ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre **3471**

THESE

En vue de l'obtention du : ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications
Discipline : Sciences de l'Ingénieur
Spécialité : Informatique

Présentée et soutenue le : 05 juin 2021
par:

Omar EL MIDAOUI

Reformulation sémantique des requêtes pour la recherche d'informations géographiques sur le Web: Usage de taxonomies, l'analyse sémantique latente, et des règles d'associations générées dans un contexte Big Data

JURY

Salma MOULINE	*PES, Université Mohammed V – Rabat, Faculté des Sciences*	*Présidente*
Majid BENYAKHLEF	*PES, USMBA– Fès, Université Sidi Mohammed Ben Abdellah*	*Rapporteur/Examinateur*
Fatima GUEROUATE	*PH, Université Mohammed V – Rabat, École Supérieure de Technologie de Salé*	*Rapporteur/Examinateur*
Hicham LAANAYA	*PH, Université Mohammed V – Rabat, Faculté des Sciences*	*Rapporteur/Examinateur*
Fadoua ATAA ALLAH	*Expert, Directrice de recherche, L'Institut Royal de la Culture Amazighe, Rabat*	*Examinatrice*
Moulay Driss RAHMANI	*PES, Université Mohammed V – Rabat, Faculté des Sciences*	*Directeur de Thèse*
M. Abderrahim EL QADI	*PES, Université Mohammed V – Rabat, École Nationale Supérieure d'Arts et Métiers*	*Co-Directeur de Thèse*

Année Universitaire : 2020-2021

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

وَقُلِ اعْمَلُوا فَسَيَرَى اللَّهُ عَمَلَكُمْ وَرَسُولُهُ وَالْمُؤْمِنُونَ

-سورة التوبة-

Dédicace

A vous ***chers parents****, vous qui m'aviez offert la vie. Vous qui étiez là pour mes premiers pas, aujourd'hui j'espère vous rendre fiers. Que Dieu vous Bénisse et vous accordes les plus haut paradis pour vos sacrifices et vos encouragements.*

A toi ***ma femme*** *et meilleur amie, tu n'as cessé de me soutenir, m'encourager et me rendre heureux dans les moments les plus difficiles de ma vie. Je voudrai que tu sois fier de moi, je t'aime.*

A vous, mes très belles sœurs, ***Nawal****,* ***Fadwa****,* ***Lamya*** *et* ***Omaima****, Je vous dédie ce travail en témoignage de ma gratitude et mon attachement. Vous qui étiez là à partager tous les moments de ma vie, les meilleurs et les plus difficiles, aujourd'hui j'espère vous rendre fiers., je vous aime trop et vous espère le meilleure dans cette vie et dans l'autre inchalah.*

A toi, ma ***grand-mère****, merci pour tes précieuses prières, J'espère te rendre fière.*

A vous ma deuxième famille: ***famille EL Ghali****. J'ai fait votre connaissance à travers la personne qui m'est la plus chère au monde, je vous remercie de m'avoir accepté entre vous et de me considérer comme votre quatrième fils.*

A toute ma famille ; mes tantes, mes oncles, mes cousines, mes cousins, mes nièces.

A vous mes meilleurs amies, ***Ahmed El Kouch****,* ***Anass Bezza****,* ***Imane Zaimi*** *et* ***Meryem EL Mouhtadi****, vos encouragements et les moments que nous avons passés ensemble ont été d'une grande aide pour moi.*

A tous mes amies et mes collègues du laboratoire LRIT,

Ce travail vous est dédié. Votre présence dans ma vie a fait de moi ce que je suis aujourd'hui.

El Midaoui Omar

Avant-propos

Les travaux présentés dans ce mémoire ont été effectués au sein du « Laboratoire de Recherche en Informatique et Télécommunications » à la faculté des sciences de Rabat, sous la direction de **M. Moulay Driss RAHMANI** (PES à la faculté des sciences de Rabat) et la co-direction du **M. Abderrahim El QADI** (PES à l'École Nationale Supérieure d'Arts et Métiers de Rabat).

Avant toute chose, Je tiens à remercier tous les membres dirigeant du **laboratoire LRIT**, de m'avoir offert cette agréable opportunité d'intégrer leur univers de la recherche scientifique, et pour tous leurs efforts indéniables qui visaient comme objectif de pouvoir procurer aux doctorants du laboratoire les meilleures conditions pour la réalisation de leurs projets de thèse.

Je voudrais remercier infiniment **M. Moulay Driss RAHMANI** (PES à la faculté des sciences de Rabat) d'avoir accepté de diriger ce projet de thèse. Un très grand merci pour sa disponibilité, sa gentillesse, son encouragement, et ses conseils.

Je tiens également à remercier **M. Abderrahim EL QADI** (PES à l'École Nationale Supérieure d'Arts et Métiers de Rabat) pour les efforts inlassables qu'il a déployés pour que ce travail soit élaboré, pour son aide précieuse et pour la qualité de son encadrement qui m'a été d'un appui considérable. Un très grand merci pour sa disponibilité, sa compréhension, et ses conseils.

Je remercie infiniment **Mme Salma MOULINE** (PES, Faculté des Sciences, Université Mohammed V de Rabat) d'avoir accepté de présider ma soutenance de thèse. Votre présence est un grand honneur pour moi. A cet effet, je tiens à vous exprimer ma grande considération et ma haute estime.

Je tiens aussi à exprimer mon profond respect et à remercier **M. Majid BENYAKHLEF** (PES, Université Sidi Mohammed Ben Abdellah Fès), d'avoir accepté de rapporter et examiner ma thèse et pour le temps qu'il a consacré à la lecture de mon mémoire de thèse et à la rédaction du rapport représentant sont avis dessus.

Je tiens également à exprimer ma profonde gratitude à **Mme Fatima GUEROUATE** (PES, EST de Salé, Université Mohammed V de Rabat), merci d'avoir accepté de rapporter et examiner ma thèse de doctorat, merci d'avoir consacré du temps pour lire et examiner mon mémoire, et pour rédiger le rapport le concernant.

Je tiens à remercier et à témoigner toute ma reconnaissance à **M. Hicham LAANAYA** (PH, Faculté des Sciences, Université Mohammed V de Rabat), merci d'avoir accepté de rapporter et examiner ma thèse de doctorat, merci pour la qualité du rapport que vous avez rédigé en ma faveur.

Je tiens également à témoigner ma profonde gratitude et à remercier **Mme Fadoua ATAA ALLAH**, (Directrice de recherche, IRCAM, Rabat). Merci pour le soutien que vous avez accordé à ma modeste personne en acceptant d'examiner ce travail de thèse.

Il ne sera jamais suffisant de dire à quel point, je suis reconnaissant envers toute personne qui m'a aidée et qui a contribuée, par son aide et ses encouragements, de près ou de loin à la bonne réalisation de ce projet.

Enfin, je souhaite une bonne continuation pour tous les doctorants du laboratoire.

Résumé

Les outils et techniques mis en œuvre par les systèmes de recherche d'information (SRI) présentent un manque d'organisation de l'information qui augmente la difficulté de distinguer les informations pertinentes des informations non pertinentes. En outre, les requêtes géographiques contiennent souvent des termes ambigus, ce qui engendre la diminution de la qualité et la dégradation de la pertinence des résultats retournés par les SRI.

Pour surmonter ces problèmes et répondre au besoin en information des utilisateurs d'une façon plus précise, nous proposons dans cette thèse deux approches de reformulation de requêtes qui se composent de trois phases : la construction d'une taxonomie afin de modéliser la notion d'adjacence spatiale entre les lieux, puis la détection et séparation des différentes composantes de la requête géographique sujette d'amélioration ; et enfin la reformulation de cette requête par l'interprétation de sa partie géographique en utilisant la taxonomie construite lors de la première phase. L'aspect de parallélisation est considéré lors du traitement sur notre deuxième contribution, afin d'augmenter le volume des données exploitées lors de la création d'une taxonomie tout en optimisant le temps de réponse de l'algorithme proposé.

Les expérimentations montrent l'efficacité des approches proposées. Ceci a été prouvé en évaluant nos approches à l'aide de mesures d'évaluation validées. Les résultats montrent que la prise en compte de l'aspect géographique et ses spécificités, augmente la pertinence des résultats et réduit le bruit.

Mots-clés : Recherche d'Information Géographique, Reformulation des requêtes, Indexation Sémantique Latente, PFP-growth, Apprentissage non-supervise, Traitement distribué.

Abstract

The tools and techniques implemented by Information Retrieval Systems (IRS) provide a lack of organization of information that increases the difficulty of distinguishing between relevant and irrelevant information. Moreover, geographic queries often contain ambiguous terms, which leads to a reduction in the quality and degradation of the relevance of the results returned by the IRSs.

To overcome these issues and meet the information needs of users more precisely, in this thesis, we propose two query reformulation approaches that consist of three phases: the first phase aims to build a geographic structure (taxonomy) in order to capture and model the notion of spatial adjacency between places. The second phase consists in detecting and separating the different components of the geographic query subject for improvement. Finally, the third phase addresses the reformulation of this query by interpreting its geographical part using a taxonomy constructed during the first phase. We also considered the aspect of parallelization of the processing on our second contribution, to increase the volume of data used when creating a taxonomy while optimizing the response time of the proposed algorithm.

The carried-out experiments show the effectiveness of the proposed approaches. This has been proven by evaluating our approaches using validated evaluation measures. The results show that taking into account the geographical aspect and its specificities when handling queries, containing spatial information, increases the relevance of the returned results and reduces noise.

Keywords: Geographic Information Retrieval, Query Reformulation, Latent Semantic Indexing, PFP-growth, Unsupervised Learning, Distributed Processing.

ملخص

أدت وفرة المعلومات وسهولة الوصول إليها عبر شبكة الإنترنت الواسعة النطاق، إلى انخفاض أهمية وجودة نتائج البحث المستردة من قبل نظم استرجاع المعلومات التقليدية، والتي لا تعتمد إلا على توفر المعلومات ومعايير اختيار مبنية على المحتوى في عملية الوصول إليها. فكلما زاد حجم المعلومات، زاد عدد المستندات التي يرجعها النظام للمستخدم، الأمر الذي يربكه ويضعه أمام حمل زائد من المعلومات فيتوجب عليه أخذ المسؤولية في تحديد ما هو ذو صلة وما دون ذلك. فمن الواضح أن المشكلة لا تكمن في توفر المعلومات وحسب، وإنما في أهمية النتائج المستردة وفق سياق استعمال معين. وبالإضافة إلى ذلك ، كثيرا ما تتضمن استعلامات المستخدمين ذات الطابع الجغرافي مصطلحات غامضة، مما يؤدي إلى انخفاض جودة وأهمية النتائج المستردة من نظام استرجاع المعلومات.

لتحقيق افضل توافق بين النظام والمستخدم وتلبيه حاجة هذا الأخير للمعلومة المطلوبة بطريقة أكثر دقة فإننا نقترح في هذه الأطروحة مساهمتان رئيسيتان. والتي تتألف من مناهج مبتكرة لإعادة الصياغة ملخصة في ثلاث مراحل : تهدف المرحلة الأولى إلى بناء بنية أو كيان جغرافي (مسمى بتصنيف) من أجل التقاط وتصميم نماذج مفاهيم الصفة المكانية التي تفيد المجاورة او التقارب بين الأماكن. بينما تشمل المرحلة الثانية كشف وفصل مختلف مكونات الاستفسار الجغرافي الخاضع للتحسين. وأخيرا ، تتناول المرحلة الثالثة إعادة صياغة هذا الطلب أو الاستفسار بتفسير جزئه الجغرافي باستخدام التصنيف الذي تم صياغته خلال المرحلة الأولى، علاوة على ذلك فقد أخذنا بعين الاعتبار استخدام تقنية التوازي الكثيف للمعالجة في مساهمتنا الثانية. والفكرة وراء ذلك هي زيادة حجم البيانات المستخدمة عند إنشاء تصنيف مع الأخذ بعين الاعتبار لتحسين وقت الاستجابة للخوارزمية المقترحة إلى أقصى حد.

وتبين التجارب التي أجريت على مجموعات البيانات فعالية المناهج المقترحة لإعادة صياغة الاستفسارات الجغرافية باستخدام تصنيف الصفات الجغرافية التي تفيد المجاورة او التقارب. وقد تجلى ذلك في تقييم نهجنا باستخدام تدابير التقييم المصادق عليها. وتبين النتائج أن مراعاة الجانب الجغرافي وخصائصه عند تناول الاستفسارات ، التي تتضمن معلومات تفيد المكان الجغرافي، تزيد من أهمية النتائج المستردة وتقلل من النتائج المشوشة.

الكلمات الجوهرية: استرجاع المعلومات الجغرافية، اعاده صياغه الاستفسارات، الفهرسة الدلالية الكامنة LSA، أنماط النمو المتوازي المتكرر PFP-Growth، التعلم الاستنتاجي، حوسبة غير ممركزة.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Contexte et problématique

Dans le cadre de la recherche d'information sur le web, la problématique principale posée est de doter le système de recherche d'information (SRI) de connaissances lui permettant d'être un intermédiaire entre l'utilisateur et les documents d'un corpus. Ce rôle d'intermédiaire se joue entre un utilisateur dont le système doit être capable d'interpréter les besoins et des documents dont le système doit pouvoir interpréter le contenu. En d'autres termes, un SRI doit utiliser des techniques d'extraction de connaissances qui lui permettront d'expliciter le contenu d'un document, d'une part. D'autre part, de connaître le contexte de la recherche et d'être capable d'inférer les intentions de l'utilisateur en fonction de la tâche visée. Cette tâche est généralement assez difficile, car la formulation de la requête par les utilisateurs provoque des déformations par rapport au besoin. En effet, l'individu n'exprime que ce qu'il pense important pour avoir une réponse, en plus qu'il a tendance à en dire le moins possible sur son besoin informationnel.

Avec l'élargissement de l'accessibilité à l'Internet et la disponibilité de données en ligne à faible coût, le nombre d'utilisateurs des moteurs de recherche a augmenté de façon spectaculaire. Toutefois, le manque d'organisation des ressources sur le World Wide Web a rendu difficile la tâche des utilisateurs pour accéder à l'information pertinente. Surtout que, s'ils soumettent des requêtes contenant des termes ambigus, celles-ci peuvent récupérer des documents qui ne correspondent pas au domaine de recherche voulu et peuvent induire ce dernier en erreur dans le cas où il n'est pas connaisseur du domaine et qu'il n'a pas d'information précise de ce qu'il cherche. De plus, ce type d'utilisateur a généralement tendance à ne pas utiliser les mêmes termes figurant dans les documents comme termes de recherche. Ajoutant à cela que, le vocabulaire peut varier considérablement d'un auteur de document à un autre.

Face à ces problématiques, l'accès à une information pertinente devient un défi pour la Recherche d'Information (RI) dite classique, qui ne propose en réponse qu'une liste de documents estimés pertinents indépendamment de la spécificité et du type de la requête, ainsi que du contexte de recherche de l'utilisateur. La stratégie de récupération de document contenant les mots de la requête, quel que soit leur ordre, est une stratégie triviale mais à moindre coût. Elle a montré des inconvénients dans plusieurs cas d'étude. Plus le volume des ressources disponibles est important, plus la liste des résultats retournés par SRI est importante. Ce qui met l'utilisateur face à une surcharge informationnelle, qui le rend incapable de distinguer ce qui est réellement pertinent de ce qui ne l'est pas (les documents non-pertinent qui forment du bruit).

Afin de gérer ces problèmes, des méthodes de « Reformulation de requêtes » sont proposées. Ce qui fait que, la stratégie de RI devient une suite de formulations et de reformulations de requêtes jusqu'à la satisfaction du besoin informationnel de l'utilisateur. Tout en considérant les relations de synonymies entre les termes pour dépasser le problème de discordance entre les termes des chercheurs d'information et ceux des auteurs de documents, ce qui permet d'enrichir et d'améliorer la requête.

Dans l'ère où nous vivons, les moteurs de recherche sont devenus le moyen principal pour la localisation de l'information pour la plupart des gens. Ainsi, les connotations géographiques sont devenues de plus en plus importantes dans la recherche sur le WEB. Pour cette raison, des chercheurs ont commencé à examiner les archives des requêtes disponibles, afin de bien comprendre les intentions des utilisateurs des moteurs de recherche, notamment comment ils expriment leurs besoins, et comment ce processus peut être amélioré. Cependant, même si le WEB a enlevé beaucoup de limitations géographiques, dans les médias, les communications et le e-commerce, plusieurs aspects géographiques du monde physique réel ne sont pas reflétés par la structure et le contenue du WEB.

Les différents outils existants qui tentent de répondre au besoin d'accès à l'information n'offrent que des services basiques et insuffisants, incapables de réponde aux problèmes de traitement de l'information spécifique, telle que l'information géographique présente dans des ressources de données non structurées comme les documents web. Or l'indexation spatiale de ces ressources nécessite une gestion de données non-structurée et du contexte dans lequel sont exprimées les informations spatiales, difficiles à interpréter automatiquement.

Les Systèmes d'Information Géographique SIGs existants ne peuvent pas être utilisés directement dans un système de recherche documentaire spatial, car ils ne sont adaptés que pour des données structurées, vu le manque des fonctions spécifiques au raisonnement spatial. Tandis que, les systèmes et approches de Recherche d'Information existantes ne tiennent pas compte à la fois de la dimension textuelle, spatiale et thématique des documents. Ils se focalisent généralement sur un seul aspect. Toutefois, il n'y a que peu de structures d'index ou algorithmes de recherche qui tiennent compte de la dimension spatiale des références géographiques intégrées dans les documents. Ce manque de structure géographique, dans les architectures de systèmes, a engendré un traitement standard des requêtes quel que soit leurs spécificités et leurs types, car même le langage d'interrogation ne supporte pas et ne considère pas ces spécificités.

Quant aux outils de système de Gestion Électronique des Documents (GED), ils n'offrent qu'une réponse générique de recherche d'information et ne tiennent pas compte des spécificités du corpus, comme les connotations géographiques par exemple. De plus ils nécessitent beaucoup de manipulations manuelles pour l'indexation, notamment pour le remplissage des notices descriptives des documents.

Pour faire face à ces limitations et pouvoir adapter les résultats de la recherche aux besoins spécifiques d'un utilisateur donné dans un contexte géographique, de nouvelles méthodes, qui sont à l'origine d'un nouveau domaine de recherche, ont été proposées. Sur la base de ces approches qui tiennent compte à la fois de l'aspect textuel et géographique des documents, le domaine de Recherche d'Information Géographique (RIG) a vu le jour. Toutefois, la plupart des travaux existant dans ce domaine ne prennent pas en compte la nature hiérarchique de l'espace géographique et les relations topologiques entre les objets géographiques pour bien définir les relations entre les documents et permettre de produire et d'injecter dans le système de nouvelles requêtes plus intéressantes et plus précises.

Afin de reprendre les différents problèmes posés, la problématique de notre thèse peut se résumer à fournir des méthodes et des outils pour répondre aux interrogations suivantes :

Comment accéder à l'information spatiale contenue dans la grande masse de documents textuels ? Et comment l'interpréter pour pouvoir l'utiliser dans un système de RI répondant aux besoins spécifiquement géographiques des utilisateurs ?

Contributions

Mon travail de thèse s'insère dans la problématique de l'accès à l'information spatiale présente dans les documents Web. Par conséquent, les travaux présentés dans ce mémoire se situent dans un contexte de développement de méthodes efficaces de reformulation de requêtes spatiales pour améliorer la qualité des résultats de recherches par l'extraction et l'interprétation de l'information géographique.

A cet effet, nous sommes allés au-delà des outils cités plus haut et des systèmes de recherche d'information classiques, basés sur l'analyse statistique des documents et peu adaptés à ce cas particulier, via un traitement linguistique ciblé afin de construire des structures géographiques. Dans l'objectif d'améliorer la performance et l'efficacité des SRIs à chaque fois que l'interrogation comporte une connotation spatiale. Ceci est réalisé en utilisant des traitements relativement peu coûteux, mais qui permettent de dégager l'essentiel de l'information et constituent une première étape dans le chemin d'interprétation des expressions géographiques extraites à partir des requêtes des utilisateurs de ces systèmes.

Nos deux contributions consistent alors à proposer des approches d'extraction de l'information géographique présente dans les documents textuels et les requêtes des utilisateurs. L'objectif étant de structurer les données extraites, en se basant sur le raisonnement spatial, sous forme de taxonomies hiérarchiques permettant, dans une deuxième phase, l'interprétation et la reformulation des requêtes géographiques des utilisateurs. Cette interprétation permet en outre de retrouver le contexte dans lequel l'information spatiale a été utilisée. En particulier, elle permet d'indexer des unités de texte en leur associant des contextes de type adjacence, inclusion ou autre.

Notre première contribution consiste en une approche de construction d'une taxonomie géographique d'adjacence pour un pays en se basant sur la méthode d'indexation sémantique latente (Latent Semantic Indexing LSI). Cette approche utilise les documents les mieux classés récupérés par le SRI lors de la soumission des entités spatiales composées de relation spatiale et du nom d'un emplacement géographique A (ville ou village). Ensuite, l'indexation des documents extraits par LSI, retourne une matrice où les vecteurs termes dont la similarité est élevée ne sont pas les termes qui apparaissent ensembles le plus, mais plutôt les termes qui partagent les mêmes cooccurrences. Ce qui permet à notre approche de trouver les lieux B_i les plus adjacents à A, et de procéder à une étape de validation de chaque lien en vérifiant si A se retrouve également dans les résultats des lieux B_i. La taxonomie construite est par la suite utilisée dans la reformulation des requêtes spatiales des utilisateurs. Cette contribution est organisée sous forme de trois phases :

- Phase 1 : Construction en mode hors ligne d'une taxonomie géographique d'adjacence ;

- Phase 2 : Détection et séparation des différentes composantes de la requête, définit par une entité thématique et une autre spatiale ;
- Phase 3 : Reformulation des deux entités de la requête. L'entité thématique en utilisant une ontologie sémantique, tandis que l'entité spatiale est reformulée en se basant sur la taxonomie construite pendant la première phase.

Notre deuxième contribution aborde une méthode d'apprentissage automatique (Machine Learning) de génération de règle d'association, nommée Frequent Pattern Growth (FP-Growth). Nous avons optimisé et personnalisé cette méthode afin de l'appliquer sur nos données. A cette fin, nous avons structuré nos données par l'extraction des noms des lieux à partir des différents documents utilisés et leur organisation sous la forme d'une base de données transactionnelle. Ensuite, nous avons appliqué l'algorithme FP-growth dans sa version parallèle (Parallel FP-Growth: PFP) afin de générer les règles d'association, qui formeront la taxonomie du pays dans un contexte Big Data. La nature récursive de l'algorithme utilisé nous a poussé vers le choix du Framework de traitement de données massives Spark. Par conséquent, nous avons utilisé la structure DataSet pour stocker nos transactions, puisque c'est une structure optimisée *très efficacement* par Spark, d'où les gains de performance sont très élevés. L'organisation de cette contribution rejoint les mêmes phases de la première contribution, en ayant également comme objectif final la reformulation de requêtes géographiques.

Plusieurs séries de tests ont été réalisées durant cette thèse. Les objectifs de ces tests sont : d'abord de spécifier les paramètres adéquats et les combinaisons adéquates pour les approches proposées, ensuite de comparer, tester et valider chacune de nos contributions. Par les mesures d'évaluation utilisées, nous avons vérifié l'effort requis de l'utilisateur pour retrouver parmi les documents retournés ceux qui répondent à son besoin, la précision du système qui s'exprime par sa capacité à sélectionner tous les documents pertinents et à rejeter tous les documents non pertinents, ainsi que le temps de réponse dans le cas de la deuxième contribution.

Organisation du mémoire

Ce mémoire est organisé en quatre chapitres : le premier présente le domaine et le contexte générale dans lequel se situent nos travaux, c'est-à-dire la recherche d'information ; le second permet de cerner le domaine spécifique de cette thèse qui est la recherche d'information géographique, tout en formant un état de l'art sur les méthodes existantes de reformulation de requêtes spatiales ; tandis que le troisième chapitre ainsi que le quatrième décrivent nos contributions dans le domaine de la RIG.

L'objectif du chapitre 1 « La recherche d'information (RI) : principes, techniques et outils », est de présenter le domaine en détails ainsi que les systèmes de recherche d'information, en exposant les étapes du processus de recherche. Ensuite, les modèles de recherche et d'indexation sur lesquels se base la RI. Enfin, de présenter les outils d'évaluation des SRIs, en décrivant la notion de pertinence et ses types, les collections de test ainsi que les mesures d'évaluation les plus utiles et les plus importantes.

Le chapitre 2 « La recherche d'information géographique », a pour but de porter la lumière sur l'émergence de la RI géographique, en s'intéressant plus précisément à la problématique de l'extraction de l'information spatiale présente dans le texte, aux approches de reformulations de requêtes géographiques proposées dans la littérature.

Le chapitre 3 « Reformulation de requêtes en construisant une taxonomie géographique par analyse sémantique latente (QRGTW) » présente notre première contribution validée, dont le but est double : premièrement de construire des taxonomies géographiques qui permettent de représenter la structure hiérarchique des données spatiales en les indexant sous une forme arborescente, deuxièmement de reformuler les requêtes spatiales après une séparation de leurs composantes. Cette approche a été testée avec variations de ses paramètres et comparées entre des méthodes de la littérature.

Le chapitre 4 « Contribution 2 : Reformulation de requêtes géographiques : approche basée sur un générateur de règles d'association parallèle » décrit notre deuxième contribution, qui se base sur l'utilisation après modification de l'algorithme d'apprentissage automatique PFP-growth. L'approche a comme objectif la reformulation de requête géographique également, par la création de taxonomie d'adjacence d'un pays en exploitant les documents les mieux classés retournés lors de la soumission de chaque entité géographique appartenant au pays en question.

Une conclusion générale qui dresse le bilan de nos travaux de thèse est présentée à la fin de ce mémoire. Cette partie résume les points essentiels de ce travail et présente quelques perspectives de recherche suggérées dans le contexte de l'amélioration de la performance des SRI lors du traitement de requêtes géographiques.

Difficultés rencontrées

Le problème qui se pose dans le domaine de recherche d'information géographique est la non-existence de collection adapté aux méthodes de recherche ou extraction de l'information géographique, ni aux techniques de reformulation de requêtes géographiques. Donc, il était de notre devoir de construire une collection constituée de requêtes spatiales et leurs documents résultants manuellement, en plus de l'ensemble de documents extraits pour la construction de taxonomie géographique spécifique à un pays donné, pour pouvoir faire les expérimentations, et tester les approches proposées. Cela a rendu le processus de validation d'une méthode plus long.

Toutefois, malgré l'absence d'une base de données standard géographique prête pour nous permettre de réaliser les expérimentations nécessaires, au début, la collection que nous avons construite, n'a pas été accepté par les examinateurs des journaux au début. Par conséquent, nous avons filtré les traces de journalisation du moteur de recherche AOL rendu public pour l'exploitation en tant que collection standard de test. Cela justifie le retard de la publication de mes travaux, car malgré les performances des approches proposées, la validation finale pour l'acceptation a été contrainte par l'utilisation d'une base de données standard qui est non existante.

Chapitre 1 – La Recherche d'Information (RI) : principes, techniques et outils

1. Introduction

La Recherche d'Information (RI) est la branche de l'informatique qui analyse les pratiques de recherche selon le besoin en information de l'utilisateur et qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information [9].

La RI a été défini de plusieurs manières dans la littérature. Selon Hernandez [8] c'est l'activité dont l'objectif est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en information. La RI est définit également comme une discipline de recherche qui intègre des techniques et des modèles dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur [10].

Du point de vue utilisateur, on parle d'une manière de recherche active de l'information, dite *recherche ad-hoc*, à travers un **Système de Recherche d'Information** (SRI). On peut également considérer une deuxième manière qui est dite passive et qui s'applique en utilisant un **Système de Filtrage d'Information (SFI)**. En définissant un **SRI**, comme un ensemble de programmes informatiques dont le but est de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents au moyen d'une fonction de correspondance (ou d'appariement document-requête). Tandis qu'un système de filtrage, consiste en un processus d'extraction des informations susceptible d'intéresser un utilisateur ou un groupe d'utilisateur à partir d'un flot d'informations (News, e-mail, actualités journalières, etc.), en considérant des besoins d'information relativement stables et commun.

Les SFI [7] sont des systèmes qui intègrent l'utilisateur d'une manière implicite dans le processus de sélection, en tant que structure informationnelle représentée le plus souvent par la notion de *profil*, comme l'indique la figure 1.1.

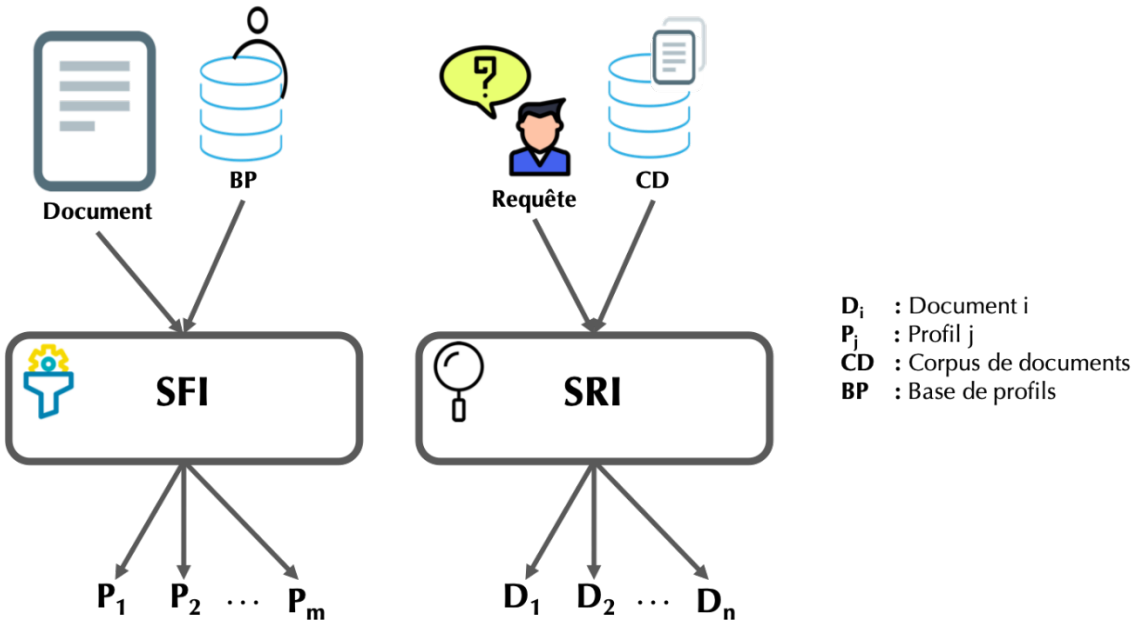

Fig. 1.1 Système de filtrage d'information *vs* Système de recherche d'information .

Dans cette thèse nous nous intéressons à la recherche active de l'information. Par conséquent, ce chapitre a pour objectif de présenter les concepts de base de la RI active, les composantes du processus de recherche d'information, les principaux modèles d'indexation et de structuration de texte utilisés en RI ainsi que les outils d'évaluation des SRIs.

2. Systèmes de Recherche d'Information

Le rôle d'un système de recherche d'information est de retourner à l'utilisateur des documents en réponse à une liste de mots-clés qu'il a soumis. Il doit veiller à ce que les contenus des documents retrouvés soient pertinents vis-à-vis du besoin initial en information exprimé par l'utilisateur.

Les SRIs développés depuis le début des années 50 reposent essentiellement sur des techniques statistiques et des techniques linguistiques de bas niveau [8]. Ces techniques prennent uniquement en compte le niveau lexical, parfois le niveau syntaxique, du contenu textuel des documents et requêtes afin d'identifier les mots permettant de retrouver les documents répondant aux besoins de l'utilisateur. Un enjeu actuel de la recherche d'information et du Web est de s'appuyer sur des connaissances afin d'enrichir les systèmes en apportant une couche sémantique.

2.1. Concepts de base de la RI

Avant d'expliquer le mode de fonctionnement d'un SRI, il faudrait clarifier quelques concept clés [11] :

Corpus de documents : le fond documentaire (ou collection de documents) d'un SRI constitue la source des données filtrée pour répondre aux besoins des utilisateurs. C'est l'ensemble des informations exploitables et accessibles facilement par le système. Dans une finalité d'optimisation du temps de réponse, cette collection peut être stockée sous forme de représentations simplifiées mais suffisantes des documents. Ces représentations permettent une gestion (ajout et suppression des documents) et une interrogation du corpus dans les meilleures conditions de coût.

Information élémentaire : généralement le document est l'information élémentaire d'un corpus de documents. L'information élémentaire est appelée également granule d'information et peut représenter tout ou une partie d'un document (un paragraphe).

Besoin en information : c'est le besoin de l'utilisateur qui est exprimé sous forme d'une liste de mots clés. Il peut être exprimé de façon incomplète et/ou ambiguë, comme il peut être stable ou variable ; l'utilisateur peut l'affiner au cours de la recherche.

Requête : représente l'expression du besoin en information de l'utilisateur. C'est l'élément déclencheur du processus de recherche d'information. Une requête est l'ensemble des termes soumis par l'utilisateur au SRI, mais elle peut être exprimée en langage naturel, booléen ou graphique. Les mécanismes de reformulation de requêtes permettent d'améliorer et d'enrichir cet ensemble à partir de connaissances extraites des contenus des granules (documents ou paragraphes du corpus) ou de ressources externes.

Modèle d'indexation (modèle de représentation) : c'est le processus permettant de structurer un document ou une requête et d'en extraire une représentation paramétrée qui couvre au mieux son contenu. Le résultat de ce processus est le *descripteur* du document ou de la requête, qui se constitue de termes pondérés significatifs, où un poids permet de différencier entre les degrés de représentativité du contenu sémantique de l'unité en question par le terme en question. Des exemples de ces modèles sont étudiés et discutés dans la section 2.3 de ce chapitre.

Modèle de recherche : c'est le modèle du noyau du SRI. Il consiste en la fonction de décision fondamentale qui permet d'associer et de classer un ensemble de documents pertinents par rapport à une requête. C'est la méthode qui réalise la tâche principale de recherche d'informations en se basant sur les descripteurs des documents et de la requête.

2.2. Processus de recherche d'information

Le processus de recherche d'information, appelé également processus en U (illustré schématiquement sur la figure 1.2), consiste en la mise en relation des éléments du corpus et des besoins de l'utilisateur grâce à un SRI. L'objectif de ce processus est de chercher et retourner à l'utilisateur le maximum de documents pertinents et le minimum de documents non-pertinents. Tel que la notion de pertinence est une notion vague et difficile à modéliser, car elle est fortement subjective. Ce processus est composé de trois étapes principales :

- L'indexation des documents de la requête de l'utilisateur ;
- L'appariement requête-document, qui permet de correspondre les documents aux requêtes et de classer les documents par rapport à leurs degrés de pertinences aux requêtes ;
- La modification, qui est une étape optionnelle et se manifeste généralement par la reformulation ou enrichissement de requête dans le cas où les résultats obtenus sont inadéquats et l'utilisateur n'est pas satisfait du résultat de sa recherche. Des modifications éventuelles peuvent concerner les documents (Ajout ou suppression de résultats).

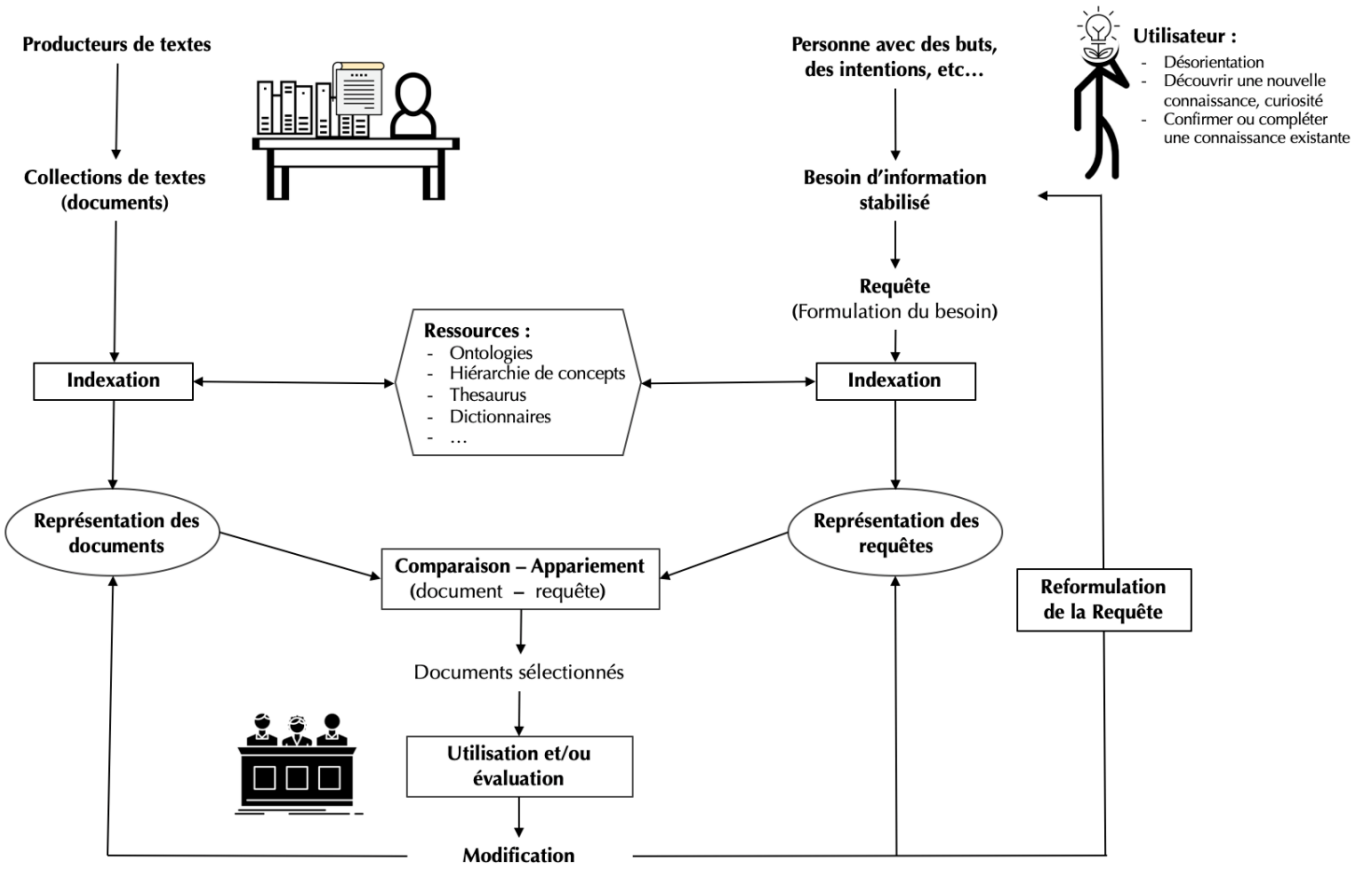

Fig. 1.2 Processus de recherche d'information (processus en U).

