UNIVERSITE KASDI MARBAH OUARGLA

Faculté des Nouvelles Technologies de l'information et de La Communication

Département d'Informatique Et des Technologies de L'information

Mémoire

MASTER ACADEMIQUE

Domaine : Mathématique Et Informatique.

Filière : Informatique

Spécialité : Informatique industriel

Présenté par :

KAFI IKRAM

Thème:

Expansion des requêtes avec optimisation des feux d'artifice multicouche

Soutenu publiquement le : 03/11/2020.

Devant le jury composé de :

Mahjoub M.Bachir	Président	UKM Ouargla
Mr. BEKKARI Fouad	Rapporteur	UKM Ouargla
Mr. MEZATI Messaoud	Examinateur	UKM Ouargla

Année universitaire : 2019/2020

Résumé

Généralement les utilisateurs écrivent des courtes requêtes quand ils recherchent à quelque information, cela affecte la qualité des résultats. Cela signifie que la précision de la requête est un facteur très important pour la récupération des informations. Pour atteindre cet objectif, nous avons eu recours à l'expansion des requêtes qui a démontré une efficacité en termes des résultats obtenus.

En revanche, les métaheuristiques sont des méthodes approchées qui ont prouvé une efficacité incontestable pour résoudre des problèmes difficiles où l'espace de solutions est très grand et le temps de la résolution du degré exponentiel.

Nous traiterons le problème de l'expansion des requêtes en tant que problème combinatoire et nous utiliserons l'algorithme des feux d'artifice avec la stratégie multi-couches pour résoudre ce problème, qui s'est avéré efficace pour résoudre bon nombre des problèmes.

Mots clé : L'expansion des requêtes, problème combinatoire, algorithme de feux d'artifice avec la stratégie multicouches

Abstract

Usually users write short queries when looking for some information, this affects the quality of the results. This means that the accuracy of the query is a very impor- tant factor in retrieving information. To achieve this goal, we have resorted to query expansion which has been shown to be effective in terms of the resultsobtained.

On the other hand, metaheuristics are approximate methods which have proved an indisputable efficiency to solve difficult problems where the space of solutions is very large and the time of the resolution of the exponential degree.

We will treat the query expansion problem as a combinatorial problem and we will use the fireworks algorithm with the multilayer strategy to solve this problem, which has been shown to be effective in solving many of the problems.

Keywords : Query expansion, combinatorial problem, fireworks algorithm with the multilayer strategy

Remerciements

En préambule à ce mémoire nous remerciant ALLAH qui nous aide et nous donne la patience et le courage durant ces langues années d'étude.

Nous souhaitant adresser nos remerciements les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire.

Ces remerciements vont tout d'abord au corps professoral et administratif de la Faculté des NTIC, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Nous tenant à remercier sincèrement Monsieur, Bekari Foued, qui, en tant que encadreur de mémoire, s'est toujours à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu nous consacrer et sans qui ce mémoire n'aurait jamais vu le jour.

On n'oublie pas nos parents pour leur contribution, leur soutien et leur patience. Enfin,

nous adressons nos plus sincères remerciements à tous nos proches et amis,
qui nous ont toujours encouragée au cours de la réalisation de ce mémoire.

Merci à tous et à toutes.

Table des matières

Table des figures

Introduction général

À l'heure actuelle, nous assistons à un grand développement dans tous les domaines scientifiques, y compris le domaine des technologies de l'information et de la communication, qui offre l'humanité des différents services. Cela fait l'objet de notre étude.

L'Internet est le facteur le plus important sur lequel se base les technologies de l'information et de la communication, car il est considéré comme le premier moyen de communication efficace en termes d'utilisation et également considéré comme une source importante contenant une énorme quantité d'informations diverses.

L'un des objectifs importants que recherchent les ingénieurs de l'informatique est de fournir les bonnes informations à la bonne personne au bon moment. Pour atteindre cet objectif, l'information doit être organisée scientifiquement, d'où l'importance des outils de recherche et des stratégies qui ont émergé du fait de l'abondance d'informations sur Internet. Parmi ces stratégies qui sont apparues : le système de la recherche d'informa- tions, car ce dernier joue un rôle important dans l'exécution de la tâche d'agencement, de modélisation et de restitution des informations à l'utilisateur.

En général, les internautes rédigent des courtes requêtes sur un sujet spécifique, et cela est dû à plusieurs raisons, notamment le manque de connaissances préalables sur ce sujet cela influe sur l'exactitude des résultats, d'où l'apparition de l'efficacité de la précision de la requête.

Ces problèmes ont conduit les ingénieurs de l'informatique à créer des nouvelles technologies pour répondre aux besoins de l'utilisateur en raison de la complexité de ses besoins et des sujets et de leur chevauchement.

Parmi ces techniques se trouve l'expansion des requêtes, cette technique est considérée comme l'un des problèmes difficiles car elle ne peut être appliquée de manière définitive

pour résoudre ce problème. Par conséquent, nous avons eu recours à la résolution au moyen de métaheuristiques, qui s'étaient avérées efficaces pour résoudre ce type de problème auparavant.

Il existe plusieurs algorithmes pour appliquer les métaheuristiques, parmi ces algorithmes : les algorithmes génétiques, algorithmes des feux d'artifices l'optimisation des essaims de particules. Dans notre étude nous avons opté l'algorithme des feux d'artifices avec la stratégie multicouches pour résoudre ce problème.

Notre mémoire est organisée en trois chapitres. Il commence par cette introduction générale et se termine par une conclusion générale : nous nous étendrons sur trois chapitres respectivement intitulés et détaillés comme suit :

—Le premier chapitre se divise en deux parties : la première c'est la définition de système de la recherche d'informations et ces concepts de base, et la deuxième nous définissons l'expansion de la requête et présentons leurs concepts.

—Le deuxième chapitre est présenté sous deux parties : la première explique la méthode de résolution que nous suivrons, qui est la méthode des métaheuristiques et leurs algorithmes. La deuxième représente l'algorithme de feux d'artifice avec la stratégie multicouches, qui fait l'objet de notre étude.

—Dans le troisième chapitre Nous expérimentons et aboutissons. nous présenterons notre moteur de recherche. Ce chapitre combine deux étapes : la première étape est l'indexation et la recherche àl'aide de modèle vectoriel. Le second est l'intro- duction de l'algorithme de feux d'artifice avec la stratégie multicouches. Nous présentons un modèle de la façon dont nous travaillons et présentons l'inter- face utilisateur avec précision et rappelons que nous obtenons et discutons du résultat.

Chapitre 1

La recherche d'information et leurs concepts de bases

1 Introduction

Avec l'augmentation de la taille et de la diversité des informations (non structurées) et l'augmentation de la base de données, il a été difficile pour l'utilisateur d'accéder aux informations qu'il recherche et dont il a besoin, car c'est une perte de temps qui va dans la recherche et ne peut pas non plus atteindre les informations dont il a besoin. Cela a conduit à l'émergence d'un système de recherche d'informations qui résout ce problème.

Ce chapitre est organisé en parties :
—La recherche d'information (définition et objectif).
—Le système de la recherche d'informations et leurs concepts de bases.
—Les phases de SRI (indexation, recherche).
—Et enfin on a définit la stratégie de l'expansion des requêtes.

2 La recherche d'informations

2.1 L'objectif principal de la recherche d'informations

L'objectif est de trouver un document pertinent à un besoin d'information à partir d'un grand ensemble de documents.

2.2 Définitions de RI

C'est l'ensemble des méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information pertinente pour un utilisateur [1].

RI est la science de la recherche de documents, d'informations dans documents, et pour les métadonnées sur les documents, ainsi que de recherche de bases de données relationnelles et du World Wide Web comme la figure 1.1 qui montre comment ça marche le processus de RI [1].

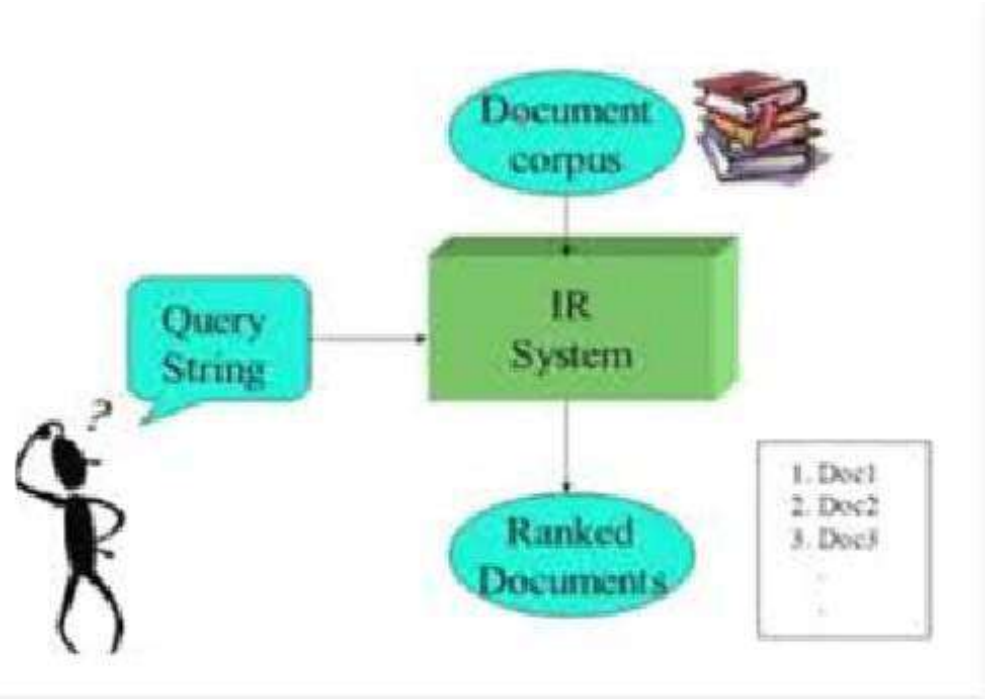


FIGURE 1.1 – Processus de la recherche d'information

3 Système de la recherche d'informations

3.1 Définition de système de la recherche d'information

—Le terme de recherche d'informations introduit pour la première fois par Calvin Mooers en 1951. Le système de recherche d'informations est une partie impor- tante du système de communication. Les principaux objectifs de la recherche d'informations sont d'offrir les bonnes informations, à la main du bon utilisateur au bon moment. Divers matériaux et méthodes sont utilisés pour récupérer les informations souhaitées [2].

