République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur
et de la **R**echerche **S**cientifique

Ecole **N**ationale **P**olytechnique

Département de Génie Electronique

PROJET DE FIN D'ETUDE EN VUE DE L'OBTENTION DU DIPLOME D'INGENIEUR D'ETAT EN GENIE ELECTRONIQUE

Présenté par: Mr KRIZOU Hocine

Thème

Les techniques SVD et traitement d'images dans la recherche intelligente d'informations en indexation sémantique latente

Proposé et Encadré par **:** Mr. LARBES Cherif et Mr. ALLALI Ali

Soutenu le : 12/10/2011, devant le jury composé de :

Président:	Mr. HADDADI Mourad
Examinateur:	Mr. AIT CHEIKH M Salah
Promoteurs:	Mr. LARBES Cherif
	Mr. ALLALI Ali

Promotion 2011

REMERCIEMENTS

Je tiens à exprimer ma reconnaissance à Mr. LARBES pour son encadrement et Mr. ALLALI, pour avoir Co-encadré mes travaux. Je les remercie pour leur disponibilité, leur écoute et leurs conseils, qui m'ont été toujours précieux, leur confiance, leur investissement scientifique et humain qui ont été essentiels à la réalisation de ce travail.

Je remercie Mr. HADDADI Mourad et Mr.AIT CHEIKH M Salah, je leur en suis reconnaissant d'avoir accepté de faire partie du jury de mon projet de fin d'étude.

Je souhaite aussi remercier tous les enseignants de l'Ecole Nationale Polytechnique, pour les connaissances qu'ils m'ont transmis, leur disponibilité et leurs efforts.

J'ai sûrement oublié de remercier beaucoup d'autres personnes méritantes, des personnes qui m'ont offert leur amitié, qui m'ont ouvert leur cœur, qui m'ont ouvert leur porte, qu'elles trouvent ici l'expression de ma profonde gratitude et de mon amitié la plus sincère.

DÉDICACES

Je dédie ce modeste travail à mon père, ma mère, ainsi qu'à mon frère et ma sœur qui m'ont tant soutenu et aidé tout au long de mon parcours et sans qui je n'en serais pas là.

A mes amis qui m'ont tant supporté, spécialement, Khaldoun et Fellah.

A tous mes camarades et amis de notre chère école avec qui j'ai passé de bons moments et appris beaucoup de choses.

A tous ceux qui ont contribué de près ou de loin à mon travail.

A tous ceux que je n'ai pas cités et qui sont présents dans mes pensées.

RESUME

L’objet des systèmes de recherche d’informations est de faciliter l’accès à un ensemble de documents, afin de permettre à l’utilisateur de retrouver ceux qui sont pertinents, c'est-à-dire ceux dont le contenu correspond le mieux à son besoin en information. La qualité des résultats de la recherche se mesure en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus le système est jugé performant.

Les premiers systèmes permettaient d’effectuer des recherches booléennes, c’est à dire, des recherches ou seule la présence ou l’absence d’un terme de la requête dans un texte permet de le sélectionner. Il a fallu attendre la fin des années 60, pour que l’on applique le modèle vectoriel aux problématiques de la recherche d’information. Dans ces deux modèles, seule la présence, l’absence, ou la fréquence des mots dans le texte est porteuse d’information.

D’autres systèmes de recherche d’information adoptent cette approche dans la modélisation des données textuelles et dans le calcul de la similarité entre documents ou par rapport à une requête. Plusieurs améliorations des systèmes de recherche d’information utilisent les relations sémantiques qui existent entre les termes dans un document. LSI (Latent Semantic Indexing), par exemple réalise ceci à travers des méthodes d’analyse qui mesurent la cooccurrence entre deux termes dans un même contexte pour créer des liens sémantiques entre les termes dans un processus de chaines lexicales.

Dans ce travail, nous étudierons la technique de LSI, du prétraitement de la base de données jusqu'à l’application d’algorithme de décomposition et la conception d’un moteur de recherche basé sur cette technique. La contribution clé du travail présenté dans ce projet est le développement d'une approche hybride et efficace de LSI pour une utilisation plus performante dans la recherche d’information, basé sur l'utilisation de techniques de traitement d'image en tandem avec les composants existants.