Comme le montre la figure 1.2, le processus en U est activé lors de l'expression du besoin d'information de l'utilisateur à travers une requête formulée dans la forme imposée par le SRI. L'indexation de la requête et des documents, la recherche dans le corpus et la présentation des résultats forment la phase suivante. Cette phase nécessite un modèle de représentation, appelé également modèle d'indexation, ainsi qu'une fonction de correspondance qui doit évaluer la pertinence des documents du corpus par rapport à la requête. La réponse du système est un ensemble de références à des documents qui obtiennent une valeur de correspondance élevée (supérieur à un seuil donné). Cet ensemble est généralement présenté sous la forme d'une liste ordonnée suivant la valeur de correspondance. La phase finale se présente comme la modification et se manifeste le plus souvent par la reformulation ou enrichissement de requête, pour une meilleure expression du besoin de l'utilisateur, dans le cas où les résultats obtenus sont non satisfaisants.

Avant de décrire en détail ces étapes, il est nécessaire de présenter l'événement déclencheur du processus qui est l'expression du besoin de l'utilisateur au travers d'une requête.

2.2.1. Expression du besoin en information : l'interrogation

L'utilisateur interroge le moteur de recherche lorsqu'il est confronté à un manque ou une inexistence de connaissances sur un sujet donné. Belkin considère que l'élément déclencheur d'une recherche d'information est un état anormal de connaissance (*Anomalous States of Knowledge*). Par conséquent, il ne faut pas perdre de vue que l'utilisateur d'un système de recherche d'information est plus concerné par retrouver les informations sur le sujet de la requête que par retrouver les données qui satisfont la requête soumise. Donc, mieux comprendre le mécanisme de satisfaction de l'utilisateur permettrait d'améliorer les performances d'un SRI.

En effet, comprendre le mécanisme de satisfaction de l'utilisateur revient à comprendre son intention lors de la soumission d'une requête et y répondre, au lieu de répondre à la requête en tant qu'une suite de mots-clés.

Une première étape dans ce chemin est de classifier la requête dans l'une des trois catégories principales que l'intention de l'utilisateur a permis de définir [12] :

- *Requête navigationnelle :* dont l'objet est un site web en particulier. L'utilisateur a un besoin vérificatif, il cherche une donnée particulière et sait même comment y accéder, puis compare le résultat avec les données connues qu'il possède déjà. Ces requêtes contiennent souvent des fragments d'URL ou des noms d'organisations. L'utilisateur clique généralement sur un seul résultat, l'amenant directement à la page désirée. Pour être considérée navigationnelle, la requête doit avoir un site web référent que l'utilisateur avait déjà en tête.
- *Requête informative :* le but à travers la soumission de ce genre de requêtes est de trouver une réponse à une question, de demander des explications ou d'explorer un sujet nouveau. Ces requêtes sont concentrées sur l'objectif de l'utilisateur de se procurer des informations sur le sujet de la requête.
- *Requête de recherche commerciale ou requête transactionnelle :* ces requêtes ayant un but de ressources ont comme objectif d'obtenir quelque chose autre que l'information disponible sur le site web. La finalité à travers leur soumission est de trouver une offre commerciale, ou une action manifestant une conversion pour le site web de destination (achat, paroles de chansons, recettes, inscription, création d'un compte, numéro de téléphone, réservation, parole …).

L'expression du besoin en information de l'utilisateur est formulée généralement sous forme de requête composée d'une liste de mots clés. Ces mots clés peuvent être reliés entre eux par des opérateurs booléens (ET, OU, NON), et/ou par des variables linguistiques (comme (plus) récent, (plus) important, ...) [14]..

De nombreux SRIs étendent ou reformulent les requêtes à travers la recherche de mots dans le contexte de la requête traitée, c'est à dire dans le voisinage de ses mots, ou bien dans le contexte de l'utilisateur, c'est-à-dire dans son profil.

Il existe également, des requêtes en texte libre dites aussi requêtes en langage naturel. Ces dernières permettent à l'utilisateur d'exprimer son besoin de façon plus naturelle qu'avec une suite

de mots-clés. Ces requêtes offrent surtout la possibilité d'utiliser un document complet ou une image pour l'expression du besoin (ce qui reviendrait à dire : trouve-moi tous les résultats semblables à ce fichier).

2.2.2. Processus d'indexation

Les requêtes et les documents sous forme de texte libre sont difficilement exploitables tel quel dans la phase de recherche. Cependant, l'étape d'indexation vient faciliter cette tâche au SRI en transformant les requêtes et les documents en une représentation exploitable automatiquement. Cette représentation est souvent une liste de mots clés significatifs pondérés, que l'on nomme descripteur de l'élément représenté. Cet élément peut être un document ou une requête.

Le processus d'indexation comporte un ensemble de traitements automatisés que nous présentons dans les sous-sections ci-après [25] : l'extraction des mots simples, l'élimination des mots vides, la lemmatisation et enfin la pondération.

2.2.2.1. Analyse lexicale : l'extraction

L'analyse lexicale est une opération qui consiste à extraire les termes ou mots simples et convertir le texte d'un document en un ensemble de termes. Ceci est réalisé grâce à la reconnaissance des espaces de séparations des termes, des chiffres, des caractères spéciaux, des ponctuations, etc.

2.2.2.2. Élimination des mots vides

La liste des mots simples extraite lors de l'étape précédente peut contenir des termes non significatifs, appelés « mots vides », tels que : les pronoms personnels, les prépositions, …, ainsi que les mots athématiques qui peuvent être contenu dans n'importe quel document, car ils exposent le sujet mais ne le traitent pas (comme par exemple contenir ou appartenir).

L'élimination de ces mots peut se faire de deux façons :

- en utilisant une liste dressée de mots vides, appelée également anti-dictionnaire ;
- en éliminant les mots qui dépassent un certain nombre d'occurrences dans la collection utilisée.

Bien que cette étape ait l'avantage évident de diminuer le nombre de termes d'indexation et spécifiquement ceux non significatifs, elle peut induire des effets de silence (par exemple en éliminant le mot « a » de « vitamine a »), et par conséquent réduire la proportion de documents pertinents renvoyés par le système par rapport à l'ensemble des documents pertinents (le rappel, notion expliqué sur la section 3 de ce chapitre).

2.2.2.3. Normalisation ou radicalisation

Un concept donné peut être représenté par différentes formes (différents mots) dans un texte, mais leur sens est le même ou très similaire. On peut donner l'exemple avec *écologie*, *écologiste*,

écologique, etc. Il n'est pas obligatoire d'indexer tous ces mots alors qu'un seul (*écologie*) suffirait pour désigner le concept véhiculé.

Généralement, quatre types de lemmatisation (en anglais, *Stemming*) sont distingués :

- par analyse grammaticale en utilisant un dictionnaire ;
- par élimination des affixes surtout pour l'anglais, en utilisant l'algorithme de Porter [26] par exemple ;
- par troncature des suffixes à X caractères (en définissant une valeur à X) ;
- par la méthode des n-grammes utilisée pour le chinois [27].

Frakes et Baeza-Yates [28] distinguent cinq types, en ajoutant la stratégie des variétés de successeurs à la liste.

Cependant, cette étape de passage à la forme canonique d'un mot n'est pas nécessaire. Ses principaux avantages étant l'amélioration de la valeur du rappel et la minimisation de la taille de l'index utilisé en radicalisant plusieurs termes à une seule racine. Ceci peut être utile par exemple dans le cas d'indexation des mots "voitures" et "voiture" de la même façon "voiture", ce qui évite à l'utilisateur de devoir entrer les formes de pluriel des noms ou les formes conjuguées des verbes lors de la soumission de la requête. Sauf que, dans certains cas, le passage vers la forme canonique du mot supprime sa sémantique originale. Par exemple, les mots "portera" (forme conjuguée du verbe porter) et "portes" seront indexés de la même façon "porte". Ainsi, lorsqu'un utilisateur soumettra une requête contenant le verbe "porter", il retrouvera certainement, parmi la liste des documents retournés, des résultats non pertinents relatifs au nom "porte".

Pour résoudre ce genre de problème, Crouch et al. [29] ont proposé une méthode en deux phases, dont les résultats ont minimisé les erreurs dues à la radicalisation des mots :

- une première recherche est réalisée en utilisant les formes normalisées des mots ;
- les documents sont par la suite réordonnés en fonction de la présence ou non des termes non-normalisés de la requête dans leurs contenus.

Ce qui a permis d'améliorer les résultats de la précision (c-à-d la proportion de documents pertinents par rapport à l'ensemble des documents renvoyés par le système) tout en gardant l'avantage de lemmatisation qui se présente dans l'augmentation du rappel.

2.2.2.4. Pondération des termes

La pondération est une étape fondamentale dans le processus d'indexation de texte puisqu'elle traduit le degré d'importance des termes dans un granule (une requête ou un document). Cette importance est souvent calculée à partir des interprétations et considérations statistiques. Toutefois, il existe des approches qui se basent sur une pondération conceptuelle, on parle alors de pondération de concepts. L'objectif de cette étape est de trouver les termes ou concepts qui représentent le mieux le contenu des granules.

Parmi les nombreuses mesures de pondération définies dans le domaine de la RI, TF-IDF est de loin la mesure la plus connue et la plus utilisée [14]. Ceci est dû au fait qu'elle a vu le jour suite à des constatations fondées, qui énoncent que si on dresse une liste de l'ensemble des termes d'un texte classés par ordre de fréquences décroissantes, on constatera que la fréquence d'un terme est

inversement proportionnelle à son rang de classement dans la liste (qui traduit son importance). Cette constatation est énoncée formellement par la loi de Zipf [33] :

$$rang * fréquence = constante \quad (1.1)$$

La loi de distributions des termes d'un texte quelconque suit alors la courbe présentée sur la figure 1.3. Selon Zipf, la courbe hyperbolique de la distribution des termes s'explique par ce qu'il a appelé le principe du moindre effort (*Principle of Least Effort*) : Ce principe considère qu'il est plus facile pour un auteur d'un document de répéter certains mots que d'en utiliser de nouveaux. Par conséquent, la relation entre la fréquence et le rang des termes permet de sélectionner les termes représentatifs d'un document. Il propose également d'éliminer respectivement les termes de fréquences très élevées car ils ne sont pas représentatifs du document (ce qui rentre dans la branche des mots outils contenus dans l'anti-dictionnaire), et les termes de fréquences très faibles (ce qui permet d'éliminer les fautes de frappes). Ce processus est illustré sur la figure 1.3. En utilisant cette approche, le nombre de termes faisant partie de l'index d'une collection peut être réduit considérablement.

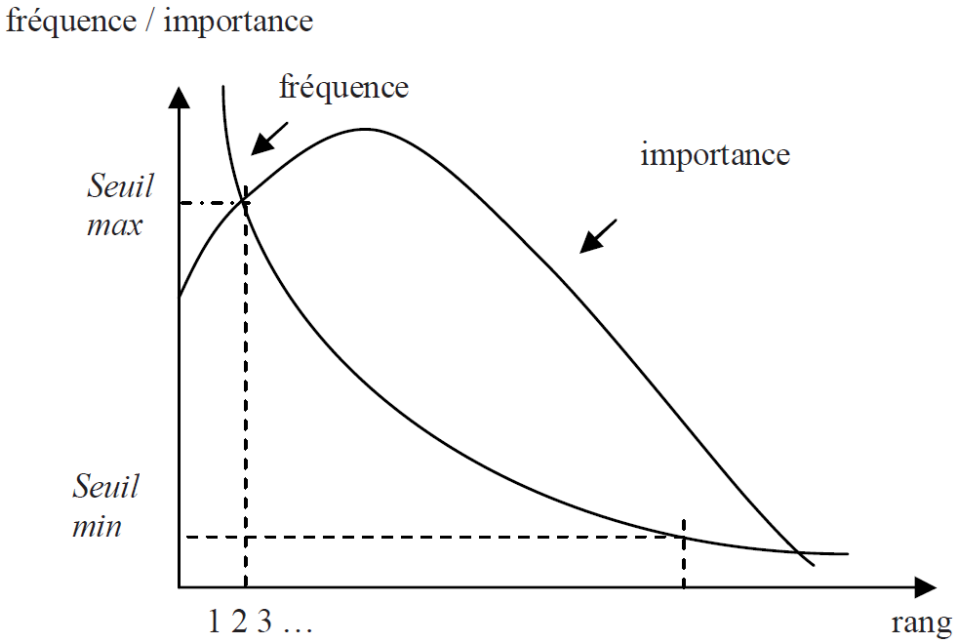

Fig. 1.3 L'importance d'un terme dans un document en fonction de sa fréquence d'apparition dans ce document

Ces à partir de ces constatations, que les formules de pondération basées sur les facteurs *TF* (Term Frequency) et *IDF* (Inverse of Document Frequency) ont vu le jour. Ces formules qui considèrent les pondérations locale et globale d'un terme :

- **Fréquence du terme (TF)** : cette mesure quantifie la représentativité locale d'un terme dans le document. L'idée sous-jacente est simple : plus un terme est fréquent dans un document plus il est important dans sa description. Ils existent plusieurs variantes de cette mesure. La fréquence TF_{ij} du terme t_i dans le document d_j est donnée selon l'une des formulations suivantes :

$$TF_{ij} = td_{ij} \quad (1.2)$$

$$TF_{ij} = log\ (td_{ij} + 1) \quad (1.3)$$

$$TF_{ij} = \frac{td_{ij}}{\sum_k td_{kj}} \quad (1.4)$$

Où td_{ij} est le nombre d'occurrences du terme t_i dans le document d_j. Les déclinaisons (équations 1.3 et 1.4) permettent de normaliser la fréquence du terme afin d'éviter les biais liés à la longueur du document et réduire les effets de grandes différences dans les fréquences.

- **Fréquence inverse du document (IDF)** : ce facteur mesure la représentativité globale du terme vis-à-vis à toute la base documentaire. Un terme qui apparaît souvent dans la collection n'a pas le même degré d'importance qu'un terme moins fréquent. La mesure IDF_i est exprimée selon l'une des déclinaisons suivantes :

$$IDF_i = log\ (N/n_i) \quad (1.5)$$

$$IDF_i = \log(N - n_i/n_i) \quad (1.6)$$

Avec N est la taille de la collection et n_i le nombre de document où apparaît le terme t_i.

La fonction de pondération TF-IDF est donnée par la multiplication de ces deux facteurs. Ils existent également d'autres approches qui se sont inspirées de cette fonction comme celles présentées dans les travaux [30,31].

Ces fonctions généralement et TF-IDF particulièrement, donnent une bonne approximation de l'importance du terme dans le document et montrent des résultats intéressants surtout dans les collections de documents de taille homogène. Cependant, elles ne tiennent pas compte d'un aspect important du document qui est sa longueur. Il est généralement constaté que les documents les plus longs ont tendance à utiliser les mêmes mots de façon répétée, ou bien à utiliser plus de mots pour décrire un sujet. Par conséquent, les fréquences des termes sont plus élevées, alors les similarités à la requête sont parfois faussement plus élevées. Pour pallier à cet inconvénient, la formule du facteur TF présentée par l'équation 1.2 est évitée ou l'intégration de la taille des documents à la formule de pondération comme facteur de normalisation est considérée, comme proposé par Singhal et al. dans [32].

Un inconvénient commun à toutes ces approches est le fait qu'elles ne considèrent que les occurrences des termes ou concepts dans les documents et ne considèrent pas l'organisation conceptuelle dont ils sont issus. La sémantique contenue dans les relations entre termes ou concepts est alors ignorée.

2.2.2.5. Indexation par la sémantique latente

La méthode d'indexation sémantique latente (LSI pour *Latent Semantic Indexing*, appelée aussi LSA *Latent Semantic Analyses*) est une méthode statistique proposée la première fois en 1990. Les auteurs de cette approche supposent que les textes sont porteurs d'une structure sémantique implicite (latente) dont ils tentent d'extraire les concepts en tant qu'unité de sens *[36]*. Cette technique a pour but d'éviter la polysémie et la synonymie des termes utilisés comme descripteurs par les techniques statistiques classiques en regroupant les termes ayant des caractéristiques communes dans leur apparition dans les documents.

LSI a été créée à l'origine pour permettre une représentation des documents pouvant s'appliquer à des collections spécifiques en s'adaptant aux variations lexicales. Cette caractéristique nous a incité à penser à l'utiliser comme méthode d'indexation dans cette thèse, puisque la majorité des données traités sont des données géographiques.

L'indexation par sémantique latente a été évaluée et validée dans la campagne d'évaluation TREC et a donné de meilleurs résultats que les approches statistiques classiques [36]. Son évaluation en utilisant la collection SMART a également prouvé ses performances [21].

La méthode utilise comme données d'entrée une matrice représentant les documents sur les colonnes et les termes sur les lignes. Pour la ligne i et la colonne j, la valeur représentée est la fréquence du terme i dans le texte j. La technique de décomposition en valeurs singulières permet d'optimiser cette matrice vers un espace réduit de dimensions orthogonales. L'originalité de la méthode est de réduire les dimensions de l'espace en modélisant les variations sémantiques significatives tout en diminuant le bruit. LSI est expliqué plus en détails dans la sous-section 2.3.2.3.

Le principal inconvénient de cette méthode est que la matrice résultante ou matrice des concepts n'est pas compréhensible par les humains. En effet, les poids des termes ne sont pas interprétables. Ceci ne limite pas le pouvoir de la méthode qui ne décèle pas explicitement les concepts et donc la sémantique associée aux documents, mais prouve ses performances d'une façon continue lors des expérimentations [37].

2.2.3. Appariement document-requête

La phase d'appariement document-requête consiste en la comparaison entre les documents du corpus et la requête et permet de mesurer le score de pertinence des documents vis-à-vis de la requête. Ceci est réalisé en calculant un score dit de correspondance entre le descripteur de chaque document et celui de la requête.

Le score calculé à ce niveau traduit le degré de pertinence système. Les SRIs essayent le plus possible de minimiser l'écart entre cette pertinence système et la pertinence de l'utilisateur, qui est représentée par son jugement de pertinence par rapport au document (voir la sous-section 3.1 de ce chapitre pour plus de détails).

La fonction d'appariement ou de correspondance, qui permet de calculer le score de pertinence permet d'ordonner les documents et d'en retourner un sous-ensemble à l'utilisateur. La qualité de cet ordonnancement est primordiale du fait que généralement, l'utilisateur se contente d'examiner les 10 ou 20 premiers documents renvoyés. Si l'information recherchée n'est pas présente dans les documents de cette tranche, l'utilisateur considèrera que sa recherche n'a pas abouti.

La fonction d'appariement est très étroitement liée aux opérations d'indexation et de pondération des termes de la requête et des documents. En général, la fonction d'appariement et le modèle d'indexation ou de représentation des documents et des requêtes utilisés permettent de caractériser le modèle de recherche d'information du SRI. Il existe un certain nombre de modèles théoriques, dans la littérature les plus connus étant le « Modèle Booléen », le « Modèle Vectoriel », et le « Modèle Probabiliste ». La sous-section suivante (Sous-section 2.3) décrit les modèles de RI.

2.2.4. Reformulation de la requête

La satisfaction de l'utilisateur concernant les résultats retournés par le SRI, n'est pas toujours acquise. Toutefois, la requête initiale est vue comme un moyen permettant d'initialiser le processus de recherche d'informations pertinentes [7]. A ce titre, les SRIs doivent incorporer des fonctionnalités permettant de prendre le relai. Ce qui est souvent réalisé par le processus de reformulation (ou enrichissement) de requêtes. Ce processus permet en fait de construire une nouvelle requête en se basant sur la requête initiale, des connaissances extraites des documents pertinents de la requête ou disponibles dans des ressources spécifiques. Néanmoins, ce processus n'est pas toujours automatique, une stratégie classique d'utilisation des SRIs consiste à reformuler manuellement la requête en tenant compte des documents pertinents et non pertinents obtenus (figure 1.4).

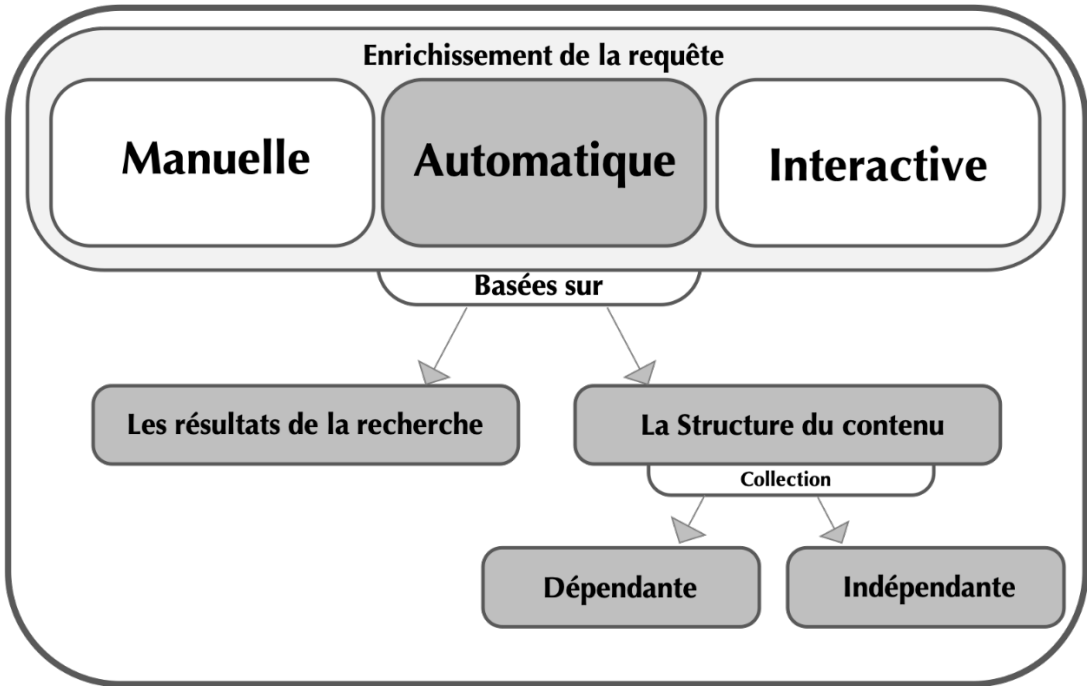

Fig. 1.4 Classification des approches de reformulation de requête.

La reformulation rentre dans le processus général d'optimisation de la fonction de pertinence qui a pour objectif de rapprocher la pertinence système de la pertinence utilisateur. Son principe consiste à modifier une requête en rajoutant des termes significatifs (dans ce cas on parle d'*Expansion de requêtes*), en remplaçant des termes par d'autres et/ou en réestimant leurs poids associés [7]. Ces termes peuvent provenir de sources différentes selon le type d'approche d'enrichissement utilisé. La figure 1.4 représente les trois classes d'approches de reformulation de requêtes existantes :

- Approches manuelles de reformlation par l'utilisateur ;

- Approches interactives, assistées par l'utilisateur, soit à travers ses jugements de pertinence (méthodes de *réinjection de pertinence*), soit en sélectionnant des termes en fonction de ses préférences à partir d'une liste de termes proposée par le système ;

- Approches automatiques, réalisées par le SRI sans avoir recours aux jugements des documents ou des termes par l'utilisateur. Leur but est de compléter la requête initiale en utilisant de l'information globale provenant du corpus entier ou éventuellement de ressources extérieures. On parle alors de reformulation par *contexte global* [18].

  Parmi ces approches il y a ceux qui se basent sur les résultats de recherche, et qu'on appel méthodes de *réinjection de pertinence implicite*. Dans cette même catégorie, il y a deux sous-catégories de méthodes qui utilisent les structures de données :

  - Collection dépendante : basée sur tous les documents du corpus (*Thésaurus*).
  - Collection indépendante : basée sur des sources externes (O*ntologie ou Taxonomie*), construite manuellement ou automatiquement à partir d'une collection.

Tel qu'un *thésaurus* est une structure de données où sont enregistrées les associations entre les termes. C'est une liste organisée de termes contrôlées et normalisés (descripteurs et non descripteurs) servant à l'indexation des documents et des questions dans un système documentaire [20]. Tandis qu'une *ontologie* est une structure qui cherche à décrire de façon formelle un domaine de connaissance, en identifiant les types d'objets de ce domaine, leurs propriétés et leurs relations. Un exemple très couramment utilisé est l'ontologie linguistique WordNet qui est développé par [38]. Cette dernière présente la particularité de couvrir la majorité des mots de la langue anglaise (noms, verbes, adjectifs et adverbes) et de rendre compte des relations sémantiques qu'ils entretiennent. Avec les relations sémantiques qui peuvent exister entre les termes et qui sont représentées par les ontologies sémantiques sont [12]: les relations de synonymie, d'hyperonymie, d'hyponymie, de méronymie et de liens sémantiques.

- Les relations de ***synonymie*** désignent des relations entre des mots ou des expressions de formes différentes dans une même langue et ayant un rapport de proximité de sens ou une signification très proche (Exemple : « habit » et « vêtement ») ;
- Les relations d'***hyperonymie*** représentent des relations sémantiques hiérarchiques d'un terme à un autre selon laquelle l'extension du premier terme, qui est plus général, englobe l'extension du second, qui est plus spécifique. (Exemple : « science » et « chimie ») ;
- Les relations d'**hyponymie** représentent des relations sémantiques d'un terme à un autre selon laquelle l'extension du premier est inclus dans l'extension du second. (Exemple : « banane » et « fruits ») ;
- Les relations de **métonymie** représentent des relations sémantiques qui consistent à remplacer le terme propre par un autre qui lui est proche ou qui en représente une

qualité *e.g.* la marque pour la chose ou l'objet. (Exemple : « automobile » peut être remplacé par « Mercedes »).

D'autre part, un autre type de structures de données est le concept de t*axonomien* qui est très utilisé par les anglo-saxons pour désigner la « science de la classifcation », c'est une structure qui contient des catégories organisés hiérarchiquement. Les taxonomies servent à classifier et à ranger des contenus ou des ressources. Une taxonomie est moyennement formelle, voir plus formelle qu'un thesaurus, dans le sens où l'hiérarchie des entrées peut suivre certaines contraintes.

2.2.4.1. Réinjection de Pertinence

Certains systèmes permettent aux utilisateurs de juger les documents résultats comme pertinents ou non pertinents. Ces jugements sont alors pris en considération afin de définir une nouvelle requête lors du processus de reformulation [11]. On parle alors de la réinjection de pertinence (RP), communément appelée Relevance Feedback ou Retour de Pertinence.

En utilisant la RP, les termes contenus dans les documents restitués par le système et jugés pertinents par l'utilisateur sont considérés uniquement. Le nombre de ces termes est assez faible par rapport au nombre des termes de la collection. Donc le gain est en pertinence et en temps de traitement. Ces méthodes permettent aux SRIs un meilleur contrôle du processus de recherche, en augmentant le poids des termes importants et diminuant celui des termes non importants [15].

Cependant, la réinjection de pertinence offre la possibilité d'avoir de très bonnes performances si les utilisateurs fournissent des jugements de pertinence suffisants et corrects [17]. Mais malheureusement, cette méthode s'est montrée peu populaire pour les utilisateurs, car dans un contexte de recherche réel, ils sont souvent peu disposés à fournir ce genre d'informations de pertinences explicites, qui sont généralement ressenties comme une charge supplémentaire lors de leurs interactions avec le SRI [16].

Pour surmonter les difficultés [17], dues au manque de jugements de pertinences suffisants, la réinjection de pertinence est remplacée par la Réinjection de Pertinence Implicite (RPI), appelée également Blind Relevance Feedback ou Réinjection Locale.

2.2.4.2. Réinjection de Pertinence Implicite

La RPI est une approche alternative de la RP, appelée également pseudo-RP car elle utilise les techniques de réinjection automatique à l'aveugle pour construire une nouvelle requête. C'est un axe de travail plus récent qui imite la « réinjection de pertinence » en considérant les documents les mieux classés par le SRI comme étant pertinents.

Plus précisément, le SRI restitue un ensemble de documents répondant à la requête initiale. Ainsi, au lieu d'attendre l'évaluation explicite des documents par l'utilisateur, les k premiers documents sont considérés comme étant pertinents (c'est des documents pseudo-pertinents). On peut également considérer les documents qui sont récupérés en fin de liste comme non pertinents. L'idée de base derrière la réinjection implicite est qu'une itération de réinjection basée sur les documents les plus similaires à la requête initiale de l'utilisateur pourrait donner une meilleure restitution des documents.

Cette catégorie de techniques est utilisée dans plusieurs systèmes de reformulation de requête [16] et apporte de nombreux avantages aux SRIs. Néanmoins, cette méthode a un inconvénient évident : si une grande partie des documents les mieux classés contiennent peu d'informations pertinentes ou aucune, alors les termes utilisés lors de la reformulation de la requête sont susceptibles de causer une dégradation des performances. Ainsi, les effets de la réinjection de pertinence implicite dépendent fortement de la qualité de la récupération initiale.

2.2.4.3. Reformulation par contexte global

Les techniques de reformulation ou plus précisément d'expansion globale correspondent à une analyse complète de corpus pour établir des liens sémantiques ou statistiques entre les termes. Ces liens peuvent être construits manuellement par un expert, ou de manière automatique à partir du corpus de documents en entier. Le but de ces approches [16] est d'enrichir la requête initiale en utilisant de l'information globale issue de ressources construites à partir de la collection de documents utilisée ou éventuellement de ressources linguistiques extérieures existantes. Cette information globale utilisée peut être sous la forme d'un thésaurus, d'une ontologie ou d'une taxonomie.

Les associations établies manuellement traduisent généralement les liens de synonymie et d'hiérarchie. Les thésaurus construits manuellement forment un moyen efficace pour la reformulation de requête [11]. Cependant, leur réalisation, le suivis et la mise à jour des informations sémantiques qu'ils contiennent sont coûteuses en temps et nécessitent l'intervention des experts des domaines considérés. Pour cette raison, ils restent peu utilisés par les SRIs. En revanche, la seconde catégorie de ressources concerne celles construites en se basant sur une analyse statistique des collections. Parmi les thésaurus automatiques, on peut citer un thésaurus basé sur les similarités [7,22] et un thésaurus statistique [23] qui peut être basé sur des techniques de clustering [24].

L'expansion de requête en utilisant ce genre de ressources consiste à chercher des associations de termes afin d'ajouter des termes voisins à la requête. Il existe aussi d'autres méthodes entièrement automatiques telles que le calcul des liens contextuels entre termes [21]. Les associations créées automatiquement sont généralement basées sur la cooccurrence des termes dans les documents et peuvent considérer le contexte de l'utilisateur également (ses préférences et son historique de recherche). Les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes.

2.3. Modèles de recherche d'information

Un modèle de RI a comme objectif de fournir une formalisation du processus de RI, ainsi qu'un cadre théorique pour la modélisation de la mesure de pertinence. Il existe un nombre important de modèles théoriques de RI textuelle dans la littérature, qui forment trois catégories (figure 1.5) : les modèles basés sur la théorie des ensembles, les modèles algébriques et les modèles probabilistes. Le point commun entre ces modèles est le vocabulaire d'indexation basé sur le formalisme mots clés, tandis qu'ils se différentient principalement par le modèle

d'appariement utilisé. Tel que, le vocabulaire d'indexation V est constitué des mots ou racines de mots contenues dans les documents : $V = \{t_1, t_2, \dots, t_n\}$.

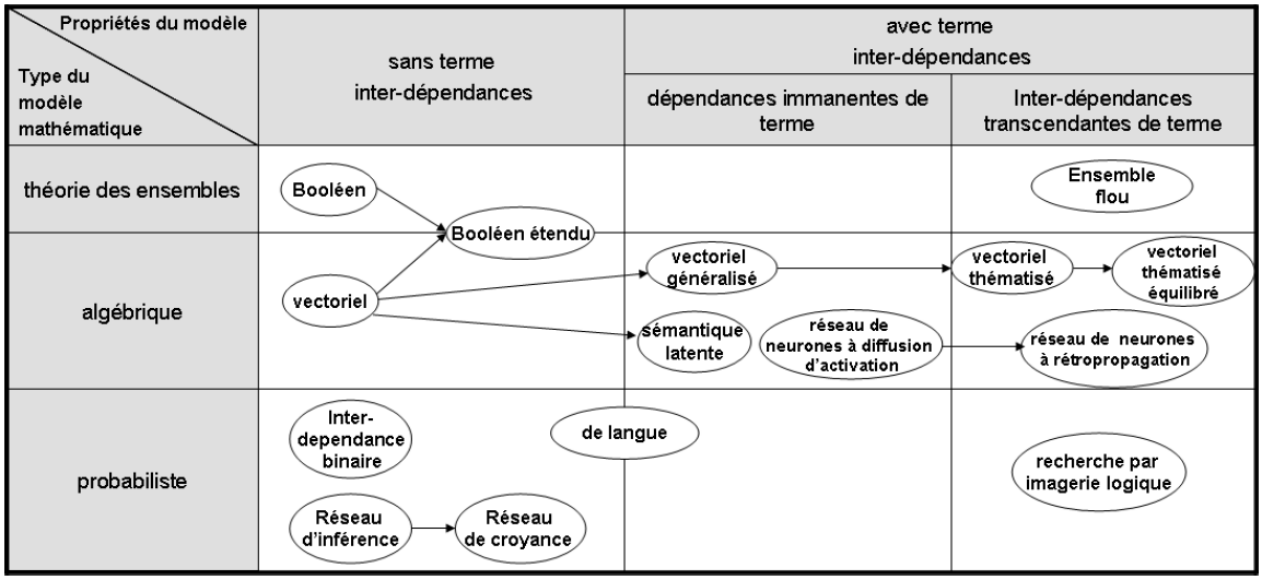

Fig. 1.5 Classification des modèles de RI selon Kuropka [39]

Un modèle de recherche d'information peut être défini par le quadruplet $[D, Q, F, R(q_i, d_j)]$ selon [38], avec D la base documentaire, Q l'ensemble de requêtes, F le schéma du modèle théorique de représentation des documents et des requêtes et $R(q_i, d_j)$ la fonction d'ordonnancement qui associe un score de pertinence à un document $d_j \in D$ par rapport à une requête $q_i \in Q$.

2.3.1. Modèles ensemblistes

Les modèles ensemblistes trouvent leurs fondements dans la théorie des ensembles et l'algèbre de Bool [14]. Le représentant le plus connu de cette catégorie est le modèle booléen (*Boolean Model*), c'est également le premier modèle qui s'est imposé dans le monde de la RI. Ils englobent également le modèle booléen étendu (*Extended Boolean Model*), qui est une fusion entre le modèle booléen et le modèle vectoriel, et le modèle flou qui se base sur la théorie des ensembles flous (*Fuzzy Set Model*).

Dans le modèle booléen, les documents et les requêtes sont représentés par des ensembles de mots clés. Les opérateurs logiques (OR, AND et NOT) séparent les termes de la requête et permettent d'effectuer des opérations d'union d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Tandis qu'un document est représenté par une conjonction logique de ses termes non pondérés, qui permet de constituer l'index du document, car ce modèle considère uniquement la présence ou absence des termes dans un document.

Un document d et une requête q sont représentés par exemple comme suit [11] :

$$d = t_1 \wedge t_2 \dots \wedge t_n \quad (1.7)$$

$$q = (t_1 \wedge t_2) \vee (t_3 \wedge t_4) \quad (1.8)$$

Dans ce modèle, la fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document dans ce modèle. Elle vérifie si l'index de chaque document implique l'expression logique de la requête. Le résultat de cette fonction est donc un score binaire défini comme suit :

$$RSV(q,d) = \begin{cases} 1 & si\ les\ termes\ de\ q\ n'appartiennent\ pas\ tous\ \text{à}\ d \\ 0 & sinon \end{cases} \quad (1.9)$$

Ainsi, le modèle booléen affirme soit la pertinence soit la non-pertinence d'un document vis-à-vis d'une requête, et ne considère pas la notion de réponse partielle à une requête. Le fait que la pertinence d'un document soit basée sur un critère binaire sans notion d'échelle de gradualité empêche le modèle booléen d'avoir de bonnes performances. De plus, le résultat binaire de la fonction d'appariement ne permet pas de fournir à l'utilisateur une liste ordonnée de résultats.

Dans la même catégorie, se trouve également le modèle flou, qui en se basant sur la théorie des ensembles flous permet de donner une certaine flexibilité. Dans ce cas, un objet x peut appartenir à un ensemble flou A avec un degré d'appartenance $\mu_A(x)$ (compris entre 0 et 1), tandis que dans le cas classique x peut soit appartenir à A (valeur 1) ou ne pas y appartenir (valeur 0). Le degré d'appartenance est utilisé pour représenter l'*incertitude* ou l'*ambigüité*. Par exemple, si l'ensemble floue « A » représente la proposition « Personne âgée » alors on pourrait dire qu'une personne âgée de 60 ans est une personne âgée à 65%, c'est-à-dire sa valeur d'appartenance à A est de 0,65.

En considérant les bases de la logique floue qui peuvent être examinées sur le document [40], les trois opérations les plus couramment effectuées sur des ensembles flous, à savoir l'intersection, l'union et le complément, sont ainsi définies :

$$\mu(A\ et\ B) = \min(\mu(A), \mu(B)) \quad (1.10)$$

$$\mu(A\ ou\ B) = \max(\mu(A), \mu(B)) \quad (1.11)$$

$$\mu(Non\ A) = 1 - \mu(A) \quad (1.12)$$

où μ est la fonction d'appartenance floue.

Sachant que la correspondance d'un document avec les termes d'une requête est approximative, une représentation en ensembles flous parait une solution plus adéquate. Ceci peut être modélisé en considérant que chaque terme de la requête définit un ensemble flou et que chaque document possède un degré d'appartenance à cet ensemble. De nombreuses expressions d'appariement requête-document ont été développées. L'une de ces expressions est celle de Paice [41], inspirée du modèle booléen, pour laquelle le score d'un document pour une requête $q = (a_1\ et\ a_2 \dots a_n)$ est calculé comme suit :

$$RSV\ (d,q) = \frac{\sum_{k=1}^{n} r^{k-1}\mu(a_k)}{\sum_{k=1}^{n} r^{k-1}} \quad (1.13)$$

où les $\mu(a_k)$ sont considérés dans un ordre décroissant pour « les requêtes ou » et croissant pour « les requêtes et ».

Cependant, il est parfois difficile pour un utilisateur d'exprimer son besoin en information avec des expressions booléennes. Par conséquent, les expressions booléennes formulées sont généralement très simples, ce qui ne permet pas d'utiliser au mieux les caractéristiques des modèles booléens.

2.3.2. Modèles algébriques

Ces modèles sont basés sur le calcul vectoriel, c'est pourquoi ils sont appelés également « Modèles Vectoriels ». Leur premier représentant a été le modèle vectoriel (Vector Model). Pour ces modèles, la pertinence d'un document vis-à-vis d'une requête est mesurée en utilisant des expressions de distance dans l'espace vectoriel. Ces modèles englobent également le modèle d'analyse sémantique latente (Latent Semantic Analyses : LSA), le modèle vectoriel généralisé (Generalized Vector Model) et le modèle connexionniste.

De nombreux modèles et méthodes d'indexation et d'ordonnancement des résultats ont été comparés au modèle vectoriel, celui-ci, s'est avéré être meilleur ou au moins aussi bon que les autres alternatives testées, malgré sa simplicité. C'est pour ces raisons que le modèle vectoriel est le plus populaire en recherche d'information [14].