—Selon Alan Smeaton : Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager [3].

—L'image suivante montre les concepts de base de SRI et leurs phases :

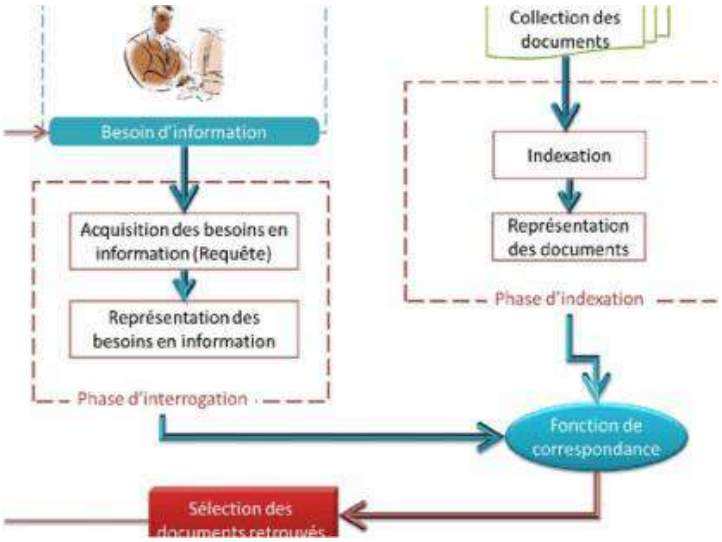


FIGURE 1.2 – Le système de la recherche d'informations

On définit un SRI comme étant un système permettant de retrouver les **documents** pertinents à une **requête** d'utilisateur écrite dans un langage libre, à partir d'une base de documents volumineuse.

3.2 Concepts de base

3.2.1 Documents

Le document est l'élément centrale du SRI, c'est un objet complexe sans cesse en évolution car il est lié aux développements des technologies de la communication .Un document peut être un texte, un morceau de texte, une page WEB, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur [4].

3.2.2 Requête

Une requête est une façon d'exprimer un besoin en information de l'utilisateur par un ensemble de mots clés, ce besoin est traduit à l'aide d'un langage naturel ou booléen [4].

3.2.3 Pertinence

La pertinence est une notion complexe et un peu floue qui dépend de l'utilisateur et de la requête mais de façon générale, le but de la RI est de retrouver seulement les documents pertinents et un document pertinent doit contenir l'information que l'utilisateur recherche. C'est sur cette notion que les SRI sont jugés [4].

3.2.4 Besoin en information

La notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur [5]. Trois types de besoin utilisateur ont été définis comme suit :

- **Besoin thématique inconnu :** cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations en dehors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète [5].
- **Besoin vérificatif :** l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit stable, c'est-à dire qu'il ne change pas au cours de la recherche [5].
- **Besoin thématique connu :** l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label [5].
- **Besoin thématique inconnu :** Cette fois, l'utilisateur cherche de nouveaux concepts en dehors des sujets ou des domaines qui lui sont habituel. Le besoin en pratique est variable et est toujours exprimé de façon incomplète [5].

3.3 Historique des systèmes de la recherche d'informations

- **1940 :** Apparition des SRI, focalisation de la RI sur les applications dans des bibliothèques.
- **1950 :** Apparition du modèle booléen et l'élaboration de petites expérimentations sur des petites collections de documents.
- **1960 et 1970** : Apparition du système SMART, Développement d'une méthodologie d'évaluation de système et conception de corpus de test(CACM).
- **1980 :** Développement de l'intelligence artificielle, ainsi on tentait d'intégrer des techniques de l'IA en RI (système expert).
- **1990 et 1995 :** L'apparition d'internet, la RI a été modifié et sa problématique plus élargie [4].

3.4 Exemples sue les systèmes de la recherche d'informations

—Internet (Web, Forum/Blog search, news)
—Entreprises (entreprise search)
—Bibliothèques numériques digital library
—Domaine spécialisé (médecine, droit, littérature, chimie, mathématique, brevets, software, …)
—Nos propres PC (Yahoo ! Desktop search)

4 Les phases de SRI

4.1 La phase d'indexation

4.1.1 Définitions

C'est un processus permettant de construire un ensemble des clés permettant de caractériser le contenu d'un document / retrouver ce document en réponse à une requête [1].

4.1.2 Comment fonctionne l'indexation?

En réalité, la table de la base de données ne se réorganise pas à chaque fois que les conditions de requête changent afin d'optimiser les performances de la requête : ce serait irréaliste. En réalité, l'index amène la base de données à créer une structure de données. Le type de structure de données est très probablement un B-Tree. Bien que les avantages du B-Tree soient nombreux, le principal avantage pour nos besoins est

qu'il est triable. Lorsque la structure des données est triée dans l'ordre, notre recherche est plus efficace pour les raisons évidentes [6].

4.1.3 Les approches d'indexation

- **Indexation manuelle :**

C'est le documentaliste ou un spécialiste du domaine qui effectue l'analyse du docu- ment, pour identifier son contenu et construire une représentation de ce contenu (choix des mots effectué par des indexeurs).Elle est basée sur un vocabulaire contrôlé (lexique, liste hiérarchiques, thésaurus, ontologie).

- **Indexation automatique**

C'est le SRI qui génère les indexes des documents. L'indexation automatique a été créée afin de remédier aux problèmes liés aux approches précédentes, elle présente l'avantage d'une régularité du processus, car l'indexation automatique fournit toujours le même index pour le même document, ce qui constitue une qualité du système. En effet, l'indexation automatique pêche par son incapacité à interpréter un texte et son manque d'adaptation à de nouveaux vocabulaires. Il est impossible de trouver dans les documents autre chose que ce que le système peut détecter [6].

4.2 La phase de la recherche (interrogation)

4.2.1 Définition

La recherche est le processus de comparaison entre la requête et tous les mots d'exis- tence en base de données avec leur document afin de récupérer des informations as- sociées à cette requête dans une base de données [7].

4.2.2 Comment fonctionne la recherche

Il est très difficile d'extraire des informations pertinentes des documents, par conséquent, ces documents doivent d'abord être représentés de manière appropriée à l'aide de n'im- porte quel modèle IR. Un tel modèle IR fournit les prémisses fondamentales et forme
la base du classement.

Les modèles de recherche d'informations définissent précisément un type de descrip- teur à partir de l'ensemble des termes utilisés pour décrire les documents, et un moyen de comparer les descripteurs pour obtenir la liste des résultats les plus similaires à la requête.

En général, les modèles RI fonctionnent sur des collections volumineuses et fixes de documents (corpus), à partir desquelles ils tentent de trouver les informations utiles qui correspondent le mieux à une requête [7].

Il existe trois catégories principales de modèle de recherche textuelle : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

5 Les modèles RI

5.1 Définition

Un modèle RI est une abstraction d'un processus de la recherche des informations [8]. Dans notre étude, nous utiliserons le modèle vectoriel.

5.2 Le modèle vectoriel

- **Idée de base**

Représenter les documents et les requêtes sous forme de vecteurs dans l'espace vec- toriel engendré par tous les termes de la collection de documents [8].

- **Les avantages**

—La pondération améliore les résultats de recherche.
—La mesure de similarité permet d'ordonner les documents selon leur pertinence vis à vis de la requête.

- **Les inconvénients**

La représentation vectorielle suppose l'indépendance entre termes

6 L'expansion des requêtes

6.1 C'est quoi une requête

Tel que défini précédemment une requête est une façon d'exprimer un besoin en information de l'utilisateur par un ensemble de mots clés, ce besoin est traduit à l'aide d'un langage naturel ou booléen [4].

6.2 Les types de requêtes

- **Requête textuelle**

L'utilisateur exprime ses besoins en fournissant un ou plusieurs mots-clés, qui peuvent être combinés à l'aide de connecteurs logiques tels que et, ou et non. Le texte peut également être des concepts d'une ontologie [9].

- **Requête par esquisse**

Dans ce cas, le système fournit à l'utilisateur des outils qui lui permettent de dessiner des images de requête [9].

- **Requête par image d'exemple**

L'utilisateur exprime ses besoins à travers une image d'exemple. Dans ce cas, il existe plusieurs manières de faire la demande : le système choisit des images aléatoires dans la base d'images et regarde l'utilisateur, l'utilisateur parcourt la base d'images et choisit une demande, l'utilisateur donne son image de demande [9].

6.3 Définition de l'expansion des requêtes

L'expansion des requêtes (QE) est un processus de recherche d'informations qui consiste à sélectionner et à ajouter des termes à la requête de l'utilisateur dans le but de minimiser la discordance requête-document et d'améliorer ainsi les performances de récupération [10].