ABSTRACT

The object of information retrieval systems is to make easy the access to documents and to allow a user to find those that are appropriate. The quality of the results of research is measured by comparing the answers of the system with the ideal answers that the user hopes to find. The system is competitive when its answers correspond to those that the user hopes.

The first retrieval systems performing Boolean researches, in other words, researches in which only the presence or the absence of a term of a request in a text allow choosing it. It was necessary to wait for the end of the sixties to apply the vector model in information retrieval. In these two models, alone presence, absence, or frequency of words in the text is holder of information.

Several Information Retrieval Systems adopt a flat approach in the modeling of data and in the counting of similarity between documents or in comparison with a request. Several improvements in information retrieval systems use the semantic relationships which exist between terms in a document. LSI (Latent Semantic Indexing), for example achieves this through analytical methods that measure co-occurrence between two terms in the same context to create semantic links between terms in a process of lexical chains.

In this work, we study the technique of LSI, the pretreatment of the database to the application of decomposition algorithm and design of a search engine based on this technique. A key contribution of the work presented in this project is the development of a hybrid and efficient approach to LSI for effective use in IR, based on the use of image processing techniques in tandem with the existing components.

ملخص

الغرض من نظم استرجاع المعلومات هو تسهيل الوصول إلى مجموعة من الوثائق التي تمكن المستخدم من العثور على تلك ذات الصلة. ويتم قياس جودة نتائج البحوث بمقارنة ردود النظام مع الأجوبة المثالية التي المستخدم يتوقع الحصول عليها. النظام ذات استجابات أكثر تتطابق مع تلك التي يقوم المستخدم بتوقعها، يعتبر نظام أكثر كفاءة.

أول الأنظمة سمحت بتنفيذ عمليات البحث المنطقية، أي البحوث أين فقط وجود أو عدم وجود مصطلح الاستعلام في نص يمكن تحديده .إلا حتى وقت متأخر من 60، حيث تم تطبيق نموذج متجه لمشاكل استرجاع المعلومات. في كلا النموذجين، إلا وجود، غياب، أو تكرار الكلمات في النص يحمل المعلومات.

غيرها من نظم استرجاع المعلومات تعتمد عن هذا المنهج في نمذجة النصوص وحساب التشابه بين المستندات أو مع استعلام .العديد من التحسينات في نظم استرجاع المعلومات تستخدم العلاقات الدلالية بين المصطلحات الواردة في الوثيقة .LSI على سبيل المثال يحقق ذلك من خلال الأساليب التحليلية التي تقيس المشترك بين مصطلحين في السياق نفسه لخلق روابط بين المصطلحات في عملية السلاسل المعجمية.

في هذا العمل، ندرس تقنية LSI، من معالجة قاعدة البيانات إلى تطبيق خوارزمية التحلل وتصميم محرك بحث على أساس هذه التقنية. مساهمة رئيسية من الأعمال التي عرضت في هذا المشروع هي تطوير نهج الهجين ولاستخدامها في استرجاع المعلومات الأكثر كفاءة، على أساس استخدام تقنيات معالجة الصور جنبا إلى جنب مع المكونات القائمة.

SOMMAIRE

SOMMAIRE

SOMMAIRE

Liste des tableaux

Liste des figures

Introduction générale

Pour qu'un savoir puisse se transmettre, il faut d'abord pouvoir le reproduire et le stocker. Dans ce contexte, l'Humanité a fait des pas de géants de la glyptique au document numérique, en passant par l'imprimerie de Gutenberg, et ce quel que soit le support utilisé (le rouleau, le codex, le numérique), ainsi que les divers agencements du texte par rapport au support. Mais ensuite et surtout, il faut pouvoir accéder aux informations stockées. Ce besoin d'accès demeure intact aujourd'hui, l'accroissement des masses d'information disponibles ne faisant qu'accentuer ce besoin ancestral, qui devient bien plus compliqué à gérer. En effet, à quoi servirait le stockage d'une information si on ne peut y accéder ? Créer une information est un travail souvent onéreux et si on ne peut y accéder, ce même travail est à refournir. Il est d'ailleurs révélateur que peu après l'invention des ordinateurs au début des années 1950, le domaine de la Recherche d'Information ait vu le jour, démontrant que le stockage et le traitement de l'information vont de pair avec les techniques d'accès qui leurs sont associées.