2.3.2.1. Modèle vectoriel

Ce modèle est basé sur une intuition géométrique, les requêtes et les documents sont représentés dans l'espace vectoriel engendré par les termes d'indexation [42]. L'espace est de dimension N (N étant le nombre de termes d'indexation de la collection de documents).

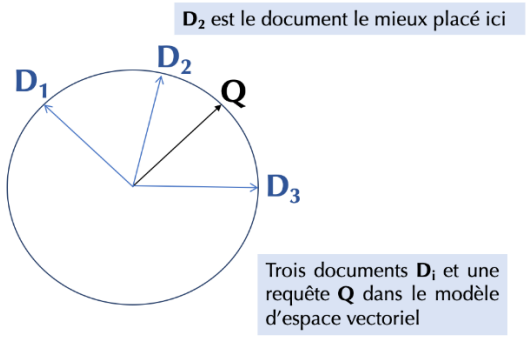

Fig. 1.6 Exemple d'une représentation du modèle vectoriel

Chaque document est représenté par un vecteur (*figure 1.6*) $D_j = (d_{1j}, d_{2j}, d_{3j}, \ldots, d_{Nj})$, et chaque requête est représentée par un vecteur : $Q = (q_1, q_2, q_3, \ldots, q_N)$, avec d_{ij} est le poids du terme t_i dans le document D_j et q_i le poids du terme t_i dans la requête Q. La pondération des composantes de la requête est soit la même que celle utilisée pour les documents, soit donnée par l'utilisateur lors de sa formulation.

Chaque poids d'un terme dans un vecteur document (respectivement requête) désigne l'importance de ce terme dans ce document (respectivement requête). Ces termes pondérés sont utilisés afin de calculer le degré de similarité entre chaque document et la requête de l'utilisateur.

Le mécanisme de recherche consiste à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête. Pour ce faire, les principales expressions de similarité utilisées dans ce modèle sont :

- Produit scalaire : $$RSV\ (d_j, q) = \sum_{i=1}^{n} q_i * d_{ij} \qquad (1.14)$$

- Mesure Cosinus : $$RSV\ (d_j, q) = \frac{\sum_{i=1}^{n} q_{i*}\, d_{ij}}{\left(\sum_{i=1}^{n} q_i^2\right)^{1/2} * \left(\sum_{i=1}^{n} d_{ij}^2\right)^{1/2}} \quad (1.15)$$
- Mesure Jaccard : $$RSV\ (d_j, q) = \frac{\sum_{i=1}^{n} q_{i*}\, d_{ij}}{\sum_{i=1}^{n} q_i^2 + \sum_{i=1}^{n} d_{ij}^2 - \sum_{i=1}^{n} q_{i*}\, d_{ij}} \quad (1.16)$$

Les documents retrouvés sont présentés dans un ordre décroissant de leur degré de similarité correspondant.

L'un des avantages du modèle vectoriel réside dans sa simplicité conceptuelle et de mise en œuvre. En outre, il permet de trier les résultats d'une recherche à travers une mesure de similarité document-requête, en plaçant en tête les documents jugés les plus similaires à la requête. Cependant, ce modèle ne permet pas de modéliser les associations entre les termes d'indexation. Chacun des termes est considéré comme indépendant des autres. Wong et al. [43] ont proposé un modèle vectoriel généralisé (Generalized Vector Space Model) qui corrige cette indépendance entre les termes. Dans ce modèle, chaque terme est représenté par un vecteur dans un espace vectoriel dont les axes sont orthogonaux par construction. Ces axes sont les produits logiques des termes d'indexation. Un document est représenté par la moyenne des vecteurs de ses termes.

Croft et al. [44] ont aussi estimé que ce modèle manque d'une base théorique saine comparé au modèle probabiliste qu'on décrira dans la suite. Sauf que, même si ce modèle est critiqué sur cette lacune, il reste l'un des modèles les plus utilisés et les plus étudiés.

2.3.2.2. Modèle connexionniste

Ce type de modèle se base sur les fondements des réseaux de neurones biologiques, tant pour modéliser des documents et leurs informations descriptives (termes, auteurs, mots clés, etc.), que pour la mise en œuvre du processus de recherche d'information [7]. Son objectif est d'imiter quelques fonctions du cerveau humain en reproduisant certaines de ses structures de base.

Le fonctionnement du réseau se fait par propagation de signaux à partir de la couche d'entrée jusqu'à la couche de sortie [14]. Chaque neurone de la couche d'entrée reçoit une valeur d'activation, calcule une valeur de sortie, puis la transmet vers les neurones qui lui sont reliés dans la couche suivante. Ce processus se reproduit jusqu'à la couche de sortie. Les sorties finales (valeurs de sortie de la dernière couche) servant de critère de décision.

Les neurones formels d'un réseau artificiel représentent des objets de la recherche d'information. Un réseau de neurone formel est construit à partir des représentations des documents et de la requête initiale. Chaque couche du réseau est un ensemble de neurones formels représentant un concept donné (requête, termes, documents, etc.) [7]. Le mécanisme de recherche d'information est fondé sur le principe de propagation de valeurs depuis les neurones descriptifs de la requête vers ceux des documents, à travers les connexions du réseau. Les résultats sont retournés à l'utilisateur selon le niveau d'activation des neurones documents.

Les modèles de recherche d'information basés sur les réseaux de neurones sont une solution pour combler les lacunes des modèles vectoriels. Les modèles connexionnistes sont connus pour leur capacité d'apprentissage, ce qui permet aux SRIs les utilisant de devenir adaptatifs. La notion de réseau est également convenable pour représenter les relations qui existent entre les termes (ex.

synonymie, voisinage), entre les documents (ex. similitude, référence), et enfin entre les termes et les document (ex. fréquence, poids). Cependant, il n'existe pas de représentation unique d'un réseau de neurones pour la recherche d'information, c'est au constructeur du modèle de spécifier ses paramètres, à savoir :

- Le nombre de couches et de neurones par couche ;
- Les différentes couches du réseau (couche d'entrée, de sortie, intermédiaires, etc.), le concept qu'elles représentes, et les neurones de chaque couche ;
- La fonction d'entrée de chaque neurone, la fonction de sortie de chaque neurone, les liens entre les neurones et leurs poids associés.

2.3.2.3. Modèle d'Analyse Sémantique Latente (LSA)

Le principe du modèle LSA (Latent Semantic Analysis Model) [14] se base sur le fait que les idées dans un texte sont plus reliées aux concepts décrits par elles que les termes de l'index utilisés pour la description de ces idées [47]. Ainsi, l'appariement entre un document et une requête donnée devrait être basé sur la correspondance entre les concepts plutôt qu'entre les termes de l'index. Donc l'idée du modèle est d'aboutir à une représentation conceptuelle des documents où les effets dus à la variation d'usage des termes dans le corpus sont nettement atténués [7]. Ainsi, les documents qui partagent des termes co-occurents ont des représentations proches, ce qui permet de retourner un document même s'il ne contient aucun mot de la requête soumisse. Pour ce faire, on se place dans un espace de moindre dimension, l'espace associé aux concepts. Les vecteurs des termes sont convertis dans ce nouvel espace.

Ce modèle se base principalement sur la décomposition en valeur singulières, désignée par SVD (Singular Value Decomposition) de la matrice documents-termes X de dimension t × d, où t est le nombre de termes distincts du corpus, et d le nombre de documents dans ce même corpus. La SVD permet d'une part de réduire l'espace des termes d'indexation, et d'autre part, de représenter les documents et les requêtes dans un espace qui ne dépend pas des termes d'indexation mais des concepts contenus dans les documents. Puisque, LSA affirme que la recherche dans l'espace réduit donne de meilleurs résultats que la recherche dans l'espace des termes de l'index.

Formellement, le modèle LSA permet de trouver une approximation X' de la matrice X en suivant les étapes suivantes :

1. Les documents sont représentés par des vecteurs de termes, qui forment ensemble la matrice terme-document pondérée $X_{t\times d}$, avec t le nombre de termes et d le nombre de documents.
2. La méthode SVD permet de décomposer toute matrice rectangulaire en un produit de trois matrices. Ainsi, la matrice X est transformée comme suit :

$$X_{t\times d} = T_0 S_0 D_0^t \qquad (1.17)$$

avec :

T_0 : une matrice orthogonale de dimension t × n ;

D_0^t : une matrice orthogonale de dimension n×d;

S_0 : une matrice diagonale de dimension n × n, les valeurs sur la diagonale sont les valeurs propres de X et sont par convention toutes positives et ordonnées par ordre décroissant sur la diagonale. La $i^{ème}$ valeur singulière indique le taux de variation tout au long du $i^{ème}$ axe;

n : est le rang de la matrice X (≤ min(t, d)).

Ainsi, les k plus grandes valeurs singulières sont supposées suffisantes pour représenter presque toute l'information de la matrice X. Concrètement, toutes les valeurs d'indice i (i > k) sont supposées nulles, et l'équation (1.17) est transformé en utilisant la matrice S de dimension k × k, qui représente l'approximation de S_0 réduite aux k premières dimensions. Le résultat de la nouvelle transformation est donné par le modèle réduit suivant :

$$X \approx X' = T.S.D^t \qquad (1.18)$$

avec :

T : une matrice orthogonale réduite de dimension t × k (k ≤ n) ;

D^t *:* une matrice orthogonale réduite de dimension k × d;

S : une matrice diagonale réduite de dimension k × k ;

3. Une requête Q est, comme tout document, un ensemble de termes. Elle peut être représentée dans le nouvel espace des documents comme suit :

$$X_q = Q.T.S^{-1} \qquad (1.19)$$

où Q est le vecteur des mots de la requête, pondéré par les termes appropriés, S^{-1} est la matrice inverse de S.

4. Une valeur de similarité est ensuite calculée entre le vecteur de la requête *Xq* et chaque document, tous deux représentés dans le nouvel espace vectoriel.

D'après [46], le principal inconvénient de cette méthode est qu'elle n'est pas souple pour certains types d'applications dont le filtrage. En effet, la performance et la stabilité du SRI dépendent largement de la quantité et de la qualité des données traitées. Si le nombre de documents utilisés est faible, le calcul de X′ ne donne pas une bonne approximation de X et le processus devient erroné.

2.3.3. Modèles probabilistes

Le principe de base de ces modèles consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Le premier modèle probabiliste a été proposé au début des années 1960 par Maron et Kuhns [48], et a connu des développements, car les approches basées sur ces modèles ont obtenu de très bons résultats, par exemple le modèle BM25 [49].

2.3.3.1. Modèle probabiliste classique

Étant donnée une requête utilisateur, il y a un ensemble des documents qui ne contient que les documents pertinents. Cet ensemble est appelé l'ensemble de réponse idéal. Répondre à une requête revient donc à spécifier les propriétés de cet ensemble idéal afin de retrouver les documents qui le composent.

Comme les propriétés de l'ensemble idéal ne sont pas connues au moment de la requête, il faut d'abord deviner ce qu'il pourrait être. Cette première tentative permet de générer une première description probabiliste de l'ensemble, qui est ensuite utilisée pour retrouver un premier ensemble de documents. Il faut ensuite une interaction avec l'utilisateur pour améliorer l'échantillon représentant cet ensemble idéal [14].

Ce modèle cherche à estimer la probabilité qu'un document d_j du corpus soit pertinent pour une requête q pour définir son appartenance ou non à l'ensemble. Pour ce faire, la requête et les documents sont représentés par des vecteurs booléens dans un espace à n dimensions. Un exemple de représentation d'une requête q et d'un document d_j est le suivant : $d_j = (w_{1,j}, w_{2,j}, w_{3,j}, .., w_{n,j})$, $q = (w_{1,q}, w_{2,q}, w_{3,q}, .., w_{n,q})$. Avec $w_{k,j} \in [0\ 1]$ et $w_{k,q} \in [0\ 1]$. La valeur du poids d'un terme $w_{k,j}$ (resp. $w_{k,q}$) indique si le terme t_k apparaît ou non dans le document d_j (resp. q) [11].

Le processus de décision vient par la suite compléter le procédé d'indexation probabiliste, en estimant la pertinence entre un document d_j et la requête q par le calcul de deux probabilités conditionnelles :

– $P(w_{ij}/Pert)$: probabilité que le terme t_i apparaît dans le document d_j sachant que ce dernier est pertinent pour la requête.

– $P(w_{ij}/NomPert)$: que le terme t_i apparaît dans le document dj sachant que ce dernier n'est pas pertinent pour la requête.

En supposant l'indépendance des termes des documents, le score d'appariement entre un document dj est une requête q est donné par :

$$RSV(d_j, q) = \sum_{i=1}^{n} \frac{P(w_{ij}/Pert)}{P(w_{ij}/NomPert)} \tag{1.20}$$

Ce modèle a donné lieu à de nombreuses extensions et sa mesure d'appariement peut se calculer par différentes formules [25]. Il est à l'origine du système OKAPI [49] proposé par Robertson et al. qui permet de se passer de l'intervention de l'utilisateur, et dont le modèle nommé BM25 calcul le poids des termes dans un document en intégrant différents aspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents :

$$w_{ij} = log\left(\frac{n-df+0.5}{df+0.5}\right) \times \frac{(k_1+1)\times TF_{ij}}{k_1 \times \left((1-b)+b\times\frac{dl_j}{avgdl}\right)+TF_{ij}} \tag{1.21}$$

où dl_j est la taille du document d_j, $avgdl$ est la moyenne des tailles des documents dans le corpus, k_1 et b sont des paramètres qui dépendent de la collection ainsi que du type des requêtes et TF_{ij} la fréquence du terme i dans le document d_j.

2.3.3.2. Les modèles de langues

Dans les modèles de probabilistes dites classiques, on cherche à estimer la probabilité que le document soit pertinent pour la requête pour qu'il puisse y répondre. Ces modèles repose sur le fait qu'un document n'est considéré pertinent que s'il ressemble à la requête. L'hypothèse de base des modèles de langage est différente, ils considèrent qu'un utilisateur en interaction avec un système de recherche fournit une requête en pensant à un ou plusieurs documents qu'il souhaite retrouver [14]. Donc, un document n'est pertinent que si la requête utilisateur peut être inférée par ce document. On cherche alors à estimer la probabilité que la requête soit générée à partir du document [51, 52].

Ainsi étant donné une requête q = t_1 t_2 t_3 …t_n, chaque document d_j est considéré comme un sous-langage pour lequel on construit un modèle de langue θ_{d_j}(noté également D_j). Le score de pertinence du document d_j face à une requête q est alors estimé par la mesure suivante :

$$RSV\left(q, d_j\right) = P\left(q\middle|\theta_{d_j}\right) = P\left(t_1 t_2 \dots t_n\middle|\theta_{d_j}\right) = \prod_i^n P(t_i|\theta_{d_j}) \quad (1.22)$$

Tel que la requête est considérée comme une séquence de termes indépendant, et $P\left(t_i\middle|\theta_{d_j}\right)$ est la probabilité du terme t_i par rapport au modèle de langue θ_{d_j} du document d_j. Cette probabilité s'appuie sur une estimation de la fréquence du terme dans le document d_j. Cette dernière est calculée en utilisant l'Estimation du Maximum de Vraisemblance (EMV), comme suit :

$$P\left(t_i\middle|\theta_{d_j}\right) = \frac{f(t_i)}{\sum_{t_k \epsilon\, Q} f(t_k)} \quad (1.23)$$

Avec $f(t_i)$ est la fréquence de t_i dans le document d_j.

Toutefois, cette probabilité est annulée pour les documents ne contenant pas tous les termes de la requête. Dans ce cas, le score de pertinence du document est nul tandis qu'il pourrait partiellement répondre au besoin en information de la requête. Afin de pallier à ce problème, des techniques de lissage [53, 54] ont été proposées. Ces dernières s'appuient sur un modèle dit de référence, en l'occurrence celui du corpus [50], et affectent une probabilité non nulle aux mots absent du document, et ce en redistribuant la masse des probabilités observées [55].

3. Évaluation des systèmes de Recherche d'Information

L'étape finale du processus de recherche d'information est l'évaluation. Cette étape détermine la réalisation ou non du but initial de recherche par le système ou l'approche utilisé. Cette activité met en rapport les moyens utilisés avec les résultats obtenus i.e. la métacognition [12]. Elle permet également de paramétrer et stabiliser les techniques utilisées lors de leurs différentes étapes.

L'évaluation d'un système de recherche d'information peut être appréhendée selon deux aspects : un aspect efficacité et un aspect efficience [7]. L'aspect efficacité concerne la capacité du système à sélectionner un maximum de documents pertinents et un minimum de documents non pertinents. L'aspect efficience dépend de l'évaluation cognitive et ergonomique de l'utilisateur, tels que la facilité d'utilisation du système, rapidité d'accès, temps de réponse à une requête, présentation des résultats, etc. Cet aspect évalue le système en fonction de la satisfaction

que retire l'usager non sur ses performances techniques. Dans cette section, nous nous intéressons principalement à présenter l'aspect efficacité.

L'évaluation de l'efficacité d'un SRI repose en général sur trois principaux éléments :

- Une collection de document de test,
- Un ensemble de requête,
- Une liste de documents pertinents pour chaque requête, produite par des connaisseurs du domaine de recherche.

En utilisant ces trois éléments, les taux de performance des SRIs sont mesurés par les différentes mesures d'évaluation que nous décrivons ci-après.

3.1. Notion de pertinence

La notion de pertinence est au centre de la problématique de la recherche d'information, donc mieux la définir permettrait certainement de mieux comprendre l'intention des individus lorsqu'ils cherchent de l'information. Depuis la fin des années 1990, ce concept a été proposé comme un candidat possible à l'élaboration d'une théorie pour unifier les travaux sur le comportement de recherche des usagers de l'information, en sciences humaines et sociales, et ceux qui cherchent à concevoir des SRIs pour les aider dans leurs quêtes, en informatique et en intelligence artificielle [56].

Le concept de pertinence dépend notamment du centre d'intérêt ou du domaine d'application, mais également du moment, du lieu et du support que l'usager a choisi pour accéder à l'information, et enfin du système qui délivre cette information [57]. Cette notion est liée au jugement individuel d'un utilisateur (pertinence utilisateur) et elle est prédite par les SRIs (pertinence système). Toutefois, pour mieux comprendre cette notion assez ambigüe il faut s'approfondir dans sa modélisation en strates proposé par Saracevic [58].

Le modèle d'interaction stratifié de Tefko Saracevic s'inspire des travaux de Schutz en philosophie et de ceux de Sperber et Wilson en pragmatique linguistique. Il souligne que le concept cognitif de pertinence implique la création interactive et dynamique de relations par inférence, et que l'utilisateur est doté d'intentions et est situé dans un certain contexte. Selon lui, la pertinence peut être caractérisée par un certain nombre de caractéristiques : la relation, l'intention, la motivation, le contexte, l'inférence – appréhendée à travers des jugements visant le plus souvent l'efficacité ou le degré de maximisation d'une relation donnée. A cet effet, il propose de modéliser l'interaction entre l'utilisateur et le SRI comme une suite de processus se produisant dans un système de strates (figure 1.7) qui distingue différents types de manifestations de la pertinence :

- ***La pertinence système ou algorithmique,*** c'est-à-dire l'évaluation par le système de l'adéquation entre documents et requête. Puisque, les SRIs doivent s'appuyer sur un modèle de pertinence qui leur permet de mesurer pour chaque document du corpus utilisé un score de pertinence. Cette pertinence est traduite par une mesure algorithmique dépendante des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte. La pertinence apparaît donc ici comme une valeur

numérique calculée par les systèmes. Cependant, cette pertinence a des limites car elle est estimée à partir d'un score de ressemblance entre les termes de la requête et ceux des documents.

- ***La pertinence-sujet ou thématique,*** basé sur l'appariement entre le sujet de la requête et celui des documents. Elle traduit le degré d'adéquation de l'information retrouvée au thème et non au contenu de la requête.
- ***La pertinence cognitive,*** adéquation d'un document ou d'un objet informationnel à l'état des connaissances et le besoin d'information de l'utilisateur ; elle est inférée à partir de la correspondance avec ses connaissances (l'information doit être compréhensible par lui et lié aux connaissances dont il dispose sur le sujet), de la nouveauté, de la qualité de l'information et des goûts de l'utilisateur ; il s'agit de la pertinence liée au thème de la requête, pondérée par la perception ou les connaissances de l'utilisateur sur ce même thème [11].
- ***La pertinence situationnelle ou utilité,*** adéquation entre la tâche ou le problème à résoudre et les textes retrouvés. Ce type de pertinence traduit principalement l'utilité de l'information relativement au but de recherche de l'utilisateur. Elle est inférée à partir de leur valeur d'usage pour la prise de décision, des préférences de l'utilisateur, de la correspondance de l'information à la solution du problème et de la réduction d'incertitude.
- ***La pertinence affective ou motivationnelle,*** un aspect subjectif de la pertinence qui représente l'adéquation entre les intentions, les buts et les motivations de l'utilisateur et les documents restitués. Elle est inférée à partir de la satisfaction, du succès, de la réussite et des préférences de l'individu. En effet, deux utilisateurs différents ayant soumis la même requête ne jugent pas les réponses du système de la même manière. Ce désaccord est dû au fait que les besoins sont différents, tandis que le même besoin peut être exprimé différemment en fonction de l'utilisateur.

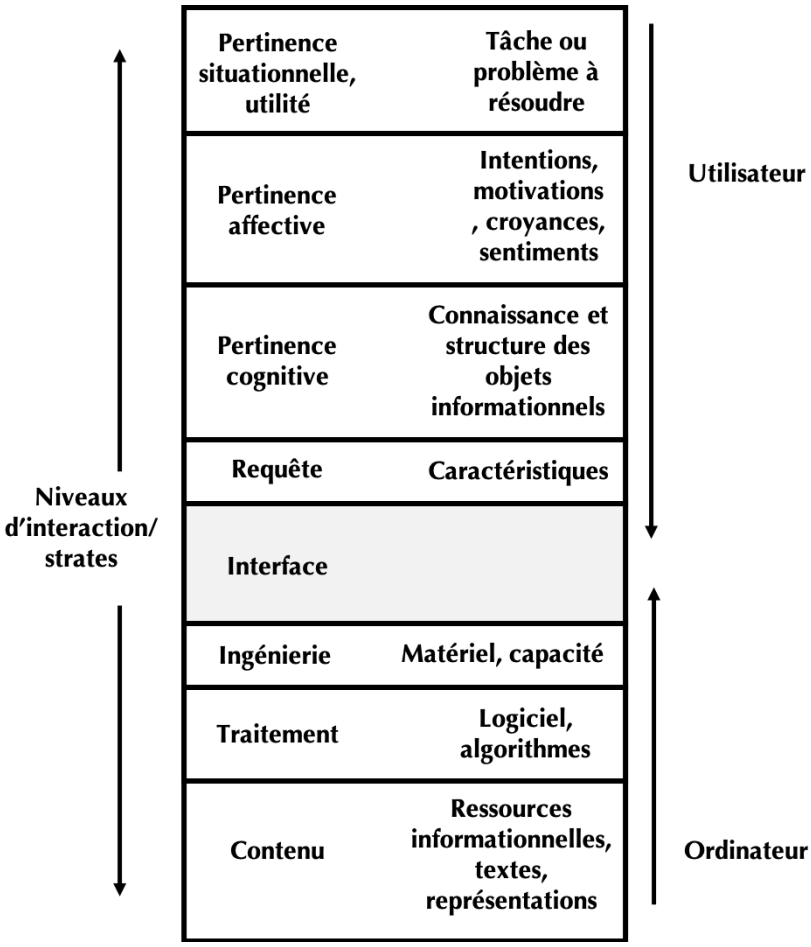

Fig. 1.7 Le système de strates de Saracevic [56]

Saracevic souligne [59,60] l'interdépendance entre ces strates de pertinences, au sens où un type de pertinence affecte les autres, et estime qu'aucune strate ne doit être ignorée dans la conception des systèmes.

Le modèle stratifié est un point de départ pour mieux modéliser l'utilisateur et améliorer les retours du système, en permettant de distinguer les différents niveaux qui interviennent dans les jugements de pertinence des utilisateurs et en modélisant leur relation avec la pertinence système.

3.2. Collections de test

Dans l'objectif d'évaluer un modèle ou une technique dans le domaine de recherche d'information, il faut évaluer le degré de pertinence des réponses du système utilisé à plusieurs requêtes. Pour ce faire, il est nécessaire d'utiliser ou de constituer des collections de test qui recoupent un corpus de documents, un ensemble de requêtes et des jugements de pertinence associés qui recensent les documents pertinents pour chaque requête. Cette procédure d'évaluation correspond au paradigme proposé par Cranfield en 1997 [61] qui a suscité le développement de plusieurs campagnes d'évaluation et qui est utilisé et jugé efficace jusqu'à maintenant. Ces campagnes présentent l'avantage de se focaliser sur une tâche particulière et d'évaluer l'efficacité des systèmes répondant à cette tâche [50].

Corpus de documents : il s'agit de l'ensemble des documents à indexer qui représentent les informations accessibles et exploitables sur lesquelles le système sera évalué. Dans le cas général et pour un souci d'optimalité, la base documentaire constitue des représentations simplifiées mais suffisantes de documents. Ces représentations sont réalisées de telle sorte que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) se font avec le moindre coût [11].

Ensemble de requêtes : Il est souhaitable que la procédure d'évaluation prenne en compte des requêtes dans la forme où elles ont été soumises au système, et non un besoin ou une question exprimée sous forme libre et détaillée.

Jugements de pertinence : ils indiquent pour chaque document s'il est pertinent ou pas, et parfois même à quel degré il l'est, pour chaque requête. C'est la tâche la plus difficile lors de la création d'une collection de test [25]. Pour établir ces listes de documents pour toutes les requêtes, les utilisateurs (ou des testeurs simulant des utilisateurs), appelés juges, doivent examiner chaque document de la base de document, et juger s'il est pertinent par rapport à une requête donnée. Dans quelques campagnes d'évaluation, les corpus de documents contiennent plus d'un million de documents, ce qui rend impossible le jugement exhaustif de pertinence. Ainsi, dans le cas de grandes collections, les jugements de pertinence sont construits selon la méthode de pooling illustrée dans la figure 1.8, qui restreint le jugement des juges au n premiers documents retrouvés par les systèmes participants (Tel que la valeur de n est définie par les organisateurs de la campagne à 100, 1000 ou autre).

Les campagnes d'évaluation les plus connues sont :

- La campagne TREC (Text REtrieval Conference), une des premières campagnes qui regroupe à un large panel de tâches, telles que la recherche ad-hoc, les taches de recherche dans les microblogs ou celles orientées pour les systèmes de questions-réponses ;
- La campagne INEX (Initiative for the Evaluation of XML Retrieval), oriente ses tâches de recherche vers des collections de documents structurés ;
- La campagne CLEF (Conference and Labs of the Evaluation Forum) propose des campagnes dans des langues différentes de l'anglais et traite majoritairement dans les campagnes TREC. En plus de proposer des tâches de recherche sur des documents, celle-ci fournit également des collections d'images associées à des annotations.

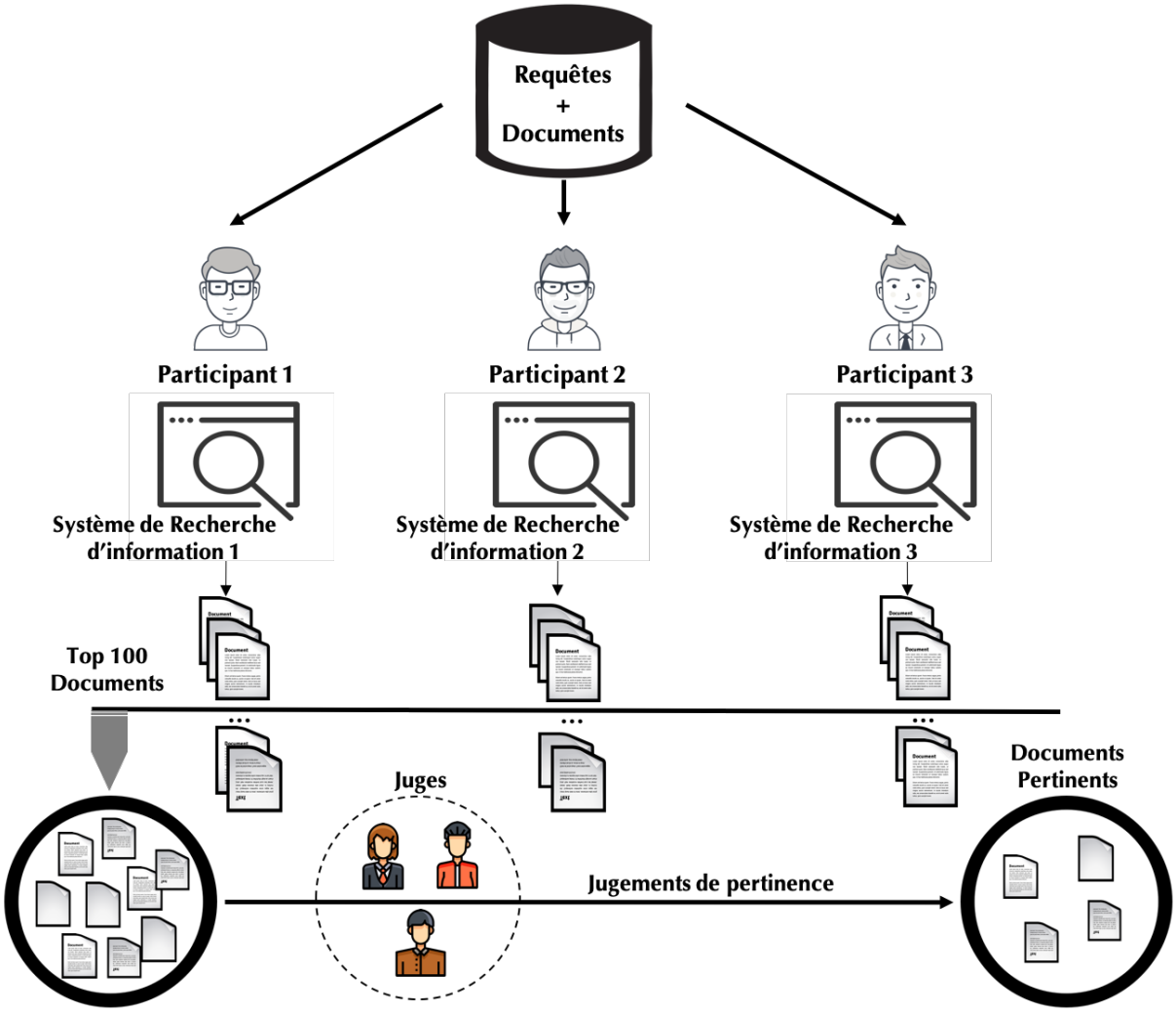

Fig. 1.8 Protocole de jugement de pertinence pour les campagnes d'évaluation officielles

Dans ces campagnes, l'évaluation est centrée principalement vers des tâches ad-hoc, sans considérer la dimension utilisateur. Avec l'émergence de la RI contextuelle et interactive, de nouvelles tâches sont apparues. Nous citons parmi d'autres :

- La tâche TREC Interactive, dont l'objectif est la résolution d'un besoin complexe. Les participants doivent alors fournir les fichiers journaux (requêtes soumises, documents visités, ...) qui recensent les interactions des utilisateurs.
- La tâche TREC Contextual Suggestion, qui consiste en la suggestion de documents à partir d'un ensemble de profils d'utilisateurs représentés par leurs préférences et d'un contexte traduit par la localisation d'utilisateurs.
- La tache TREC Session Search, qui consiste en le tri des documents vis-a-vis d'une requête spécifique, qui a été soumise à un moment donné de la session, à partir de l'historique de recherche d'un utilisateur (requêtes reformulées antérieurement et leurs ordonnancements et jugements de pertinence associés).

En dehors de ces campagnes d'évaluation, d'autres protocoles d'évaluation ont été proposés. Nous dénombrons trois procédures d'évaluation autre que celle de Cranfield [62] :

- ***Les expérimentations basées sur la simulation*** [63] qui permettent de simuler le comportement des utilisateurs en construisant le scenario de recherche afin d'évaluer le modèle selon des cas d'utilisation bien particuliers.
- ***Les expérimentations basées sur les fichiers journaux*** de systèmes de recherche (*log study)* [64], qui permettent d'évaluer un modèle de RI sur des données réelles. Ces expérimentations ne reposent pas sur des jugements explicites mais infèrent implicitement le comportement des utilisateurs à partir des fichiers logs.
- ***Les expérimentations utilisateurs*** (*user study*) [65], qui sont généralement effectuées en laboratoire où les participants sont confrontés au modèle de RI en situation réelle de recherche.

3.3. Mesures d'évaluation

L'objectif ultime des systèmes de recherche d'information est de minimiser la distance entre la pertinence utilisateur et la pertinence système. Plusieurs mesures standards ont été proposées pour évaluer les performances des SRIs. Dans ce qui suit, nous introduisons les principales mesures les plus utilisées qui permettent d'estimer quantitativement l'efficacité d'un système. Le but étant d'identifier la capacite du système à retourner des documents pertinents pour chaque requête soumise, ainsi qu'à ne pas retourner les documents non pertinents. Ces mesures permettent également la comparaison de plusieurs modèles ou la mise au point de leurs paramètres.

La Figure 1.9 illustre les différents ensembles manipulés lors de l'évaluation de l'efficacité d'un SRI : l'ensemble des documents retournés par le système et l'ensemble des documents pertinents pour la requête. Les documents pertinents non retournés forment l'ensemble ***silence*** tandis que les documents non pertinents retournés constituent le ***bruit*** [50]. Donc, la principale difficulté d'un système de RI est de reposer sur des outils qui lui permettront de minimiser le silence sans augmenter le bruit, mais plutôt le minimiser également.

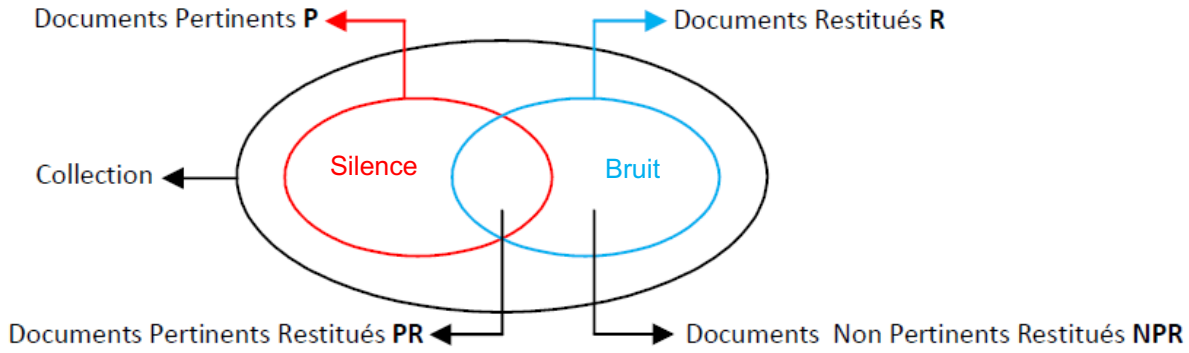

Fig. 1.9 Partition du corpus des documents pour une requête

Afin d'évaluer la qualité du système, chaque requête qi est évaluée individuellement au moyen d'une mesure statistique estimée. La mesure est ensuite agrégée sur l'ensemble des requêtes du corpus de test afin d'obtenir la mesure d'efficacité moyenne. Deux catégories de mesures sont

identifiées dans la littérature : des mesures orientées rappel et précision et d'autres mesures orientées rang.

3.3.1. Mesures orientées rappel et précision

Ces mesures évaluent l'efficacité du tri des documents retournés en se basant sur le nombre de documents pertinents restitués par le système. Nous distinguons trois principales mesures :

- Le rappel mesure la capacité d'un système à retourner tous les documents pertinents, et par conséquent, sa capacité à minimiser le silence, illustré sur la Figure 1.9. Pour une requête donnée, il est estimé par le ratio du nombre de documents pertinents retournes par le système par rapport au nombre de documents pertinents pour cette même requête dans tout le corpus. Ce ratio est ensuite agrégé sur l'ensemble des requêtes $q_i \in Q$:

$$Rappel = \frac{1}{|Q|}\sum_{q_i \in Q} Rappel(q_i) = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|R_{q_i} \cap P_{q_i}\right|}{\left|P_{q_i}\right|} \quad (1.24)$$

  ou R_{q_i} est l'ensemble des documents sélectionnés par le système pour la requête q_i et P_{q_i} l'ensemble des documents pertinents pour la requête q_i.

- La précision mesure la capacité d'un système à ne retourner que des documents pertinents, et par conséquent, sa capacité à minimiser le bruit. Pour une requête donnée, la précision est estimée par le ratio du nombre de documents pertinents restitués par le SRI par rapport au nombre total de documents restitués pour cette même requête. Ce ratio est agrégé sur l'ensemble des requêtes $q_i \in Q$ pour calculer la précision du système évalué :

$$Presicion = \frac{1}{|Q|}\sum_{q_i \in Q} Precision(q_i) = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|R_{q_i} \cap P_{q_i}\right|}{\left|R_{q_i}\right|} \quad (1.25)$$

- La F-mesure, une mesure qui combine le rappel et la précision estimées afin de mesurer la capacite du système à retourner le maximum de documents pertinents, et seulement ces documents-ci, en se focalisant sur le double objectif d'un SRI, c'est-à-dire minimiser le bruit et le silence. La F-mesure est calculé en utilisant l'expression suivante :

$$F_\beta = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{(1+\beta^2) * Rappel(q_i) * Precision(q_i)}{\beta^2 * Rappel(q_i) + Precision(q_i)} \quad (1.26)$$

  Avec β un paramètre qui traduit l'importance donnée au rappel par rapport à la précision.

- Un cas particulier de la F-mesure est la moyenne harmonique (Dans le cas particulier où β=1). Cette mesure permet de considérer une importance égale pour le rappel et la précision. L'expression de la moyenne harmonique est la suivante :

$$F = \frac{1}{|Q|}\sum_{q_i \in Q} 2 * \frac{Rappel(q_i) * Precision(q_i)}{Rappel(q_i) + Precision(q_i)} \quad (1.27)$$

3.3.2. Mesures orientées rang

Ces mesures ajoutent un niveau d'analyse supplémentaire en mesurant la fiabilité des ordonnancements en limitant le calcule aux documents les mieux classés par le système mais surtout grâce à la considération des rangs de classement des documents pertinents [50]. Ces mesures sont dites également de haute précision, car l'idée derrière est qu'un système qui retourne en tête de liste un grand nombre de documents pertinents obtient des valeurs supérieures pour ces mesures par rapport à un autres système pour lequel les documents pertinents sont dispersés dans la liste restituée. Les trois mesures principales les plus utilisées de cette catégorie sont :

- Precision@x ou P@x : c'est la précision à un rang x, c'est-ç-à-dire à un niveau de coupe précis x. Cette précision mesure la proportion des documents pertinents restitués parmi les *x* premiers documents retournés par le SRI.

$$Presicion@x = \frac{1}{|Q|}\sum_{q_i \in Q} Precision(q_i)@x = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|R_{q_i}@x \cap P_{q_i}\right|}{\left|R_{q_i}@x\right|} \quad (1.28)$$

- La Moyenne des précisions moyennes MAP (Mean Average Precision) correspond à la moyenne des valeurs de précision moyenne non interpolées obtenue à chaque rang calculé en considérant les x premiers résultats de la liste l_i retournée à partir du corpus C pour la requête q_i :

$$MAP@x = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{1}{x} \sum_{X=1}^{x} Precision(q_i)@\mathrm{X} \quad (1.29)$$

- Le rang inverse moyen au rang x, MRR (Mean Reciprocal Rank) également permet de favoriser l'hypothèse que le système doit retourner les documents pertinents en premier sur la liste d'ordonnancement par rapport aux x premiers documents retournés. Ainsi, cette métrique estime le rang moyen $Rang(l_i)@x$ du premier document pertinent dans les listes de résultats l_i parmi les x premiers documents retournés en réponse à la requête q_i :

$$MRR = \frac{1}{|Q|}\sum_{q_i \in Q} \sum_{l_i} \frac{1}{Rang(l_i)@x} \quad (1.30)$$

4. Conclusion

Ce chapitre présente, dans un premier temps, les concepts de bases de la recherche d'information ainsi que les principaux modèles de recherche et d'indexation. Dans un second temps, nous avons présenté les composantes principales du mécanisme d'évaluation des systèmes de recherche d'information.