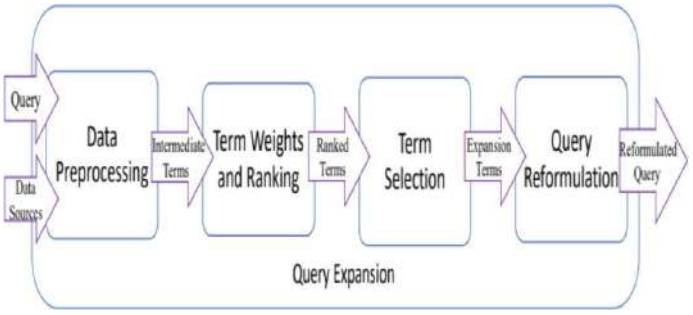


FIGURE 1.3 – Modèle de travail d'expansion de requête

6.4 Les modèles d'expansion de requête

Les modèles d'expansion de requête sont les techniques, algorithmes ou méthodologies qui reformulent la requête d'origine en ajoutant de nouveaux termes dans la requête,

afin d'obtenir une meilleure efficacité de récupération [11].

Les modèles d'expansion de requêtes peuvent être classés en trois catégories : ma- nuels, automatiques et interactifs.

- **L'expansion manuelle des requêtes**

L'expansion manuelle des requêtes repose sur les connaissances et l'expérience du chercheur dans la sélection des termes appropriés à ajouter à la requête développés.

- **L'expansion automatique de la requête**

L'expansion automatique de la requête pondère les termes candidats pour l'expansion en traitant les documents renvoyés par la récupération de premier passage et étend la requête d'origine en conséquence.

- **L'expansion interactive des requêtes**

L'expansion interactive des requêtes automatise le processus de pondération des termes, mais c'est l'utilisateur qui décide quels sont les termes de requête.

6.5 Importance de l'expansion des requêtes

L'un des principaux aspects de l'expansion des requêtes est qu'elle améliore les chances de récupérer les informations pertinentes sur Internet, qui ne sont pas récupérées autrement à l'aide de la requête originale. Bien que cela améliore le taux de rappel, tenter de récupérer un grand nombre de documents pertinents nuit à la précision [12].

6.6 Les approches de l'expansion des requêtes

Plusieurs méthodes ont été proposées où toutes ces méthodes peuvent être classées en deux groupes principaux : (1) analyse globale et (2) analyse locale.

Les techniques globales examinent les occurrences de mots et les relations dans le corpus dans son ensemble, et utilisent ces informations pour étendre toute requête particulière, étant donné leur concentration sur l'analyse du corpus, ces techniques sont des extensions de Sparck Jones (approche originale) [13]. L'analyse locale, en revanche, n'implique que les documents les mieux classés récupérés par la requête d'origine, nous l'avons appelée locale car les techniques sont des variations du travail original sur le feedback local [13].

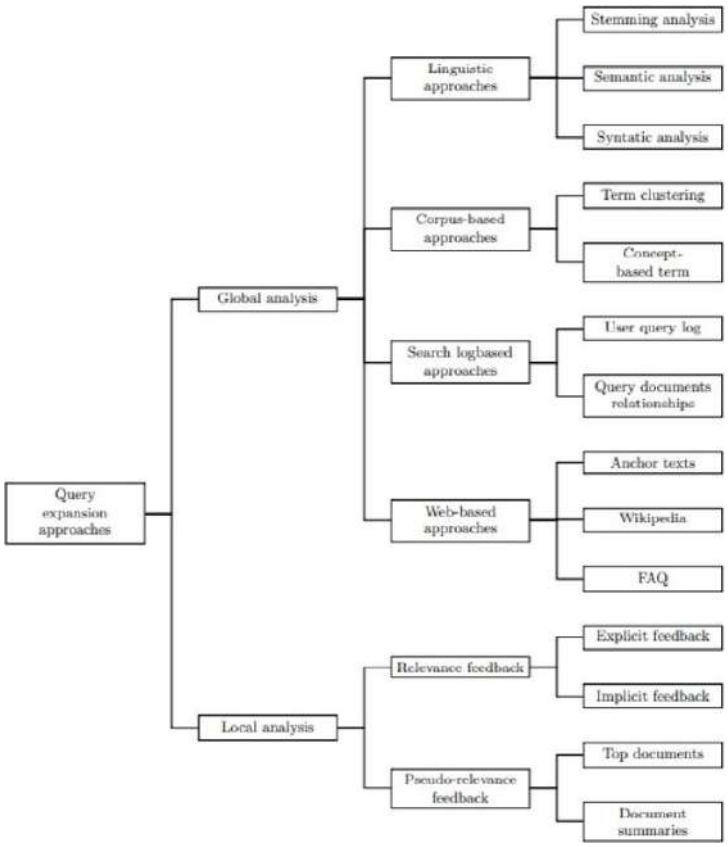


FIGURE 1.4 – Taxonomies des approches d'expansion des requêtes

6.6.1 Analyse globale

Dans l'analyse globale, les techniques d'expansion des requêtes sélectionnent impli- citement des termes d'expansion à partir de ressources de connaissances construites à la main ou de grands corpus à étendre, que l'analyse peut être classée en quatre catégories sur la base de termes de requête et de sources de données :

- **Approches linguistiques**

Les approches de cette catégorie analysent les fonctionnalités d'expansion telles que les relations lexicales, morphologiques, sémantiques et syntaxiques des termes pour reformuler ou développer les termes de la requête initiale.

- **Approches basées sur le corpus**

Examiner le contenu de l'ensemble du corpus de texte pour reconnaître les fonction- nalités d'extension utilisées pour l'expansion des requêtes.

- **Approches basées sur la recherche des journaux**

Ces approches sont basées sur l'analyse des journaux de recherche. Les commentaires des utilisateurs, qui sont une source importante pour suggérer un ensemble de termes similaires basés sur la requête initiale de l'utilisateur, sont généralement explorés sur la base de l'analyse des journaux de recherche.

- **Approches basées sur le Web**

Ces approches incluent les textes d'ancrage et Wikipedia pour étendre la requête d'origine de l'utilisateur. Le texte d'ancrage a d'abord été utilisé pour associer des hyper-liens avec des pages liées, ainsi qu'avec les pages dans lesquelles les textes d'ancrage se trouvent.

6.6.2 Analyse locale

L'analyse locale inclut des techniques d'expansion de requête qui sélectionnent les termes d'expansion à partir de la collection de documents récupérés en réponse à la requête initiale de l'utilisateur. Par conséquent, les termes présents dans ces docu- ments devraient également être pertinents pour la requête initiale. Par une analyse locale, il existe deux façons d'étendre la requête d'origine de l'utilisateur : (1) retour d'information sur la pertinence et (2) rétroaction de pseudo-pertinence [10].

- **Retour d'information sur la pertinence (Relevance Feedback)**

Si les commentaires des utilisateurs sur les documents récupérés en réponse, la requête initiale est collectée, puis la requête est reformulée en fonction des documents jugés pertinents selon les commentaires de l'utilisateur.

Les commentaires peuvent en outre être classés en deux types : les commentaires explicites et les commentaires implicites : dans les commentaires explicites, l'utilisateur évalue explicitement la pertinence des documents récupérés, tandis que dans les com- mentaires implicites, l'activité de l'utilisateur sur l'ensemble des documents récupérés en réponse à la requête initiale est utilisée indirectement déduire les préférences de l'utilisateur [11].

- **Rétroaction de pseudo-pertinence (Pseudo Relevance Feedback)**

Le processus de collecte des commentaires est automatisé en utilisant directement les documents les mieux classés (ou leurs extraits (principaux documents)), récupérés en réponse à la requête initiale, pour l'expansion de la requête. Ensuite, PRF commence

par attribuer une étiquette à chaque terme des conditions d'extension candidates en fonction de la coexistence des conditions de requête, puis sélectionnez les conditions candidates qui ont les scores les plus élevés pour réécrire la requête. Pour cette étape, la corrélation caractéristique de chaque terme d'extension est définie à l'aide de règles ambiguës, car les termes précis avec les scores les plus élevés sont sélectionnés pour étendre la requête [11].

7 Problématique

La plupart des méthodes de création de requêtes étendues permettent de trouver les bons mots à ajouter à la requête d'origine. Pour cela, nous voulions obtenir les meilleures requêtes possibles à partir d'un grand nombre de requêtes provenant d'une source, qui dans notre cas est le meilleur ensemble de mots récupérés à partir de la requête d'origine. Pour cela, nous considérerons le problème comme un problème de combinaison, et nous utiliserons l'algorithme de feux d'artifice avec la stratégie d'explosion multicouche comme approche de la solution.

8 Conclusion

Ce premier chapitre s'est basé sur l'étude de la recherche d'informations et des SRI de manière générale, nous avons présenté leurs principales notions et concepts.

D'après ce qui précède, nous comprenons que l'objectifs de RI et SRI est la permission aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information, il s'agit donc de retourner l'ensemble de documents pertinents. Nous avons également évoqué la stratégie de l'expansion des requêtes.

Dans le deuxième chapitre, nous examinerons l'approche de l'algorithme de feux d'artifice et son fonctionnement.

Chapitre 2

optimisation des algorithmes de feux d'artifice multicouches

1 Introduction

De nos jours l'optimisation est devenue un domaine indispensable pour résoudre plusieurs problèmes que se soit dans l'industrie ou d'autres secteurs. En effet nous avons assisté ces dernières années à une croissance très rapide des travaux utilisant les méthodes d'optimisation. Cette tendance peut être observée dans tous les domaines de la science.