Aujourd'hui, toutes les données méritant publication sont destinées -à terme- à être numérisées, le problème du stockage et de la pérennisation des informations tend ainsi à être résolu. Nous le voyons bien, nous assistons à une époque, où le savoir individuel est théoriquement, dès sa publication, universel. Notre civilisation peut désormais prétendre à la capitalisation synchronisée du savoir, à l'accélération des avancées technologiques en mutualisant les efforts et en évitant le gaspillage et la redondance. Mais aujourd'hui, la masse de connaissances stockées est tellement immense, que nous assistons au phénomène inverse : l'information n'est désormais plus une denrée rare et l'instantanéité de sa disponibilité est assurée grâce à Internet. C'est désormais au niveau individuel que nous nous posons des questions. En effet, pour un individu, un système ou une organisation, rechercher une information précise dans l'amas des données en croissance exponentielle sur le Net serait comme chercher une aiguille dans une botte de foin.

A la naissance du domaine de la Recherche d'Information, les chercheurs s'enthousiasmaient à l'idée d'utiliser les ordinateurs, pour la recherche des informations dont la taille dépassait les capacités calculatoires humaines. Dès les premiers Systèmes de Recherche d'Information (SRI), les modèles de RI sont construits autour du triplet <document, besoin, correspondance>, ces modèles constituent encore aujourd'hui la base autour de laquelle sont développés les moteurs de recherche sur leWeb. Ainsi, un SRI est un

système qui stocke un ensemble de *documents* sous une forme électronique (corpus ou base documentaire), dans le but de permettre aux utilisateurs de retrouver ceux dont le contenu correspond le mieux à leur *besoin d'information*. Une phase d'indexation permet de stocker une abstraction des contenus des documents. Ces abstractions sont ensuite comparées à la représentation des besoins de l'utilisateur (la requête) à la phase d'interrogation (ou de recherche) grâce à une fonction de *correspondance*.

Très vite les problèmes inhérents à la richesse des langues, se sont imposés. Les SRI doivent traiter les problèmes de synonymie et de polysémie des termes.

Problématique

Problématique

En indexation classique, les entités textuelles (documents et requêtes) sont représentées par des mots clés issus de leurs contenus. L'utilisation des mots pour représenter le contenu des documents et requêtes pose deux problèmes, l'ambiguïté des mots et leur disparité.

L'ambiguïté des mots, dite ambiguïté lexicale, se rapporte à des mots lexicalement identiques et portant des sens différents. Elle est généralement divisée en deux types l'ambiguïté syntaxique et l'ambiguïté sémantique.

L'ambiguïté syntaxique se rapporte à des différences dans la catégorie syntaxique. Par exemple, « *play* » peut apparaître en tant que nom ou verbe. L'ambiguïté sémantique se rapporte à des différences dans la signification, et est décomposée en homonymie et polysémie selon que les sens sont liés ou non.

Le problème d'ambiguïté implique que des documents non pertinents, contenant les mêmes mots que la requête sont retrouvés.

La disparité des mots (*word mismatch*) se réfère à des mots lexicalement différents mais portant un même sens. Ceci implique que des documents, pourtant pertinents, ne partagent pas de mots avec la requête, ne sont pas retrouvés.

Les travaux du domaine ont d'abord adressé ces problèmes séparément en apportant des solutions spécifiques à chacun d'eux, puis une solution globale s'est dégagée.

(1) Solutions spécifiques

- Une réponse au premier problème, en l'occurrence l'ambiguïté des mots, est d'utiliser les expressions ou termes composés, pour réduire l'ambiguïté.

Cependant, il n'est pas toujours possible de fournir une expression dans laquelle le mot apparaît seulement avec le sens désiré, et la formulation des expressions exige un effort cognitif de la part de l'utilisateur.

- Une réponse au second problème, en l'occurrence la disparité des mots, consiste à étendre la requête à l'aide de mots synonymes d'un thésaurus. Cette extension n'est pas aléatoire. Pour enrichir un mot dans la requête par ses synonymes, on doit non seulement connaître le sens du mot dans la requête, mais aussi le sens du mot qui est utilisé pour l'étendre.