La plupart des activités humaines sont situées géographiquement. Il n'est donc pas surprenant qu'une grande quantité de documents Web et des requêtes soumissent au moteur de recherche contiennent des références géographiques. Une étude réalisée sur le moteur de recherche Excite montre que sur cinq requêtes, il y a une requête qui a un contexte géographique [66]. Dans le chapitre suivant, nous avons axé notre état de l'art sur la recherche d'information géographique (RIG) qui, en complément des approches précédentes, considère des ressources supplémentaires généralement hiérarchiques afin de permettre une représentation géographique.

Chapitre 2 – La Recherche d'Information Géographique (RIG)

1. Introduction

Étant donné que la plupart des activités humaines sont situées géographiquement, il n'est donc pas surprenant que la plupart des documents sur le Web contiennent des références spatiales. Ce fait a été prouvé par une étude menée par Aloteibi et Sanderson en 2014 [77] sur le moteur de recherche Excite, qui montre qu'un utilisateur sur cinq ayant reformulé ses requêtes recherchait des informations situées dans l'espace géographique.

En outre, le traitement des requêtes géographique en tenant compte de la portée géographique, est très utile pour affiner la requête de l'utilisateur et améliorer les performances des SRIs. C'est la raison pour laquelle la plupart des critiques des moteurs de recherche étaient liées à leur ignorance des contraintes géographiques sur les requêtes des utilisateurs [78]. Par conséquent, les résultats récupérés sont moins pertinents. Cela pourrait s'expliquer par le fait que les moteurs de recherche traitent généralement les requêtes en adoptant une approche généraliste de correspondance de mots clés sans passer par l'interprétation des parties spatiales des requêtes.

Cependant, les moteurs de recherche renvoient de meilleurs résultats quand ils analysent des fonctionnalités telles que les termes géographiques dans les pages Web, les requêtes des utilisateurs ou l'emplacement de l'utilisateur lors de recherche [79]. Cela pourrait même être d'une grande valeur commerciale car il permet une publicité spécifique à l'emplacement cherché et améliore la recherche d'entreprises locales. De plus, il peut également promouvoir des régions géospatiales culturelles à l'aide d'approches telles que les systèmes de recommandation [80], qui mesure la similitude entre les utilisateurs en termes d'historiques de localisation et recommande à chaque utilisateur un groupe de lieux potentiels et d'amis dans une communauté géographique.

En général, ce genre d'approches vont de l'intégration des répertoires d'entreprises (pages jaunes) pour répondre à des requêtes simples mais lucratives (par exemple pour les hôtels, boutiques et restaurants) à une analyse plus détaillée des requêtes cherchant à interpréter et analyser le contenu des pages dans le but de faciliter et améliorer les retours des requêtes géographiques plus générales.

En raison du rôle primordial de la géographie dans les requêtes de recherche et du potentiel commercial significatif de ces requêtes (par exemple pour l'immobilier, les entreprises locales ou les hôtels), notre thèse se focalise sur des méthodes qui permettent d'abord d'extraire l'information géographique du texte puis celles qui visent à apporter des meilleures réponses aux requêtes géographiques. Plus précisément dans ce chapitre, nous explorant le processus du traitement de l'information géographique présente dans le texte, de l'extraction à la recherche et du classement de l'information géographique ; en passant par l'indexation géographiquement spécifique et l'étape de désambiguïsation des noms de lieux ; puis ; enfin à l'analyse des méthodes de reformulation de requêtes géographiques présentes dans la littérature.

2. Le Système de Recherche d'Information Géographique

L'architecture d'une approche typique d'un système GIR comporte quatre étapes comme le montre le processus de recherche présenté sur la figure 2.1. Ces quatre étapes sont :

- La reconnaissance et la classification des entités nommées (NERC pour Named Entity Recognition and Classification) ;
- Résolution toponymique (TR pour Toponym Resolution) ;
- Indexation géographique ;
- Recherche d'information en utilisant une requête géographique.

Par rapport à un système de recherche d'information traditionnel, NERC et TR sont utilisés pour identifier les emplacements géographiques d'une façon unique dans la collection et les requêtes, de sorte que leur présence puisse être prise en compte lors de la récupération et le classement des documents retournés.

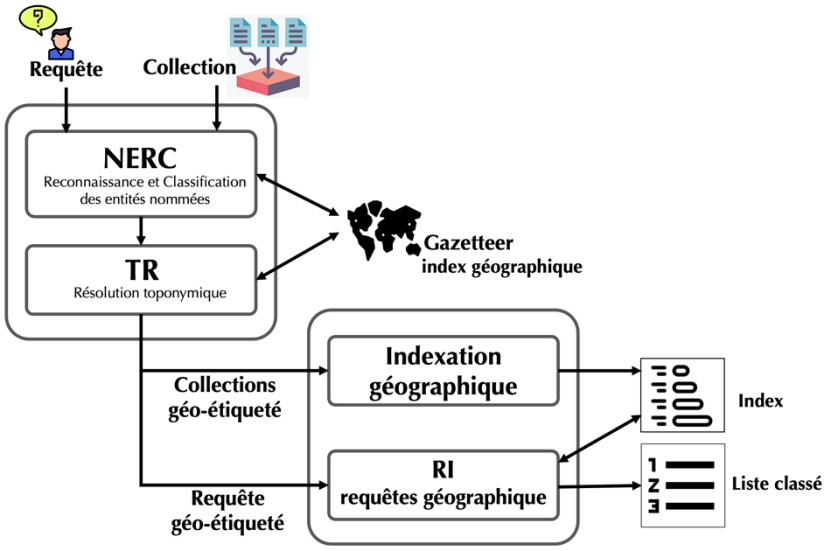

Fig. 2.1. Composantes d'un système de recherche d'information géographique.

La tâche NERC se préoccupe d'identifier les toponymes, en détectant d'abord les entités nommées, puis en différenciant entre ceux qui référencent des noms de lieux (qui nous intéressent) et ceux qui représentent des noms de personnes ou d'organisations (ce qui ne nous intéressent pas). La résolution toponymique, également connue sous le nom « Désambiguïsation de Toponyme », est la tâche suivante consistant à attribuer un identifiant d'emplacement unique à chaque entité étiquetée comme un nom de lieu. La TR est similaire à la désambiguïsation du sens des mots, dans le sens où le contexte autour du nom de lieu dans le texte est utilisé pour déterminer le lieu géographique exacte indiqué par l'entité nommée détectée.

Ces deux étapes peuvent être désignées collectivement comme une annotation. Une fois l'annotation de la collection terminée, la tâche suivante consiste à créer un index spatial qui permet d'exploiter les relations spatiales lors d'une phase de requêtage ultérieure. Des approches du GIR ont été suggérées par divers auteurs dans ce contexte, et il existe également des exercices d'évaluation coopérative, notamment la piste de recherche géographique *GeoCLEF* du Cross Language Evaluation Forum (CLEF) [83].

Une exemple réel implémenté de Système de RI-géographique est le système SINAI-GIR. La recherche d'information par ce système se fait en deux étapes :

- ***Etape 1 :*** une étape de prétraitement de la collection des documents et des requêtes. Une analyse et interprétation des requêtes est faite pour l'identification des différentes parties de la requête en utilisant l'étiqueteur morpho-syntaxique[1] « TreeTagger », le lemmatiseur « SNOWBALL Stemmer » et un éliminateur de mots vides.
- ***Etape 2 :*** une étape d'indexation des documents de la collection qui permet de générer deux index, un index textuel des termes mots-clés des documents et un autre index géographique, après annotation de toutes les entités géographiques présentes dans chaque document de la collection. L'outil GéoNER est utilisé pour la détection des entités géographiques. Ce dernier se base sur des ressources externes comme GEONAMES, une base de données contenant tous les pays et à peu près 8 millions emplacements géographiques.

À la fin de chaque prétraitement, chaque requête « incluant ses entités spatiales » est exécutée sur le moteur de recherche (TERRIER), et les documents récupérés sont par la suite filtrés et classés, en mettant en premier les documents entrant dans la portée géographique détectée dans la requête. La figure 2.2 décrit le processus de recherche par le système SINAI-GIR.

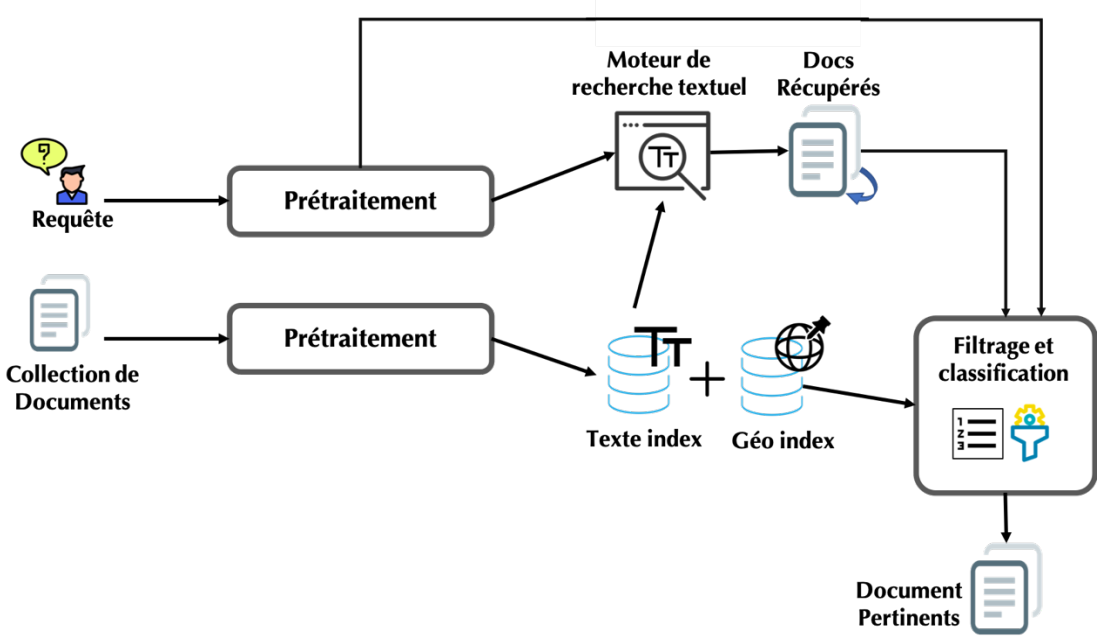

Fig. 2.2. Description du Système RI-Géo SINAI-GIR

L'évaluation de ce processus a été faite à l'aide du Framework GeoClef et par la méthode d'évaluation de TREC « Text Retrieval Conference ».

Le manuscrit [110] également décrit l'architecture d'un système GIR qui utilise une structure d'index, combinant un index textuel et un index spatial, et définit un flux de travail pour extraire les références géographiques des documents afin de résoudre l'inadéquation entre les requêtes et

[1] Un étiqueteur morpho-syntaxique, aussi appelé étiqueteur grammatical, POS tagger (part-of-speech tagger) en anglais, est l'outil informatique qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc.

les documents. Ce système prend en charge trois types de requêtes : les requêtes purement spatiales, les requêtes purement thématiques et les requêtes avec des composants spatiaux et thématiques. Les tests élaborés à l'aide de la collection de documents TREC FT-91 (Financial Times, 1991) ont montrés que ses performances sont jugées acceptables par rapport à une structure d'index utilisant uniquement des indices spatiaux purs.

3. La reconnaissance et désambiguïsation des références géographiques

Les noms de lieux ou toponymes peuvent être utilisés pour désigner des lieux sur Terre, mais ils peuvent représenter également les noms d'organisations et quelques toponymes faisant partie des noms de personnes. Il est également possible que les noms de lieux soient utilisés de manière métonymique[2], par exemple pour faire référence à des entités administratives comme dans les « *talks with Washington* » [81]. Le processus d'analyse géographique (geo-parsing) consiste à explorer le texte pour identifier la présence de noms de lieux et d'autres éléments spatiaux et à distinguer leurs cas d'utilisation pour des intentions géographiques, des cas où ils sont utilisés pour désigner une autre entité. Ce processus est souvent traité comme une extension de la reconnaissance d'entité nommée qui fait partie intégrante de l'analyse linguistique dans le traitement du langage naturel (NLP pour Natural Language Processing). A cet effet, la détection et la résolution de toponymes sont des tâches de traitement du langage naturel, qui font partie intégrante de la plupart des architectures de recherche d'informations géographiques. Sans ces composantes, la détection des synonymes, la résolution des ambiguïtés et l'expansion précise des toponymes ne seraient pas possibles.

Ce sujet est exploré par Stokes et al. [82] avec un regard critique sur les approches de détection et homonymie (« résolution » dans leur terminologie) de toponyme et par conséquent leur annotation dans les documents. En utilisant un ensemble de données GeoCLEF annoté manuellement pour évaluer certaines méthodes NLP standard, ils ont constaté que leurs performances étaient inférieures à celles annoncées précédemment. Ils ont également constaté lors de leurs expérimentations, qu'une approche de détection simple des toponymes, utilisant un répertoire géographique (Gazetteer) surpassait les méthodes NLP. Afin de mapper les noms désambiguïsés face à un identifiant d'emplacement unique ils ont utilisé le thesaurus des noms géographiques Getty (http://www.getty.edu/). Leur étude examine l'interaction entre les erreurs dans les deux étapes du traitement (détection et résolution), constatant que dans certaines situations, les erreurs de détection peuvent être annulées au stade de la désambiguïsation, si ce dernier ne reconnaît pas le nom identifié en tant que toponyme après l'avoir traité en utilisant ses ressources de connaissances toponymiques.

3.1. Extraction de l'information géographique

[2] Métonymie : figure de style par laquelle on désigne le tout par la partie, le contenu par le contenant, etc. Exemple : "*talks with Washington*", pour désigner la communication d'une entité avec le gouvernement des états unis.

La détection automatique de requêtes géographiques ou d'expressions géographiques est une tâche d'extraction d'informations à partir de corpus textuels qui se fait à l'aide des techniques de reconnaissance d'entités nommées (*NER* : Named Entity Recognition). Dans ces approches, les étendues de texte de l'entité mentionnée sont détectées et étiquetées avec des types tels que : emplacement ; organisation ; personne ; etc. La tâche NER permet également de distinguer entre les références métonymiques et les noms de lieux. Ces références sont des occurrences de noms de lieux dans le texte qui ne sont pas utilisés dans leur sens littéral. Par exemple, dans la phrase « Washington défend son invasion de l'Irak », le mot « Washington » fait référence à une entité politique, le gouvernement américain, alors que « Irak » est une référence à une entité géographique.

D'une part, cette tâche a été effectué en 2005 par [84] en utilisant les champs aléatoires conditionnels (en anglais Conditional Random Fields : CRFs), où les CRFs sont des modèles d'apprentissage statistique permettant de prendre en compte l'interaction de variables « voisines ». Ces modèles ont été appliqués en chaîne linéaire sur des séquences de mots.

L'approche se basant sur les CRFs a également été améliorée et réutilisé dans les travaux de Ratinov et Roth en 2009 [85] en ajoutant des caractéristiques provenant de sources lexicales externes (comme les répertoires géographiques) et en agrégant les contextes des différentes occurrences du même terme. Tel que le contexte d'un terme est représenté par une fenêtre de taille 2 (deux termes significatifs avant et deux après). Lors de leurs tests, Ratinov et Roth ont utilisé une collection de 14 listes à haute-précision et faible-rappel extraites du Web qui couvrent les noms communs, les pays, les unités monétaires, les expressions temporelles, etc. Si ces gazetteers utilisés ont une excellente précision, ils ne fournissent pas une couverture suffisante. Donc pour améliorer encore la couverture, ils ont extrait 16 gazetteers supplémentaires de Wikipédia, contenant collectivement plus de 1,5 million d'entités. Au total, ils ont eu 30 gazetteers tel que les correspondances avec chacun d'entre eux sont pondérées comme une caractéristique distincte de la tâche NER. Cela leur a permis d'atteindre un score F1 de 90.8 sur la tâche NER avec la collection CoNLL-2003.

D'autre part, il existe la méthodologie d'annotation sémantique des entités nommées comme sur le travail présenté dans [89]. Ce qui fait la force de cette approche c'est qu'elle prend en compte le coté sémantique spatial. Ce qui a permis de l'implémenter dans un projet de grande envergure tel que le projet « Pyrénées Itinéraires Virtuels » (PIV).

Cette approche utilisée pour l'extraction des marqueurs spatiaux est basée sur un modèle spatial, qui supporte deux types d'entités spatiale (ES) : des Entités Spatiales Absolues (ESA) et des Entités Spatiales Relative (ESR). Les ESA sont des ES nommées et connues comme par exemple : "Paris". Les ESR sont des ES complexes qui doivent être interprétées et traitées selon un raisonnement spatial, comme par exemple : « à l'ouest de la frontière Franco-Espagnol ».

L'approche utilisée est une approche de parcours active. Contrairement aux approches standards de Traitement Automatique de Langue (TAL), la chaîne de traitement linguistique utilisée est appliquée localement aux entités spatiales. C'est-à-dire, les ESAs (noms de villages,

de forêts, etc.) sont d'abord détectées et marquées en se basant sur un lexique composé notamment d'introducteurs spatiaux. Puis, les ESRs sont construites à partir de ces ESAs.

Plus en détail, premièrement un étiqueteur et un diviseur (splitter) parcourent le flux de texte et insèrent des marqueurs de structures logiques et des séparateurs de mots avec leur lemme. Les entités spatiales candidates sont détectées dans une seconde étape : tous les termes commençant par une majuscule et précédés d'un introducteur d'entité spatiale appartenant au lexique prédéfini ("dans", "proche de", etc.) sont marqués. Puis, un étiqueteur morpho-syntaxique (POS tagger) classe ces entités nommées (noms propres, etc.) pour effectuer une analyse basée sur les grammaires à clause définie (DCG pour Definite Clause Grammar) permettant l'interprétation des syntagmes extraits (inclusion, adjacence, distance par rapport à une autre entité spatiale, etc.). Tel que une grammaire à clause définie est un moyen d'exprimer la grammaire, pour les langages naturels ou formels, dans un langage de programmation logique tel que Prolog[3]. Enfin, une étape de confirmation des entités géographiques détectées fait appel à des services externes (gazetteers) pour confirmer chaque ESA candidate. Et toutes les ESR candidates associées à une ESA non valide sont automatiquement supprimées. Une représentation géo-localisée de chaque ES validée est ensuite ajoutée à l'index géographique.

L'évaluation de l'approche est faite à travers un cas d'étude du processus de RI spatiale sur un corpus de test fourni par la Médiathèque Intercommunale à Dimension Régionale (MIDR). Cette évaluation est faite en comparant l'approche sémantique proposée aux approches statistiques, en fonction de la portée de la requête, et selon les critères de précision moyenne sur toutes les requêtes utilisées, et du nombre total de documents trouvés. L'approche sémantique spatiale, intégrant une interprétation de la sémantique des ESs pour récupérer les documents pertinents semble parfaite pour les requêtes purement spatiales. Tandis que la combinaison des deux approches : Approche sémantique spatiale pour le traitement de la sous requête portant sur l'entité spatiale et l'approche classique basée sur les mots-clés pour le traitement de la sous-requête contenant l'entité thématique, améliorent de manière significative « la précision », surtout dans le cas d'une requête incluant les deux types de sous-requêtes (spatiale et thématique).

Plus récemment, l'apprentissage en profondeur (Deep Learning), qui est une nouvelle branche de l'apprentissage automatique ayant été largement appliquée dans les travaux de recherche géospatiale depuis 2016, a fait des progrès significatifs dans le domaine du traitement du langage naturel. Des tâches tels que le géoparage (geoparsing) à partir de texte non-structuré [86] et la correspondance d'adresse sémantique, qui est également une tâche NLP essentielle dans l'apprentissage [87] ont été développées. Cette dernière vise à faire correspondre une adresse, soumisse sous forme d'une requête, avec les enregistrements d'adresses correspondants stockés dans une base de données urbaine. Considérant une adresse comme étant une chaîne formée d'éléments d'adresse consécutifs, la disposition de ces éléments est contrainte par leur hiérarchie spatiale [88], ce qui rend la correspondance entre adresses différente de la correspondance entre d'autres types de séquences de mots.

[3] Prolog : un langage de programmation logique qui définit les règles logiques attendues d'une solution afin de laisser le compilateur les transformer en séquence d'instructions.

3.2. Désambiguïsation des noms de lieux

En règle générale, chaque entité nommée extraite et étiquetée pendant l'étape de reconnaissance d'entité nommée est soumise à la désambiguïsation d'entité nommée (*NED* : Named Entity Disambiguation). En d'autres termes et spécifiquement à l'entité nommée géographique, lorsqu'un nom de lieu est reconnu comme toponyme utilisé dans un sens géographique, la tâche suivante nécessaire est de déterminer de manière unique le lieu auquel le nom fait référence [81]. Car de nombreux noms de lieu sont partagés entre plusieurs endroits, comme Richmond et Springfield par exemple. Le cerveau humain résout ce genre d'ambiguïté en utilisant les connaissances acquises à partir d'indices contextuels dans le texte lu.

La résolution automatique de la portée géographique, également connue en tant que « Résolution de toponyme », dont l'objectif est la clarification des noms de lieux ambigu, a été traité par trois stratégies principales dans la littérature. La première, tente d'imiter les méthodes utilisées par les humains, en considérant ensemble tous les noms de lieux dans le texte traité. Pour ce, elle exploite le contexte linguistique où le toponyme sujet de résolution est mentionné dans le document traité.

Le voisinage du toponyme contient souvent des indices qui aident les lecteurs à l'interpréter. Ces indices peuvent être d'autres entités nommées [97], d'autres toponymes [98], ou encore plus généralement des sujets spécifiques associés plus souvent à un toponyme candidat particulier qu'à d'autres [99, 100]. Généralement, Si un nom de lieu apparaît en association avec un ensemble d'autres noms, dont plusieurs sont des lieux voisins ou sont des instances de lieux dans la même région parente, alors cela fournit des preuves pour distinguer la signification implicite. De même, si le texte mentionne une région parente ou enfant d'une instance du nom, cela peut aider à déterminer le sens particulier souhaité. Concrètement, cela se fait en mappant l'entité détectée aux entités dans une base de connaissances en utilisant la cohérence sémantique des entités sélectionnées et une mesure de similarité contextuelle.

Dans ce sens, Adams et McKenzie [99] évaluent l'utilisation de la modélisation de sujets sur un ensemble d'entrées de blogs de voyage afin de résoudre ce problème par l'identification des thèmes les plus étroitement associés à des lieux précis à travers le monde. En utilisant ces représentations, ils ont pu calculer la similitude des lieux. De plus, en se concentrant sur des sujets individuels ou des ensembles de sujets, ils ont identifié de nouvelles régions où les sujets sont les plus marquants. Ils ont également évalué à l'aide de ces méthodes la manière dont les changements temporels du sens du même toponyme peuvent avoir lieu. Ju et al. [100] ont également abordé la même stratégie en combinant la modélisation de sujets et la cooccurrence d'entité. Leurs méthodes visent principalement à améliorer la désambiguïsation des noms de lieux dans les textes courts.

La plupart des heuristiques de désambiguïsation des toponymes supposent une proximité géographique entre les toponymes d'un même contexte. Toutefois, cette proximité peut être en termes de distance spatiale ou en termes de position dans l'arbre hiérarchique des lieux du monde. Le deuxième cas peut être présenté par le travail de Bensalem et Kholladi [96], qui ont proposé une heuristique contextuelle de désambiguïsation des toponymes basée sur la quantification de la

proximité arborescente entre toponymes. Cette quantification est à appliquer par une mesure de corrélation géographique qu'ils appellent la densité géographique, se basant sur le chemin hiérarchique des toponymes candidats (référents) obtenu à partir de WordNet. Cette approche désigne la proximité arborescente comme la distance de proximité entre les référents toponymes. Les résultats des tests réalisés sur cette dernière ont montré que les toponymes d'un même contexte sont beaucoup plus proches en termes de relations arborescentes qu'en termes de relations spatiales.

La seconde stratégie de résolution, s'appuie sur les propriétés physiques des toponymes pour lever l'ambiguïté de leurs mentions dans les documents. L'heuristique de population ou l'heuristique de distance minimale sont des heuristiques populaires utilisant de telles propriétés. L'heuristique de population désambiguïse les toponymes en prenant, parmi les candidats ambigus, le candidat ayant la plus grande population, tandis que l'heuristique de distance minimale prend l'ensemble des candidats les plus proches les uns des autres [101]. Une heuristique plus récente calcule à partir de Wikipédia un réseau exprimant des toponymes importants et leur relation sémantique avec d'autres entités. Le réseau est ensuite utilisé pour désambiguïser conjointement tous les toponymes d'un document, en utilisant une mesure pour calculer la cohérence sémantique entre entités [102] ou bien en se basant sur la distance textuelle des toponymes qui permet d'obtenir une relation qui encode l'importance des co-occurrences des lieux mentionnés [103].

Overell et Rüger [95] ont également utilisé Wikipédia pour générer un corpus d'apprentissage étiqueté et un modèle de cooccurrence. Leur méthode vise à utiliser les catégories des articles et des liens vers d'autres articles de Wikipédia pour améliorer la connaissance contextuelle et lever l'ambiguïté sur les toponymes d'un texte non structuré.

Stokes et al. [82] ont proposé une approche combinant les deux premières stratégies de résolution, en utilisant la connaissance contextuelle en conjonction avec les deux sources de connaissances externes : Wikipedia et le Thesaurus Getty. L'algorithme proposé consiste en un premier parcours sur les données, où tous les toponymes qui ont des sens candidats dans Getty représentant une grande masse terrestre (telle qu'une nation ou un continent) se voient attribuer un identifiant Getty. Les deux passages suivants utilisent les informations à partir du contexte local pour déterminer le sens correct du toponyme. L'algorithme recherche d'abord les mots déclencheurs géographiques tels que « État », « Comté », « Montagne » ou « Rivière », puis attribue le Getty Id du sens candidat au type géographique défini par Getty qui correspond au mot déclencheur. Ensuite, l'algorithme identifie les toponymes adjacents à l'aide d'un algorithme de traversée de chemin qui trouve le chemin le plus court entre leurs emplacements candidats dans la hiérarchie Getty. Par exemple, bien qu'une relation hiérarchique existe entre Washington, D.C. et Seattle à travers leur holonyme commun « États-Unis », le chemin le plus court entre ces deux noms de lieux est entre les candidats « Washington State » et « Seattle city ». À ce stade du processus de désambiguïsation, il y aura encore un certain nombre de toponymes ambigus dans le texte ; par conséquent, sans informations contextuelles suffisantes, l'algorithme doit recourir à une stratégie de désambiguïsation en *back-off* (retrait) ou *best guess* (meilleure supposition). Ces deux stratégies ont été exploré par les auteurs, sur la base des informations de localisation prédominantes

de Wikipédia et des statistiques démographiques de World Gazetteer. En guise d'étape finale, pour tous les noms de lieux ambigus restants, l'algorithme proposé applique une technique de classement des sens candidats qui attribue des scores de probabilité aux candidats d'emplacement en fonction de leur profondeur dans la hiérarchie Getty, c'est-à-dire que les emplacements de niveau inférieur dans la hiérarchie sont des entités géographiques plus petites et sont donc considérés comme moins probable (par exemple, un village versus un état).

La dernière stratégie de résolution est moins fréquemment utilisée car elle dépend des métadonnées décrivant les documents où les toponymes sont mentionnés. Ces métadonnées sont de natures diverses, mais elles indiquent toutes, directement ou indirectement, des zones géographiques afin d'aider à interpréter les toponymes traités. Ces métadonnées peuvent être la géolocalisation de publication sur les réseaux sociaux, comme sur le travail de Zhang et Gelernter [104] qui pour désambiguïser un toponyme contenu dans un message sur Twitter ont pondéré les différents champs du message en utilisant l'apprentissage automatique supervisé. Ces caractéristiques ainsi que celles extraites à partir d'un index géographique mondial leurs ont permis de créer un modèle de résolution de l'expression extraite. Les métadonnées considéré peuvent également être des bases de données externes structurant les informations détaillées dans un document [105].

Les trois stratégies de résolution de toponyme, présentées ci-dessus, sont complémentaires [82] et peuvent être unifiées avec des algorithmes d'apprentissage automatique comme le montrent Santos et al. dans [106], qui trient les toponymes candidats en fonction de la probabilité d'être le bon référent, à l'aide de l'algorithme d'apprentissage d'ordonnancement (Learning to Rank) LambdaMART. Ainsi que Kamalloo et Rafiei dans [107] qui proposent une méthode non supervisée permettant d'estimer la portée géographique des documents et d'exploiter les connexions entre les noms de lieux proches comme caractéristiques afin de résoudre les toponymes. La méthode proposée explore les interactions entre plusieurs interprétations de mentions et les relations entre différents toponymes dans un document pour créer un modèle qui trouve la résolution la plus cohérente.

4. L'indexation de l'information Spatial à partir des textes

L'objectif principal de l'indexation de l'information spatiale en RI est que les documents soient indexés de manière à pouvoir être trouvés rapidement en réponse aux requêtes géographiques correspondantes. Par conséquent, faciliter le classement des documents en fonction de leur contexte géographique lors de l'étape de recherche. Les techniques d'indexation textuel des documents en fonction des mots qu'ils contiennent sont bien établies. En règle générale, dans un fichier inversé de documents chaque mot est associé à une liste des documents qui le contiennent. Cette indexation de texte peut être combinée avec un index spatial qui associe les documents relativement à des régions particulières dans l'espace.

La création d'un index spatial peut être effectuée si chaque document possède une ou plusieurs portées géographiques qui représentent les régions de l'espace géographique auxquelles le

document fait référence. Chaque portée de document peut correspondre à l'étendue spatiale d'une référence géographique qui apparaît dans ce document. Sur l'identification des portées géographiques des pages Web, Silva et al. [108] définissent la portée géographique d'une page comme la région, si elle existe, où plus de personnes que la moyenne trouverait cette page pertinente. Une fois les portées attribuées, elles sont intégrées sur la recherche par le biais d'informations contextuelles. S'il existe de nombreuses références de ce type, il faudra établir une liste des principaux foyers géographiques du document, représentés par un plus petit nombre de portées ou ce qui est également appelé empreintes (*footprints*). Ces empreintes peuvent ensuite être indexées de la même manière que tout autre élément de géométrie serait indexé dans un SIG conventionnel.

Les requêtes géographiques peuvent être caractérisées comme un triplet de <thème> <relation spatiale> <emplacement> composé d'un sujet d'intérêt en combinaison avec un nom de lieu qualifié par une préposition spatiale telle que *près de*, *dans* ou *au nord de*. Sachant que la récupération des documents pertinents nécessite de faire correspondre la spécification de la requête aux caractéristiques des documents indexés. Dans les moteurs de recherche conventionnels, cela commence par rechercher les documents contenant les termes de la requête, avant de classer les documents résultants. En ce qui concerne la recherche géographique, il est également nécessaire de faire correspondre la composante géographique de la requête avec le contexte géographique des documents tel que représenté par les empreintes de document. La combinaison du nom de lieu (après désambiguïsation) et de la préposition spatiale peut être utilisée pour générer une empreinte de requête représentant, par exemple, l'interprétation d'une expression telle que « près de Bristol ». Cette empreinte de requête peut ensuite être utilisée pour accéder à la partie appropriée de l'index spatial et donc trouver des empreintes de document qui croisent l'empreinte de la requête soumisse. Les documents récupérés seront alors les membres de ce dernier ensemble de documents géographiquement pertinents qui contiennent également les termes de la composante thématique de la requête [109].

Le défi reste cependant de trouver des moyens efficaces de combiner les indexes textuels et les index spatiaux. Sans ignorer le fait qu'un index spatial doit prendre en compte la nature hiérarchique de l'espace géographique et les relations topologiques entre les objets géographiques, afin de bien définir les relations entre les documents et d'améliorer les retours du système face à des requêtes contenant une composante spatiale et une autre thématique. Dans ce sens, l'approche de Luaces et al. [110] présente une structure d'index tenant compte de tous ces aspects en combinant un index textuel inversé, un index spatial basée sur ontologie et une table de hachage des toponymes. Etant donnée qu'une ontologie est une ressource de connaissances largement utilisée pour améliorer la recherche d'information par l'intégration de l'aspect sémantique. Les trois composants principaux de la structure d'index proposée, peuvent être détaillé comme suit :

- *Index textuel :* un index inversé construit à l'aide de ***Lucene*** [113].
- *Index Spatial :* un index basé sur une structure d'ontologie et construit en utilisant le répertoire géographique « GeoNames ». Puisque GeoNames ne fournit pas la géométrie des noms de lieux (par exemple la frontière des pays), mais seulement un seul point

géographique représentatif. Cette ontologie est complétée par un service fournisseur de la géométrie, basé sur la cartographie fournit par The Vector Map (VMAP), et est décrite en utilisant le langage standard OWL-DL du W3C [111].

- *Table de hachage des noms de lieux :* stocke dans l'index spatial, pour chaque nom de lieu, sa position correspondante. Ce qui permet un accès direct à un seul nœud à l'aide d'un mot clés retourné par le service géographique d'ontologie spatial, si le terme à traiter est un nom de lieu.

Cette approche d'indexation comprend trois étapes : une étape extraction des noms de lieux à partir du texte, après une analyse grammaticale des documents par l'outil LingPipe [112], dont l'apprentissage est fait par le corpus MUC6 marqué avec les lieux, les personnes et les organisations. A la fin de cette étape, un filtre est appliqué sur les entités nommées résultantes afin de sélectionner que les lieux. Une seconde étape de traitement des noms de lieux trouvés afin de déterminer leur nature : s'ils sont des noms de lieux réels, dans ce cas un calcul de leurs emplacements géographiques est fait. La réalisation de cette étape a nécessité le développement d'un module appelé « Service géographique d'ontologie spatiale », qui renvoie pour un nom de lieu candidat un graphe d'ontologie contenant l'instance de classe qui représente ce nom de lieu, et toutes les objets liés à cette instance par la relation « Spatialement contenu par » et au cas où l'ontologie n'a pas d'instance pour le nom de lieu candidat, il est ignoré, sinon il est considéré comme nom de lieu réel. Enfin, la troisième étape permet la construction de l'index spatial : à l'aide des graphes d'ontologie des lieux géo-référencés calculés dans la deuxième étape, et en ajoutant des références aux documents qui les contiennent. L'index spatial est alors sous forme d'un arbre des noms de lieux, connectés par le biais de relations d'inclusion. Dans chaque nœud est stocké : [le mot clés : nom de lieu, sa super Classe, Tableau des documents incluant des références géographiques correspondant à ce nom de lieu, et une liste des sous-classes].

Comprendre les textes sur le Web, qui est un domaine ouvert, est très difficile. Leur indexation aussi, ce qui a donné lieu à d'autre formes d'indexation. La diversité et la complexité du langage humain nécessitent des structures comme les taxonomies pour capturer des concepts avec diverses granularités dans chaque domaine [114]. Par exemple, une taxonomie spatiale capture les concepts géographiques et les organise en une hiérarchie d'objets connectés par des relations topologiques telles que la contiguïté.

Les taxonomies jouent également un rôle important dans de nombreuses applications de recherche d'informations telle que la recherche sur le Web en organisant les requêtes spécifiques au domaine dans une hiérarchie qui peut aider à mieux comprendre les requêtes et à améliorer les résultats de recherche [115] ou à reformuler les requêtes [116]. Dans la publicité en ligne, les taxonomies de domaine spécifiques (telle que l'assurance) sont utilisés pour décider de la connexité entre une requête donnée et un mot-clé propre à une offre publicitaire.

Plus récemment, une autre façon d'indexation et de stockage des données dans des documents semi-structurés a apparu en se basant sur le langage de balisage géographique GML (Geography Markup Language). GML est une spécification de codage d'information géographique qui a été

implémentée sur la base du langage de balisage extensible (XML : Extensible Markup Language) et normalisée (ISO 19136-2007). Les données GML sont utilisées dans le modèle de RIG proposé par [117] comme une ressource permettant de simplifier le processus d'acquisition d'information d'utilisateur par l'analyse et l'extraction des caractéristiques d'attributs, des caractéristiques spatiales et des caractéristiques de structure des données. Cependant, selon une étude de synthèse réalisée en 2020, la structure hiérarchique des taxonomies géographiques et des ontologies a fait sa force notamment lorsque l'on considère le lien d'adjacence entre les lieux [118].

5. Recherche l'information géographique : Geo-querying

Parmi les principales approches de recherche d'information parues qui traite la donnée géographique, l'approche proposée par [89] qui se base sur l'intersection des Entités spatiales, la technique de classement par appariement des empreintes géographiques des documents et de la requête proposée par [109] et la plateforme de recherche d'information Terrier [91] sur laquelle se base le système de recherche d'information géographique SINAI-GIR exploité et présenté par [90].

L'approche de RI géographique proposé par [89] est basée sur l'intersection des entités spatiales contenus dans la requête à traiter et dans un index construit précédemment après une application des étapes d'annotation sur un corpus de textes. La recherche d'information pour répondre à une nouvelle requête soumise commence alors par une reconnaissance et extraction des informations géographiques contenus dans la requête, avec un traitement similaire à la détection et l'extraction d'entité spatiale (ES) fait par l'approche sémantique spatiale qui a permis de construire l'index géographique utilisé (expliqué sur la sous-section 3.1 de ce chapitre). Suivie d'un appariement basé sur le calcul d'intersections entre les zones géoréférencées des entités spatiales de la requête et de celles contenues dans l'index.

Pour chaque requête, la pertinence d'un document de la collection est calculée en fonction de la surface d'intersection et suivant les trois paramètres illustrés sur la figure 2.3 :

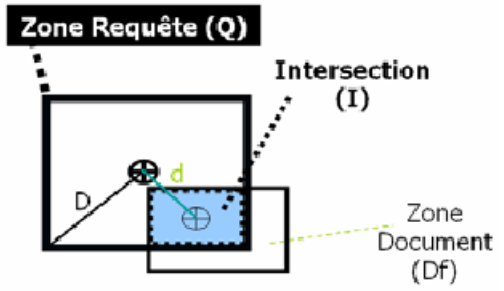

Fig. 2.3. Les zones de la requête et du document et leur intersection.