Dans ce chapitre, on donne une étude générale sur les méthodes de résolution et leurs algorithmes dans la première partie. Et dans la deuxième partie nous nous sommes spécialisés sur l'algorithme des feux d'artifices et leurs stratégies et particulièrement la stratégie multicouches.

2 Classification des méthodes de résolution

Les méthodes de résolution peuvent être réparties en deux grandes classes :

2.1 Méthodes exactes (complètes)

Elles se basent généralement sur une recherche complète de l'espace des combinaisons afin de trouver une solution optimale. En pratique, l'application de ces méthodes est limitée, pour de petits cas des problèmes difficiles.

2.2 Méthodes approchées (incomplètes)

Elles permettent de trouver une bonne solution (pas forcément optimale) dans un temps raisonnable basées sur des méthodes stochastiques appelées heuristiques. Parmi ces heuristiques, nous trouvons des méthodes assez générales à appliquer sur un large spectre de problèmes, ces méthodes sont appelées méta-heuristiques [14].

3 Méta-heuristiques

3.1 Définition

Un méta-heuristique est formellement défini comme une génération itérative processus qui guide une heuristique subordonnée en combinant intelligemment différents concepts pour explorer et exploiter espace de recherche, des stratégies d'apprentissage sont utilisées pour structurer l'information afin de trouver efficacement des solutions quasi optimales [15].

3.2 Principales caractéristiques

—Les méta-heuristiques sont des stratégies qui permettent de guider la recherche d'une solution.
—Le but visé par les méta-heuristiques est d'explorer l'espace de recherche efficacement afin de déterminer des points (presque) optimaux.
—Les techniques qui constituent des algorithmes de type méta-heuristiques vont de la simple procédure de recherche locale à des processus d'apprentissage com- plexes.
—Les concepts de base des méta-heuristiques peuvent être décrits de manière abstraite, sans faire appel à un problème spécifique.
—Les méta-heuristiques peuvent contenir des mécanismes qui permettent d'éviter d'être bloqué dans des régions de l'espace de recherche [16].

3.3 Classification des méthodes métaheuristiques

Il existe plusieurs critères de classification des méthodes méta-heuristiques [14], selon le critère de fonctionnement on a :

3.3.1 Méthodes de trajectoire

Manipulent un seul point à la fois et tentent itérativement d'améliorer ce point. Elles construisent une trajectoire dans l'espace des points en tentant de se diriger vers des

solutions, par exemple La recherche Tabou [16].

3.3.2 Méthodes basées sur une population

Méthodes qui travaillent avec une population de points, en tout temps on dispose d'une base de plusieurs points, appelée population [16].

Il y a deux types de méthodes qui sont basées sur la population :

—Méthodes basées sur une population en utilisant le mécanisme d'évolution na- turelle (algorithmes génétiques) [16].

—Méthodes basées sur une population en utilisant l'intelligence collective (Opti- misation des essaims de particules) [16].

4 Comparaison entre les algorithmes d'intelligence

4.1 La recherche Tabou

La recherche tabou est l'une des méta-heuristiques les plus utilisées dans le domaine de l'optimisation de problèmes difficiles, elle est introduite par Glover (1986) sur la base de ses premières idées formulées dans (1977). La méthode de recherche tabou utilise l'historique de recherche pour implémenter un mécanisme pour échapper à l'optimum local, et aussi pour éviter les cycles de recherche aboutissant à une bonne exploration de l'espace de recherche.

D'une manière générale la méthode de recherche tabou se comporte comme la des- cente récursive, elle applique une recherche locale basée sur la règle d'amélioration de la fonction objectif, c'est-à-dire de choisir parmi tous les voisins de la solution commune, la meilleure solution et que elle doit être meilleure que la solution actuelle ; Si nous atteignons un optimum local, la recherche tabou nous permet d'accepter la meilleure solution qui appartient à tous les voisins même si elle a une qualité inférieure à la solu- tion actuelle, autrement dit nous avons choisi la moins mauvaise ; La recherche taboue ne s'arrête pas au premier optimum trouvé [17].

Pour résoudre le problème de la recherche cyclique, une mémoire à court terme est implémentée, cette mémoire stocke les dernières meilleures solutions déjà visitées sous la forme d'une liste appelée liste tabou, où tout mouvement vers une solution appartenant à cette liste n'est pas autorisé. A chaque itération, la meilleure solution de toutes les solutions autorisées (solutions proches de la solution actuelle et n'appartenant

pas à la liste tabou) est choisie comme nouvelle solution courante. De plus, cette solution est ajoutée à la liste des tabous et l'une des solutions qui figuraient déjà dans la liste des tabous est supprimée (généralement dans un ordre FIFO) [18].

4.2 Les algorithmes génétiques (AG)

Les algorithmes génétiques (AG) ont été inventés par John Holland dans les années 1960 et développés par Holland et ses étudiants et collègues de l'Université du Mi- chigan dans les années 1960 et 1970. Contrairement aux stratégies d'évolution et de programmation, l'objectif initial de Holland, évolutionnaire, n'était pas de concevoir des algorithmes pour résoudre des problèmes spécifiques, mais plutôt d'étudier formel- lement le phénomène d'adaptation, et comment il se produit dans la nature ; et de développer des méthodes dont les mécanismes d'adaptation normale pourraient être importés dans les systèmes informatiques. En 1975, Holland présente dans son livre Adaptation in Natural and Systèmes artificiels les algorithmes génétiques comme abs- traction de l'évolution biologique, et il a donné un cadre théorique pour l'adaptation basé sur des algorithmes génétiques [19].

Dans les algorithmes génétiques, nous essayons de simuler le processus d'évolution d'une population. Dans un premier temps, nous générons une population de solutions de manière aléatoire, cette population forme la population initiale. Le degré d'adap- tation de chaque individu à l'environnement (exprimé par la valeur de la fonction de coût appelée fonction de fitness) est calculé ; après, des couples d'individus P1 et P2 (appelés parents) sont sélectionnés en fonction de leurs adaptations ; puis un opérateur de croisement est appliqué sur P1 et P2 avec une probabilité Pc et génère des couples C1 et C2 (appelés enfants). Les autres individus P sont sélectionnés en fonction de leur adaptation ; un opérateur de mutation est appliqué sur P avec une probabilité Pm (Pm est généralement très inférieure à Pc) et génère des individus mutés P0.

Le niveau d'adaptation des enfants (C1, C2) et des individus mutés P0 est ensuite évalué avant de les insérer dans la nouvelle population ; l'insertion se fait par le choix de N individus parmi la population des parents et la population des enfants. En répétant ce processus jusqu'à ce qu'un critère d'arrêt soit atteint. Un critère d'arrêt peut être un nombre fixe d'itérations ou lorsque la population n'évolue pas plus ou moins rapidement [20].

4.3 L'Optimisation des essaims de particules (PSO)

Inventée par Russel Eberhat (ingénieur en électricité) et James Kennedy(Socio-psychologue) en 1995 [21].

Une technique évolutionnaire qui utilise une population (essaim) de solutions candi- date pour développer une solution optimale au problème d'optimisation [21].

La technique est basée sur un ensemble d'individus qui explorent l'espace de re-cherche avec un mécanisme d'interaction ; chaque individu est appelé une particule, chaque particule est caractérisée par deux propriétés principales :

—Sa position actuelle (la solution actuelle appartient à l'espace de recherche).

—Sa vitesse Chaque particule vient toujours dans deux positions, la meilleure position qu'elle ait connue dans son évaluation et la position la plus connue dans son voisinage [21].

5 Algorithmes des feux d'artifice

5.1 Définitions de l'algorithme des feux d'artifice

—L'algorithme des feux d'artifices est un nouveau groupe d'algorithmes intelli- gents développés ces dernières années basé sur le phénomène naturel de simu- lation d'étincelles de feux d'artifice, et peut résoudre une certaine optimisa-tion problèmes efficacement. Comparé à d'autres algorithmes intelligents tels que l'optimisation des essaims de particules et des algorithmes génétiques, la FWA adopte un nouveau type de mécanisme de recherche explosive, pour cal- culer l'amplitude de l'explosion et le nombre d'étincelles explosives à travers le mécanisme d'interaction entre les feux d'artifice [22].

—L'algorithme de feux d'artifice est une optimisation méta-heuristique basée sur la population algorithme qui simule le processus d'explosion de vrais feux d'ar- tifice à plusieurs reprises afin de trouver le global optimum [22].

5.2 Le cadre général de l'algorithme des feux d'artifices

Quand un feu d'artifice est déclenché, une pluie d'étincelles remplira l'espace local autour du feu d'artifice. Selon les auteurs de cet algorithme le processus d'explosion d'un feu d'artifice est considéré comme étant une recherche locale dans l'espace autours d'un point bien déterminé où le feu d'artifice est déclenché par les étincelles générées

dans l'explosion. La figure 2.1 ci-dessous présente le Framework de l'algorithme des feux d'artifice [23].