(2) Solution globale

La solution globale permettant de répondre à ces deux problèmes consiste en l'indexation sémantique. L'indexation sémantique tente d'apporter des solutions au niveau de la représentation des documents et des requêtes. L'objectif est d'indexer par les sens des mots plutôt que par les mots. Dans un contexte où l'ambiguïté est présente, l'indexation sémantique est sensée améliorer les performances du SRI.

L'indexation sémantique s'intéresse à deux principaux points : d'abord retrouver le sens correct de chaque mot dans le document (respectivement de la requête), ensuite représenter ce document (respectivement cette requête).

Chapitre 1
Recherche d’information : concepts de base

I.1 Introduction

Le monde assiste depuis ces dernières décennies, à une production massive d'informations dans tous les domaines d'intérêt. De multiples directions de recherche ont tenté de mettre en œuvre des processus automatiques d'accès a l'information. L'objectif est d'exploiter au mieux les bases volumineuses de ces informations.

Un Système de Recherche d'Information (SRI), nécessite la combinaison de modèles et algorithmes. Ces derniers permettent la représentation, le stockage, la recherche et la visualisation des informations. L'objectif principal de ce système est de mettre en œuvre un processus de comparaison entre besoin utilisateur et documents d'une collection dans le but de retrouver ceux qui sont pertinents. L'élaboration d'un mécanisme de recherche d'information pose alors des problèmes lies tant à la représentation qu'à la localisation de l'information pertinente.

L'objet d'un système de recherche d'information est de faciliter l'accès à un ensemble de documents, afin de permettre à l'utilisateur de retrouver ceux qui sont pertinents, c'est-à-dire ceux dont le contenu correspond le mieux à son besoin en information. La qualité des résultats de la recherche se mesure en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus le système est jugé performant.

Tout au long de ce chapitre, notre intérêt se porte ainsi sur les principes de la recherche d'information. Les section 1.2 à la section 1.5 en décrivent ses concepts de base ainsi que les différents modèles. La section 1.6 est consacrée à l'évaluation de ces systèmes; nous représentons les mesures utilisées pour comparer les performances des SRI.

I.2 Un survol de l'histoire de la Recherche d'Information

Les sociétés et les entreprises ont toujours essayé de mieux préparer leur avenir en se dotant d'outils et de méthodes afin de se rendre le plus compétitifs vis-à-vis de leurs voisins et concurrents, en utilisant les techniques de renseignement, d'espionnage et des stratégies prévisionnelles, c'est-à-dire différentes formes de veille.

La stratégie de ces organismes consiste à recueillir l'information, la synthétiser et tirer les conclusions pouvant orienter leur développement. Mais toute information ne peut contribuer à l'amélioration de la productivité et à la compétitivité d'une organisation que lorsqu'elle répond aux vrais besoins des responsables, à savoir progresser, moderniser, innover et diversifier.

Toutefois, la recherche de cette information plus qu'indispensable pour toutes les fonctions d'une organisation se heurte généralement à des obstacles de nature à réduire son efficacité, notamment :

- L'abondance des supports d'information réels et potentiels sur le marché de la communication,
- Le flot de l'information pouvant entraîner l'inopportunité et la non pertinence des données lorsqu'elles ne répondent pas aux besoins précis des décideurs, alors que ces derniers ont besoin d'une information précise, analysée, filtrée et condensée.

Il s'agit ainsi, d'une information sur mesure, personnalisée et gérée pour répondre aux besoins spécifiques et de plus en plus exigeants des décideurs, or « sans gestion d'information, pas d'organisation viable ».

Des outils d'observation et de mesure ont été crées tout au long de l'histoire pour aider les sociétés à mieux mesurer leur environnement. Les Grecs ont développé des mécanismes d'observation très complexes capables de prédire les cycles de la terre. Ces mécanismes seront transmis aux horlogers européens via les arabes. Ils donneront naissance à différentes machines de calculs (machine à calcul de Pascal, les cartes perforées de Jacquard) pour arriver à la création des premiers ordinateurs.