La pertinence d'un document par rapport à une requête est calculée alors par la formule :

$$Score\ Df = \frac{Précision\ de\ Df + Importance\ de\ Df}{2 + Distance\ de\ Df} \qquad \text{(Équation 2.1)}$$

Tel que :

$$Précision\ de\ Df = \frac{surface\ I}{surface\ Df} \quad \text{(Équation 2.2)}$$

$$Importance\ de\ Df = \frac{surface\ I}{surface\ Q} \quad \text{(Équation 2.3)}$$

$$Distance\ de\ Df = \frac{d}{D} \quad \text{(Équation 2.4)}$$

Par conséquent, plus les centroïdes de I et Q sont proches, plus la pertinence du document par rapport à la requête est grande.

Cette approche a été validée par l'implémentation d'un prototype de RI spatiale dans le projet Pyrénées Itinéraires Virtuels, soutenu par la Communauté d'Agglomération Pau-Pyrénées (CDAPP) et la Médiathèque MIDR. Son évaluation a permis de mettre au clair : sa supériorité par rapport à l'approche classique dans le cas des requêtes purement spatiales, ainsi que la nécessité de combiner l'approche avec l'approche statistique classique dans le cas de requêtes générales, traitant aussi bien de l'espace géographique que d'autres thèmes.

Frontiera et al. [109] ont également contribué par une approche de classement des documents par pertinence géographique. Leur travail se concentre spécifiquement sur les aspects spatiaux du classement de pertinence en examinant comment les documents peuvent être classés en fonction du degré de similitude spatiale entre l'empreinte du document et l'empreinte de la requête. Sachant que les méthodes existantes de classement par pertinence spatiale précédentes étaient généralement basées sur une seule mesure de similitude, comme, par exemple, le degré de similarité entre la requête et les empreintes de document divisé par la zone de l'empreinte de la requête. L'innovation présenté par ce travail réside dans l'introduction de mesures probabilistes de pertinence spatiale par la méthode de la régression logistique. Cela permet de prendre en compte plusieurs mesures de similitude, d'une manière qui exploite un ensemble de données d'apprentissage pour déterminer les valeurs des coefficients de régression logistique pour chacune des mesures de similarité. Les mesures de similarité utilisées sont : l'intersection entre la requête et l'empreinte du document divisé par l'empreinte de la requête pour la première mesure et divisé par l'empreinte du document pour la seconde mesure. Une troisième mesure est également introduite, celle-ci est liée aux proportions sur la terre des empreintes de la requête et du document respectivement.

La collection utilisée pour tester ce travail de recherche y-est spécifique. Elle est constituée d'empreintes de différentes régions de Californie, dont certaines correspondent à des zones administratives bien définies tandis que d'autres sont définies de manière moins précise avec uniquement des cadres de délimitation. Les résultats montrent que cette méthode a pu surpasser une variété de méthodes précédemment publiées en termes de précision et de rappel.

Une autre approche intéressante et performante de recherche d'information est celle implémentée par le moteur de recherche Terrier (Figure 2.4). Cette approche n'est pas personnalisée pour le traitement des données spatiales, sauf que son architecture diversifiée a montré ses performances également dans le domaine de recherche d'information géographique [90].

Pour une requête donnée, Terrier est capable de sélectionner automatiquement le modèle de pondération de document optimal et/ou les approches de recherche appropriées (par exemple, expansion de requête, analyse de lien ou autre), en utilisant d'autres fonctionnalités, afin de prédire les performances de requête avant l'application de l'étape de récupération des documents. Si l'expansion de requête (QE) est à appliquer, un modèle de pondération des termes approprié est sélectionné et les termes les plus informatifs des documents les mieux classés sont ajoutés à la requête. De plus, Terrier permet d'adapter facilement la sortie récupérée aux exigences de l'application (par exemple les formats TREC ou XML) et fournit des techniques d'évaluation standard.

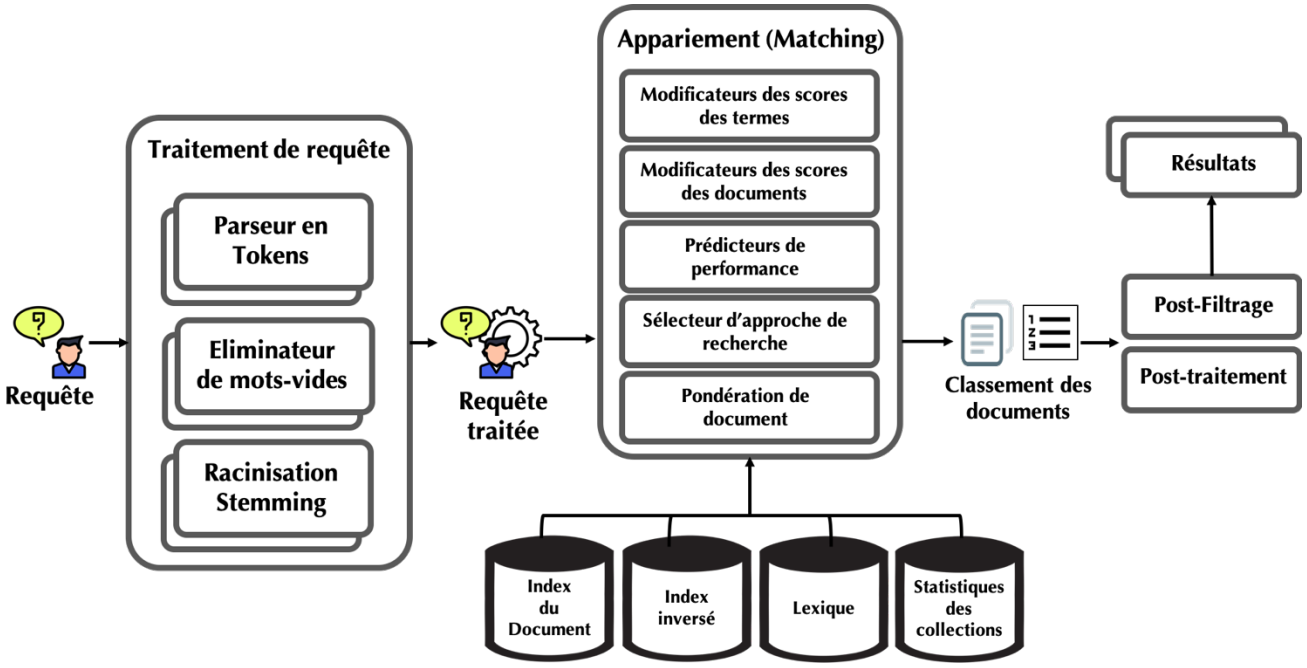

Fig. 2.4. Le processus de recherche d'information par Terrier.

Les fonctionnalités de récupération de Terrier ont été sélectionnées afin d'être utiles pour un large éventail de domaines de RI et non pas uniquement le domaine de RIG. En effet, Terrier offre une grande flexibilité dans le choix d'un modèle de pondération, ainsi que dans la modification du score des documents récupérés [92]. De plus, Terrier propose un langage de requête avancé. Une autre fonctionnalité très importante est l'expansion automatique des requêtes.

La fonctionnalité principale de mise en correspondance des documents avec les requêtes et le classement des documents a eu lieu dans le module de mise en correspondance (Matching). Ce module utilise un modèle de pondération pour attribuer un score à chacun des termes de la requête dans un document. Afin de faciliter la comparaison entre les modèles de pondération, un ensemble de modèles de pondération est fourni, y compris BM25, TF-IDF et des modèles de pondération de documents issus du cadre Divergence From Randomness (DFR) [93]. L'approche DFR fournit des modèles probabilistes sans paramètre, basée sur une idée simple : « Plus la divergence de la fréquence intra-document de terme par rapport à sa fréquence au sein de la collection est importante, plus est l'information portée par le terme t dans le document d ».

Afin de pondérer les termes et les documents, plusieurs fonctionnalités du module Matching rentre dans l'équation. Le score d'un terme individuel dans un document peut être modifié en appliquant un *TermScoreModifier*. Par exemple, un *TermInFieldModifier* peut être appliqué afin de garantir que les termes de la requête apparaissent dans un champ particulier d'un document. Si un terme de requête n'apparaît pas dans un champ particulier, TermInFieldModifier réinitialise le score du document. De plus, un FieldScoreModifier augmente le score d'un document dans lequel un terme de requête apparaît dans un champ particulier. De la même manière, la modification du score d'un document récupéré est obtenue en appliquant un DocumentScoreModifier. Le modificateur en question est PhraseScoreModifier, qui utilise les informations de position enregistrées dans l'index de Terrier, et réinitialise le score des documents récupérés dans lesquels les termes de la requête n'apparaissent pas sous forme de phrase ou dans un nombre donné de blocs. En règle générale, un DocumentScoreModifier est idéal pour appliquer des preuves indépendantes de la requête, telles que des preuves provenant de l'analyse de la structure des liens hypertexte ou de l'URL des documents. De plus, l'application sélective de différentes techniques de recherche basées sur des preuves issues de la structure d'hyperliens [94] peut être appliquée en tant que DocumentScoreModifier.

Après l'application des TermScoreModifiers ou DocumentScoreModifiers nécessaires, l'ensemble des documents récupérés peut être modifié davantage en appliquant un post-traitement ou un post-filtrage. Le post-traitement est approprié pour implémenter des fonctionnalités qui nécessitent de modifier la requête d'origine. Un exemple de post-traitement est l'extension de requête (QE pour Query Expansion). L'application de QE pourrait être activée par une requête, avant la récupération, en fonction de la sortie d'un prédicteur de performance de pré-récupération. Un autre exemple possible de post-traitement pourrait être l'application du clustering. Tandis que le post-filtrage est la dernière étape du processus de récupération de Terrier, où une série de filtres peut supprimer des documents déjà récupérés, qui ne satisfont pas à une condition donnée. Par exemple, dans le contexte d'un moteur de recherche Web, un filtre de publication pourrait être utilisé pour réduire le nombre de documents récupérés sur le même site Web, afin d'augmenter la diversité des résultats.

6. Reformulation des requêtes géographiques

Plusieurs études ont été menées par différents auteurs afin de comprendre ce que les utilisateurs recherchent lorsqu'ils soumettent une requête géographique et quelle est leur flexibilité en ce qui concerne la reformulation des requêtes [81]. L'un des principaux résultats de ces études a montré que les requêtes géographiques visent plus fréquemment à localiser des biens et des services, tandis que les requêtes non géographiques visent plus probablement le divertissement, les téléchargements ou les listes de pages contenant des informations supplémentaires [79].

Lors de l'analyse d'un journal de requêtes d'un moteur de recherche Web commercial, Jones et al. [119] ont révélé qu'environ 13% des requêtes contenaient des noms de lieux. Ils ont observé que les noms de lieux et les qualificatifs spatiaux associés apparaissent en conjonction avec les

emplacements probables des utilisateurs comme indiqué par leurs adresses IP (Internet Protocol). Ils ont examiné les distances entre la localisation des utilisateurs et les lieux pour lesquels ils recherchaient des informations et ont constaté que les gens sont, plus susceptibles de rechercher des informations locales et que les informations sur les villes sont recherchées plus fréquemment que celles sur les pays. Lors de l'examen du sujet de recherche, ils ont trouvé des corrélations entre différents types de phénomènes et les distances aux lieux d'intérêt. Les recherches de restaurants, par exemple, ont tendance à être plus proches de l'emplacement de l'utilisateur que celles des hôtels. Les auteurs se sont également penchés sur la manière dont les utilisateurs reformulent les requêtes en distinguant les changements dus aux corrections orthographiques et les changements entre l'emplacement de présence lui-même, ce qui fournit des informations importantes sur la variabilité des besoins géographiques. L'étude a révélé des différences dans le degré de changement selon l'emplacement de l'utilisateur. Les utilisateurs de Californie étaient plus susceptibles de limiter les changements dans leur recherche à des distances plus courtes de leur emplacement que ceux de Vermont, ce qui reflétait vraisemblablement la différence de densité et de disposition spatiale entre ces deux États.

Les méthodes de reformulation de requêtes géographiques quant à elles, ont été étudiées en 2003 par Kohler [120], qui a conclu que la forme la plus courante de réécriture est l'expansion en utilisant des entités géographiques par l'ajout de plus de termes géographiques afin de différentier entre les endroits partageant le même nom. Toutefois, en 2005, Fu et al. [121] ont proposé une méthode d'expansion de requête spatiale basée sur une ontologie supportant la recherche des documents considérés spatialement pertinents. Et cela a dû améliorer les résultats de recherche en réponse aux requêtes contenant une connotation ou une relation spatiale. Jusqu'en 2007, Cardoso et Silva [122] ont expérimenté l'expansion des requêtes en utilisant des caractéristiques géographiques uniquement. L'approche proposée lie les caractéristiques géographiques et les types de caractéristiques aux concepts d'une ontologie géographique. Ce qui permet une expansion plus polyvalente et ciblée vers le besoin d'information géographique de l'utilisateur. L'étude de Leveling [123] suggère également que l'expansion des requêtes en utilisant uniquement des termes géographiques fonctionne mieux.

En 2008, Stokes et al. [82] décrivent l'expansion géographique par le fait qu'une requête est complétée par des toponymes supplémentaires qui peuvent être considérés comme équivalents en un certain sens à ceux de la requête. Ils présentent et démontrent les avantages d'une méthode de normalisation de similarité qui surmonte les problèmes pouvant survenir lorsque l'expansion entraîne un très grand nombre de termes de requête supplémentaires. Ils discutent des effets de l'expansion des requêtes et mettent en évidence le potentiel d'amélioration des performances GIR en utilisant des approches plus sophistiquées. Selon leur analyse, ils ont conclu que pour avoir des gains significatifs il faut prendre en compte l'enrichissement à la fois des concepts géographiques et non géographiques.

Tenant compte les particularités de l'espace géographique, des concepts tels que sa nature hiérarchique et les relations topologiques entre les objets géographiques sont à ne pas ignorer. Dans ce sens, plusieurs structures ont été développées afin d'améliorer la performance de la

recherche d'information géographique en général et la reformulation de requêtes spatiales en particulier. A cet effet, les ontologies qui sont des ressources de connaissances largement utilisées pour améliorer les SRIs, jouent un rôle clé dans la reformulation géographique. Leur utilisation se distinguent des méthodes conventionnelles par le fait qu'ils prennent en compte des facteurs tels que les types de termes spatiaux et non spatiaux tels que codés dans les ontologies, la sémantique des relations spatiales, leur contexte d'utilisation, et la satisfiabilité des résultats de la recherche par la requête initiale (avant reformulation). Un exemple de ce genre de structure est GeoWordNet, une ressource sémantique qui a été créée à partir de l'intégration complète de GeoNames, d'autres ressources de haute qualité et de WordNet [124]. Cette ontologie a été testé et recommandé par Buscaldi et Rosso [75].

Une dernière approche qui peut être cité dans ce contexte, est celle proposé par Perea-Ortega et al. [6] en 2012. Cette technique reformule et/ou étant à la fois les parties thématiques et géospatiales détectées sur la requête géographique originale. La reformulation est réalisée en utilisant respectivement les synonymes extrait de Wordnet et les concepts spatiaux liées aux mots-clés thématiques et aux concepts spatiaux trouvées dans la requête. Comme base de connaissances géographiques ils ont utilisé GeoNames. Les résultats obtenus par ce travail montrent que toutes les reformulations de requêtes proposées récupèrent des documents pertinents qui n'ont pas été récupérés à l'aide de la requête originale.

7. Conclusion

Ce chapitre représente un état de l'art du domaine de Recherche d'Information Géographique, portant essentiellement sur les principaux concepts du domaine. A cet effet, nous avons décrit l'architecture typique d'un SRI géographique, dans un premier temps. Ensuite, nous avons détaillé les composantes de cette architecture, qui représentent les étapes de RIG : la reconnaissance des toponymes, leur résolution, l'indexation des documents et requêtes d'un point de vue géographique et l'indexation de l'information spatiale en générale sous forme de structures exploitable par les étapes suivantes. Les approches de recherche les plus significatives ont été par la suite analysées, enfin une étude chronologique des méthodes de reformulation des requêtes géographique a été présentée.

L'analyse effectuée nous a permis de conclure qu'il y a un manque en termes de méthodes spécifiques de reformulation de requêtes contenant une composante géographique. Par conséquent, afin de tenir compte de la nature géographique et les relations topologiques entre les objets géographiques, nous avons proposé deux approches de construction de structure taxonomique permettant une meilleure reformulation des requêtes géographiques. La partie suivante de ce mémoire décrit en détails nos deux contributions dans ce domaine.

Chapitre 3 *:*

Reformulation de requêtes en construisant une taxonomie géographique par analyse sémantique latente (QRGTW)

---

1. Introduction
2. GATB : Constructeur de taxonomie géographique d'adjacence
3. Extraction de l'information géographique
4. Méthode d'exploitation de cette taxonomie pour la reformulation de requêtes géographique
5. Résultats d'expérimentations :
   5.1. La 1ère série de tests : Taxonomie du Maroc
   5.2. La 2ème série de tests : Reformulation de requêtes contenant des villes anglaises
   5.3. La 3ème série de tests : Cas d'utilisation des requêtes géographiques des journaux log de AOL
6. Conclusion et limitations

---

1. Introduction

Ce chapitre décrit la première contribution qui consiste en une approche de construction d'une taxonomie géographique d'adjacence pour un pays afin de l'utiliser dans la reformulation des requêtes spatiales. L'approche proposée utilise les documents les mieux classés récupérées par le moteur de recherche lors de la soumission des entités spatiales composées de relation spatiale et du nom d'un emplacement géographique A (ville ou village). Puis, la méthode d'indexation sémantique latente (Latent Semantic Indexing LSI) est appliquée sur les documents extraits, afin de trouver les villes B_i (ou villages) les plus proches de A, et de procéder à une étape de validation de chaque lien en vérifiant si A se retrouve également dans les résultats des emplacements B_i.

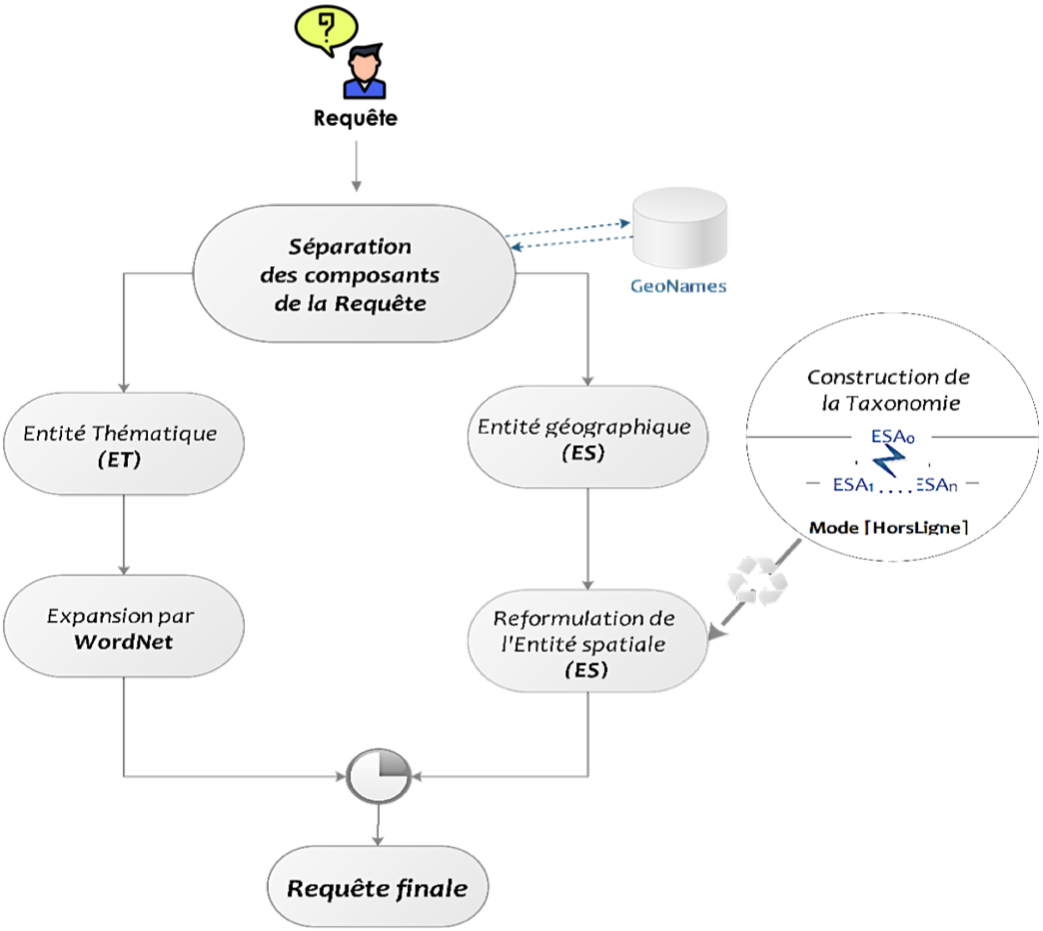

Fig. 3.1 La reformulation de requête en utilisant une taxonomy géographique et le sens des mots (QRGTW).

Comme l'indique la figure 3.1, l'approche proposée est organisée sous forme de deux phases :

- Phase 1 : Construction d'une taxonomie géographique d'adjacence (En mode hors ligne).

- Phase 2 : Détection et séparation des différentes composantes de la requête, définit par une entité thématique et une autre spatiale, en utilisant la base de données **GeoNames**[4].
- Phase 3 : Reformulation des deux entités de la requête. L'entité thématique est étendue en utilisant un thesaurus ou ontologie sémantique, tandis que l'entité spatiale est reformulée en se basant sur la taxonomie construite pendant la première phase.

Dans nos expérimentations, on a commencé par construire une taxonomie géographique d'adjacence pour le Maroc. On a fait varier la relation spatiale utilisée dans l'étape de récupération de documents pour comparer les résultats des différentes relations spatiales, et on a utilisé les services Web Google comme moteur de recherche pour comparer les résultats retournés dans tous les cas. Puis on a utilisé la taxonomie construite résultante pour la reformulation de requêtes géographiques. On a également construit des parties de la taxonomie de l'Angleterre, selon le besoin, afin de tester notre approche. Lors de ces deux premières séries d'expérimentations, nous avons reformuler des requêtes rédigées dans un contexte de recherches simulé avec l'aide de 10 utilisateurs volontaires. Enfin, nous nous sommes servis des journaux archivés du moteur de recherche AOL, afin de valider notre approche de reformulation dans un contexte réel de recherche.
On a utilisé la mésure MAP (Mean Average Precision) afin de comparer les documents retournés avant et après reformulation. D'après nos résultats, nous notons que la reformulation de requêtes géographiques basées sur la taxonomie construite améliore largement la précision des requêtes.

2. GATB : Constructeur de taxonomie géographique d'adjacence :

Notre approche de construction de taxonomie géographique, nommée Geographical Adjacency Taxonomy Builder (GATB), se base sur la méthode d'indexation Sémantique latente (LSI), appelée aussi analyse sémantique latente (LSA) qui est une technique qui projette les éléments donnés en entrée (les documents dans notre cas) dans un espace avec des dimensions sémantiques « latentes » [36]. Dans cet espace, les noms de deux emplacements géographiques peuvent avoir une grande valeur de similitude, même s'ils ne partagent pas les mêmes documents, et partagent les mêmes termes (emplacements) co-occurrents avec eux, on parle dans ce cas d'une similarité sémantique entre les termes.
LSI assume qu'il y a une certaine structure latente (sous-jacente) dans l'utilisation du mot qui est partiellement masquée par la variabilité dans le choix des mots. Une technique mathématique, appelée décomposition en valeurs singulières (Singular Value Decomposition SVD) est appliquée à une matrice terme-document pour estimer cette structure de mots à travers les documents.
Sur la suite du chapitre nous parlerons d'Entité Spatial Absolu ou relative (ESA ou ESR). Vue qu'on utilise le modèle de requête spatiale (expliqué précédemment sur l'état de l'art). C'est-à-dire que les entités nommées spatiales tel que la ville de "Paris" sont des entités bien connus nommées et sont définis comme une ESA. Tandis que les entités spatiales complexes comme "à Côte de Paris" sont définies comme une ESR.

[4]GeoNames est une base de données géographique disponible en téléchargement gratuit sous licence de type « creative commons attribution ». Elle contient plus de 10 millions de noms géographiques et comprend plus de 9 millions de caractéristiques uniques avec 2,8 millions de lieux habités et 5,5 millions de noms alternatifs. [http://www.geonames.org/about.html]

Afin de construire une taxonomie géographique d'adjacence, nous indexons les documents en construisant une matrice terme-document à partir des documents récupérés lors de la soumission d'une ESR, dont l'ESA est une ville ou un village du pays pour lequel nous construisons la taxonomie. Ensuite, la projection SVD est calculée par la décomposition de la matrice terme-document $A_{t\times d}$ en produit de trois matrices $T_{t\times n}$, $S_{n\times n}$ et $D_{d\times n}$.

$$A_{t\times d} = T_{t\times n}\, S_{n\times n}\, (D_{d\times n})^T = T_{t\times n}\, S_{n\times n}\, D'_{n\times d} \qquad (3.1)$$

Où t est le nombre de termes, d le nombre de documents, n=min(t,d) est le nombre de dimensions pour la matrice A, appelé aussi le rang de A et D' est la transposé de D.

Les matrices T et D représentes les termes et les documents dans le nouvel espace. La matrice S est une matrice diagonale et elle contient les valeurs singulières de A dans un ordre décroissant. La $i^{ème}$ valeur singulière indique la quantité de variation tout au long du $i^{ème}$ axe.

Dans l'étape suivante, on tronque la SVD en réduisant le rang n de la matrice A. L'objectif de la méthode est de trouver le rang k < n qui donne une nouvelle matrice A' qui est la meilleure approximation de A.

A' est construite en restreignant les matrices T, S et D à leurs premières k rangées et en les multipliant comme suite :

$$A'_{t\times q} = T_{t\times k}\, S_{k\times k}\, (D_{q\times k})^T \qquad (3.2)$$

Le choix du nombre de dimensions k pour A' est un problème intéressant pour la méthode LSI. Alors que la réduction de n peut éliminer une grande partie du bruit, en gardant trop peu de dimensions des informations importantes peuvent être perdues. Cependant, on observe que LSI fonctionne bien avec un nombre relativement restreint de dimensions k. Cette observation montre que ces dimensions capturent une partie majeure de la structure significative.

En se basant sur la matrice A' résultante de l'application de la méthode, nous restreignons A' aux termes qui représentent des ESAs du pays pour lequel on construit la taxonomie uniquement. nous considérons que la matrice résultante de cette restriction est nommée A''.

Enfin, nous calculons les similarités entre l'ESA soumise au moteur de recherche (pour extraire les documents) et les autres ESAs présentes sur la matrice A'', en utilisant la mesure de similarité Cosinus [35], dont l'expression est la suivante :

$$Simc(ESA_i, ESR_j) = |\cos(\overrightarrow{ESA_i}, \overrightarrow{ESR_j})| = \left|\frac{\overrightarrow{ESA_i}\times\overrightarrow{ESR_j}}{||\overrightarrow{ESA_i}||\times||\overrightarrow{ESA_i}||}\right| \qquad (3)$$

Un seuil minimisant l'erreur est défini lors des expérimentations. Dans le but de construire la taxonomie en utilisant les ESAs les mieux classées, dont la valeur de similarité avec l'ESA initial dépasse ce seuil.

Le processus de construction d'une partie d'une taxonomie d'adjacence en utilisant l'approche GATB est présenté étape par étape sur la figure 3.2.

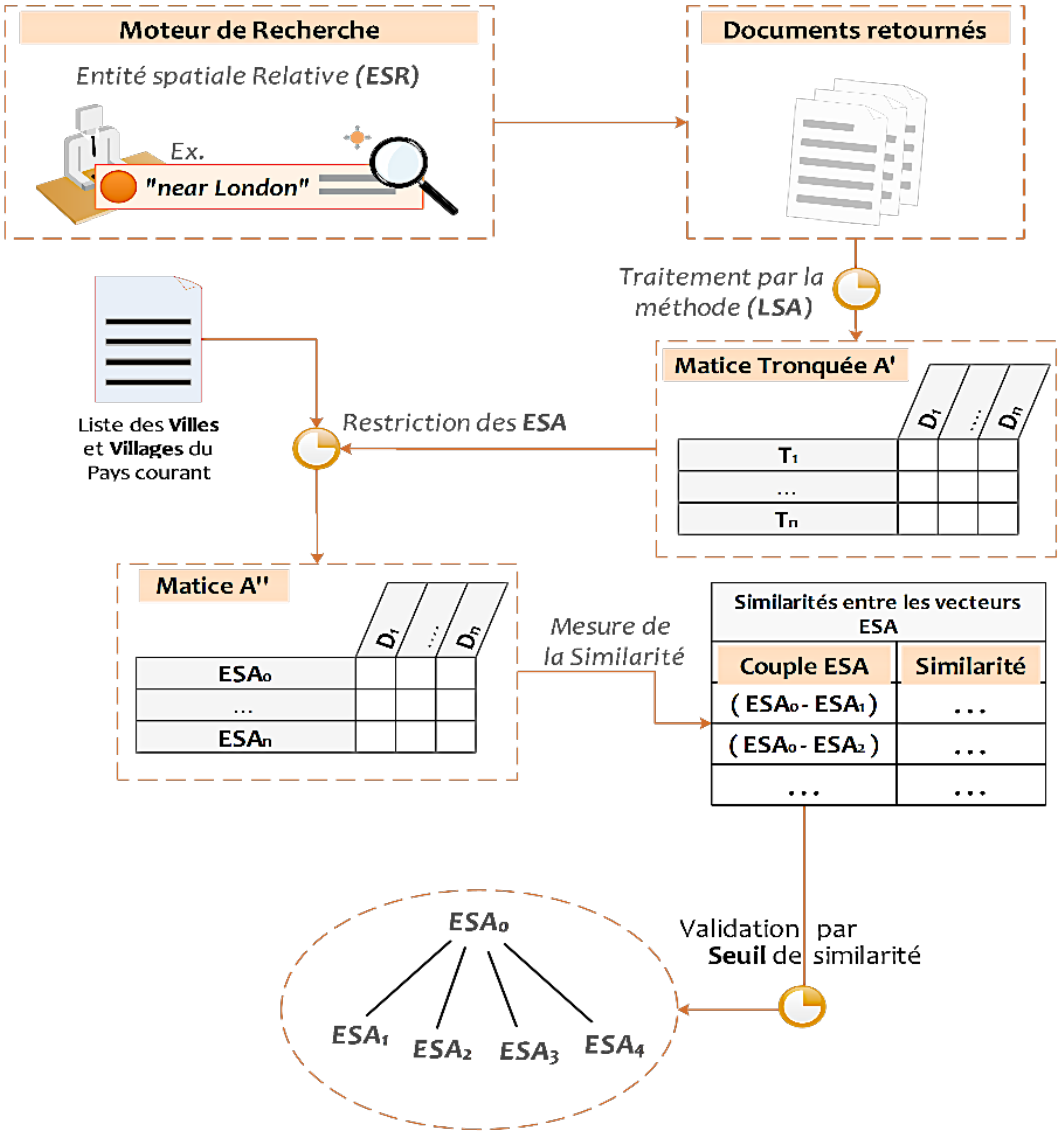

Fig. 3.2 Construction d'une partie de taxonomie d'adjacence en utilisant l'approche GATB (exemple avec ESA_0=London).

La sortie de cette étape est une taxonomie de niveau 1 comme le montre la figure 3.2. Pour laquelle nous devons appliquer une étape de validation.

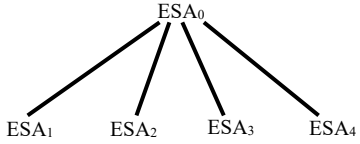

Fig. 3.3 Taxonomie de niveau 1 pour la ESA_0.

Pour valider chaque arc de cette taxonomie nous soumettons l'ESA au bout de cet arc avec la même relation spatiale utilisé précédemment. Ensuite, nous appliquons la méthode LSI de la même façon pour construire une taxonomie de niveau 1 pour l'ESA à valider. Par exemple, pour ESA_1 si parmi les résultats nous trouvons l'ESA_0 alors nous gardons cet arc et il est dit « lien validé ». De la sorte,

la taxonomie évolue vers le niveau 2 comme sur la figure 3.4. Dans le cas contraire, si l'ESA_3 n'est pas validé l'arc en question est supprimé de la taxonomie de l'ESA_0.

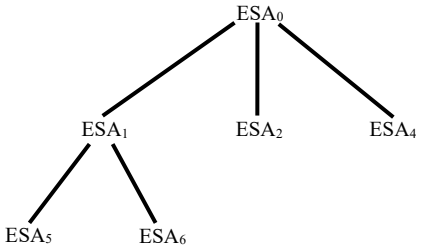

Fig. 3.4 Taxonomie de niveau 2 pour la ESA_0.

Le même processus est appliqué à chaque nouvelle ESA qui apparait sur la taxonomie, afin de construire la taxonomie géographique d'adjacence pour un pays.

3. Extraction de l'information géographique :

Dans le but de reformuler une nouvelle requête géographique, nous procédons premièrement par la séparation des différentes composantes de la requête en se basant sur l'approche d'extraction d'informations géographiques (GIE). Cette approche utilise une méthodologie d'annotation sémantique pour l'extraction de marqueurs spatiaux : d'abord, l'entité spatiale absolue (noms de la région, des villes ou villages) est détectée et marquée. Puis l'entité spatiale ES est construite à partir de cette ESA basée sur un lexique de relations spatiales. Les termes restant de la requête sont considérés comme une entité thématique (ET).

Notre contribution sur cette étape se base sur une hypothèse, qui nous a permis de faire quelques modifications sur la méthode GIE cité ci-dessus.

Hypothèse. Si la relation spatiale n'existe pas dans la requête, la présence d'un ESA ne veut pas dire que c'est une requête géographique. Par exemple, une requête contenant "George Washington". Nous pouvons aussi considérer l'exemple de la requête de recherche pour "Hôtel de Paris". Dans ce contexte, "Paris" est le nom d'un hôtel dont l'emplacement est à Monaco, Monte Carlo ou Tanger. Donc malgré la nature géographique de l'expression, elle ne référence pas l'emplacement géographique "Paris".

Notre méthodologie consiste en la détection de la relation spatiale (RS) en utilisant un lexique de RSs que nous avons préparé d'abord, puis la vérification du terme suivant, si c'est une ESA valide. Dans le cas de validité, nous marquons la relation spatiale et l'ESA en ESR valide. Les autres termes sont considérés comme faisant partie de l'entité thématique (TE). Tel que la validité de l'ESA dépend de sa présence dans la base de données GeoNames.

Puisque GeoNames contient des noms alternatifs pour décrire le même terme en différentes formes et langues, notre procédure de correspondance prend en charge les correspondances avec des noms de lieux similaires et pas seulement des correspondances exactes.

Cette méthode peut être perçu comme une technique de séparation des composantes de requête géographique et au même temps une sorte de processus de désambiguïsation Géo/non-Géo. Tel que l'ambiguïté peut être comprise comme un mot ou une expression qui a plus d'un sens [2] et

l'ambiguïté Géo/non-Géo est considéré quand un mot ou une expression peut avoir un sens non-géographique comme dans le cas du terme "Turkey" (Table 3.1).

Tableau 3.1. Example de séparation des composants d'une requête géographique

Requête	*Turkey cooking recipe in Turkey*
Etape 1	La détection de la relation spatiale : "*in*"
Etape 2	Vérification du terme suivant : "*Turkey*" est un nom d'emplacement, stocké sur la base de données GeoNames en tant que Pays.
Etape 3	Validation de l'entité spatiale relative (ESR) : "*In Turkey*"
Etape 4	Considération des termes comme composants de l'entité thématique : "*Turkey cooking recipe*"

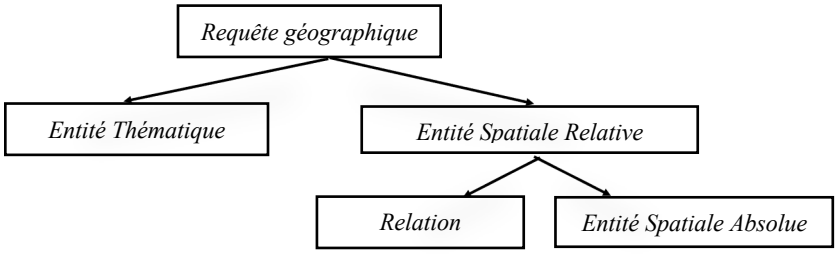

Fig. 3.2 Les composantes d'une requête géographique.

À la fin de cette phase (voir la **figure 3.5**), la requête est divisée en deux parties : l'ET et l'ESR. L'entité spatiale relative contient également deux parties : la relation spatiale RS et l'entité spatiale absolue l'ESA.

4. Méthode d'exploitation de cette taxonomie pour la reformulation de requêtes géographique

Dans cette approche, nous avons focalisé notre attention sur l'amélioration de requête contenant une Entité Spatial Relative en la reformulant à travers l'interprétation de la relation spatiale qu'elle englobe, ainsi qu'en étendant l'entité thématique.

Nous avons proposé de reformuler l'entité spatiale en utilisant la taxonomie construite à l'aide du GATB décrit dans la section 2. Notre taxonomie est créée hors-ligne et sera utilisé dans la reformulation de chaque entité spatiale détectée lors de l'étape d'extraction de l'information géographique. Tandis que l'entité thématique est étendue à travers l'utilisation d'ontologie sémantique. Lors de nos expérimentations nous avons exploité l'ontologie WordNet pour chercher les synonymes des termes composants l'ET puis les intégrer dans la nouvelle requête afin d'enrichir l'ET de manière purement sémantique.

Une requête contenant une relation d'adjacence dans son ES, signifie que l'utilisateur cherche le sujet indiqué dans la partie thématique dans les lieux qui entourent l'ESA de cette requête. Ainsi, nous proposons que l'entité spatiale doit être éliminée, et remplacée par les noms de lieux qui représentent les nœuds fils de l'ESA dans la taxonomie utilisée comme suit :

- **Requête Utilisateur** = ET RS ESA
- **Requête Reformulée** = ET étendue + ("lieu adjacent 1" OR " lieu adjacent 2" OR …)

Dans la requête reformulée, les guillemets sont utilisés pour rechercher l'endroit désiré et ne pas chercher séparément les termes que le non de l'emplacement contient si l'ESA est composé de plus d'un terme (par exemple en soumettant New York sans guillemets, le moteur de recherche va considérer New et York comme deux mots clés). Tandis que, l'opérateur booléen "OR" est utilisé, pour garantir que les résultats retournés par le moteur de recherche contiennent par exemple "lieu adjacent 1" ou "lieu adjacent 2" ou les deux en même temps et ainsi de suite pour tous les nœuds fils ajoutés à la requête.