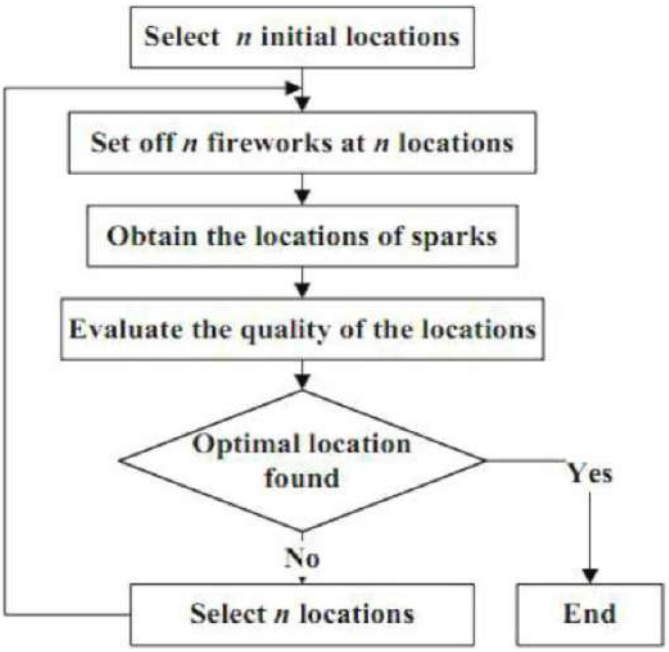


FIGURE 2.1 – Framework de l'algorithme des feux d'artifice

—Au départ, les N feux d'artifice sont configurés de manière aléatoire et leur qualité (condition physique) est évaluée pour déterminer la taille de l'explosion et le nombre d'étincelles. Pour tous les feux d'artifice.

—Par la suite, les feux d'artifice explosent et génèrent différents types d'étincelles (Sparks) dans leur espace local.

—Enfin, N feux d'artifice candidat sont sélectionnés parmi l'ensemble des candi- dats (après l'évaluation de leurs qualités), ce qui inclut les étincelles nouvelle- ment générées ainsi que les N feux d'artifice originaux.

Afin d'assurer la diversité et d'équilibrer la recherche globale et locale, l'amplitude de l'explosion et la population des étincelles d'explosion nouvellement générées diffèrent parmi les feux d'artifice. [27]

5.3 Définition d'un feu d'artifice

5.3.1 Un bon feu d'artifice (Good Firework)

Un feu d'artifice avec une meilleure forme physique peut générer une plus grande population d'étincelles d'explosion dans une gamme plus petite, c'est-à-dire avec une petite amplitude d'explosion.

5.3.2 Un mauvais feu d'artifice (Bad Firework)

Au contraire, les feux d'artifice ayant une forme physique inférieure ne peuvent générer qu'une population plus petite dans une plage plus large, c'est-à-dire avec une amplitude d'explosion plus élevée.

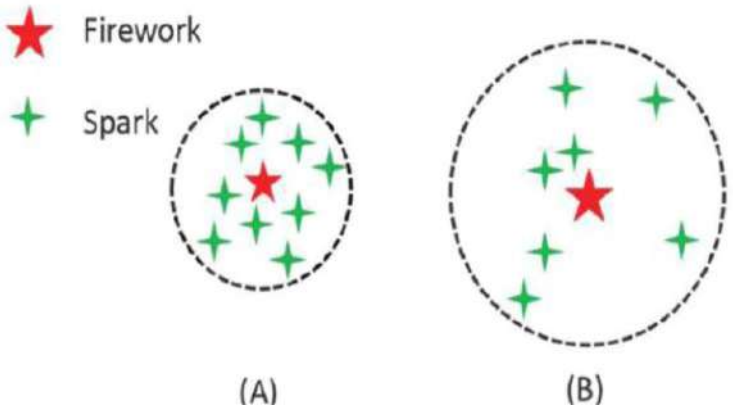


FIGURE 2.2 – Les bonnes et les mauvaises explosions d'un feu d'artifice

La figure 2.2 montre les bonnes et les mauvaises explosions d'un feu d'artifice tell que A représente une bonne explosion et B une mauvaise explosion.

Cette technique permet d'équilibrer les capacités d'exploration et d'exploitation de l'algorithme. Alors que l'exploration fait référence à la capacité de l'algorithme à ex- plorer différentes régions de l'espace de recherche afin de localiser de bonnes solutions prometteuses, l'exploitation fait référence à la possibilité de mener une recherche ap- profondie dans une zone plus petite reconnue comme prometteuse [24].

5.4 Comment ça fonctionne

Après l'explosion d'un feu d'artifice, des étincelles sont apparues autour d'un endroit. Le processus d'explosion peut être traité comme une recherche dans la zone voisine

autour d'un emplacement spécifique. Inspiré des feux d'artifice dans le monde réel, l'algorithme de feux d'artifices.

Chaque individu de la population explose et génère des étincelles autour de lui / elle. Le nombre d'étincelles et l'amplitude de chaque individu sont déterminés par certaines stratégies. De plus, une explosion gaussienne est utilisée pour générer des étincelles afin de conserver la diversité de la population. Enfin, l'algorithme conserve le meilleur individu de la population et sélectionne les autres (N-1) individus en fonction de la distance pour le prochain ; la stratégie des étincelles d'explosion imite l'explosion des feux d'artifice et constitue la stratégie de base de l'algorithme des feux d'artifice. Lorsqu'une étincelle éclate, l'étincelle disparaît et de nombreuses étincelles apparaissent autour d'elle. La stratégie des étincelles d'explosion imitant ce phénomène est utilisée pour produire de nouveaux individus par explosion. Dans cette stratégie, deux pa- ramètres doivent être déterminés. Le premier est le nombre d'étincelles.

Calculez le nombre d'étincelles

La première étape consiste à déterminer le nombre d'étincelles, son nombre est en relation directe avec la qualité du feu d'artifice, il est calculé avec la formule :

$$s_i = m.\frac{y_{max} - f(x_i) + \xi}{\sum_{i=1}^{n}(y_{max} - f(x_i)) + \xi}$$

Dans la formule s_i représente le nombre d'étincelles générées par un individu de la population, où varie de 1 à N. En tant que paramètre de contrôle du nombre total d'étincelles générées, m est défini comme une constante. Supposons que le but soit de trouver le minimum d'une fonction. La variable y_{max} représente la pire valeur de fitness de la génération actuelle, tandis que $f(x_i)$ est la valeur de fitness pour un individu x_i.

Le dernier paramètre exprimé par ξ est utilisé pour empêcher le dénominateur de devenir nul.

Calculez l'amplitude des étincelles

Le deuxième paramètre de cette stratégie est l'amplitude des étincelles, ce paramètre est inversement lié à la qualité du feu d'artifice, il est calculé avec la formule :

$$A_i = \hat{A}\frac{f(x_i) - y_{min} + \xi}{\sum_{i=1}^{n}(f(x_i) - y_{min}) + \xi}$$

La variable A_i donne l'amplitude pour qu'un individu x_i génère les étincelles d'explosion et $\hat{A}$ est une constante pour contrôler les amplitudes. La meilleure valeur de fitness y_{min} est utilisée pour calculer les amplitudes. Dans cette formule, le dernier paramètre permet d'éviter l'erreur d'avoir le dénominateur égal à zéro.

Si un individu est proche de la limite, les étincelles générées peuvent se trouver hors de l'espace réalisable. Par conséquent, une méthode de cartographie est utilisée pour garder les étincelles à l'intérieur de l'espace réalisable.

Générer des étincelles

En cas d'explosion, les étincelles peuvent subir les effets d'une explosion provenant de directions z aléatoires (dimensions). Dans l'algorithme des feux d'artifices, nous obtenons le nombre de directions affectées au hasard comme suit :

$$z = round(d.round(0, 1))$$

Où d est la dimensionnalité de l'emplacement x, et $round(0, 1)$ est une distribution uniforme sur [0, 1].

La localisation d'une étincelle du feu d'artifice x_i est obtenue à l'aide de l'algorithme 1.

Imitant le processus d'explosion, l'emplacement d'une étincelle x_j est d'abord généré. Ensuite, si l'emplacement obtenu tombe hors de l'espace potentiel, il est mappé sur l'espace potentiel.

La stratégie de cartographie garantit que tous les individus restent dans l'espace réalisable. S'il y a des étincelles éloignées de la limite, elles seront mappées à leurs portées autorisées.

$$x_j = x_{min} + |x_i|\%(x_{max} - x_{min})$$

Où x_i représente les positions de toutes les étincelles qui se trouvent hors des limites, tandis que x_{min} et x_{max} représentent la limite maximale et minimale d'une po- sition d'étincelle. Le symbole% représente l'opération arithmétique modulaire. Outre la stratégie des étincelles d'explosion, une autre façon de générer des étincelles est proposée en tant que stratégie d'étincelles gaussiennes.

5.5 Les principaux processus dans l'algorithme des feux d'ar- tifice

5.5.1 Le processus d'exploration

L'exploration est réalisée par les bons feux d'artifice qui ont une grande amplitude d'explosion, puisqu'ils ont la capacité d'échapper aux minima locaux.

5.5.2 Le processus d'exploitation

L'exploitation est réalisée par les mauvais feux d'artifice qui ont une faible ampli- tude d'explosion, puisqu'ils renforcent la capacité de recherche locale dans des zones prometteuses.

5.5.3 Le processus de mutation

Après l'explosion, un autre type d'étincelles est généré sur la base d'une mutation gaussienne des feux d'artifice choisis au hasard. L'idée derrière ceci est d'assurer davantage la diversité de l'essaim.

Afin d'améliorer la lisibilité, les fondateurs de l'algorithme assignent de nouvelles notations aux deux types d'étincelles distinctes : les étincelles d'explosion sont générées par le processus d'explosion, et les étincelles gaussiennes sont générées par une mutation gaussienne.

6 Stratégie d'explosion multicouche

Avec le développement de techniques de production, divers exquis les formes d'explosion de feux d'artifice peuvent être personnalisées, généralement là-bas, en forme de cœur explosion, explosion multicouche et explosion de zone spécifique, etc. Inspiré par diverses formes et formes d'explosion, nous introduisons d'abord un modèle d'explosion différent, une explosion multicouche pour améliorer l'utilisation du paysage local du fitness, tandis que FWA conventionnel génère des étincelles les individus autour d'un individu de feu d'artifice à la fois.

