* (3), 3056-3061.

Simonnot, B. (1996). Modélisation multi-agents d'un système de recherche d'information multimédia à forte composante vidéo. thèse de doctorat, Université Henri-Poincaré , France.

Singh, V., Garg, S., & Kaur, P. (2016). Efficient Algorithm for Web Search Query reformulation Using Genetic Algorithm. *Springer India 2016, Computational Intelligencein Data Mining, part of the Advances in Intelligent Systems and Computing book series (AISC, 410) , 1*, 459-470.

Song, F., & Croft, W. (1999). A General Language Model for Information Retrieval. *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management* (pp. 316–321). Kansas City, Missouri, USA: ACM.

Song, M., Song, B., Hu, B., & Robert, A. (2007). Integration Of Association Rules And Ontologies For Semantic Query Expansion. *Data & Knowledge Engineering , 63*, 63-75.

Soulier, L. (2014). Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs. thèse de doctorat, université de Toulouse, spécialité: Image, Information, Hypermedia, France.

Squared, A. (2009). *ZoomText 9.1 "User's Guide".*

Tamine, L. (2000). Optimisation de requêtes dans un système de recherche d'information approche basée sur l'exploitation de techniques avancées de l'algorithmique génétique. thèse de doctorat, Université Paul Sabatier - Toulouse III, Spécialité: Informatique, France.

TAW - Servicios de accesibilidad y movilidad web. (s.d.). Consulté le 11 17, 2016, sur http://www.tawdis.net/

Text Retrieval Conference (TREC). (s.d.). Consulté le 3 26, 2020, sur https://trec.nist.gov/data/test_coll.html

Thirugnanasambandam, K., Anitha, R., Enireddy, V., Raghav, R. S., Anguraj, D. K., & Arivunambi, A. (2021). Pattern mining technique derived ant colony optimization for document information retrieval. *Journal of Ambient Intelligence and Humanized Computing* .

Tibbitts, B., Crayne, S., Hanson, V., Brezin, J., Swart, C., & Richards, J. (2002). HTML parsing in Java for accessibility transformation. *XML Conference and Exposition*. Baltimore, MD U.S.A.

TotalValidator. (s.d.). Consulté le 9 27, 2016, sur https://www.totalvalidator.com/

TreeTagger - a part-of-speech tagger for many languages. (s.d.). Consulté le 8 5, 2017, sur http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Troiano, L., Birtolo, C., & Miranda, M. (2008). Adapting palettes to color vision deficiencies by genetic algorithm. *GECCO '08: the 10th Annual Conference on Genetic and Evolutionary Computation* (pp. 1065–1072). New York, NY, USA: ACM.

Tulasi, R., Meda, S., Ankita, K., & Hgoudar, R. (2017). Ontology Based Automatic Annotation: An Approach for Efficient Retrieval of Semantic Results of Web Documents. *Advances in Intelligent Systems and Computing , 507*, 331-340.

Turtle, H., & Croft, W. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems , 9* (3), 187–222.

Turtle, H., & Croft, W. (1989). Inference networks for document retrieval. *SIGIR '90: Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 1–24). New York, NY, USA: Association for Computing Machinery.

Turtle, H., & Croft, W. (1990). Inference networks for document retrieval. *ACM SIGIR Forum , 51* (2), 124–147.

User Agent Accessibility Guidelines (UAAG) 2.0. (s.d.). Consulté le 7 14, 2021, sur https://www.w3.org/TR/UAAG20/

Vaidyanathan, R., Das, S., & Srivastava, N. (2016). Query Expansion based on Central Tendency and PRF for Monolingual Retrieva. *International Journal of Information Retrieval Research (IJIRR) , 6* (4), 30-50.

Valcarce, D., Parapar, J., & Barreiro, Á. (2019). Document-based and term-based linear methods for pseudo-relevance feedback. *ACM SIGAPP Applied Computing Review , 18* (4).

Veningston, K., & Shanmugalakshmi, R. (2014). Efficient Implementation of Web Search Query Reformulation Using Ant Colony Optimization. *International Conference on Big Data Analytics* (pp. 80-94). LNCS 8883.

Vogt, C. (1999). Adaptive combination of evidence for information retrieval. thèse de doctorat, Université de California, San Diego, Spécialité: informatique.