5. Résultats d'expérimentations :

5.1. La 1ère série de tests : Taxonomie du Maroc

Pour tester et vérifier la performance de l'approche proposée, en premier lieu nous avons réalisé une partie de la taxonomie du Maroc. Lors de ce test, pour pouvoir exploiter les pages web réalisés par les marocains eux même on a réalisé nos expérimentations en langue française. Nous avons considéré la capitale du Maroc, l'ESA « Rabat » comme la racine de la taxonomie et nous avons commencé le processus de construction avec.
Lors de ce test, nous avons essayé de vérifier si la relation spatiale utilisée, influence la qualité des résultats. Les relations spatiales utilisées sont présentées sur le tableau 3.2.

Tableau 3.2. Les relations spatiales utilisées dans la construction d'une taxonomie du Maroc

Annotation	Expression
RS 1	à côté de
RS 2	à la périphérie de
RS 3	à proximité de
RS 4	Aux alentours de
RS 5	aux environs de
RS 6	les environs de
RS 7	près de

Les documents auxquels nous avons appliqué l'approche GATB sur cette étape d'expérimentation contiennent le texte des 5 premières pages web retournées lors de la soumission de la relation spatial et l'ESA pour laquelle on cherche les ESAs adjacentes. Pour l'extraction de texte utile pour notre méthode, nous mettons les conditions suivantes :

- Ne pas inclure les liens hypertextes qui permettent d'aller vers une autre page car c'est la plupart du temps des publicités ou des propositions qui forme un bruit.
- Désactiver la recherche basée sur l'historique de recherche.
- Utiliser les services de recherche de Google sans utilisation d'un profil spécifique.

Premièrement nous avons varié le seuil de similarité entre ESAs de 0,5 à 0,9 sans l'étape de validation.

Tableau 3.3. Taux d'erreur dans la taxonomie lors de la variation du seuil de similarité et de la relation spatiale utilisée

RS\Seuil	0,5	0,6	0,7	0,8	0,9
RS 1	57,14	57,14	50	60	0
RS 2	50	50	0	0	0
RS 3	80	80	100	100	100
RS 4	46,15	62,5	62,5	71,42	71,42
RS 5	41,66	41,66	50	50	66,66
RS 6	38,46	38,46	50	50	0
RS 7	11,11	0	0	0	0

Selon, les résultats illustrés par le tableau 3.3, nous remarquons que pour les seuils 0,7 à 0,9 l'utilisation de la RS 3 induit à un taux d'erreur de 100% et pour les autres RS malgré le fait que pour quelques RS nous avons un taux d'erreur 0 pour ces seuils-là, nous remarquons qu'en augmentant le seuil le bruit diminue mais on a également la perte d'un nombre important de ESAs adjacente. Donc ces seuils sont à ignorer.

En ce qui concerne la comparaison entre le seuil 0,6 et le seuil 0,5 nous remarquons que pour le cas de la RS 4 le taux d'erreur pour le seuil 0,6 est le plus élevé, tandis que pour la RS 7 c'est le contraire, en plus du fait que le seuil 0,6 nous donne un taux d'erreur égale à 0 en perdant une seul ESA adjacente. Tandis que pour les autres RS le taux d'erreur est le même. Cela permet de dire que le seuil 0,6 nous donne un résultat optimal.

En comparant le résultat des 7 RS nous remarquons que la 7ème RS « près de » donne le taux d'erreur le moins élevé.

Nous continuons donc en considérant le seuil de similarité 0,6 et nous évaluons l'importance d'ajout de l'étape de validation pour les 7 relations spatiales en analysant le tableau 3.4.

Tableau 3.4. Influence de l'étape de validation sur le taux d'erreur dans la taxonomie

N° relation spatiale	RS 1	RS 2	RS 3	RS 4	RS 5	RS 6	RS 7
Sans validation	57,14	50	80	62,5	41,66	38,46	0
Avec Validation	50	33,33	50	60	33,33	30	0

Nous remarquons sur le tableau 3.4 que pour les différentes relations spatiales le taux d'erreur diminue en appliquant l'étape de validation. Et que la RS 7 donne un taux d'erreur nul contrairement à toutes les autres relations spatiales qui ont un taux d'erreur d'au moins 30% même après validation. Sur la base de ce résultat, nous décidons de continuer la construction de notre taxonomie du Maroc en utilisant la relation spatiale « près de ».

Après plusieurs itérations récursives de calcul, une partie de la taxonomie résultante est comme présenté sur la figure 3.6.

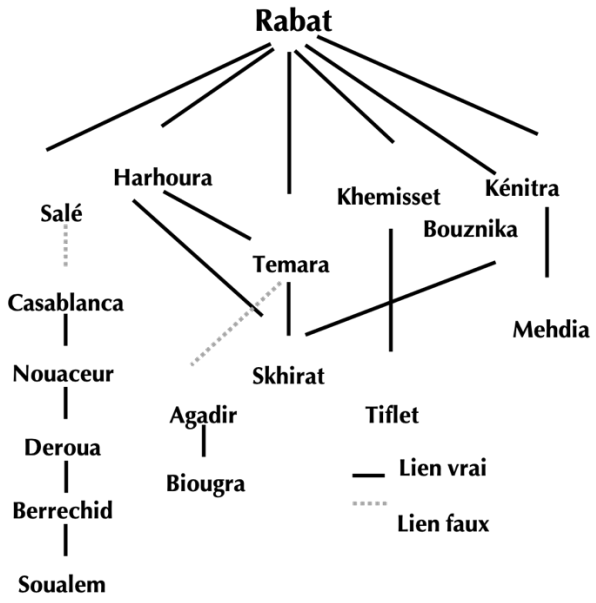

Fig. 3.6 Une partie de la taxonomie d'adjacence marocaine.

Nous remarquons sur la figure 3.6, que parmi 19 liens retrouvées, 2 sont faux. Donc, le taux d'erreur sur cette partie de la taxonomie est de 10,53%.

Afin d'évaluer plus précisément les performances de notre approche, nous avons réalisé notre premier test complet (incluant la reformulation de requête) en utilisant une collection de 50 requêtes géographiques d'adjacence (dont quelques exemples sont présentés par le tableau 3.5, contenant des ESAs marocaines, que nous avons créées à partir d'une simulation réalisée à l'aide de 10 utilisateurs. Pour chaque requête, nous avons comparé la valeur de la MAP (Mean Average Precision) et de la précision à 10 documents (P@10) dans les deux cas : avant et après reformulation en utilisant l'approche QRGTW. Dans cette étape d'expérimentation, toutes les requêtes proposées étaient géographiques, ce qui nous a permis de tester notre approche en utilisant la taxonomie construite avant de généraliser la simulation à une collection plus large (qui contient des requêtes géo et non-géo).

En utilisant la taxonomie du Maroc que nous avons construit pour la reformulation des 50 requêtes proposées, nous avons appliqué la technique de reformulation expliquée dans la section 4 de ce chapitre. Ensuite, nous avons comparé les 10 premiers URL retournées dans chaque cas en utilisant les mesures MAP et P@10 (Fig. 3.7 et tableau 3.6).

Tableau 3.5. Quelques exemples des requêtes contenant des ESAs marocaines utilisées

Requêtes
Appartement à vendre près de skhirat
Lieux touristiques aux alentours de Meknès
Villas à louer aux environs de Temara
Catastrophe naturelle à proximité de Rabat
Vente de terrain à la périphérie de Marrakech

Le jugement de la pertinence de chaque URL retourné est manuel, collectif et est réalisé par nous-même et des volontaires externes à l'étude, en vérifiant si l'URL résultante est une bonne réponse à la requête pour l'entité thématique (ET) et l'entité spatiale (ES) à la fois. Par exemple, si l'ET de la requête est "appartement à vendre" et l'URL propose une villa, un terrain ou un bien immobilier à louer pas à vendre, l'URL est jugée non pertinente. De la même manière, si l'ES est "à la Périphérie de Marrakech", un URL qui propose un bien immobilier au centre de Marrakech est considérée comme non pertinent.

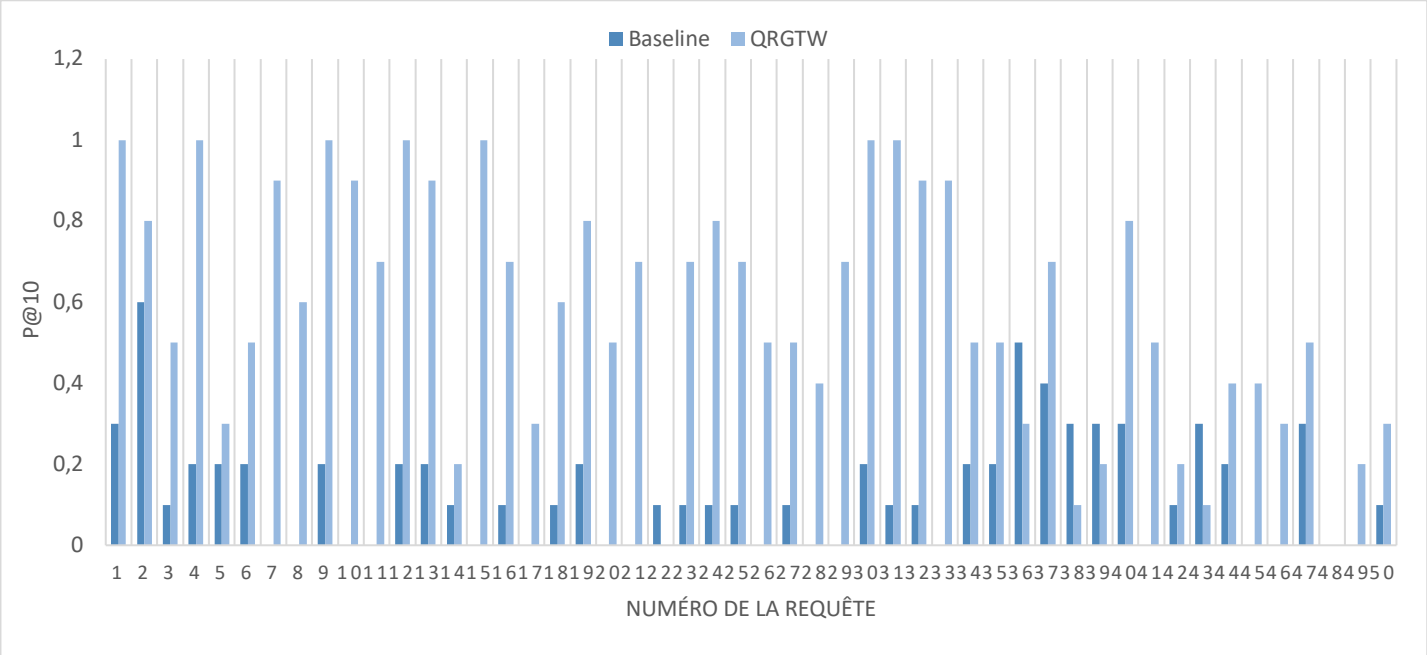

Fig. 3.7 Les valeurs de P@10 avant et après reformulation de 50 requêtes contenant des ESAs marocaines.

La figure 3.7 montre que la valeur de P@10 de la plupart de ces 50 requêtes géographiques a été amélioré de manière significative et atteint la valeur 1 dans le cas de onze requêtes. Toutefois, la valeur 1 ne signifie pas toujours que les 10 premiers documents récupérés sont pertinents pour la requête, mais que les documents pertinents figurent en haut de la liste renvoyée.

Cette figure montre également que les valeurs de P@10 de 4 requêtes uniquement ont diminué après reformulation. Après une analyse des résultats de recherche de ces requêtes, nous concluons que la dégradation de la précision est due aux URL qui ne sont pas adaptées à l'ET.

En ce qui concerne les requêtes que nous avons reformulé en utilisant des liens vrais et faux (comme le lien entre "**Témara**" et "**Agadir**" dans la Figure 3.6), nous avons remarqué que les URL non pertinentes sont des URL de vente, alors que la requête recherche des offres de location par exemple. Les liens incorrects, dans ce cas, n'ont pas influencé les résultats de la recherche vue qu'ils sont accompagnés par un pourcentage plus important de liens corrects.

Tableau 3.6. La moyenne de la P@10 et la MAP pour les requêtes contenant des ESAs marocaines

	Baseline (Originale)	QRGTW
P@10	0,136	0,580
MAP	0,229	0,744

Les résultats de la figure 3.7 sont confirmés dans le tableau 3.6 par la moyenne des deux mesures : la P@10 et la MAP. Ce tableau compare notre approche QRGTW à la précision des mêmes 50 requêtes avant reformulation (Baseline).

Le tableau 3.6, montre que l'approche proposée améliore les résultats significativement comparés à la base de référence qui est formée par les requêtes originales des utilisateurs, dans le cas de notre collection de test contenant des ESAs marocaines.

5.2. La 2ème série de tests : Reformulation de requêtes contenant des villes anglaises

Lors de la 2ème série de simulations, nous avons proposé d'appliquer l'approche QRGTW pour construire des taxonomies géographiques des villes d'Angleterre afin de définir les paramètres adéquats à utiliser et puis pouvoir reformuler 52 requêtes, contenant des ESAs anglaises. Ainsi, la langue utilisée est l'anglais, et nous avons commencé ces expérimentations par la variation des relations spatiales (tableau 3.7). Au même temps, nous avons fait une comparaison entre le cas d'utilisation ou non de l'étape de validation (Figure 3.8).

Tableau 3.7. Les relations spatiales anglais utilisées sur la 2ème série de tests

Annotation	Expression
SR 1	around
SR 2	at the periphery of
SR 3	close to
SR 4	nearby
SR 5	in the proximity of
SR 6	in the surroundings of
SR 7	in the vicinity of
SR 8	near
SR 9	next to

Les résultats de la figure 3.8 représentent la variation du taux d'erreur lors de la création de taxonomies de niveau 1 pour Londres associés à chaque expression du tableau 3.7. Les dix premiers documents sont extraits à chaque cas et nous avons fixé le seuil de similarité à 0,5.

Les liens corrects sont considérés comme les liens entre deux villes, qui sont éloignées l'une de l'autre avec un maximum de 113 kilomètres, ce qui est équivalent à une heure de voyage dans les autoroutes anglaises. Comme la limite de vitesse dans les autoroutes du **Royaume-Uni** est de 113 km/h (70 mph). Tel que le point de départ du calcul de distance est le centre de **Londres**.

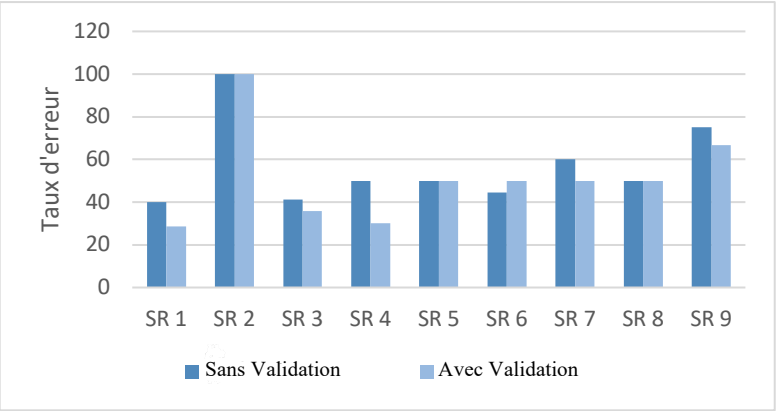

Fig. 3.8 L'impact de l'étape de validation sur la valeur du taux d'erreur pour les taxonomies de Londres.

La figure 3.8 montre que pour les différentes relations, le taux d'erreur est soit stable soit en diminution après l'application de l'étape de validation, à l'exception de la relation spatiale *"in the surrounding of"* pour laquelle le taux d'erreur augmente de 5,56%. En moyenne, la performance de l'étape de validation est de 5,52%.

En comparant les RSs, nous remarquons que les relations spatiales *"around"*, "close to" et "*nearby*" (SR 1, 3 et 4) donnent les plus faibles valeurs de taux d'erreur avec une différence minime. Ainsi, nous avons utilisé ces trois RSs, qui paraissent assez prometteuse, dans la création de taxonomies niveau 1 d'autres villes afin de découvrir laquelle parmi ces relations est la plus utilisée dans les documents web en association avec les villes d'Angleterre.

Par la suite, nous avons appliqué l'approche GATB proposée pour construire les taxonomies des cinq plus grandes villes d'Angleterre (tableau 3.8) en utilisant les trois relations spatiales citées ci-dessus. Le taux d'erreur de chaque taxonomie construite est donné dans le tableau 9.

Tableau 3.8. Les villes les plus grandes (en matière de population) de l'Angleterre – Statistiques correspondantes à l'année 2015 (Source GeoNames)

Villes	**Population**
London	8,173,941
Birmingham	1,073,045
Leeds	751,485
Sheffield	552,698
Liverpool	468,945

Tableau 3.9. Les taux d'erreurs des taxonomies des cinq plus grandes villes d'Angleterre en utilisant trois relations spatiales différentes

Taux d'error (%)	**London**	**Birmingham**	**Leeds**	**Liverpool**	**Sheffield**
around	28,57	33,33	0	12,50	0
close to	33,33	25	0	NR	0
nearby	30	0	0	0	100

Nous remarquons à partir du tableau 3.9 que "nearby" donne une taxonomie parfaite pour "*Birmingham*", "*Leeds*" et "*Liverpool*". Tandis que la taxonomie résultante pour "*Sheffield*" est

pleine de faux liens (100% d'erreur). En revanche, "close to" ne donne aucun résultat (NR) dans le cas de "Liverpool". Alors que "around" donne des résultats optimaux dans tous les cas. Pour plus de détails et afin de trancher par rapport au choix de la relation spatial à utiliser dans la suite des simulations, nous présentons dans la figure 3.9 le nombre de liens corrects dans tous les cas.

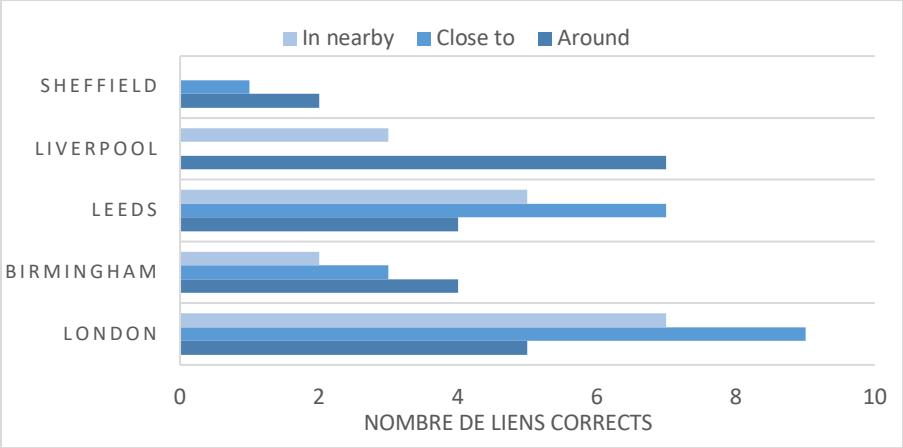

Fig. 3.9 Le nombre de liens corrects dans les taxonomies en variant la relation spatiale utilisée.

La figure 3.9 montre que la relation spatiale "around" donne un nombre optimal de liens corrects pour les cinq villes. Tandis que, les deux autres relations spatiales, retourne un nombre très élevé ou très faible de liens corrects. Ainsi, nous avons choisi de continuer nos tests en utilisant la relation spatiale "around".

La série d'expérimentations qui suit concerne la variation du nombre de documents utilisés dans la construction de taxonomies entre 2 et 20 documents.

Tableau 3.10. L'impact du nombre de documents utilisés dans la construction des taxonomies sur le taux d'erreur - Cas des cinq plus grandes villes d'Angleterre.

Nbr Docs	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**	**18**	**20**
London	100	50	50	47,37	28,57	27,27	22,22	28,57	30	62,50
Birmingham	NR	NR	33,33	50	33,33	20	33,33	37,50	66,67	77,78
Leeds	NR	NR	0	0	0	20	28,57	36,36	30	33,33
Sheffield	NR	NR	NR	0	0	0	0	0	33,33	50
Liverpool	50	40	16,67	14,29	12,50	0	0	0	0	0

En examinons le tableau 3.10, Nous remarquons qu'en utilisant 2, 4 ou 6 documents, nous avons des cas "NR" (Aucune ESA trouvé), donc ces valeurs doivent être ignorées. Une analyse détaillée du tableau, nous permet de conclure que les valeurs les plus faibles du taux d'erreur sont regroupées dans l'intervalle d'utilisation de [10-14] documents.

De plus, Nous remarquons que lors de la construction de la taxonomie d'une plus grande ville (tableau 8), généralement nous avons besoin de plus de documents pour atteindre un taux d'erreur plus faible. Pour Londres, la valeur la plus basse est atteinte par 14 documents, pour Birmingham la

valeur la plus basse est donnée en utilisant 12 documents, alors que l'utilisation de 10 documents est suffisante pour trouver un bon résultat pour des villes plus petites. Ceci peut être expliqué par le fait que les villes peuplées sont sujet de plus de documents sur le net, mais la quantité qui est en augmentation dans ce cas induit à la diminution de la qualité et la véracité des informations.

Du fait que les résultats du tableau 3.10 n'était pas concluant, et afin de trancher par rapport au nombre optimal de documents à utiliser, on a testé le résultat d'exploitation des taxonomies créées dans la phase de reformulation de 10 requêtes géographiques contenant les noms des villes d'Angleterre (Tableau 3.11).

Tableau 3.11. Les 10 premières requêtes géographiques proposées pour tester l'approche QRGTW sur les villes de l'Angleterre.

Id. query	**Expression**
Q1	Apartment for sale near Oxford
Q2	Land for sale in the proximity of London
Q3	Touristic places close to Nottingham
Q4	Apartment to rent around Liverpool
Q5	Best places for picnic at the periphery of Sheffield
Q6	Rent cottage in the vicinity of Manchester
Q7	Hotels at a good price in the surroundings of Derby
Q8	Office to rent nearby Birmingham
Q9	historical monuments next to York
Q10	Buying villa in the vicinity of Leeds

On a calculé la MAP en utilisant les 10 documents récupérés les mieux classés, pour les requêtes avant et après reformulation avec des taxonomies construites avec 10 documents (*Ref 10 Docs*), 12 documents (*Ref 12 Docs*) et 14 documents (*Ref 14 Docs*). Les résultats sont représentés sur la figure 4.10.

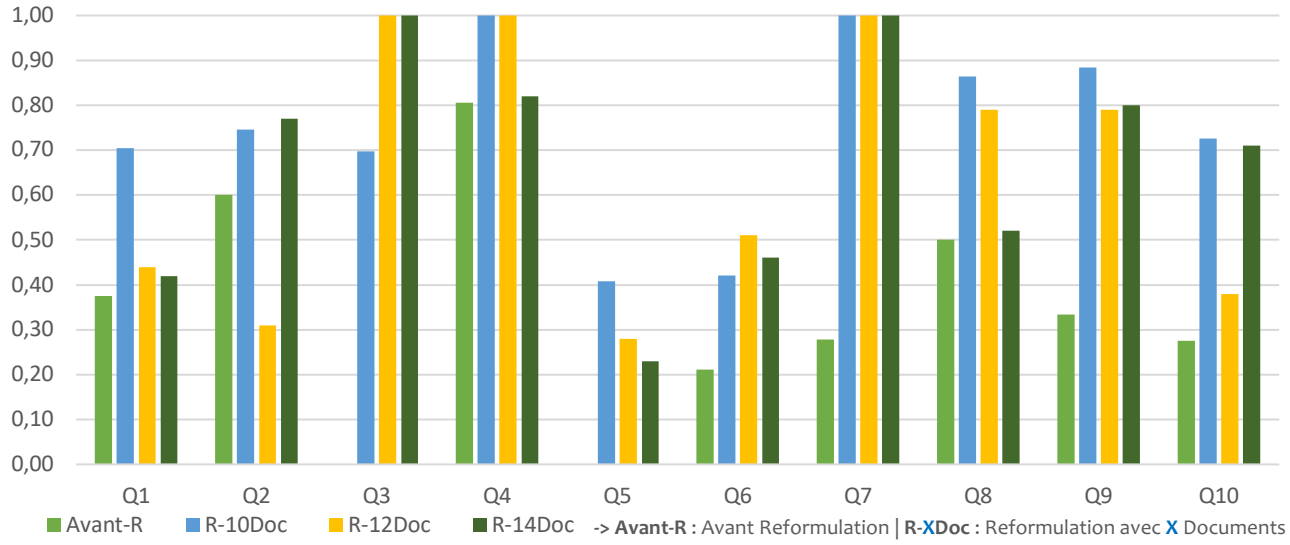

Fig. 3.10 Comparaison de la valeur de MAP des requêtes initiales avec le résultat de reformulation en utilisant des taxonomies construites sur la base de 10/12 et 14 documents

Tableau 3.12. La moyenne des valeurs de MAP de l'approche de reformulation en utilisant différents nombres de documents.

	Baseline (Originale)	Ref 10 Docs	Ref 12 Docs	Ref 14 Docs
MAP	0,34	0,75	0,65	0,67

D'après la figure 3.10, nous remarquons que la "Ref 10 Docs" est celle qui donne les meilleures valeurs de MAP avec une nette différence. La performance de chaque cas de reformulation est présentée dans le tableau 3.12 et montre que la reformulation utilisant les taxonomies construites à partir de 10 documents donne en moyenne une amélioration de 0,41 par rapport aux requêtes initiales.

A titre d'exemple, nous présentons dans la figure 3.11 une partie de la taxonomie de l'Angleterre contenant la ville de Londres qui est utilisée pour la reformulation de la requête Q2, et qui représente une amélioration de 24,23% par rapport à la valeur de MAP.

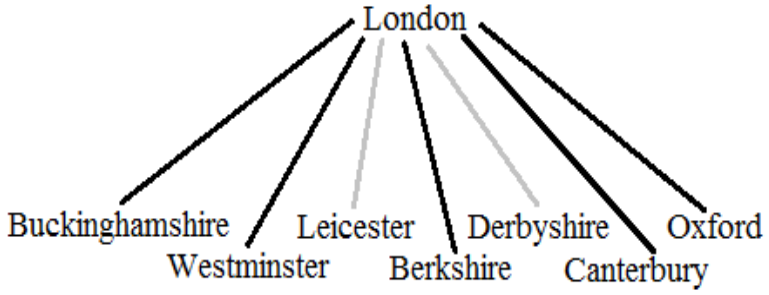

Fig. 3.11 La taxonomie d'adjacence de Londres construite en utilisant les dix documents récupérés en soumettant l'entité spatiale relative "*around London*".

Nous avons reformulé la requête Q_2 : "*Land for sale in the proximity of London*", en utilisant la taxonomie de niveau 1 présentée dans la Figure 3.11. La requête reformulée par l'approche QRGTW est comme suite :

(Land or ground or "landed estate" or acres or demesne) + for sale + (Buckinghamshire OR Westminster OR Leicester OR Berkshire OR Derbyshire OR Canterbury OR Oxford)

Afin de conclure concernant la performance de l'approche QRGTW, nous avons inclus 42 autres requêtes contenant des ESAs d'Angleterre à l'étude, en plus des requêtes présentées dans le tableau 3.11. Ainsi, les résultats qui suivent (Fig. 3.12 et Tableau 3.13) correspondent aux moyennes des 52 requêtes anglaises, en utilisant les deux mesures d'évaluation : MAP et précision à 10 (P@10). Les résultats ci-après compare notre approche (QRGTW) à la précision des mêmes requêtes dans les cas suivants :

- Sans reformulation (Baseline).
- Reformulé par l'algorithme de reformulation Rocchio (RRA) [1] appliqué en utilisant les entités spatiales absolues contenues dans les documents seulement. Ce qui donne de meilleurs résultats que l'algorithme connu de Rocchio appliqué tel qu'il est aux requêtes géographiques.

- Reformulé en utilisant la méthode de reformulation de requêtes en utilisant une taxonomie géographique (QRGT) [3], sans expansion de l'entité thématique.

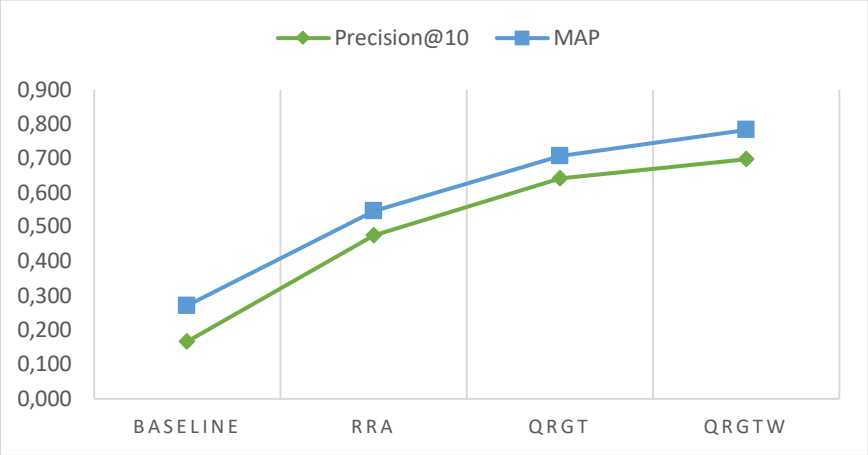

Fig. 3.12 Comparaisons des valeurs moyennes de P@10 et MAP de l'approche QRGTW aux requêtes originales, requêtes reformulées par RRA et QRGT.

La figure 3.12 montre que l'approche proposée améliore significativement la précision des requêtes utilisées lors de nos simulations, par rapport aux trois référentiels (cas de bases) utilisés. Le tableau 3.13, vient confirmer ce résultat et donner plus de détails, en présentant le taux de performance de l'approche QRGTW par rapport à la reformulation par l'algorithme Rocchio modifié et l'approche QRGT.

Tableau 3.13. Le taux de performance de l'approche QRGTW par rapport aux approches RRA et QRGT.

Measures	**RRA**	**QRGT**
Precision@10	46,37%	8,68%
MAP	43,11%	10,83%

À partir du tableau 3.13, nous pouvons conclure que l'approche QRGTW proposée donne une amélioration très importante de la précision des requêtes selon la reformulation en utilisant l'algorithme de Rocchio même après sa modification pour mieux s'adapter aux requêtes géographiques.

Dans la sous-section suivante, nous présentons notre 3ème série de tests de validation de l'approche QRGTW. D'après les résultats présentés jusqu'à ce point-là, nous remarquons que l'approche est prometteuse et que pour une validation finale nous utiliserons une autre méthode de la littérature comme référentiel de base en plus de l'algorithme Rocchio, que nous n'excluons pas de l'étude malgré l'écart de performance observé. Car c'est une approche communément utilisé qui appliqué à des requêtes réelles, a toujours montré des performances intéressantes.

5.3. La 3ème série de tests : Cas d'utilisation des requêtes géographiques des journaux log de AOL

Lors de cette série de tests, nous avons utilisé un journal de requêtes web qui est couramment utilisé comme collection de test.

- **Description des données :**

Nous avons utilisé une trace (journaux de requête), qui enregistre les requêtes des utilisateurs du moteur de recherche AOL sur les trois premiers mois de 2006. La collection se compose d'environ 36 millions lignes de données, chaque ligne contient cinq champs :

- **AnonID :** un user-ID anonyme
- **Query :** les termes de la requête
- **QueryTime :** le temps de soumission de la requête
- **Item-Rank :** le classement du résultat cliqué
- **ClickURL :** le résultat sur lequel l'utilisateur a cliqué au niveau hôte. (S'il existe)

Dans le cas où l'utilisateur a cliqué sur plusieurs résultats pour une seule requête, ces événements sont enregistrés sous forme de lignes supplémentaires (une ligne par résultat) et les trois premières données sont redondant.

Les principales remarques qui ont été notées par rapport à cette trace sont [4] :

- La longueur moyenne des requêtes de la collection est de 3.5 termes et la longueur maximale est de 9 termes
- 28% des requêtes sont des reformulations d'une requête précédente et en moyenne une requête est reformulée 2.6 fois.
- Les requêtes de navigation[5] représentent 21% de la fréquence totale des requêtes.

D'un point de vue géographique [5] :

- Les requêtes géographiques ont tendance à avoir plus de termes que les requêtes non géographiques. Par conséquent, la probabilité qu'une requête soit géographique augmente avec l'augmentation du nombre de termes. Une requête très courte est moins susceptible d'être reconnue comme requête géographique même si l'intention sous-jacente de l'utilisateur est géographique.
- Il a également été rapporté que 12,7% des réécritures de requêtes ajoutent un terme géo-spécifique. Ainsi, la requête d'origine avait probablement une intention géographique non exprimée.

- **Prétraitement de la collection et classification des requêtes :**

Lors de ces dernières simulations, nous avons utilisé la collection-01 des journaux de requêtes AOL. Cette collection comprend 3 558 412 lignes. Les statistiques de l'étude et résultat de prétraitement (nettoyage) effectués sur cette collection sont reportées dans le tableau 3.14.

[5]Requêtes de navigation : dans le cas où l'utilisateur a un site Web ou une page en tête qu'il sait ou dont il suppose l'existence. Les requêtes de navigation contiennent souvent des fragments d'URL ou des noms d'organisations. L'utilisateur clique généralement sur un seul résultat, l'amenant directement à la page désirée.
Les autres types de requêtes sont : requêtes informatives et requêtes transactionnelles (ressources).

Avant la classification des requêtes de la collection en requêtes géographiques et requêtes non géographiques, nous avons d'abord supprimé les requêtes dupliquées et/ou mal formées (contenant des fautes de frappes). Cela nous a conduit à une liste de 344 794 requêtes.

Nous avons également supprimé toutes les requêtes constituées exclusivement d'URL (celles de navigation), qui représentent 19,44% de la fréquence de requêtes dans la collection-01.

Après une analyse manuelle effectuée sur la liste de requêtes résultante de ce prétraitement, nous avons décidé qu'afin de détecter une entité spatiale, nous considèrerons les relations de distance et de direction[6] uniquement (comme : *at; nearby; in the vicinity; west; ...).*

Enfin, pour chaque requête restante, nous avons décidé si elle avait une intention géographique ou non (Tableau 3.14). Ceci est réalisé sur la base de la phase 2 de notre approche qui consiste en l'extraction de l'information géographique et qui a été expliquée dans la section 3 de ce chapitre.

Cette phase a permis de différentier entre les requêtes comme "*birds with rings **around** neck*" ou "wild **turkey** eggs" qui contiennent des termes géographiques mais ne concernent pas une intention géographique, et les requêtes similaires à "*auto mobile accidents **around boston***", contenant des termes géographiques et ayant une intention géographique.

Ainsi, à partir des requêtes restantes (après prétraitement), nous avons extrait les requêtes géographiques (requêtes avec des termes géographiques et des intentions géographiques). Ces requêtes représentent 18,65% des requêtes de la collection, comme indiqué dans le tableau 3.14.

Tableau 3.14. Statistiques de la collection-01 des journaux de requêtes AOL.

	Fréquence
Enregistrements	3 558 412
Requêtes	374 851
Requêtes après pré-traitement	344 794
Requêtes de navigation (Supprimées)	67 411
Requêtes restantes	277 383
Requêtes géographiques	64 289

Avant de construire la taxonomie et afin de restreindre notre champ d'expérimentations, nous avons recherché quels sont les pays concernés par les requêtes géographiques contenues dans les journaux de requêtes utilisés dans notre étude.

Le résultat trouvé, indique que 49,29% des entités spatiales des requêtes géographiques contiennent le nom de lieux contenus dans les États-Unis d'Amérique. Ainsi, nous avons continué nos tests avec ces requêtes et nous avons utilisé le générateur de taxonomie géographique de l'adjacence GATB afin de construire la taxonomie des États-Unis.

[6] https://en.wikipedia.org/wiki/Spatial_relation

L'expérimentation qui suit concerne l'étude de l'impact de la variation du nombre de documents utilisés sur le taux d'erreur dans la taxonomie américaine construite. Nous avons varié le nombre de document de 5 à 30 et les résultats de la simulation sont illustrés par le tableau 3.15.

Tableau 3.15. L'impact de la variation du nombre de documents utilisés dans la construction de la taxonomie des États-Unis sur le taux d'erreur.

Documents number	5	10	15	20	25	30
Error rate	20	7,67	12,50	14,29	22,22	33,33

Nous remarquons sur le tableau 3.15 que 5 documents sont insuffisants pour créer une bonne taxonomie. Les résultats montrent également que la plus faible valeur du taux d'erreur est atteinte en utilisant 10 documents, puis la valeur augmente continuellement. L'augmentation est graduelle mais la performance du cas d'utilisation de 10 documents est significative, avec une amélioration minimum de 4,83%.

Conformément aux résultats du tableau 3.15, qui vient confirmer les résultats illustrés par la figure 3.10 et le tableau 3.12, le nombre de documents utilisés dans le processus de construction de la taxonomie exploitée dans l'étape de reformulation de requête est de dix.

- **La reformulation de requêtes :**

L'objectif principal de la 3ème série de tests et qui a permis de valider notre travail, est réalisé lors de cette dernière simulation. Cet objectif peut être exprimé comme la reformulation des requêtes géographiques extraites de la collection-01 des journaux de requêtes AOL et qui concernent des lieux aux États-Unis d'Amérique.

Les résultats qui suivent (figure 3.13 et tableau 3.16) correspondent à la moyenne des valeurs des mesures d'évaluation calculées pour les 31 688 requêtes extraites des traces AOL (Original), et les requêtes reformulées par différentes méthodes de reformation.

Les mesures d'évaluation utilisées sont la MAP et la précision à 20 documents (P@20).

Afin de conclure sur la performance de notre approche nommée QRGTW et la valider, nous avons comparé ses résultats, une fois de plus, à la précision des requêtes reformulées en utilisant :

- L'algorithme de Rocchio présenté dans [1] et appliqué dans un contexte purement géographique. Puisque nous avons limité les données d'entrée de l'algorithme aux termes géographiques des requêtes et des documents uniquement : RAGR.
- La méthode d'expansion de requêtes géographiques proposée par Perea-Ortega et al. dans [6]. Cette approche qui est une méthode de traitement du langage naturel (NLP pour Natural Language Processing) reformule et/ou étend à la fois les entités thématique et géospatiale : GQEM.
- Notre approche se basant sur le générateur GATB mais qui ne reformule que l'entité spatiale des requêtes : QRGT.

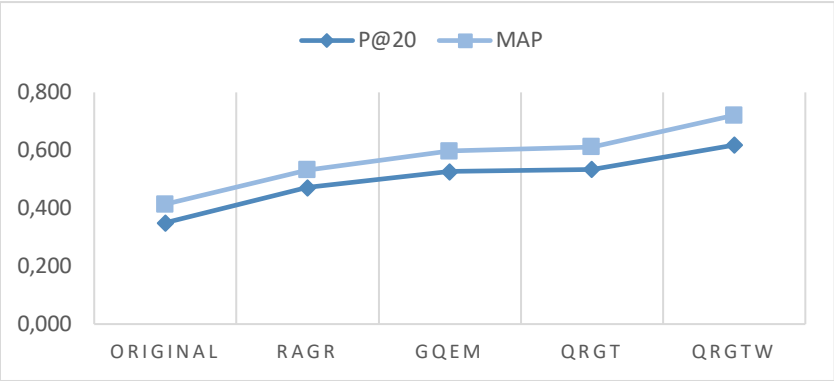

Fig. 3.13 Comparaisons des valeurs moyennes de P@20 et MAP de l'approche QRGTW aux quatre référentiels de bases : les requêtes originales et les requêtes reformulées par RAGR, GQEM et QRGT.