La figure 2.3 illustre le processus de génération d'étincelles en utilisant notre stratégie proposée. Nous avons fixé le nombre de couches à 2 dans cette pensée papier. Il peut être fixé à n'importe quel entier positif en théorie [25].

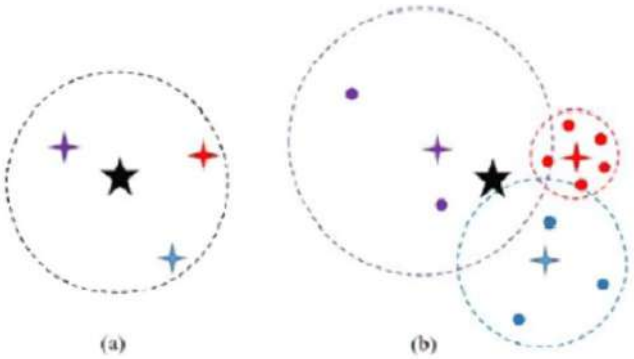


FIGURE 2.3 – Le cadre général de nos multicouches proposées stratégie d'explosion Dans la figure 2.3 (a)et (b) sont les première et deuxième explosions, respectivement

Explosion de la première couche

L'explosion d'un feu d'artifice est déterminée par deux facteurs, le nombre d'étincelles et amplitude de l'explosion. Ils déterminent individuellement comment de nombreux individus d'étincelles peuvent être générés par un feu d'artifice et sa recherche rayon (amplitude). Dans la première couche, chaque feu d'artifice détermine son rayon de recherche alloué de manière adaptative en fonction de son aptitude, qui est le identique au FWA conventionnel.

Cependant, nous traitons chaque feu d'artifice de la même manière dans la première couche, et tous les feux d'artifice génèrent le même nombre de quelques étincelles individus pour enquêter sur son paysage de fitness environnant dans son propre rayon de recherche. Ces individus ont généré des étincelles dans la première couche sont évalués avec une fonction objective et utilisés pour déterminer le forme explosive de la deuxième couche et répartition de l'étincelle individus de la deuxième couche. Le nombre d'étincelles générées les individus pour chaque feu d'artifice de la première couche sont décidés de manière adaptative selon sa forme physique [25].

Explosion de couche ultérieure

Définissons les symboles temporellement pour en expliquer l'idée principale papier, explosions multicouches. N est le nombre d'individus de feux d'artifice ; $\boldsymbol{f}_i$ est le i-ème individu de feu d'artifice ; m est le nombre total d'étincelles. m est le nombre maximum de couches d'explosion ; m_i est le nombre total des étincelles dans toutes les explosions de couche sous le $\boldsymbol{f}_i$; m^k est le nombre des étincelles dans la k-ème couche d'explosion ;

i

$s_{i,j}^{(k)}$ Est le j-ème individu d'étincelle ($j = 1 \sim m^{(k)}$) généré dans l'explosion de la k-ième couche 180 sous le f_i. Les relations entre eux sont $m = \sum_{i=1}^{N} m_i$ et $m_i = \sum_{k=1}^{i} m_i^{(k)}$

Le i-ème sous-groupe est formé par le i-ème feu d'artifice individus f_i , et toutes ses étincelles $s_{i,j}^{(k)}$, Figure 2.4. Puisque les paramètres de recherche dans chaque couche d'explosion sont décidés indépendamment dans chaque sous-groupe, nous expliquons le processus de multi-explosion au i-ème sous-groupe dans le premier problème clé est de savoir comment distribuer le nombre restant d'étincelles $m_i - N * m^{(1)}$, aux couches d'explosion restées. Le total nombre d'étincelles dans les couches d'explosion suivantes est $m_i - m_i^{(1)} = \sum_{k=2}^{l} m^{(k)}$. Nous distribuons simplement un nombre égal des étincelles à chaque couche $\frac{(m_i - m_i^{(1)})}{l-1}$. Après tous les individus de feu d'artifice terminent leurs premières explosions, leurs secondes explosions sont déclenchées par pas les individus de feu d'artifice mais leurs individus d'étincelle générés $s_{i,j}^{(1)}$. Le deuxième problème clé est de savoir comment décider du rayon de recherche autour $s_{i,j}^{(k)}$ et comment diviser $\frac{(m_i - m_i)}{l-1}$ explosion étincelle à chacun $s_{i,j}^{(k)}$ dans chaque couche.

Ces deux paramètres sont déterminés par l'aptitude du $s_{i,j}^{(k)}$ de manière adaptative. Cette explosion de couche est répétée jusqu'à ce que les explosions se répètent l fois [25].

6.1 Le cadre général de l'explosion multicouche proposée stratégie

For i = 0; i < n; i + + do
Decide the number of generated sparks, $m_i^{(1)}$, in the first layer for each firework.
Decide a search radius around the i-th firework according to its fitness.
Conduct the first layer explosion for each firework.
While the number of explosions does not reach a pre-defined maximum layer do
For j = 0; $j < m_i^{(k)}$; j + + do
Decide the number of sparks generated by the *j*-th spark in the previous layer.
Decide a search radius of around the *j*-th spark in the previous layer.
End for
Generate the next explosion sparks for
Each spark in the previous layer.
End while
End for
End of explosion.

FIGURE 2.4 – Le cadre général de l'explosion multicouche proposée stratégie

Notre stratégie proposée divise les étincelles en explosion multicouche, et les explosions séquentielles élargissent les zones de recherche vers de meilleures directions progressivement couche par couche en fonction de l'aptitude des étincelles dans chaque couche explosion, tandis que l'aptitude d'un individu de feu d'artifice décide de toutes les étincelles immédiatement. Il est important de noter que notre stratégie d'explo- sion proposée il suffit de changer la couche des explosions sans changer aucune autre opérations, c'est-à-dire génération d'individus étincelles et mutation opération. La fi- gure 5 montre le déroulement de notre explosion proposée stratégie. Lorsqu'il est com- biné avec d'autres versions du FWA, uniquement leur opération d'explosion correspon- dante est remplacée [25].

7 Conclusion

Au début de ce chapitre, nous avons introduit les méthodes de résolution qui nous a conduit à aller vers les méta-heuristiques que nous connaissions et mentionné leurs classes, nous sommes passés directement à la zone qui nous intéresse l'algorithme des feux d'artifices qui contient plusieurs stratégies pour ouvrir plusieurs problèmes, y compris la stratégie multicouches dont nous proposerons des travaux pour étendre la requête standard l'utilisant. Dans le chapitre suivant, nous appliquerons l'algorithme de feux d'artifice avec la stratégie multicouches dans le processus de l'expansion des requêtes.

Chapitre 3

Expérimentation et résultat

1 Introduction

Dans ce chapitre, nous avons d'abord une explication détaillée des données où nous avons donné une explication simplifiée de tous les outils utilisés : langage de program- mation, bibliothèques que nous avons saisies, nous sommes passés à la définition de la base de données utilisée dans ce travail, puis l'étape de traitement passe par deux phases : Nous avons d'abord indexé les documents en utilisant Modal Victor où les documents de ce modèle sont placés dans Victor afin que chaque mot dans une boîte de celui-ci et ces mots soient connus par indexation spéciale, à la fin de l'indexation vient l'étape de recherche où nous saisissons la requête et effectuons une recherche dans le dictionnaire, qui est un grand ensemble de mots placés dans Victor sous Vic- tor, comparant les mots entre la requête et les mots du dictionnaire et récupérons le résultat.

La seconde consiste à extraire les meilleurs mots (poids ou plus important) des docs récupérés de la première phase et de les mettre en préparation pour l'algorithme de feux d'artifice, ce dernier est basé sur quatre étapes (nous avons expliqué dans le deuxième chapitre en détail). Nous sommes ensuite passés à la visualisation de l'interface utilisateur et à la discussion des résultats, et en conclusion, nous avons mentionné ce que nous fournirons pour le développement de ce travail à l'avenir.

2 Implémentation

2.1 Langage de programmation

Python est un langage de programmation interprété de haut niveau pour la program- mation à usage général. Créé par Guido van Rossum et publié pour la première fois en 1991, Python a une philosophie de conception qui met l'accent sur la lisibilité du code, notamment en utilisant des espaces blancs importants. Il fournit des constructions qui
permettent une programmation claire à la fois à petite et à grande échelle [26]. La grande bibliothèque standard de Python, communément citée comme l'une de ses plus grandes forces, fournit des outils adaptés à de nombreuses tâches. Le référentiel officiel des logiciels Python tiers contient plus de 130 000 packages avec un large éventail de fonctionnalités [27]. Nous avons utilisé des packages de collecte TKinter, math, numpy, nltik.

De plus, nous utilisons des listes, des tuples :

- **Listes :** Une liste est une structure de données en Python qui est une séquence ordonnée et mutable d'éléments. Chaque élément ou valeur qui se trouve à l'intérieur d'une liste est appelé un élément. les listes sont définies par des valeurs entre crochets [28].
- **Tuples :** Un tuple est une séquence d'objets Python immuables. Les tuples sont des séquences, tout comme les listes. Les différences entre les tuples et les listes sont que les tuples ne peuvent pas être modifiés contrairement aux listes et les tuples utilisent des parenthèses, tandis que les listes utilisent des crochets [29].
- **Dictionnaire :** tableau d'éléments indexés par types immuables auxquels des éléments peuvent être ajoutés ou supprimés [30].
- **Ensemble :** tableau d'éléments uniques non indexés [30].