Depuis l'avènement d'Internet qui facilite l'accès à une grande masse de données et le développement des nouvelles technologies, la veille est à la mode, elle s'élargit, de l'entreprise privée, elle devient une affaire d'état nommée Intelligence économique. Avec la chut de l'URSS, les agences des services secrets et les militaires se sont converti au civil utilisant les moyens légaux pour la cueillette d'informations.

Les nouvelles technologies de l'information et de la communication ont ainsi conduit :

1- A une transformation des pratiques de gestion de l'information : les fichiers manuels se transforment en fichiers informatisés, en banques de données,

2- à la législation et la gestion électronique des documents qui amènent à une véritable ingénierie informationnelle,

3- au besoin d'être tenus correctement informés qui devient une nécessité vitale de toutes les catégories d'utilisateurs, notamment des entrepreneurs, des chercheurs, etc.

Ceci a engendré de nouvelles pratiques, entre autres :

- La veille stratégique qui consiste à surveiller l'environnement externe de l'entreprise par le service d'information afin de recueillir l'information nécessaire à la prise des décisions stratégiques et aux actions au sein de toute organisation,

- la veille technologique qui consiste à observer l'environnement technologique et suivre les évolutions qu'il subit afin de dégager les opportunités et les menaces qu'il offre et que le service d'information doit prendre en considération,

- la veille concurrentielle qui consiste à suivre de prêt et de manière systématique les concurrents réels et potentiels du service d'information, leur expansion dans le temps et dans l'espace, leurs produits, leurs services, leurs innovations,

- la veille commerciale qui, pour rationaliser les achats et ventes, consiste à suivre les marchés de matières premières, la situation des circuits commerciaux, etc.

Bref, l'intelligence économique, qui est une démarche globale qui vise à inclure tous les types de veille en une approche globale permettant non seulement de surveiller mais aussi de prévoir toutes les menaces et opportunités relatives au contexte concurrentiel, juridique, technologique, commercial, sociétal, etc. de l'organisation.

Etant donné ces nouvelles pratiques, le professionnel de l'information est tenu de :

- Savoir et pouvoir maîtriser l'information de veille, de découverte, d'innovation et d'ouverture sur le monde,

- savoir et pouvoir développer et exploiter l'information utile qui rend possible l'activité quotidienne des individus, des centres et des laboratoires de recherche, des entreprises,…

- valoriser l'information auto produite, tenir compte de l'information vivante, de l'information de communication, etc.…

- raisonner en terme de différentiation fonctionnelle multidimensionnelle et donc de richesse d'intervention potentielle avec autant de compétences spécifiques à développer.

La veille suppose une maitrise de l'information nécessaire à la surveillance des environnements précis (sociopolitiques et économiques). C'est un processus continu et systématique de gestion de l'information stratégique.

Un processus de veille comporte en général trois étapes essentielles :

1- La cueillette : il s'agit de bien rassembler les données pour dresser un bilan sur le contexte donné, ses principaux défis sont :

- Le traitement d'un très grand volume de données dans un temps assez court,
- la classification des données.

Dans cette étape, la recherche s'effectue dans les bases de données, les sites Web et l'échange entre veilleurs, à l'aide des répertoires, des annuaires, des bases de connaissances commerciales sur le Web, des outils linguistiques, … etc.

2- l'analyse et la synthèse : cette étape sert à synthétiser les données rassemblées afin de découvrir les principales tendances qui serviront à convertir certaines stratégies en scénarios,

3- la diffusion : il s'agit de présenter aux décideurs divers scénarios qui faciliteront leur prise de décision. Les principaux défis se résument en :

- La pertinence des choix en fonction du long terme,
- le développement de stratégies conduisant aux innovations.

Pour exécuter un tel processus, les outils de veille se divisent le plus souvent en 2 catégories :

- Les outils de recherche d'information,
- les outils de surveillance.

La *recherche d'information* concerne les mécanismes qui facilitent l'accès à une base d'informations. Il existe un grand nombre de *modèles* de recherche d'information. Ces modèles diffèrent principalement sur la façon dont les informations disponibles sont représentées et sur la façon d'interroger la base. Notre projet présenté ci-après porte sur le point particulier des outils de la recherche d'information.