La figure 3.13 montre que l'approche proposée améliore significativement la précision des requêtes par rapport à tous les référentiels de bases utilisés pour les deux mesures utilisées. Nous remarquons aussi que la méthode d'expansion de requête géographique (GQEM) et la méthode QRGT ont des performances assez proches de l'approche QRGTW. Par ailleurs, nous remarquons que la technique de reformulation géographique basée sur l'algorithme Rocchio (RAGR) n'a pas amélioré la précision des requêtes de la trace AOL de façon considérable.

Tableau 3.16. Le taux de performance de l'approche QRGTW par rapport aux approches RAGR, GQEM et QRGT.

	RAGR	**GQEM**	**QRGT**
P@20	31,21	17,27	15,73
MAP	35,52	20,77	17,81

A partir de la figure 3.13 et le tableau 3.16, nous concluons que l'approche proposée apporte une amélioration importante à la précision des requêtes comparée aux méthodes QRGT et GQEM [6]. Ce fait peut être expliqué comme suite :

- QRGTW dépasse QRGT en raison de la reformulation des entités thématiques et spatiales, en utilisant les sens des mots extraits de la base de données sémantique WordNet afin d'étendre l'entité thématique de la requête.
- QRGTW surpasse GQEM, car au lieu d'étendre l'entité spatiale avec plus de noms de lieux géographiques et de surcharger la requête d'informations, ce qui peut conduire à des ambiguïtés. Notre approche remplace l'entité spatiale de la requête par des endroits qui expliquent plus précisément l'intention de la requête à l'aide d'une taxonomie géographique construite en se basant sur une méthode d'indexation qui a été validée et recommandée par plusieurs travaux.

6. Conclusion :

Dans ce chapitre, nous avons présenté notre première contribution, qui consiste en une nouvelle technique efficace de reformulation de requête géographique utilisant une taxonomie géographique et les sens ou synonymes des termes, appelés QRGTW. Notre approche sépare automatiquement les composants d'une requête géographique et reformule l'entité spatiale à l'aide du GATB (Geographical Adjacency Taxonomy Builder), qui utilise la méthode LSI (Indexation Sémantique

Latente). Tandis que, la partie thématique de la requête est enrichie en utilisant les sens des mots extraits de l'ontologie sémantique WordNet.

Nous avons mené plusieurs séries de tests, pour vérifier si notre approche améliore la précision des requêtes, avec trois collections différentes. Nous avons construit la première et deuxième collections basée sur une simulation réalisée avec 10 utilisateurs volontaires et construit une taxonomie géographique du Maroc ainsi que des parties de la taxonomie d'Angleterre pour reformuler les requêtes de ces collections. La troisième collection de requêtes est composée de requêtes géographiques extraites d'un journal de requête du moteur de recherche AOL. Une étude statistique nous conduit à la construction de la taxonomie des états unis d'Amérique pour reformuler les requêtes de cette collection. Notre méthode QRGTW, a été comparé à une technique de reformulation de requête géographique basée sur l'algorithme de Rocchio (GQRA), une technique d'expansion de requête géographique (GQEM) et notre approche QRGT qui ne considère pas la reformulation de l'entité thématique. Les résultats de nos simulations montrent que l'approche proposée surpasse les différents référentiels de bases considérés en termes de toutes les métriques utilisées.

Malgré les valeurs satisfaisantes de précision que nous a donné l'approche QRGTW, nous ne pouvons ignorer le fait que le taux d'erreur des taxonomies exploitées dans les différentes simulations reste non négligeable. De ce fait, nous avons eu recours à d'autre test qui ont mené vers une deuxième contribution qui suit la structure et le processus de la première. Toutefois, celle-ci permettra d'améliorer la précision des requêtes à intention géographique, toute en minimisant le taux d'erreur de la taxonomie construite en première phase. Un autre inconvénient, qui est spécifique à l'utilisation de la méthode d'indexation Sémantique latente est la complexité de l'algorithme qu'elle sous-entend. Elle est difficilement applicable à de larges collections car la taille des matrices augmente considérablement plus le nombre de documents et le nombre de termes sont élevés.

Chapitre 4 :
Reformulation de requêtes géographiques : approche basée sur un générateur de règles d'association parallèle

1. Introduction

Malgré l'efficacité de plusieurs algorithmes séquentiels, les performances de ces derniers se dégradent lorsque la taille des données traitées augmente. Pour maintenir les performances de ces algorithmes, le développement d'algorithmes distribués et parallèles est une solution pouvant aider à accélérer la vitesse de traitement et réduire la taille d'espace mémoire utilisée.

Dans ce chapitre, nous présentons une autre approche automatique de construction de taxonomie spatiale. Celle-ci modélise la notion de contiguïté dans un contexte distribué, qui sera utilisée dans la reformulation de la partie spatiale d'une requête géographique. Cette approche exploite les documents les mieux classés sur la liste récupérée lors de la soumission d'une entité spatiale, qui est composée d'une relation spatiale et d'un nom de ville. Ces documents sont utilisés afin de créer une base de données transactionnelle. Nous considérons chaque document extrait comme une transaction contenant les noms des villes du pays dont nous allons construire la taxonomie.

Dans cette contribution, nous nous sommes basés sur une méthode d'apprentissage automatique (Machine Learning) et précisément de génération de règle d'association, nommée Frequent Pattern Growth (FP-Growth). Cette méthode a été optimisée et appliquée à la base de données transactionnelle, que nous avons créé, dans sa version parallèle (Parallel FP-Growth: PFP) afin de générer les règles d'association, qui formeront la taxonomie du pays dans un contexte Big Data. Ainsi, ceci nous a mené également vers l'évaluation de la rapidité de réponse du système pour valoriser le choix d'un algorithme parallèle. Pour le calcul de la performance en terme de rapidité, nous avons utilisé la loi d'Amdahl que nous expliquons également dans ce chapitre.

Toutefois, l'approche de reformulation proposée suit le processus et les étapes de réalisation de l'approche présentée dans le chapitre 3, seules les méthodes de construction de taxonomies diffèrent. Donc, nous n'allons pas nous attarder sur la ré-explication du modèle de requête utilisé ni la façon de reformulation des requêtes géographique. Nous allons nous focaliser plutôt sur les détails de création de la taxonomie en se basant sur l'algorithme FP-growth et sa parallélisation.

2. L'algorithme FP-growth

L'algorithme FP-growth est une technique d'apprentissage non-supervisée avec FP est l'acronyme pour Frequent Pattern, Motif fréquent en français. Cet algorithme diffère de l'algorithme Apriori [67, 68], conçu pour le même but, par sa deuxième étape qui utilise une structure arborescente, appelé FP-tree, pour coder des transactions sans la génération explicite des ensembles candidats, qui sont généralement coûteux à générer lors de l'utilisation de *Apriori*. Par contre, FP-growth adopte une stratégie de découpage pour décomposer les tâches d'exploration de données. Il utilise la méthode « pattern fragment growth » pour éviter le coûteux processus de génération et de test des ensembles candidats.

Cette technique est représentée par un algorithme de deux phases. La première consiste en la construction d'une structure d'arbre (FP-tree), « Arbre des motifs fréquents » en français, pour

stocker une forme compressée d'une base de données. Et la seconde génère les motifs fréquents à partir d'une structure arborescente conditionnelle générée élément par élément, appelée « conditional FP-tree ».

2.2. Phase 1 : Construction du FP-Tree

Cette phase consiste en deux scans sur la base de données utilisée, représentant par deux étapes :

- **Étape 1 :** le premier scan permet de sélectionner les items (éléments) fréquents, qui sont ordonnés en se basant sur leurs valeurs de support $Supp(E_i)$ (équation 4.1) dans l'ordre descendant, afin de former une nouvelle structure appelée F-list. Les items non fréquents sont éliminés de la F-list en se basant sur un seuil spécifié noté minsup.
  Soit un ensemble d'éléments A, avec N est le nombre de transactions dans la base de données, et f(A) la fréquence de A dans la base. Le support de A est la probabilité que l'itemset A se vérifie dans la base de données :
  $$Supp(A) = P(A) = f(A) / N \quad (4.1)$$
- **Étape 2 :** le deuxième balayage de la base compresse les ensembles (itemsets) fréquents représentés dans la base de données à l'aide d'un FP-tree (frequent-pattern tree) dont les branches contiennent les associations possibles des items [69]. Le FP-tree est construit en modélisant les transactions de la base une par une sous forme d'arbre, en insérant successivement chaque élément à partir du plus fréquent. Cet ordre fixe est utilisé de sorte que les transactions se superposent lorsque les items partagent des éléments communs (ils ont alors le même préfixe).
  Chaque nœud de l'arbre correspond à un item auquel est associé un compteur. Chaque compteur est incrémenté lorsque le chemin est revisité, c-à-d que la transaction en cours de modélisation partage les mêmes nœuds d'une ou de plusieurs transactions précédentes. Plus il y a de chemins superposés, plus la compression est grande.
  Pour compléter le FP-tree, des pointeurs (en pointillé) sont ajoutés afin d'indiquer les liens entre les différents nœuds des mêmes items et afin d'accélérer la recherche dessus.

2.2. Phase 2 : Génération des motifs fréquents (Frequents patterns)

L'algorithme FP-Growth extrait les ensembles fréquents à partir du FP-tree, par un parcours ascendant des feuilles vers la racine. L'arbre est parcouru nœud par nœud en commençant par l'élément le moins fréquent sur la F-list. FP-growth utilise la stratégie "diviser et dominer" (divide-and-conquer) en collectant les éléments sur le chemin du nœud à la racine de l'arbre. Ces éléments collectés constituent les éléments de la base de motif conditionnel de l'élément courant dans la F-list.

La base de patterns conditionnels d'un élément est définie comme une petite base de données de motifs qui coexistent avec cet élément. Une base de patterns conditionnels d'un élément E_i est extraite à partir de la base de données initiale, en éliminant les transactions où E_i n'apparait pas,

puis en éliminant E_i des transactions restantes. Ensuite, FP-growth crée de petits FP-Trees en se basant sur les bases de patterns conditionnels et réexécute l'algorithme de manière récursive sur les nouveaux FP-Trees jusqu'à ce qu'aucune base de modèle conditionnelle ne puisse être générée.

3. Génération des règles d'association

Les règles d'association à utiliser pour l'aide à la décision sont générées à partir des patterns fréquents résultants.

Une règle est valide lorsque sa confiance (équation 4.2) est supérieure ou égale à la valeur de confiance minimale fixé *minconf*.

Soit A et B des ensembles d'éléments formant les transactions de la base de données. La confiance que A peut impliquer B est la probabilité conditionnelle que si A est trouvé, alors B aussi est inclut dans la même transaction :

$$Confiance(A \Rightarrow B) = P(A|B) = Supp\ (A \cup B)/Supp(A) \qquad (4.2)$$

Il faut également noter que si la règle (A-B ⇒B) est éliminée car sa confiance est trop faible, alors toutes les règles (A-C ⇒ C), avec B ⊂ C, seront éliminées.

4. L'algorithme Parallel FP-Growth

Avec la disponibilité des Framework de programmation performants comme MapReduce [70] et Spark [71], le traitement des données massives est devenu une tâche facile à accomplir. Cependant, la plupart des algorithmes parallèles d'apprentissage automatique souffrent encore de plusieurs problèmes. En effet, les algorithmes parallèles d'extraction des motifs fréquents gardent les mêmes limitations que leurs versions séquentielles. Toutefois, l'algorithme FP-Growth, a été considéré comme l'algorithme le plus efficace pour l'extraction des motifs fréquents avec un très grand volume de données. Son implémentation parallèle PFP-Growth, qui a été proposée la première fois par [72] dans un contexte MapReduce et par [73] avec Spark, est considérée comme un des algorithmes parallèles d'extraction des motifs fréquents les plus performants dans un environnement massivement distribué.

Malgré ses avantages, avec un seuil de support minimum très petit et de grands volumes de données, PFP-Growth ne passe pas à l'échelle avec son implémentation MapReduce [74]. La raison de cette limitation est justement liée au fait qu'il travaille en mémoire, ce qui rend les performances dépendantes des capacités mémoire de la machine. Par conséquent, nous avons choisi d'adapter et d'optimiser son implémentation sur le moteur de traitement de données Apache Spark dont la caractéristique principale est le traitement In-memory, ce qui permet d'augmenter les performances d'algorithme similaire à PFP-growth.

Sur la littérature, le PFP-Growth a été appliqué avec succès pour extraire efficacement les itemsets fréquents à partir des données massives. Précisément en utilisant Spark, le processus de fouille de PFP-Growth se déroule en mémoire.

L'algorithme FP-growth parallélisé œuvre sur des machines distribuées [76]. Sa tâche de partitionnement est effectuée de telle manière que chaque machine exécute un groupe indépendant de tâches. Cette méthode de partitionnement élimine les dépendances de calcul entre les machines, et par la suite la communication entre elles. La figure 4.1, illustre le workflow de notre version de l'algorithme PFP-growth, celle que nous avons considéré pour notre approche en se basant sur l'architecture du Framework Spark.

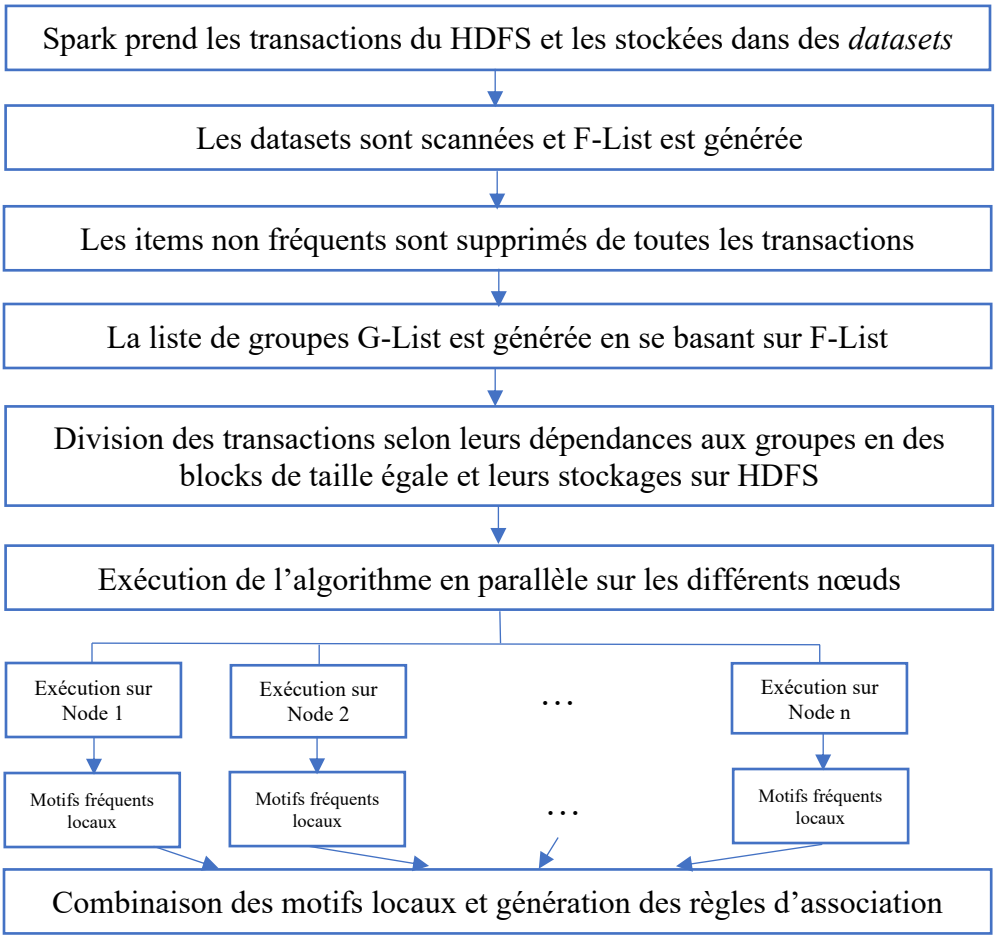

Fig. 4.1 Algorithme PFP-growth amélioré basé sur Spark.

Étant donné une base de données de transactions DB, les étapes de l'algorithme PFP-growth utilisé dans la création de nos taxonomies ultérieurement sont les suivantes :

- **Partitionnement (Sharding)** : diviser DB en parties successives et mémoriser les parties sur *n* machines différentes. Chaque partie résultante est appelée un fragment (Shard).
- **Calcul parallèle :** les blocs sont lus du HDFS et stockés sur des structures *datasets*, en suite, les valeurs de support de tous les éléments apparaissant dans chaque fragment sont calculées. Cette étape permet de découvrir implicitement le vocabulaire des items, qui est normalement inconnu pour une énorme base de données. Le résultat de cette étape est une F-list.

- **Réduction de taille** : en se basant sur la F-list résultante de l'étape précédente, les items non fréquents sont supprimés de toutes les transactions de DB.
- **Regroupement des éléments**: Considérant I le vocabulaire de DB, diviser les | I | éléments apparaissant dans la F-List en Q groupes. La liste des groupes est appelée G-list. La F-list et la G-list étant toutes les deux petites, cette étape peut être exécutée sur un seul nœud du cluster en quelques secondes.
- **Parallélisation:** division des transactions selon leurs dépendances en groupes en des blocks de taille égale et leurs stockages sur le HDFS. L'algorithme FP-Growth est appliqué sur les blocks générés et des FP-Trees locaux sont construites en parallèle et leurs FP-trees conditionnelles sont créées de manière récursive, jusqu'à génération des motifs fréquents locaux.
- **Agrégation:** agrégation des résultats générés lors de l'étape de parallélisation et génération des règles d'association.

Puisque, PFP distribue le travail de croissance des FP-trees en fonction des groupes de transactions. Par conséquent, cette approche est plus évolutive qu'une implémentation à nœud unique. Il est également implémenté sur la bibliothèque Machine Learning (Mllib) de Spark et prend trois paramètres notamment, le support minimum afin d'identifier les ensembles fréquents, la confiance minimale pour générer des règles d'association et enfin le nombre de fragments utilisés pour distribuer le job.

5. Taxonomie géographique d'adjacence

Considérant une base de données où les documents sont les transactions et les éléments sont les villes du pays pour lequel nous voulons construire une taxonomie. Nous proposons de construire une taxonomie géographique d'adjacence en utilisant l'algorithme PFP-growth. Par exemple, si la taxonomie que nous voulons créer est celle du Maroc, les éléments de la base de données sont les villes du Maroc, les transactions (Table 4.1) sont les documents les mieux classés extraits du moteur de recherche lors de la soumission de chaque Entité Spatial Relative ESR (ex. "près de Casablanca"), et la racine de la taxonomie est l'ESA initiale "Rabat", la capitale.

Tableau 4.1 Exemple de base de données créée dans le but de construction d'une taxonomie spatiale

ID Transaction	***Transaction***
Document1	Rabat Salé Temara Kénitra
Document2	Skhirate Rabat Kénitra
Document3	Bouznika Casablanca
Document4	Rabat Bouznika Marrakech

Notre méthode ne tient pas compte de tous les k-ensembles, mais seulement des ensembles qui contiennent l'ESA en cours, pour laquelle nous cherchons les règles valides dans le moment courant. Par conséquent, à partir des règles générées nous ne retenons que celles dont la racine est l'ESA courante (ESA_i). Les règles résultantes ont la forme "$ESA_i \Rightarrow$ {ensemble d'ESA}", où l'ensemble d'ESA contient un seul ESA ou plus. Ainsi, nous gardons les règles dont la confiance

est élevée, sachant qu'à partir de chaque k-ensembles fréquent nous n'avons besoin que d'une seule règle spécifique dont la racine est ESA_i. Ces conditions réduisent le temps consommé pour le traitement de l'algorithme, et conduisent vers une vitesse d'exécution plus élevée de l'algorithme PFP-growth. Enfin, la fusion des règles générées forme une taxonomie dont la racine est l'ESA_0 (figure 4.2), avec laquelle nous choisissons de commencer notre processus de construction de taxonomie.

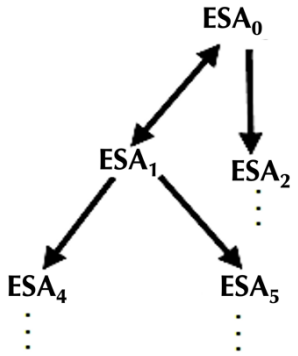

Fig. 4.2 Une taxonomie exemple commençant par l'ESA_0.

Les tests réalisés sur cette approche ont également confirmé le gain en précision lors de l'intégration de l'étape de validation que nous avions déjà proposé d'utiliser sur notre première contribution. Ainsi, chaque arc de la taxonomie est sujet à une étape de validation par la recherche de la règle inverse correspondante sur la liste des règles générées. Par exemple, pour ESA_1, si elle implique également ESA_0 ($ESA_0 \Rightarrow ESA_1$ et $ESA_1 \Rightarrow ESA_0$) alors on garde cet arc et il est dit « lien validé ». Dans le cas contraire, si l'ESA_3 n'est pas validé ($ESA_0 \Rightarrow ESA_3$ mais $ESA_3 \nRightarrow ESA_0$) l'arc en question est supprimé de la taxonomie. La double implication est considérée comme une confirmation des informations générées par l'implication descendante. Donc, nous assumons que les ESAs en question se génèrent mutuellement et sont hautement liées et adjacentes par conséquent.

A la fin de ce processus, une taxonomie géographique d'adjacence de pays est complétée. L'objectif principale de cette dernière est de permettre la reformulation de requêtes d'adjacences par l'interprétation de leurs entités spatiales relatives. Sachant qu'une requête contenant une relation d'adjacence signifie que l'utilisateur cherche les lieux qui entourent l'ESA de cette requête. Ainsi, toute l'entité spatiale relative doit être éliminée et remplacée par les nœuds fils de l'ESA en question dans la taxonomie créée (voir la section 4 du chapitre 3 pour les détails de l'étape de reformulation).

6. Résultats des expérimentations

6.1. Construction de taxonomies

L'approche de construction de taxonomie géographique, se basant sur l'algorithme parallèle de recherche des motifs fréquents, est appliquée en utilisant un lexique de relations spatiales et une base de données d'ESAs associées à leurs pays.

Afin de tester et de vérifier la performance de notre approche, nous avons commencé par réaliser la taxonomie du Maroc. Donc, pour pouvoir exploiter les pages web réalisées par les marocains eux même, les expérimentations ont été réalisées en langue françaises. "Rabat", la capitale, est considérée comme la racine de la taxonomie. C'est avec cette ESA que nous avons commencé le processus de construction.

Nous appliquons notre méthode en utilisant une base de données de transactions qui est construite en itérant sur la liste des ESAs du Maroc (une liste sélectionnée de 50 villes et villages du Maroc). Nous avons sélectionné les trente premières pages Web récupérées lors de la soumission d'une entité spatiale relative contenant l'ESA courante sur la liste. Toutefois, pour l'extraction de texte utile uniquement, on a également spécifié les conditions suivantes :

- Ne pas inclure les liens hypertexte qui permettent d'aller vers d'autres pages pouvant être généralement des pages publicitaires ou des propositions qui forment un bruit ;
- Désactiver la recherche basée sur l'historique de recherche ;
- Utiliser les services de recherche de Google sans avoir besoin d'un profil spécifique.

En guise de prétraitement des textes extraits, nous avons unifié les noms des ESAs en raison des différentes manières d'écrire les noms de villes par les personnes qui ont rédigé le contenu des documents. En effet, nous avons remarqué précédemment que le problème de la correspondance se pose particulièrement dans le cas des noms qui contiennent des accents ou bien qui sont écrits en dialectes marocains. L'unification des noms des ESAs est réalisée en utilisant la base de données GeoNames qui contient des noms alternatifs pour décrire le même terme en différentes formes et langues.

Lors du premier test, nous avons varié la relation spatiale (RS) sur les entités relatives soumises pour vérifier si la variation de RS influence la performance de l'approche proposée. Les relations spatiales utilisées dans cette étape de test sont listées sur le tableau 4.2. Les cinq documents les mieux classés ont été extraits sur ce test en utilisant l'ESA Rabat et variant la relation d'adjacence utilisée.

Tableau 4.2 Relations Spatiales utilisées lors du premier test

Annotation	Expression
RS 1	à côté de
RS 2	à la périphérie de
RS 3	à proximité de
RS 4	aux alentours de
RS 5	aux environs de
RS 6	les environs de
RS 7	près de

L'algorithme PFP-growth est appliqué à cette première base construite (nommée BD1), puis les règles d'association sont générées entre Rabat et les ESAs marocaines qui coexistent avec elle dans BD1. Lors de ce test, nous faisions varier le support minimum de 0,2 à 0,8 (tableau 4.3), tout en fixant la confiance minimale à 0,6. Tandis que l'étape de validation des règles extraites n'a pas été considérée. Ensuite, nous avons calculé le taux d'erreur et le nombre de règles générées dans chaque cas.

Tableau 4.3 Le taux d'erreur et le nombre de règles générées en variant le minsup et la relation spatiale utilisée pour les motifs fréquents contenant l'ESA "Rabat"

	Taux d'erreur (%)				**Nombre de règles générées**			
RS\ minsup	0,2	0,4	0,6	0,8	0,2	0,4	0,6	0,8
RS 1	72,73	28,57	33,33	0	22	7	3	1
RS 2	42	0	0	0	5	2	1	1
RS 3	25	0	0	-	4	1	1	0
RS 4	40	33,33	0	0	10	3	1	1
RS 5	33,33	0	0	0	9	2	2	1
RS 6	40	50	0	0	10	4	2	2
RS 7	0	0	0	-	6	6	1	0

D'après le tableau 4.3, nous constatons qu'en utilisant le minsup = 0,8, l'algorithme ne retournent aucun résultat dans certains cas, sinon il ne donne que 1 ou 2 réponses. La même chose pour minsup = 0,6 qui ne dépasse pas les 2 bonnes réponses. En ce qui concerne la valeur 0,2, il donne généralement un taux d'erreur élevé et retourne parfois un nombre très élevé de réponse jusqu'à 22 ESA résultant dans le cas de la RS 1 avec 6 ESA adjacentes correctes. Ainsi, nous avons privilégié la valeur du support minimal égal à 0,4, car elle donne le meilleur rapport entre un taux d'erreurs minimum et un nombre acceptable de réponses.

Les tests suivants, dont les résultats sont présentés dans le tableau 4.4, sont réalisés afin de comparer les cas où nous utilisons ou non l'étape de validation pour la construction de la taxonomie géographique d'adjacence. Tel que, les paramètres considérés pour l'application de notre approche sont fixés à 0,4 pour le support minimum et 0,6 pour la confiance minimum. Nous avons également ajouté un cas supplémentaire lors des tests, c'est le cas de calcul du support moyen (SM). Son expression est définie par l'équation (4.3) :

$$SM(A \Leftrightarrow B) = Supp(A \Rightarrow B) + Supp(B \Rightarrow A) \quad (4.3)$$

Tableau 4.4 Le taux d'erreur et le nombre de règles correctes générées en utilisant l'étape de validation ou non, et en utilisant la moyenne du support avec variation de la relation spatiale

	Taux d'erreur (%)			Nombre de règles correctes		
Relation Spatiale	**SV**	**AV**	**SM**	**SV**	**AV**	**SM**
RS 1	28,57	0	0	7	2	2
RS 2	0	0	0	2	1	1
RS 3	0	-	0	1	0	1
RS 4	33,33	50	33,33	3	2	3
RS 5	0	0	0	2	2	2
RS 6	50	0	0	4	2	2
RS 7	0	0	0	6	5	6

SV : Sans validation AV : Avec validation SM : Support moyen

En comparant les résultats avec la validation et les résultats sans la validation, nous remarquons que le taux d'erreur diminue lors de l'utilisation de l'étape de validation, à l'exception de la RS 4 où à partir de 3 résultats incluant 2 ESA correctes, la validation a éliminé l'un des ESA correctes et elle a gardé l'une des ESAs erronées. Concernant la RS 3, nous remarquons que la seule ESA qui a été récupérée sans validation a été éliminée à l'étape de la validation. Toutefois, en général, nous concluons que l'étape de validation minimise suffisamment le taux d'erreur.

Afin de minimiser le taux d'erreur tout en gardant autant que nous pouvons de résultats corrects (éliminer seulement les ESA erronées par l'étape de validation), nous proposons de calculer la moyenne des deux supports des règles opposées (i.e. $ESA_1 \rightarrow ESA_2$ and $ESA_2 \rightarrow ESA_1$). Le tableau 4.4 montre que les résultats donnés par le cas d'utilisation de la moyenne des supports résout les problèmes mentionnés ci-dessus pour la RS 3 et la RS 4.

En comparant les sept relations spatiales, nous avons favorisé la RS 7 "près de" qui donne le meilleur résultat avec 0% d'erreur et six ESA correctes qui seront les nœuds fils de l'ESA Rabat dans la taxonomie d'adjacence (Figure 4.3). Nous avons également essayé de varier la valeur de la confiance minimale. A cet effet, nous avons constaté que les meilleurs résultats sont donnés par la valeur 0,6 fixé apriori.

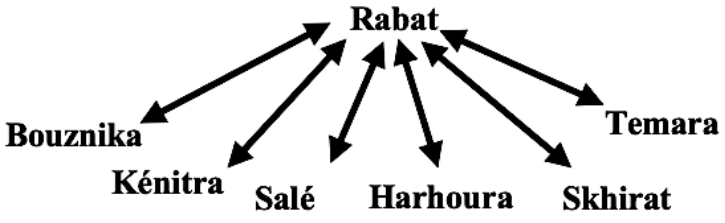

Fig. 4.3 Taxonomie niveau 1 pour l'ESA Rabat en utilisant la relation spatiale « près de ».

En utilisant les conditions favorables représentées ci-dessus, nous continuons la construction de la taxonomie du Maroc (figure 4.4) avec 0,4 comme support minimum, une valeur de 0,6 comme confiance minimum et en utilisant le support moyen pour la validation des liens.

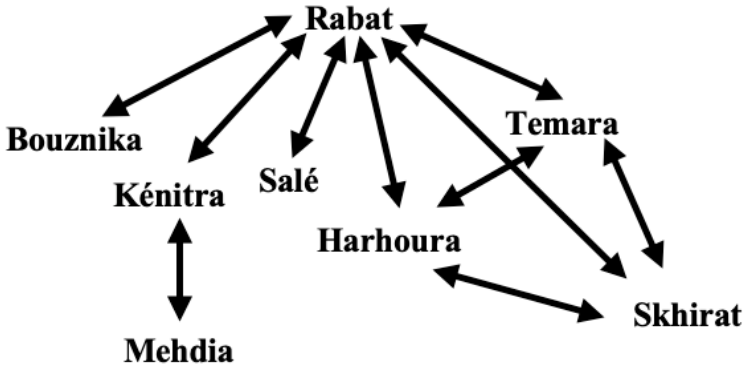

Fig. 4.4 Taxonomie géographique d'adjacence du Maroc (niveau 2).

En examinant, la figure 4.4 nous remarquons qu'elle contient 10 liens corrects. Tel que, la taxonomie dont une partie est représentée par cette figure est construite en utilisant une liste de 50 ESAs marocaines et en sélectionnant pour chaque ESR les 30 premiers documents retournés par le service de recherche google. Donc, 1500 documents ont été utilisés pendant ce test avec un support minimum de 0,4 et une confiance minimum de 0,6. La partie restante de la taxonomie contient quelques erreurs que nous pouvons ignorer, vu que le taux d'erreur globale est de 8,73% (11 liens faux sur 126 liens générés).

Nous avons également varié le nombre de nœuds (machines) utilisés afin d'évaluer l'efficacité de notre algorithme parallèle. Ainsi, nous avons pris la ligne de base de cette comparaison, comme l'utilisation d'un nœud avec deux cœurs de 2,3 GHz et 8G de mémoire, ce test a été réalisé en 1,69 secondes. Les caractéristiques des autres nœuds ajoutés pour former notre cluster sont :

- 2ème nœud : 2 cœurs de 2,3 GHz et 16 Go de mémoire ;
- 3ème nœud : 2 cœurs de 2,8 GHz et 4G de mémoire ;
- 4ème nœud : 2 cœurs de 3,1 GHz et 16 Go de mémoire.

Le tableau 4.5, montre le temps d'exécution de notre technique sur une branche de la taxonomie en variant le nombre de nœuds machines utilisés de 1 à 4, par conséquent, le nombre de cœurs de 2 à 8 et la capacité des RAMs combinées de 8G à 44G.

Tableau 4.5 Temps d'exécution (/secondes) de l'algorithme PFP sur un cluster Spark

# de nœuds	**1**	**2**	**3**	**4**
# de cœurs	2	4	6	8
RAM (Go)	8	8+16	8+16+4	8+16+4+16
Temps d'exécution	1.69	0.85	0.71	0.54

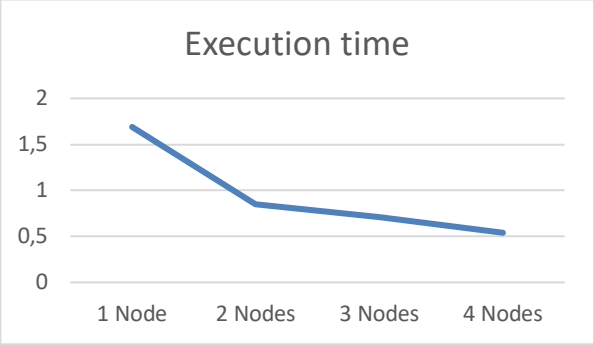

Fig. 4.5 Temps d'exécution (Illustration des résultats du tableau 4.5)

Le tableau 4.5 montre qu'en terme de temps d'exécution, l'efficacité de notre technique parallèle augmente tout en augmentant le nombre de nœuds utilisés. Compte tenu du cas de base, nous avons eu une accélération de 3,13 lors de l'utilisation de 4 nœuds. Cependant, nous remarquons sur la figure 5 que cette amélioration n'est pas proportionnelle au nombre de nœuds utilisés. Ainsi, nous nous attendons à ce que l'efficacité de l'algorithme stagne à un certain niveau.

6.2. Reformulation des requêtes

Afin d'évaluer la performance de notre approche d'une manière plus précise, nous avons commencé par proposer dix requêtes géographiques d'adjacence que nous avons soumis au service de recherche Google avant et après reformulation en utilisant la taxonomie construite avec l'algorithme PFP-growth. L'évaluation de ces tests est réalisée en comparant les valeurs de précision dans les deux cas (avant et après l'application de notre approche).

Toutes les requêtes géographiques utilisées lors de ce premier test sont du domaine de l'immobilier (tableau 4.6). L'objectif est de pouvoir tester notre taxonomie dans un domaine spécifique avant de la généraliser.

Tableau 4.6 Requêtes à reformuler sur le premier test.

Id requête	***Expression***
Q1	Appartement à vendre près de skhirat
Q2	Terrain à vendre à proximité de Rabat
Q3	Appartements à louer aux environs de Temara
Q4	Louer chalet aux alentours de Harhoura
Q5	Vente terrain à la périphérie de Marrakech
Q6	Bureau à louer dans les environs de Nouaceur
Q7	Appartement à vendre aux alentours de Kénitra
Q8	Acheter villa à proximité de Berrechid
Q9	Terrain à vendre près de Khemisset
Q10	Appartement à vendre à proximité de Casablanca

La technique de la reformulation expliquée sur la section 4 du chapitre 3 a été appliquée aux requêtes du tableau 4.6. Puis, nous avons soumis les requêtes avant reformulation et après reformulation au service de recherche. Enfin, nous avons mesuré la précision au rang 10 dans chaque cas, comme illustré sur la figure 4.6.

Le jugement de la pertinence de chaque document résultant se fait manuellement en vérifiant s'il représente une bonne réponse à la requête entière, l'entité thématique (ET) et l'entité spatiale (ES) au même temps. Par exemple, si l'ET de la requête est "appartement à vendre" et le document propose une villa, un terrain ou un bien immobilier à louer et non pas à vendre, ce document est jugé non pertinent. De la même manière, si l'ES est "à la Périphérie de Marrakech", un document qui propose un bien immobilier à Marrakech est considéré comme non pertinent.

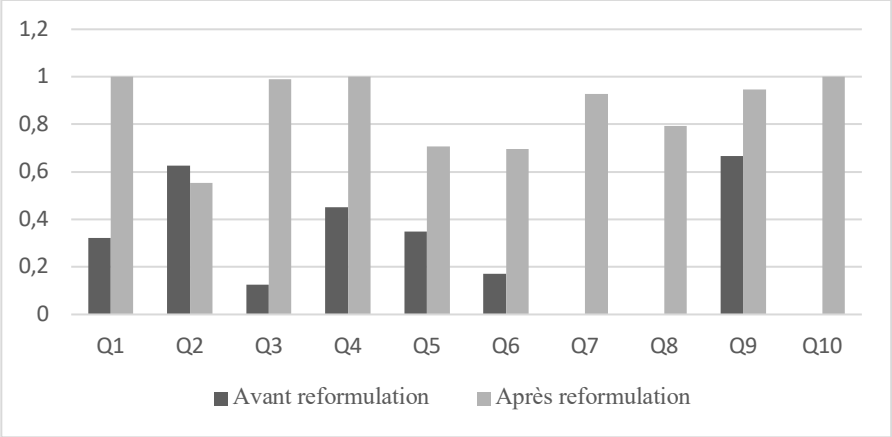

Fig. 4.6 La précision avant et après reformulation des requêtes du tableau 6

D'après les résultats de la figure 4.6, nous remarquons que la précision de neuf requêtes géographiques a été nettement améliorée et a même atteint la valeur maximale de 1 dans le cas des requêtes Q1, Q4 et Q10. Toutefois, la valeur 1 ne signifie pas que les 10 premiers documents retournés sont tous pertinents, mais que les documents pertinents sont au top de la liste (sont les mieux classés). Par contre, la seconde requête montre une légère diminution de la valeur de précision.

Sachant que, toutes les requêtes avaient été reformulées en utilisant des liens corrects, c'est-à-dire que les ESAs qui ont été ajoutées sont réellement adjacentes de l'ESA initiale des requêtes. Nous nous sommes penchés vers l'analyse du nombre d'ESAs utilisées dans la reformulation de chaque requête, dont les résultats sont présentés par la figure 4.7.

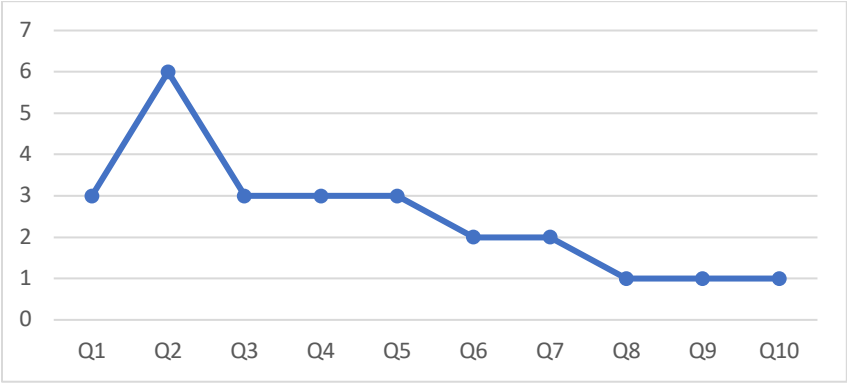

Fig. 4.7 Nombre d'ESAs utilisées pour reformuler chaque requête géographique du tableau 4.6.