2.2 La base de données

Le 20 groupes de discussion base de données est une collection d'environ 20 000 documents de groupes de discussion, répartis uniformément sur 20 groupes de discussion différents. Tel que je l'ai trouvé, il a été recueilli à l'origine par Ken Lang, bien qu'il ne soit pas explicitement mentionné.

Il est devenu un ensemble de données commun pour les expériences dans les applica- tions textuelles des techniques d'apprentissage automatique, telles que la classification de texte et la synthèse de texte. Chaque sous-répertoire de l'ensemble représente un

groupe de discussion, chaque fichier d'un sous-répertoire est le texte d'un document de groupe de discussion qui a été publié dans ce groupe de discussion.

Le premier (”19997”) est l'original, le second (”bydate”) est trié par date dans les ensembles de formation (60%) et de test (40%), le troisième (”18828”) n'inclut pas les cross-posts et inclut uniquement les en-têtes De et Objet [31].

Dans ce travail, nous utilisons le premier ensemble de données qui est l'original (”19997”), Contient du document 1554 de type (.TXT).

3 La configuration de notre travail

3.1 Première phase : La recherche

se compose de deux étapes :

3.1.1 Le stage de l'indexation

Sur la première étape, l'algorithme forme ce genre d'index, ou plus précis pour dire une concordance car il contient le terme avec le renvoi pour les retrouver dans le texte.

Normalisation

Une partie importante de l'indexation est la normalisation. Il s'agit d'un traitement de texte, qui met le texte source sous une forme canonique standard. Cela signifie que les mots vides et les articles sont supprimés, les signes diacritiques (comme dans les mots pâté , naïf , zloty) sont supprimés ou remplacés par des signes alphabétiques standards.

En outre, un seul cas est choisi (uniquement supérieur ou inférieur).

Un autre élément important de la normalisation est la lutte. Il s'agit d'un processus de réduction d'un mot en forme de racine ou de base. Par exemple, pour les mots manger , manger , la forme de la tige mangé est manger . Comme si la demande de recherche les végétaliens mangeant du pâté de viande capturé sur bande se transforme en bande de pâté de viande végétalienne . De plus, il est très important de spécifier la langue dans laquelle l'algorithme fonctionne.

3.1.2 Le stage de la recherche

Une fois cet index créé, l'algorithme de recherche analyse l'index au lieu de l'ensemble de documents d'origine et expose les résultats. Comme vous l'avez remarqué, cette approche demande beaucoup de temps pour créer un index, mais il est alors beaucoup plus rapide de rechercher des informations dans les documents à l'aide d'index que de simples méthodes de recherche de chaînes.

Dans le domaine de la recherche d'informations, cette étape est basée sur des modèles pour atteindre l'objectif, où les modèles travaillent sur des ensembles de documents volumineux et fixes (corpus), à travers lesquels nous pouvons trouver des informations utiles qui correspondent le mieux à la requête de recherche.

Dans ce travail, nous nous sommes appuyés sur un modèle d'espace vectoriel, qui est un modèle algébrique, impliquant deux étapes, dans un premier temps nous représentons les documents texte en vecteur de mots et dans un second temps nous nous transfor- mons en format numérique afin que nous puissions appliquer toutes les techniques de **text mining** comme la recherche d'informations, l'extraction d'informations, le fil- trage d'informations, etc. Le vecteur de document s'écrit, où W_i est le poids du terme T_i qui indique son importance. Si le document D ne contient pas de terme T_i, alors le poids W_i est nul.

Dans ce modèle, il existe plusieurs méthodes de mesure, que nous avons utilisé mesure $tf - idf$.

Dans cette approche, les termes reçoivent une pondération basée sur la fréquence à laquelle un terme apparaît dans un document particulier et la fréquence à laquelle il apparaît dans l'ensemble de la collection de documents.

La première partie du schéma tf-idf est appelée le terme fréquence, le nombre d'occurrences du terme dans le document D. La deuxième partie est appelée la fréquence inverse du document et est calculée comme suit :

$$Idf_i = \log[\frac{n}{df_i}]$$

Où n est le nombre total de documents de la collection et df_i le nombre de documents dans lesquels le terme apparaît au moins une fois. Le facteur de pondération W_i du document i est déterminé par le produit du terme fréquence et de la fréquence inverse

du document :

$$W_i = tf_i - idf_i$$

Les hypothèses derrière $tf - idf$ sont basées sur deux caractéristiques des documents texte. Premièrement, plus un terme apparaît dans un document, plus il est pertinent pour le sujet du document. Deuxièmement, plus un terme apparaît dans tous les do- cuments de la collection, plus il discrimine les documents.

3.2 Seconde phase : Implémentation de l'algorithme des feux d'artifices avec la stratégie multicouches

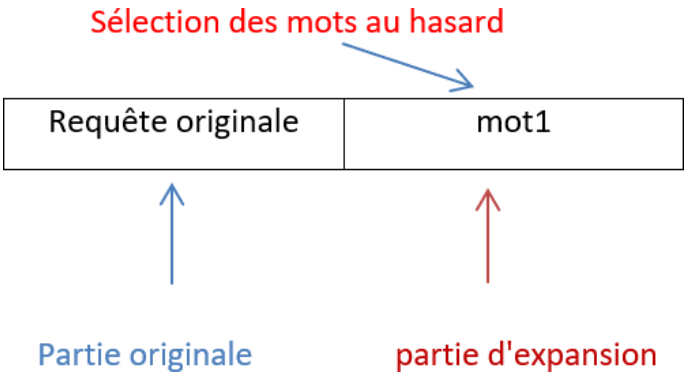
Lorsqu'un feu d'artifice est déclenché, une pluie d'étincelles remplira l'espace local autour du feu d'artifice. Ici, ces étincelles passent par quatre étapes, qui sont les étapes de l'algorithme :

3.2.1 Initialisation

Pour notre travail, nous avons utilisé l'approche de rétroaction de pseudo perti- nence, en récupérant les meilleurs documents de la requête d'origine, puis en extrayant les meilleurs mots de ces documents et en les utilisant comme point de départ de l'algorithme.

Exemple :

Dans ce travail, nous avons ajouté un mot à titre d'exemple :



3.2.2 Explosion et évaluation

- **Explosion de la première couche**

Nous avons sélectionné un nombre fixe de mots que nous ajouterons à la requête originale en utilisant une fonction pour extraire des combinaisons aléatoires. Après la sélection fixé des mots et ajoutés à la requête, où les étincelles apparaissent dans leur site au début puis éclatent des feux d'artifice où la transaction est une recherche dans le voisinage autour d'un site spécifique.

Cette explosion s'applique à chaque membre de la population, entraînant la génération d'étincelles autour de chaque individu, avec cela, les deux sont calculés :

Le nombre d'étincelles :

$$s_i = m.\frac{y_i - f(x_i) + \xi}{\sum_{i=1}^{n}(y_{max} - f(x_i)) + \xi}$$

L'amplitude de chaque individu :

$$A_i = \hat{A}\frac{f(x_i) - y_{min} + \xi}{\sum_{i=1}^{n}(f(x_i) - y_{min}) + \xi}$$

Exemple :

Par exemple, supposons que nous obtenions 4 nouvelles étincelles et 1 amplitude, ce qui signifie que nous avons produit quatre requêtes, mais chacune a un mot différent de la première requête étendue.

Requête originale	Mot 1

Partie d'expansion 1

Requête originale	Mot 2

Partie d'expansion 2

Requête originale	Mot 3

Partie d'expansion 3

Requête originale	Mot 4

Partie d'expansion 4

- **Explosion de couche ultérieure**

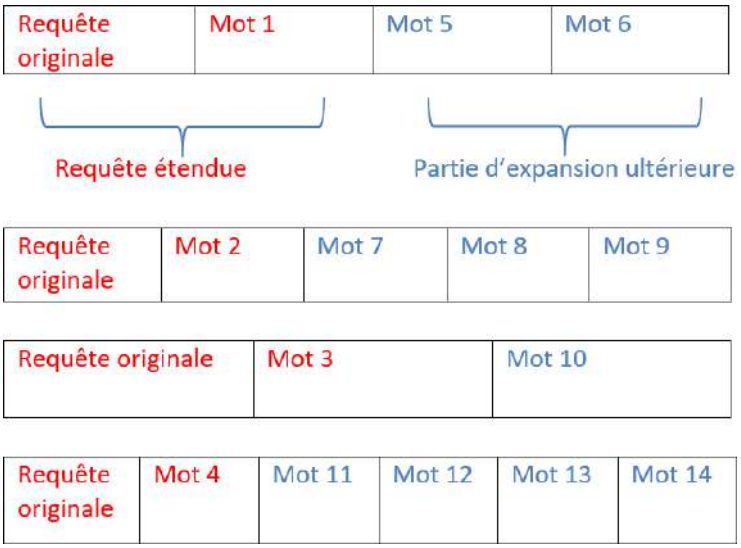
Dans la couche ultérieure nous avons sélectionné un nombre aléatoire de mots que nous ajouterons à la requête étendue d'après la première couche en utilisant une fonc- tion pour extraire des combinaisons aléatoires.

Après la sélection aléatoire des mots et ajoutés à la requête étendue, où les étincelles apparaissent dans leur site pour chaque individu puis éclatent des feux d'artifice où la transaction est une recherche dans le voisinage autour d'un site spécifique.

Cette explosion s'applique à chaque membre de la population, entraînant la génération d'étincelles autour de chaque individu, et on calcule le nombre d'étincelles et l'ampli- tude de chaque individu.