I.3 La naissance de la recherche d'information

Le domaine de la recherche d'information remonte au début des années 1950, peu après l'invention des ordinateurs, les chercheurs voulaient les utiliser pour automatiser la recherche des informations, qui dépassaient les capacités humaines à cause de l'explosion de la quantité d'information après la deuxième guerre mondiale.

Le terme de recherche d'information '*Information Retrieval*' fut donné par Calvin N. Mooers en 1948 pour la première fois dans son mémoire de maîtrise [1] et la première conférence dédiée à ce thème – International Conference on Scientific Information - s'est tenue en 1958 à Washington.

Les premiers problèmes qui intéressaient les chercheurs portaient sur l'indexation des documents.

Déjà à la 'International Conference on Scientific Information', Luhn avait fait une démonstration de son système d'indexation 'KWIC' qui sélectionnait les index selon la fréquence des mots dans les documents et filtrait des mots vides. C'est à cette période que le domaine de la recherche d'information est né.

I.4 Ère Internet

Le domaine de recherche d'information fut créé à cause de l'explosion de l'information dans les années 1950. Mais cette explosion apporte de nouveaux problèmes dans le domaine de la recherche d'information.

- Sur le Web, on ne peut plus créer une collection statique. La collection (qui est le Web au complet) est une collection gigantesque qu'il est impossible (au moins pour le moment) de couvrir au complet,

- un système de recherche propose toujours des documents. Certains sont pertinents, mais noyés parmi beaucoup d'autres documents non pertinents. Plus notre collection contient des documents, plus ce problème devient aigu. Il est de plus en plus demandé que la recherche soit plus précise, même si on doit accepter que certains documents pertinents ne soient pas retrouvés,

- l'existence des documents non textuels (image, son, vidéo, etc.) nécessite de nouvelles façons pour les indexer et les rechercher. Les méthodes traditionnelles de recherche sont surtout destinées aux textes et ne sont pas directement applicables à d'autres médias,

- l'utilisation des langues différentes pose un autre problème. Avec une requête en français, on ne peut retrouver que des documents en français. Or, la pertinence d'un document est souvent indépendante de la langue utilisée. Ainsi, nous avons besoin d'outils pour la recherche d'information translinguistique ou multilingue.

I.5 Généralités sur les Systèmes de Recherche d'Information(SRI)

I.5.1 Définition

La recherche d'information [1] est l'ensemble des techniques permettant de gérer des textes. Gérer des textes ou des documents implique stocker, rechercher et explorer des documents pertinents.

Un système de recherche d'information intègre un ensemble de techniques et de processus permettant de sélectionner dans une collection de documents ceux qui sont susceptibles de répondre au besoin d'un utilisateur. Ces processus permettent :

– La représentation des informations et des besoins,

– L'interrogation, la recherche et la sélection des informations pertinentes répondant aux besoins d'un utilisateur.

La problématique majeure émanant de tout système de recherche d'information est de retrouver les quelques dizaines ou milliers de documents pertinents parmi des millions de documents. Cet écart de cardinalité rend cette tache encore plus difficile.

I.5.2 Concepts clés de la recherche d'information

Un système de recherche d'information, intègre un ensemble de modèles pour la représentation des unités d'informations (documents et requêtes). Il intègre également un mécanisme de recherche/sélection. Ce dernier permet de sélectionner l'information pertinente en réponse aux besoins exprimes par l'utilisateur a l'aide d'une requête.

Il peut être représenté par le processus en U de recherche d'information.

La figure 1 illustre l'architecture générale d'un système de recherche d'information.