Voorhees, E. (1994). Query Expansion using Lexical-Semantic Relations. Dans B. Croft, & C. van Rijsbergen, *SIGIR '94.* London: Springer.

VoxiWeb. (s.d.). *« VoxiWeb », Présentation.* Consulté le 3 18, 2016, sur https://www.voxiweb.com

Wang, X., & Zhai, C. (2008). Mining term association patterns from search logs for effective query reformulation. *CIKM '08: the 17th ACM conference on Information and knowledge mining* (pp. 479–488). Napa Valley, California, USA: ACM.

WAVE. (s.d.). Consulté le 11 21, 2016, sur http://wave.webaim.org/

WebbIE - le navigateur web gratuit pour les personnes aveugles ou malvoyantes. (s.d.). Consulté le 5 29, 2016, sur http://www.webbie.org.uk/fr/

Wilkinson, R., & Hingston, P. (1991). Using the cosine measure in a neural network for document. *SIGIR '91: the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 202–210). Chicago, Illinois, USA: ACM.

Wong, S., Ziarko, W., & Wong, P. (1985). Generalized vector space model in information retrieval. *the 8th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 18-25). NewYork, NY, USA: ACM.

Xu, B., Lin, H., Lin, Y., Yang, L., & Xu, K. (2018). Improving Pseudo-Relevance Feedback With Neural Network-Based Word Representations. *IEEE Access , 6*, 62152-62165.

Yang, X. (2008). *Nature-Inspired Metaheuristic Algorithms.* UK: Luniver Press.

Yang, X. S. (2010). Firefly Algorithm. Dans X. Yang, *Nature-Inspired Metaheuristic Algorithms* (pp. 81-89). United Kingdom: Luniver Press.

Yang, X. S. (2020). Firefly Algorithm: Variants and Applications. Dans A. Slowik, *Swarm Intelligence Algorithms* (p. 12). CRC Press.

Yang, X., & He, X. S. (2018). Why the firefly algorithm works? Dans X. S. Yang, *Nature-Inspired Algorithms and Applied Optimization* (pp. 245-259). Springer International Publishing AG.

Yue, X., & Zhang, H. (2020). Modified hybrid bat algorithm with genetic crossover operation and smart inertia weight for multilevel image segmentation. *Applied Soft Computing , 90*.

Yurtay, N., Bicil, Y., Çelebi, S., Çit, G., & Dural, D. (2011). Library automation design for visually impaired people. *Turkish Online Journal of Educational Technology , 10* (4), 255-260.

Zahera, H., El Hady, G., & Abd El-Wahed, W. Query Recommendation for Improving Search Engine Results. *WCECS 2010: World Congress on Engineering and Computer Science 2010 Vol I.* San Francisco, USA.

Zeboudj, M., & Belkadi, K. (2022). A Firefly Algorithm-Based Approach for Web Query Reformulation. *International Journal of Information Retrieval Research (IJIRR), 12 (2), 1-16.*

Zhang, J., Deng, B., & Li, X. (2009). Concept Based Query Expansion Using WordNet. *International e-Conference on Advanced Science and Technology*, (pp. 52-55).

ANNEXE 1 – Les algorithmes génétiques pour la reformulation des requêtes web

Les AGs sont des algorithmes d'optimisations utilisés pour obtenir la meilleure solution possible dans un espace vaste de solution. L'algorithme génétique se compose de trois opérateurs : la sélection, le croissement et la mutation.

Le principe d'un AG permet de simuler l'évolution d'une population d'individus représentant l'espace des solutions jusqu'à aboutir à un critère d'arrêt. Initialement, la population d'individus est générée de manière aléatoire. Ensuite, à chaque génération, des individus sont sélectionnés, cette sélection s'effectue suivant une fitness (fonction d'adaptation). Puis, les opérateurs de croisement et de mutation sont appliqués et une nouvelle population est créée. Ce processus est itéré jusqu'à un critère d'arrêt.

Eléments de l'AG

Inspirant par les travaux de Tamine et Boughanem (Tamine, 2000), nous avons utilisé l'algorithme génétique pour la reformulation des requêtes web (GA-QR). Dans cette section, nous présentons les éléments caractéristiques de GA-QR proposé.