Exemple :

De l'exemple précédent, nous avons 4 requêtes étendues, mais chacune a un mot différent. Dans cette couche nous avons sélectionné un nombre aléatoire de mots que nous ajouterons à la requête étendue d'après la première couche



3.2.3 La sélection

Le sens de la sélection : Nous sélectionnons les membres de la prochaine génération qui résultent de l'explosion d'étincelles, où initialement le meilleur individu est retenu, tandis que le reste, n - 1 individu est sélectionné en fonction de la distance entre chaque individu et les autres, où l'opportunité prendre chacun dépend de l'individu le plus éloigné des autres individus.

4 Tests et résultats

Pour la requête as on a avec notre modèle le résultat suivant :

```
**********************************************
0.6061599417222193
['aa', 'nobodi', 'keith', "'m", 'live', 'you']
The docid is 0007 and the weight is 0.6061599417222193
The docid is 0014 and the weight is 0.1462215349710298
The docid is 0016 and the weight is 0.0922971481168074
The docid is 0024 and the weight is 0.0667980267930268
The docid is 0003 and the weight is 0.0566155645385623
The docid is 0000 and the weight is 0.0553714660796463
The docid is 0008 and the weight is 0.0506249340321034
The docid is 0001 and the weight is 0.0471103139877056
The docid is 0002 and the weight is 0.0469020180594595
The docid is 0004 and the weight is 0.0468548653189391
The docid is 0005 and the weight is 0.0360474568883082
The docid is 0021 and the weight is 0.0221376151068236
The docid is 0015 and the weight is 0.0194760080673000
The docid is 0011 and the weight is 0.0068132756325607
The docid is 0013 and the weight is 0.0065968850919868
>>>
```

FIGURE 3.1 – Résultat d'exécution (1)

Mais avec le modèle standard on a :

```
['as', 'by', 'be', 'violat', 'avoid']
The docid is 0015 and the weight is 0.4725576054930076
The docid is 0009 and the weight is 0.2106832451970357
The docid is 0000 and the weight is 0.1891183899929213
The docid is 0001 and the weight is 0.1215690902837797
The docid is 0002 and the weight is 0.1210266006452746
The docid is 0014 and the weight is 0.0338602460107407
The docid is 0016 and the weight is 0.0276823801297472
The docid is 0008 and the weight is 0.0221068067719196
The docid is 0020 and the weight is 0.0205900437985397
The docid is 0011 and the weight is 0.0175810684331462
The docid is 0007 and the weight is 0.0127786449038104
The docid is 0012 and the weight is 0.0117851375138893
The docid is 0023 and the weight is 0.0092154557314862
The docid is 0013 and the weight is 0.0075358753856305
The docid is 0024 and the weight is 0.0065508777956277
```

FIGURE 3.2 – Résultat d'exécution (2)

5 Conclusion

Nous avons trouvé quelques difficultés dans la construction de l'algorithme des feux d'artifice avec la stratégie multicouches, mais les résultats sont encourageants, sa- chant que le processus de développement doit se poursuivre afin d'obtenir de meilleurs modèles.

Conclusion général

La simple utilisation des outils de recherche conduit à l'émergence de nombreux problèmes liés à la qualité des informations, y compris l'obtention d'information qui n'a pas de relation par rapport aux besoins de l'utilisateur, ce qui signifie que l'ob- jectif principal des outils de recherche n'est pas atteint. Ces problèmes ont incité les ingénieurs de l'informatique à utiliser des technologies modernes pour atteindre cet objectif, y compris les systèmes de recherche d'informations.

Dans notre travail nous avons expliqué les concepts de la recherche d'informations et nous avons abordé la technique de l'expansion de la requête avec l'approche rétroaction de pseudo-pertinence (Pseudo Relevance Feedback), qui est un moyen d'augmenterla qualité des informations récupérées, d'où nous avons considéré comme un problème difficile dans sa solution (application) et ne peut être résolu de manière définitive, et à partir de là nous avons eu recours à la méthode métaheuristique et nous avons donné quelques exemples (les algorithmes génétiques, la recherche Tabou, PSO).

Parmi ces exemples qui nous intéressent est l'algorithme de feux d'artifice avec la stratégie des multicouches. Dans cette section, nous expliquons les différentes étapes qu'un algorithme traverse pour le mettre en œuvre. D'après les résultats obtenus, on peut dire que : cette méthode est assez efficace car dans tous les cas, elle augmente le degré de corrélation entre la requête étendue et les résultats obtenus.

Il est à noter que la méthode utilisée est sensible à de nombreux facteurs, dont la qualité du prétraitement des documents, point où on n'est pas allé plus loin et aussi le modèle de recherche où la plupart des travaux reposent sur le modèle Victor. Les autres facteurs qui influencent l'algorithme sont les mêmes facteurs qui peuvent être contrôlés par de nombreuses expériences et analyses.

Étant donné que l'omission des relations sémantiques entre les mots peut être une rai- son de réduire l'efficacité de la méthode, l'ajustement des paramètres fait une différence de telle manière que nous cherchons à améliorer, nous pouvons également souligner la

nécessité de l'étape tribale, qui est le traitement des textes que nous espérons faire dans nos futurs travaux car il augmente l'efficacité du chemin.

Bibliographie

[1] Mohand boughanem, introduction à la recherche d'information 2014, université paul sabatier de toulouse.

[2] www.lisbdnet.com/information-retrieval-syste/# :˜ :text=informationvisite le 12/10/2020.

[3]A.f. smeaton. "information retrieval and natural language processing". in procee- dings of a conference jointly sponsored by aslib, university of york, page 2, march 1989.

[4]La recherche d'information-introduction -master micr paris 13-recherche et extraction d'information- a.rozenknop source : Romaric besançon cea-list/lic2m.

[5]P. ingwersen. "polyrepresentation of information needs and semantic entities : elements of a cognitive theory for information retrieval interaction". in proceedings of the seventeenth annual international acm sigir conference on research and de- velopment in information retrieval., pages 101-110, 1994.

[6]https ://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2002-1-page-2.htm# visité le 2/5/2020.

[7]C.d. manning, p.raghavan and h.schütze, introduction to information retrieval, cambridge university press. 2008.

[8] https ://www.irit.fr/ mohand.boughanem/slides/ri/chap4-mod-bool-vect.pdf vi- site le 22/10/2020.

[9] Melle amira laoubi. recherche d'images sémantique basée sur la sélection des concepts 2014.

[10]He b. (2018) query expansion models. in : Liu l., Özsu m.t. (eds) encyclopedia of database systems. springer, new york.

[11]M. m. rahman, s. k. antani, and g. r. thomas, a query expansion framework in image retrieval domain based on local and global analysis, inf process manag. 2011 sep 1 ; 47(5) :676–691.

[12]H.kumar, a.akshay deepak, query expansion techniques for information retrieval : a survey,1 aug 2017.

[13]M.song (new jersey institute of technology, usa) and y.brook wu (new jersey insti- tute of technology, usa), handbook of research on text and web mining technologies (2 volumes), september, 2008.

[14]https ://www.slideshare.net/sanaaroussi3/chapitre-4-heuristiques-et-mta-heuristiques. visit le 5/10/2020.

[15]S. voß, s. martello, i.h. osman and c. roucairol (eds), meta-heuristics - advances and trends in local search paradigms for optimization. kluwer academic publishers, dordrecht, the netherlands, (1999).

[16]Introduction aux métaheuristiques /s.le digabel ecole polytechnique de montréal 2018.

[17]J.-k hao, p. galinier and m. habib. méthaheuristiques pour l'optimisation combi- natoire et l'affectation sous contraintes. revue d'intelligence artificielle vol : No. 1999.

[18]C. blum and a. roli. metaheuristics in combinatorial optimization : Overview and conceptual comparison. acm computing surveys, vol. 35, no. 3, september 2003, pp. 268–308.

[19]M. melanie. an introduction to genetic algorithms. a bradford book the mit press cambridge, massachusetts. london, england fifth printing, 1999.

[20]J.-m. alliot and n. durand. algorithmes génétiques. march 14, 2005.

[21]M. settles an introduction to particle swarm optimization department of computer science, university of idaho, moscow, idaho u.s.a 83844 november 7,2005.

[22]Tan, y. ; zhu, y. fireworks algorithm for optimization. in advances in swarm in- telligence, proceedings of the 2010 international conference in swarm intelligence, beijing, china.

[23]Das, s., abraham, a., konar, a. : Swarm intelligence algorithms in bioinformatics. studies in computational intelligence 94, 113–147 (2008).

[24]Y. tan and y. zhu, "fireworks algorithm for optimization," in advances in swarm intelligence, vol. 6145, y. tan, y. shi, and k. c. tan, eds. berlin, heidelberg : Springer berlin heidelberg, 2010, pp. 355–364.2010; springer : Berlin/heidelberg, germany, 2010.

[25]International journal of swarm intelligence and evolutionary computation, multilayer-explosion-based- fireworks algorithm 2018.

[26]Python. "https ://www.python.org/about/ ", visité le 20 /10 /2020.

[27]Open classroom. "https ://openclassrooms.com/courses/apprenez-a-programmer-enpython/qu-est-ce-que-python", visité 20/ 10/ 2020.

[28]https ://www.digitalocean.com/community/tutorials/understanding-lists-in- python-3 visité 20/ 10/ 2020.

[29]https ://www.tutorialspoint.com/python/python _tuples visité 20/ 10/ 2020.

[30]http ://qwone.com/ ~jason/20newsgroups/ visité 20/ 10/ 2020.

[31]http ://www.xavierdupre.fr/app/ensae teaching cs/helpsphinx/notebooks/ code liste tu ple.html visité 20/ 10/ 2020.