A partir de la figure 4.7, nous remarquons que Q2 dont la précision a diminué après reformulation, a été reformulé par six entités spatiales absolues. A cet effet, nous estimons que l'utilisation d'un certain nombre modérément élevé d'ESA pour la reformulation engendre un conflit dans la requête. Ce qui peut induire vers une diminution de la performance du système.

Dans l'objectif de confirmer les résultats du test précédent, nous avons proposé 50 requêtes géographiques. Ces requêtes ont été soumises au service de recherche avec et sans reformulation en utilisant la taxonomie construite par l'approche basée sur le PFP-growth. Nous avons également comparé ces résultats aux mêmes requêtes reformulées avec la méthode d'expansion de requête géographique GQEM proposée par [6]. Cette dernière est une méthode de traitement de langage naturel (NLP : Natural Language Processing) qui reformule et/ou étend les deux composantes, thématique et géospatiale, des requêtes géographiques.

Lors de cette série de tests, nous avons comparé les valeurs de la précision au rang 10 (P@10), la moyenne des précisions moyennes et le temps d'exécution de la nouvelle approche proposée avec les deux cas suivants :

- Requêtes sans reformulation ;
- Requêtes reformulées avec l'approche GQEM.

La performance des résultats retrouvés est présentée en pourcentage afin de montrer l'amélioration de la précision et de l'efficacité de l'approche proposée. Tel que, le pourcentage d'amélioration du temps d'exécution a été mesuré en se basant sur la loi d'Amdahl.[7]

Tableau 4.7 Le taux de performance de la technique présentée appliquée à 4 nœuds.

	Performance de notre approche		
Comparé à	**P@10**	**MAP**	**Temps d'exécution**
Requêtes sans reformulation	10,56%	12,92%	43,24%
GQEM	5,30%	7,68%	68,05%

Sur le tableau 4.7, nous remarquons que l'approche proposée apporte une amélioration intéressante de la précision des requêtes géographiques utilisées lors de nos expérimentations. Cependant, l'amélioration des performances en termes de temps d'exécution est tout à fait normale, ceci est due à la parallélisation du processus.

Afin d'évaluer notre approche sur d'autres lieux géographiques que le Maroc, nous avons mené des tests sur des villes de France en utilisant les mêmes valeurs des paramètres validées lors de nos test précédents. Nous avons considéré la capitale Paris comme point de départ.

Tableau 4.8 Nœuds enfants d'une taxonomie de niveau 1 de la France

Nœuds enfants	**Distance avec Paris**	**Validé ?**
Aubervilliers	6,57 km	Oui
Bobigny	8,85 km	Oui
Les-Mureaux	34,54 km	Oui
Ligné	**319 km**	**Non**

[7] https://www.techno-science.net/definition/303.html

Liré	**3539,32 km**	**Non**
Livry-Gargan	15,13 km	Oui
Noisy-le-Grand	16,38 km	Oui
Pantin	6,47 km	Oui
Saint-Denis	8,22 km	Oui
Saint-Ouen	5,78 km	Oui
Villepinte	18,72 km	Oui

L'application de notre approche de construction de taxonomie d'adjacence, en utilisant 120 documents afin de récupérer des lieux adjacents à Paris, nous a conduit vers les résultats illustrés sur le tableau 4.8. Comme le montre ce dernier, le pourcentage d'exactitude des résultats est significatif. Ceci est dû à la haute précision et la double validation réalisée lors de l'application de la technique proposée.

7. Conclusion

La contribution présentée sur ce chapitre, représente une nouvelle méthode de construction de taxonomies géographiques d'adjacence en utilisant l'algorithme parallèle PFP-Growth, et une technique de reformulation des requêtes géographiques qui contiennent une relation spatiale d'adjacence.

Nous avons mené plusieurs tests sur l'approche proposée, dont le principal test a donné lieu à une taxonomie d'adjacence du Maroc. Lors de nos expérimentations, nous avons varié le support minimum et la relation spatiale utilisée afin de rechercher les paramètres de l'approche qui permettent de générer les règles les plus appropriées. Ceci, nous a permis de construire la taxonomie marocaine en utilisant les meilleures conditions sélectionnées. Tandis que, la technique de reformulation proposée a été testée sur 50 requêtes avec des intentions géographiques et des entités thématiques de différents domaines. Ces requêtes ont été évaluées en termes de la précision au rang 10, la MAP et le temps d'exécution.

Les résultats retrouvés montrent que la reformulation basée sur notre proposition et en utilisant un petit nombre de termes de reformulation a considérablement amélioré la valeur des indicateurs utilisés. Compte tenu des résultats expérimentaux, nous concluons que la méthode présentée est un travail efficace qui permet d'interpréter et d'améliorer les résultats et l'efficacité des requêtes contenant une entité spatiale d'adjacence.

La performance de la méthode proposée est due à l'utilisation d'une approche de fouille de données basée sur une technique de génération de règles d'association. De plus, compte tenu de la structure hiérarchique des données géographiques, leur traitement de manière personnalisée à l'aide de taxonomies a augmenté la précision des résultats. Sans ignorer le point fort de la parallélisation du processus dans l'objectif de minimiser le temps d'exécution de l'approche.

Conclusion générale

Synthèse

Cette thèse vise à prendre compte des spécificités des données géographiques afin de contribuer à l'amélioration des performances des SRIs lors de la soumission des requêtes contenant des connotations spatiales d'adjacence.

Les travaux présentés dans ce mémoire rentrent dans le cadre de la recherche d'information géographique et s'étalent sur trois étapes du processus de recherche : La reconnaissance des toponymes, l'indexation de l'information géographique, et la reformulation de la requête de l'utilisateur. Nos contributions dans ce domaine ont porté sur les deux composantes des requêtes géographique : l'entité spatiale et l'entité thématique. Le but étant de retourner à l'utilisateur des documents plus adaptés à son besoin en prenant en compte la sémantique de l'entité thématique et l'interprétation de l'entité spatiale de la requête originale qu'il a soumis. Cette tâche est réalisée par une taxonomie qui représentent les toponymes extraits du corpus utilisé sous formes d'une structure hiérarchique modélisant la relation d'adjacence entre ses éléments.

Les taxonomies résultant de l'application de nos approches d'indexation sont construites en parcourant une liste des noms de lieux à inclure dessus, par exemple, une liste des villes et villages d'un pays si la taxonomie à construire concerne un pays précis. Chaque élément de la liste est soumis au SRI en combinaison avec une relation d'adjacence afin d'extraire les documents qui forment l'entrée de l'itération permettant de former une branche de la taxonomie dont le nœud source est l'élément courant dans la liste. Ce qui fait que, nos approches sont également utilisables en temps réel sans pré-création d'une taxonomie complète car ils ne nécessitent aucun apprentissage, les ensembles de documents à utiliser par branche sont indépendants. Elles peuvent donc être facilement intégrées dans un moteur de recherche ayant un cache de réponse. Toutefois, la création de la taxonomie complète et son stockage en tant qu'index géographique permettra de l'intégrer dans l'architecture du système de recherche géographique comme base de connaissance.

Nos travaux sont diversifiés, se basant sur deux types de méthodes différentes : L'Indexation Sémantique Latente et la technique non-supervisée de fouille de données Frequent Pattern Growth. Les approches proposées séparent automatiquement les composants d'une requête géographique, reformulent l'entité spatiale à l'aide d'une taxonomie, et étendent la partie thématique de la requête en l'enrichissant en utilisant l'aspect sémantique. Les tests que nous avons réalisés en utilisant plusieurs collections ont prouvé et confirmé que les approches de reformulation de requêtes proposées sont pertinentes et améliorent la performance de la recherche géographique d'une façon significative. L'indexation des toponymes sous formes de taxonomie permet une bonne interprétation des requêtes géographiques ce qui adapte les résultats retournés aux besoins informationnels de l'utilisateur soumettant une requête à intention géographique.

La première contribution, qui se présente par l'approche QRGTW, a donné des résultats satisfaisants, du fait qu'elle prend en compte la structure hiérarchique des données géographiques. Leur traitement de manière personnalisée à l'aide de taxonomies a augmenté la précision des résultats. Ceci est également dû au fait que l'approche reformule les entités thématiques et spatiales des requêtes, et qu'au lieu d'étendre l'entité spatiale avec plus de noms de lieux géographiques et de la surcharger d'informations, ce qui peut conduire à des ambiguïtés. Notre approche remplace l'entité spatiale de la requête par des endroits qui expliquent plus précisément l'intention de la requête.

Les résultats que nous avons eu pendant les tests élaborés sur la première contribution, nous ont également permit de tirer profit des limitations qui ont été remarqué dessus afin d'améliorer plus la précision des requêtes reformulés, minimiser le taux d'erreur des taxonomies construites et optimiser la complexité et le temps d'exécution de l'approche afin de pouvoir l'appliquer sur un nombre de données plus grands. Par conséquent, notre deuxième contribution suit la structure et le processus de la première tout en corrigeant ses faiblesses. Cette dernière, qui se base sur une technique de génération de règles d'association, propose une version optimisée et parallèle de l'algorithme FP-growth qui a augmenté la performance des résultats.

Perspectives

A court terme, notre perspective porte sur l'enrichissement et l'extension des structures créées par la considération des autres types de relations topologiques, ce qui permettra d'évoluer d'une taxonomie géographique vers une ontologie de domaine géographique. Cette structure va améliorer les capacités à interroger un système de recherche par rapport aux travaux existants dans ce domaine. Ce qui nous permettra à moyen terme de concevoir un système de recherche d'information géographique répandant à un besoin commercial spécifique comme le domaine de l'immobilier au Maroc par exemple. Ce SRIG aura une architecture utilisant la nouvelle structure d'indexation pour guider l'utilisateur dans sa recherche, et offrir des résultats plus pertinents répondant au besoin en information géographique de l'utilisateur selon son champ d'intérêt, dans le cas où la partie thématique de sa requête porte sur l'immobilier.

A long terme, ces travaux ouvrent également la voie vers diverses perspectives dans le domaine de la recherche d'information géographique. Nous citons :

1. Données spatio-temporelles :
   L'exploitation et la valorisation des grandes masses de données spatio-temporelles disponibles dans différents domaines est devenu crucial. Ceci nécessite l'utilisation d'approches innovantes aptes à traiter conjointement les aspects spatiaux et temporels, ce qui n'est pas le cas de la plupart des méthodes existantes. Ces méthodes de fouille, d'analyse et d'extraction de connaissances n'exploitent, en général, qu'une seule dimension de l'information, spatiale ou temporelle. Cela induit le plus souvent à une perte de précision et possibilité

d’interprétation des résultats. C’est pourquoi nous nous proposons d’exploiter un modèle de graphe spécifique intégrant différents types de relations, les graphes spatio-temporels. Par contre les graphes sont des outils puissants qui peuvent être utilisés pour modéliser des phénomènes temporels ou spatiaux.

2. Deep Learning :
   Depuis plusieurs années, nous assistons à une utilisation croissante des réseaux de neurones profonds dans la génération de représentation intégrant des concepts de haut niveau. Depuis, ces modèles suscitent un réel intérêt dans la communauté scientifique travaillant sur des graphes. Ces algorithmes qui étaient généralement utilisés dans des technologies comme la reconnaissance d'images ou la vision robotique, ont été également proposés pour résoudre des tâches de RI géographique comme la reconnaissance et résolution de toponymes.
   L’ouverture vers les techniques de Deep Learning pour la construction de graphe hiérarchique d’indexation des données géographique permettra une amélioration en termes de précision et afin d’optimiser en temps de traitement et espace de stockage, ces techniques peuvent être utilisées dans un contexte Big Data.

Bibliographie :

[1] C. C. Jordan and C. Watters, 2004. Extending the Rocchio Relevance Feedback Algorithm to Provide Contextual Retrieval. in Advances in Web Intelligence, Second International Atlantic Web Intelligence Conference, AWIC 2004. pages 135-144.

[2] R.N.P. Vargas, M.F. Moura, E.A. Speranza, E. Rodriguez and S.O. Rezende, 2012. Discovering the Spatial coverage of the documents through the SpatialCIM Methodology. In AGILE'2012 : International Conference on Geographic Information Science. Avignon. April 24-27, 181-186.

[3] O. E. Midaoui and A. E. Qadi and M. D. Rahmani and D. Aboutajdine (2015), "A new approach to build a geographical taxonomy of adjacency automatically using the latent semantic indexing method", in Intelligent Systems and Computer Vision (ISCV). 1-6. DOI=10.1109/ISACV.2015.7105551

[4] Pass, G., Chowdhury, A. and Torgeson, C. (2006) "A picture of search", in Proceedings of the 1st International Conference on Scalable Information Systems (InfoScale '06), Hong Kong ACM, New York, NY, USA, Article No. 1.

[5] Jones, R., Zhang, W.V., Rey, B., Jhala, P. and Stipp. E. (2008), "Geographic intention and modification in web search", International Journal of Geographical Information Science, Vol. 22 No. 3, pp. 229-246.

[6] Perea-Ortega, J.M., García-Cumbreras, M.A. and Ureña-López, L.A. (2012), "Applying NLP techniques for query reformulation to information retrieval with geographical references", in Proceedings of the 2012 Pacific-Asia conference on Emerging Trends in Knowledge Discovery and Data Mining (PAKDD'12), Springer-Verlag, Berlin, Heidelberg, pp. 57-69.

[7] H. Tebri. "Formalisation et spécification d'un système de filtrage incrémental d'information", thèse de doctorat en informatique, Université Paul Sabatier, Toulouse, 2004.

[8] Hernandez N. "Ontologie de domaine pour la modélisation du contexte en recherche d'information", thèse de doctorat en informatique, Université Paul Sabatier, 2006.

[9] Boubekeur F. "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets", thèse de doctorat en informatique, Université Paul Sabatier, 2008.

[10] Daoud M. "Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche", thèse de doctorat en informatique, Université Paul Sabatier, 2009.

[11] Bouramoul A. " Recherche d'information contextuelle et sémantique sur le web", thèse de doctorat en informatique, Université MENTOURI de Constantine, 2011.

[12] LATOUR M. " Du besoin d'informations à la formulation des requêtes : étude des usages de différents types d'utilisateurs visant l'amélioration d'un système de recherche d'informations", thèse de doctorat en informatique, Université de Grenoble, 2014.

[13] Belkin N. J., Oddy R., and Brooks H. " Ask for information retrieval: Part I background and theory", Journal of Documentation, 38(2), pp. 61–71, 1982.

[14] Sauvagnat K. " Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés ", thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2005.

[15] HLAOUA Lobna, " Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurés ", Thèse de l'Institut De Recherche En Informatique, Université Paul Sabatier de Toulouse, 14 Décembre 2007.

[16] Aurélien Saint Requier, Gérard Dupont, Sébastien Adam, Yves Lecourtier, " Évaluation d'outils de reformulation interactive de requêtes ", Conférence en Recherche d'Information et Applications, Mar 2010, Tunisie. pp. 223-238, 2011.

[17] Hang Cui, Ji-Rong Wen, Jian-Yun Nie and Wei-Ying Ma, " Probabilistic Query Expansion Using Query Logs ", WWW2002, Honolulu, Hawaii, USA, May 7-11, 2002.

[18] GOMEZ CARPIO Guillermo Valente, " Enrichissement de requêtes et visualisation sémantique dans une coopération de systèmes d'information : méthodes et outils d'aide à la recherche d'information ", thèse de doctorat en informatique, Université de Bourgogne, 2010.

[19] Mustapha BAZIZ, " Indexation conceptuelle guidée par ontologie pour la recherche d'information", Thèse d'informatique de l'Institut De Recherche En Informatique, l'Université Paul Sabatier de Toulouse, 14 décembre 2005.

[20] Dégez Danièle et Ménillet Dominique, " Thésauroglossaire des langages documentaires ", Bulletin des bibliothèques de France (BBF), n° 1, pp. 123-123, 2002.

[21] El Ghali Btihal et El Qadi Abderrahim, " Context-aware query expansion method using Language Models and Latent Semantic Analyses ", Knowledge and Information Systems, 50 (3), pp. 751-762, 2017.

[22] Hazra Imran and Aditi Sharan, "*Thesaurus and Query Expansion*", International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009.

[23] Crouch C. J. et Yang B., "Experiments in automatic statistical thesaurus construction", In Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhage, Denmark, pp. 77–88, 1992.

[24] Attar R. and Fraenkel A., "Local feedback in full-text retrieval systems", Journal of the ACM, 24(3): pages 397–417, 1977.

[25] RESSAD-BOUIDGHAGHEN Ourdia, "Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes", Thèse de Doctorat, l'Université Paul Sabatier de Toulouse, 12 décembre 2011.

[26] Porter M. F., "An algorithm for suffix stripping", Program 14, 1980.

[27] Adamson G. and Boreham J., "The use of an association measure based on character structure to identify semantically related pairs of words and document titles", Information Storage and Retrieval, 10(1), pp. 253–60, 1974.

[28] Frakes W. B. and Baeza-Yates R., "Stemming Algorithms", Frakes W B, Baeza-Yates R (eds) Prentice Hall, New jersey, pp. 131–160, 1992.

[29] C. J. Crouch, D. B. Crouch, Q. Chen, and S. J. Holz, "Improving the retrieval effectiveness of very short queries", Information Processing and Management, 38(1), pp. 1–36, 2002.

[30] Vallet D., Fernández M. and Castells P., "An Ontology-Based Information Retrieval Model", In Proceedings of the 2nd European Semantic Web Conference, pp. 455-470, 2005.

[31] Baziz M., Boughanem M., Aussenac-Gilles N. and Chrisment C., "Semantic Cores for Representing Documents in IR", In Proceedings of the 20th ACM Symposium on Applied Computing, ACM Press ISBN: 1-58113-964-0, pp. 1020-1026, 2005.

[32] Singhal A., Salton G., Mitra M. and Buckley C., "Document length normalization. Information Processing and Management", 32(5), pp. 619–633, 1996.

[33] Zipf G., "Human Behaviour and the Principle of Least Effort", Addison-Wesley, 1949.

[34] Aas K. and Eikvil L., "Text Categorisation: A survey", Technical Report, Norwegian Computing Center, Norway, June 1999.

[35] Slimani, T., Yaghlane B.B. and Mellouli, K., "Une extension de mesure de similarité entre les concepts d'une ontologie", in Proceedings of the 4rth International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, (SETIT), TUNISIA, pp. 25-29, 2007.

[36] Deerwester S., Dumais S., Landauer T., Furnas G. and Harshman R., "Indexing by latent semantic analysis", Journal of the American Society for Information Science, 41(6), pp. 391-407, 1990.

[37] Manning, C.D., Raghavan, P. and Schütze, H., "Matrix decompositions and latent semantic indexing", in an Introduction to Information Retrieval, Online edition (c) Cambridge University Press., pp. 403-419, 2009.

[38] Baeza-Yates R. and Ribeiro-Neto R. A., "Modern Information Retrieval", New York: ACM Press; Harlow England: Addison-Wesley, cop., 1999.

[39] Kuropka D., editor. "Modelle zur Repräsentation natürlichsprachlicher Dokumente- Information-Filtering und -Retrieval mit relationalen Datenbanken", volume 10 of Advances in Information Systems and Management Science. Logos Verlag, Berlin, 2004.

[40] Zadeh L., "Fuzzy sets", Information and control, 8, pp. 338–353, 1965.

[41] Paice C. P., "Soft evaluation of boolean search queries in information retrieval systems", Information Technology: Research and Development, 3(1), pp. 33–42, 1984.

[42] Salton, G., "The SMART Retrieval System - Experiment in Automatic Document Processing", Englewood Cliffs, NJ : Prentice-Hall, 1971.

[43] Wong, S., Ziarko, W., and Wong, P., "Generalized vector space model information retrieval", In Proceedings of the 8th annual international ACM SIGIR Conference on Research and development in Information Retrieval, pp. 18–25, ACM, 1985.

[44] Croft, W.B., Belkin, N., Bruandet, M.-F., Kuhlen, R., and Oren, T., "Hypertext and information retrieval: what are the fundamental concepts?", In Hypertext: concepts, systems and applications, N. Streitz, A. Rizk, and J. André (Eds.). Cambridge University Press, New York, NY, USA, pp. **36**2-366, 1992.

[45] Rosario B., "Latent Semantic Indexing: An overview", INFOSYS 240, Final Paper; Spring, 2000.

[46] Tmar M., "Modèle auto-adaptatif de filtrage d'information : apprentissage incrémental du profil et de la fonction de décision", Thèse de Doctorat, Toulouse : Université Paul Sabatier, 2002.

[47] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harsh- man, R., Indexing by latex semantic analysis, Journal of the American Society for Information Science, 41(6), pp. 391–407, 1990.

[48] M. Maron and J. Kuhns, "On relevance, probabilistic indexing and information retrieval", Journal of the Association for Computing Machinery, 7, pp. 216–244, 1960.

[49] Robertson Stephen E., Walker Steve, Jones Susan, Hancock-Beaulieu Micheline, and Gatford Mike, "Okapi at trec-3", In Text Retrieval Conference (TREC-3), pp. 109-126, 1996.

[50] SOULIER Laure, "Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs", Thèse de Doctorat, l'Universit. Toulouse 3 Paul Sabatier, 01 décembre 2014.

[51] J. Ponte and W. Croft, "A language modeling approach to information retrieval", In Proceedings of the 21st ACM conference on research and development in information retrieval (SIGIR 98), 1998.

[52] M. Boughanem, W. Kraaij, and J.-Y. Nie, "Modèles de langue pour la recherche d'information", In Les systèmes de recherche d'informations, pp. 163–182. Hermes-Lavoisier, 2004.

[53] Jelinek, F. and Mercer, R. L., "Interpolated estimation of Markov source parameters from sparse data", In Proceedings of the Workshop on Pattern Recognition in Practice, pp. 381–397, 1980.

[54] MacKay, D. J. C. and Peto, L. C. B., "A Hierarchical Dirichlet Language Model", Natural Language Engineering, 1(3), pp. 1–19, 1994.

[55] Chen Stanley F. et Goodman Joshua, "An Empirical Study of Smoothing Techniques for Language Modeling", Technical Report TR-10-98, Computer Science Group. Harvard University, Cambridge, Massachusetts, pp. 310-318, 1998.

[56] Simonnot B., "La pertinence en sciences de l'information : des modèles, une théorie?", pp. 161-182, ch 6 in Papy F. (dir.), Problématiques émergentes dans les Sciences de l'Information, Paris : Hermes-Lavoisier, 2008.

[57] Zemirli, W.N., "Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif ", Thése de doctorat en informatique, Université Paul Sabatier de Toulouse III, 2008.

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

[59] Saracevic, T., "Relevance: a review of the literature and a framework for thinking on the notion in Information Science. Part II:Nature and manifestations of relevance", Journal of the American Society for Information Science and Technology, vol.58 (13), p. 1915-1933, 2007.

[60] Saracevic, T., "Relevance: a review of the literature and a framework for thinking on the notion in Information Science. Part III: Behavior and Effects of Relevance", Journal of the American Society for Information Science and Technology, vol.58 (13), p. 2126-2144, 2007.

[61] Cleverdon, C., "Readings in Information Retrieval", chapter The Cranfi, pages 47–59. Morgan Kaufmann Publishers Inc., 1997.

[62] Kelly, D., Dumais, S., and Pedersen, J. O., "Evaluation Challenges and Directions for Information-Seeking Support Systems", IEEE Computer, Vol. 42(3), p. 60–66, 2009.

[63] Sieg, A., Mobasher, B., and Burke, R., "Web Search Personalization with Ontological User Profiles", In Proceedings of the Conference on Information and Knowledge Management, CIKM'07, p. 525–534, ACM, 2007.

[64] Anick, P., "Using Terminological Feedback for Web Search Refinement: A Log-based Study", In Proceedings of the Annual International SIGIR Conference on Research and Development in Information Retrieval, SIGIR'03, p. 88–95, ACM, 2003.

[65] Ringel, M., Cutrell, E., Dumais, S. T., and Horvitz, E., "Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores", In Proceedings of the International Conference on Human-Computer Interaction, INTERACT'03, pages 184–191. IOS Press, 2003.

[66] Sanderson, M., Kohler, J., "Analyzing geographic queries," Proceedings of SIGIR the Workshop on Geographic Information Retrieval, Sheffield UK, pp. 8-10, 2004.

[67] M. Al-Maolegi, B. Arkok, "An improved apriori algorithm for association rules," International Journal on Natural Language Computing, vol. 3, issue 1, pp. 21-29, February 21-29, 2014.

[68] H. M. Najadat, M. Al-Maolegi, B. Arkok, “An improved Apriori algorithm for association rules,” International Research Journal of Computer Science and Application, vol. 1, issue 1, pp. 01-08, 2013.

[69] Liwu, ZOU, Guangwei, REN, “The data mining algorithm analysis for personalized service,” Fourth International Conference on Multimedia Information Networking and Security, 2012.

[70] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Journal of Commun. ACM, pp. 107–113, 2008.

[71] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster computing with working sets,” in HotCloud 2010, (Boston, USA), 2010.

[72] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang, “Pfp: parallel fp-growth for query recommendation,” in Proceedings of the 2nd International Conference on Recommender Systems, (Lausanne, Switzerland), pp. 107–114, 2008.

[73] Y. Miao, J. Lin and N. Xu, "An improved parallel FP-growth algorithm based on Spark and its application," 2019 Chinese Control Conference (CCC), Guangzhou, China, pp. 3793-379, 2019.

[74] M. Zitouni, "Parallel Itemset Mining in Massively Distributed Environments", Information Theory [cs.IT], Université de Tunis El Manar, Inria, 2018.

[75] Buscaldi, D. and Rosso, P. (2009), “Using GeoWordNet for Geographical Information Retrieval”, in Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access, Aarhus, Denmark, Springer-Verlag, Berlin, Heidelberg, pp. 863-866.

[76] L. Deng, Y. Lou, “Improvement and research of FP-growth algorithm based on distributed spark,” Proceeding of International Conference on Cloud Computing and Big Data (CCBD’2015), Shanghai, pp. 105-108, 4-6 November, 2015.

[77] Aloteibi, S. and Sanderson, M. (2014), “Analyzing Geographic Query Reformulation: An Exploratory Study”, Journal of the Association or Information Science and Technology, Vol.65 No. 1, pp. 13-24.

[78] Gravano, L., Hatzivassiloglou, V. and Lichtenstein, R. (2003), “Categorizing web queries according to geographical locality”, in Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA, ACM, New York, pp. 325–333.

[79] Gan, Q., Attenberg, J., Markowetz, A. and Suel, T. (2008), “Analysis of geographic queries in a search engine log”, in Proceedings of the first international workshop on Location and the web (LOCWEB '08), Beijing, China, ACM, New York, NY, USA, pp. 49-56.

[80] Zheng, Y., Zhang, L., Ma, Z., Xie, X. andMan W.-Y. (2011), Recommending Friends and Locations Based on Individual Location History, ACM Transactions on the Web, Vol. 5 No. 1, pp. 1-44.

[81] Jones, C. B. and Purves, R. S. (2008), “Geographical information retrieval”, International Journal of Geographical Information Science, Vol. 22, No. 3, pp. 219-228.

[82] Stokes, N., Li, Y., Moffat, A. and Rong, J. (2008), “An empirical study of the effects of NLP components on Geographic IR performance”, International Journal of Geographical Information Science, Vol. 22, No. 3, pp. 247-264.

[83] Gey, F., Larson, R. and et al., 2006, GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Peters.

[84] Finkel, J. R., Grenager, T., and Manning, M. (2005), “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, in Proceedings ACL’05, ACL, pp. 363–370.

[85] Ratinov, L. and Roth, D. (2009), “Design Challenges and Misconceptions in Named Entity Recognition”, in CONLL’09, ACL, pp. 147–155.

[86] Karimzadeh, M., Pezanowski, S., MacEachren, A. M., &Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. Transactions in GIS, Vol. 23, No. 1, pp. 118-136.

[87] Yue, L., Mengjun, K., Yuyang, W., Qingyun, D., Tao, L. (2019), "A deep learning architecture for semantic address matching", International Journal of Geographical Information Science, Vol. 34, No. 3, pp.559-576.

[88] Li, L., Wang, W., He, B., Zhang, Yu. (2018), "A hybrid method for Chinese address segmentation", International Journal of Geographical Information Science, Vol. 32, No. 1, pp. 30-48.

[89] Sallaberry, M. Baziz, J. Lesbegueries, et Gaio, M. (2007), "Une approche d'extraction et de recherche d'information spatiale dans les documents textuels," Proceeding de la « Conférence en Recherche d'Informations et Applications ». CORIA 2007, 4th French Information Retrieval Conference, Saint-Étienne, France, March 28-30.

[90] Perea-Ortega, J. M., Garcia-Cumbreras, M. A., et Urena-Lopez, L. A. (2012), "Evaluating different query reformulation techniques for the Geographical Information Retrieval task considering geospatial entities as textual terms," In «The Pacific-Asia Conference on Knowledge Discovery and Data Mining» (PAKDD). Kuala Lumpur, Malaysia.

[91] Ounis, I., Amati, G., Plachouras, V. He, B., Macdonald, C., Johnson, D. (2005), "Terrier Information Retrieval Platform", Lecture Notes in Computer Science, 3408, pp. 517-519.

[92] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C. (2006), "Terrier: A High Performance and Scalable Information Retrieval Platform", In Proceedings of ACM SIGIR'06 Workshop on Open-Source Information Retrieval (OSIR 2006). 10th August, Seattle, Washington, USA.

[93] Amati G. (2003), "Probabilistic Models for Information Retrieval based on Divergence from Randomness", Thèse de doctorat, Department of Computing Science, University of Glasgow.

[94] Plachouras, V., Ounis, I. (2004), "Usefulness of hyperlink structure for query-biased topic distillation", In SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in information retrieval, pp. 48–455 Sheffield, UK.

[95] Overell, S., Ruger, S. (2008); "Using co-occurrence models for placename disambiguation", International Journal of Geographical Information Science, Vol. 22, No. 3, pp. 265–287.

[96] Bensalem, I., Kholladi, M.-K. (2010), "Toponym Disambiguation by Arborescent Relationships", Journal of Computer Science, Vol. 6, No. 6, pp. 653-659.

[97] Kirk E. Roberts, Cosmin A. Bejan, and Sanda M. Harabagiu. (2010), "Toponym disambiguation using events", In FLAIRS Conference.

[98] Tobin, R., Grover, C., Byrne, K., Reid, J., Walsh, J. (2010), "Evaluation of georeferencing", In Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10, pages 7:1–7:8.

[99] Adams, B., McKenzie, G. (2013), "Inferring Thematic Places from Spatially Referenced Natural Language Descriptions", Crowdsourcing Geographic Knowledge, Springer Netherlands, pp. 201-221.

[100] Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., Mckenzie, G. (2016), "Things and strings: Improving place name disambiguation from short texts by combining entity cooccurrence with topic modelling", In 20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024, EKAW 2016, pp. 353–367. Springer-Verlag New York Inc.

[101] Leidner, J. L. (2007), "Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names", Thèse de Doctorat, Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh.

[102] Hoffart, J., Weikum, G. (2013), "Discovering and disambiguating named entities in text", In Proceedings of the 2013 SIGMOD/PODS Ph.D. Symposium, SIGMOD'13 PhD Symposium, pp. 43–48. ACM.

[103] Spitz, A., Geiß, J., Gertz, M. (2016), "So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks", In Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, GeoRich '16, ACM.

[104] Zhang, W., Gelernter, J. (2014), "Geocoding location expressions in twitter messages: A preference learning method", Journal of Spatial Information Science, Vol. 9, pp. 37–70.

[105] Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M., and Gonzalez, G. (2015). "Knowledge-driven geospatial location resolution for phylogeographic models of virus migration", Bioinformatics, Vol. 31, No. 12, pp. i348–i356.

[106] Santos, J. Anastacio, I., Martins, B. (2015), "Using machine learning methods for disambiguating place references in textual documents", GeoJournal, Vol. 80, No. 3, pp. 375–392.

[107] Kamalloo, E., Rafiei, D. (2018), "A coherent unsupervised model for toponym resolution", In Proceedings of the 2018 World Wide Web Conference, WWW '18, pages 1287–1296.

[108] Silva, M., Martins, B., Chaves, M., Cardoso, N. Afonso, A. P. (2006), "Adding geographic scopes to web resources", Computers, Environment and Urban Systems, Vol. 30, pp. 378–399.

[109] Frontiera, P., Larson, R., Radke, J. (2008), "A comparison of geometric approaches to assessing spatial similarity for GIR", International Journal of Geographical Information Science, Vol. 22.

[110] Luaces, M.R., Places, A.S., Rodriguez, F.J., Seco, D. (2008), "Retrieving Documents with Geographic References Using a Spatial Index Structure Based on Ontologies," In Proceedings of the ER 2008 Workshops (CMLSA, ECDM, FP-UML, M2AS, RIGiM, SeCoGIS, WISM) on Advances in Conceptual Modeling: Challenges and Opportunities.

[111] World Wide Consortium: Owl web ontology language reference. Retrieved March 2008 from http://www.w3.org/TR/owl-ref/

[112] Alias-i: LingPipe, Natural Language Tool. Retrieved March 2008 from http://www.alias-i.com/lingpipe/

[113] Apache: Lucene. Retrieved March 2008 from http://lucene.apache.org.

[114] Wu, W., Li, H., Wang, H. and Zhu, K.Q. (2012), "Probase: A probabilistic taxonomy for text understanding", in Proceedings of ACM SIGMOD International Conference on Management of Data, Scottsdale, Arizona, USA, ACM, New York, pp. 481-492.

[115] Yangqiu, S.,Shixia, L.,Xueqing L. and Haixun, W. (2015),"Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees", IEEE Transactions on Knowledge and Data Engineering, Vol. 27 No. 7, pp. 1861-1874.

[116] Sadikov, E., Madhavan, J., Wang, L. and Halevy, A.Y. (2010), "Clustering query refinements by user intent".in Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, ACM, New York, pp. 841–850.

[117] Fang, C., Zhang, S. (2018), "Geographic Information Retrieval Method for Geography Mark-Up Language Data", ISPRS International Journal of Geo-Information, Vol 7, Issue 3.

[118] Kokla, M., Guilbert, E. (2020), "A Review of Geospatial Semantic Information Modeling and Elicitation Approaches", ISPRS International Journal of Geo-Information, Vol. 9, March, page 149.

[119] JONES, C.B., PURVES, R.S., CLOUGH, P.D AND JOHO, H., (2008), "Modelling Vague Places with Knowledge from the Web", International Journal of Geographical Information Science, Vol. 22, pp. 1045 – 1065.

[120] Kohler, J. (2003), "Analysing search engine queries for the use of geographic terms", Thèse de master, Université de Sheffield - United King.

[121] Fu, G., Jones, C.B., Abdelmoty, A.I. (2005), “Ontology-based spatial query expansion in information retrieval”, In: Lecture Notes in Computer Science, Springer, Vol. 3761, pp. 1466–1482.

[122] Cardoso, N. and Silva, M. J. (2007), “Query expansion through geographical feature types”, in Proceedings of the 4th ACM workshop on Geographical information retrieval (GIR '07), Lisbon, Portugal, ACM, New York, NY, USA, pp. 55-60.

[123] Leveling, J. (2007), “Exploring term selection for geographic blind feedback”, in Proceedings of the 4th ACM workshop on Geographical information retrieval (GIR '07), Lisbon, Portugal, ACM, New York, NY, USA, pp. 43-48.

[124] Giunchiglia, F., Maltese, V. Farazi, F., Dutta, B. (2010), "GeoWordNet: A Resource for Geo-spatial Applications", in proceedings of the 7th Extended Semantic Web Conference (ESWC) 2010.

Liste des publications et communications

▪ Publications:

- Omar El Midaoui, Btihal El Ghali, Abderrahim El Qadi, Moulay Driss Rahmani. **« Geographical query reformulation using a geographical adjacency taxonomy builder and word senses ».** Journal of Systems and Information Technology, Volume 23, No 1, 2021. **(indexé Scopus)**

- Omar El Midaoui, Btihal El Ghali, Abderrahim El Qadi, Moulay Driss Rahmani. **« Geographical queries reformulation using a parallel association rules generator to build spatial taxonomies »**. International Journal of Electrical and Computer Engineering, Volume 11, No 3, 2021, pp. 2586–2595. **(indexé Scopus)**

- Btihal El Ghali, Abderrahim El Qadi, Omar El Midaoui, Mohamed Ouadou. **« Query Recommendation based terms and relevant documents using language Models »**. the WSEAS Transactions on Information Science and Applications. Volume 12, 2015, pp. 112-119.

- Btihal El Ghali, Abderrahim El Qadi, Omar El Midaoui, Mohamed Ouadou, Driss Aboutajdine. **« Probabilistic Query Expansion Method based on a Query Recommendation Algorithm »**. International Journal of Web Applications (IJWA), March 2013. **(indexé dblp)**

▪ Communications:

- Omar El Midaoui, Btihal El Ghali, Abderrahim El Qadi**. « Geographical queries reformulation using parallel fp-growth for spatial taxonomies building ».** IoTBDS 2020, Proceedings of the 5th International Conference on Internet of Things, Big Data and Security, 2020, pp. 375–381. **(Indexé Scopus et dblp)**

- Omar El Midaoui, Btihal El Ghali, Abderrahim El Qadi, Moulay Driss Rahmani. **« Geographical Query reformulation using a Geographical Taxonomy and WordNet ».** Procedia Computer Science, Volume 127, 2018, pp. 489–498. **(indexé Scopus et ACM)**

- Omar El Midaoui, Abderrahim El Qadi, Moulay Driss Rahmani, Driss Aboutajdine. **« A New Approach to build a Geographical Taxonomy of Adjacency Automatically Using the Latent Semantic Indexing Method »**. The International Conference on Intelligent Systems and Computer Vision (ISCV'15), 25-26 Mars 2015 à Fès. **(indexé Scopus)**

- Omar El Midaoui, Abderrahim El Qadi, Moulay Driss Rahmani, Driss Aboutajdine. **« Geographical Query Reformulation Based on Spatial Taxonomies Constructed Using the Apriori Algorithm »**. Poster à The international Conference on NETworked sYStems (NETYS'15), 13-15 Mai 2015 à Agadir.

- Omar El Midaoui, Abderrahim El Qadi, Moulay Driss Rahmani and Driss Aboutajdine. **«A query reformulation technique in the Geographical Information Retrieval field »**. International Workshop on Innovation and New Trends in Information Systems (INTIS'2013), 29-30 Novembre 2013 à Tanger.

- Omar El Midaoui, Abderrahim El Qadi, Moulay Driss Rahmani and Driss Aboutajdine. **« Extraction et Recherche d'Information géo-spatiale dans les documents Web »**. Les 5èmes Journées Doctorales en Technologies de l'Information et de la Communication (JDTIC'13), 29-30 Novembre 2013 à Kénitra.

- Omar El Midaoui, Abderrahim El Qadi. **« Extraction et Recherche d'Information géo-spatiale dans les documents Web »**. 2ème Journées URAC'13 du Laboratoire de Recherche en Informatique et Télécommunication (LRIT), 31 Mai 2013 à Rabat.