- Individu : c'est une requête représenté par :

$$(\ t_1 \quad t_2 \quad \dots \ t_n)$$

$$Q_u = (\ q_{u1} \quad q_{u2} \quad \dots \ q_{un}) \qquad \text{Équation 72}$$

  Avec :

  Q_u : individu requête ;

  t_n : liste des termes (mots-clés) ;

  q_{un} : poids du terme t_i dans la requête individu Q_u ;

  Et $1 \leq n \leq 5$.

- Population : la population initiale est constituée des termes (items fréquents) générés par FP_Growth. De plus, on ajoute à chaque nouvelle génération, l'individu requête le mieux adapté résultant de la génération courante.
- Fonction d'adaptation : la performance de notre approche de reformulation est exprimée dans le cadre de l'algorithme génétique par la fitness qui permet

d'obtenir un classement suffisamment discriminant des individus; Pour cela, la mesure OkapiBM25 a été choisie comme fitness dans notre cas (Equation 47).

– Opérateurs génétiques :
  - La sélection permet d'identifier les individus qui doivent se reproduire. Cette opération ne génère pas de nouveaux individus mais identifie les individus sur la base de leur fonction d'adaptation, les individus les mieux adaptés sont sélectionnés alors que les moins bien adaptés sont écartés.
  - Le croisement : il consiste à combiner les individus sélectionnés (dits parents). Nous utilisons un croisement basé sur le poids des termes suivants:

$$Q_u = (\ q_{u1} \quad q_{u2} \quad \dots\ q_{un}) \qquad Q_v = (\ q_{v1} \quad q_{v2} \quad \dots\ q_{vn}) \longrightarrow Q_p = (\ q_{p1} \quad q_{p2} \quad \dots\ q_{pn}) \quad \text{Équation 73}$$

q_{pj}= max (q_{uj}, q_{vj}) Si Poids(t_j, D_r) ≥ Poids(t_j, D_{nr})

Sinon min (q_{uj}, q_{vj})

Avec :

Poids(t_j, D)= $\sum_{d_k \in D} d_{kj}$;

D_r : ensemble des documents pertinents;

D_{nr} : ensemble des documents non pertinents;

d_k : poids du terme t_i dans D_k.

  - La mutation : détermine la probabilité d'un gène d'individu muté après une fusion. Dans notre cas, nous avons trié par ordre décroissant les poids des termes de l'agent. Le terme de poids fort occupant la 1er position, etc.
  - Paramètres de contrôle : Les valeurs des paramètres de contrôle sont la probabilité de croisement (p_c), la probabilité de mutation (p_m) et le nombre de générations (T). Ces valeurs sont respectivement les suivantes: 0.7, 0.07 et 20.

Liste des Publications

Zeboudj, M., & Belkadi, K. (2018). Design and development of an accessibility environment for visually impaired people. *The 1st International Conference on Artificial Intelligence and its Applications AIAP'2018*, (pp. 14-19). El Oued, Algeria.

Zeboudj, M., & Belkadi, K. (2019). Development approach of an accessibility environments for the Visually Impaired. *Proceedings of the International Conference on Artificial Intelligence and Information Technology 2019 (ICA2IT'2019)*, (pp. 101-108). Ouargla, Algeria.

Zeboudj, M., & Belkadi, K. (2019). Reformulation de la Requête Web par l'algorithme FireFly. *16ème Colloque sur l'Optimisation et les Systèmes d'Informations (COSI'2019)*, Tizi-Ouzou, Algeria.

Zeboudj, M., & Belkadi, K. (2020). Web Query Reformulation Using FireFly Algorithm. *2020 Second International Conference on Embedded & Distributed Systems (EDiS 2020)* (pp. 87-90). Oran, Algeria: IEEE.

Zeboudj, M., & Belkadi, K. (2022). A Firefly Algorithm-Based Approach for Web Query Reformulation. *International Journal of Information Retrieval Research (IJIRR), 12 (2), 1-16.*

Zeboudj, M., & Belkadi, K. (2022). Designing a Web Accessibility Environment for the Visually Impaired. *3rd International Conference on Embedded & Distributed Systems (EDiS 2022)*. Oran, Algeria: IEEE. (Article Accepté)