Plusieurs éléments clés y sont distingues :

– La collection de documents,

– Les documents,

– Les langages d'interrogation,

– La représentation des documents et des requêtes (indexation ou analyse),

– L'appariement requête-document,

– Le besoin en information (requête),

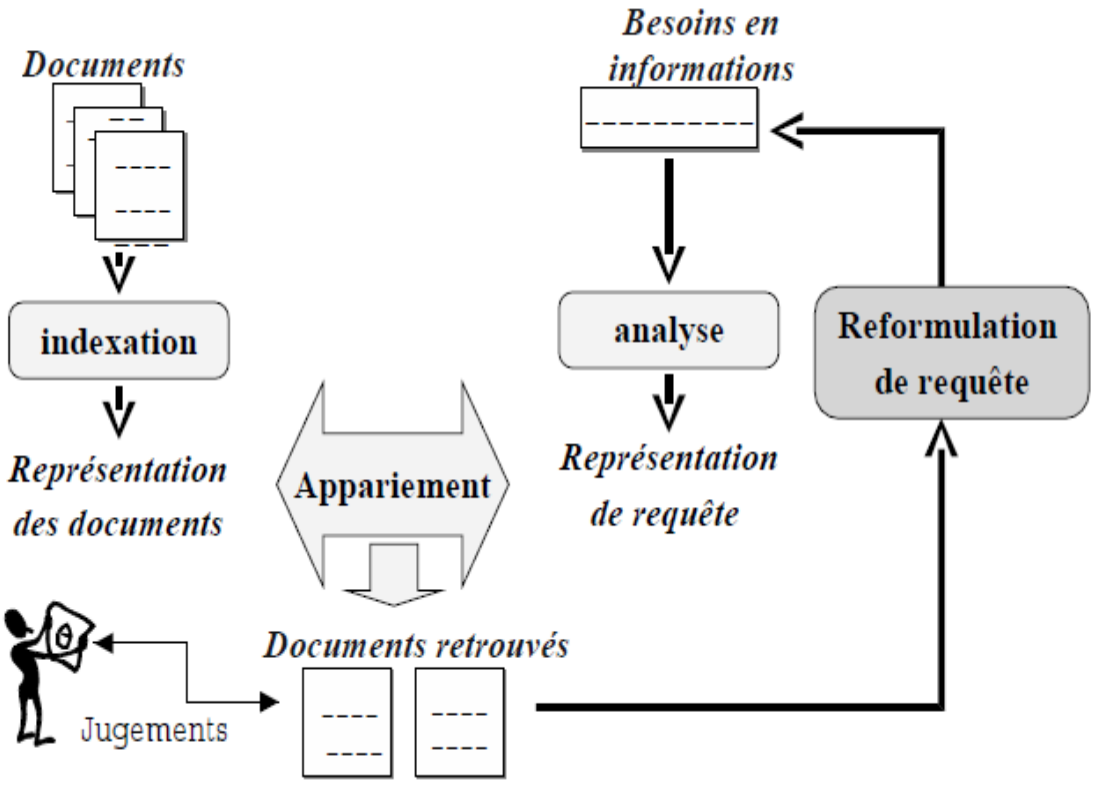


Figure. 1 – *Le processus de recherche d'information*

Dans la section qui suit, nous allons définir ces éléments séparément.

I.5.2.1 La collection de documents

La collection de documents constitue l'ensemble des informations exploitables et accessibles par l'utilisateur. Elle est constituée d'un ensemble de documents. Dans le cas général et pour des raisons d'optimalité, la collection constitue des représentations très simplifiées mais suffisantes de ces documents. Ces représentations sont étudiées de telle sorte que la gestion (ajout, suppression d'un document) et l'interrogation (recherche) de la collection se font dans les meilleures conditions de cout.

I.5.2.2 Le document :

Le document constitue l'information élémentaire d'une collection documentaire.

L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document. Dans la suite de ce rapport, nous utiliserons indifféremment les termes document ou information pour designer un granule documentaire.

I.5.2.3 Les langages d'interrogation :

Un besoin en information d'un utilisateur est exprimé par une requête. La littérature propose divers types de langages d'interrogation pour formuler cette requête. Nous citons les plus répandus :

– Interrogation en langage booléen,

– Interrogation en langage naturel ou quasi naturel,

– Interrogation en langage graphique.

Détaillons à présent ces différents langages.

1. **Interrogation en langage booléen :**

   L'utilisateur exprime sa requête sous forme d'un ensemble de termes relies entre eux par des operateurs booléens. Beaucoup de moteurs de recherche, se basent sur ce mode d'interrogation, citons les plus connus : Altavista, Google, etc.

2. **Interrogation en langage naturel ou quasi naturel :**

   L'utilisateur exprime sa requête en langage libre (langage naturel) sous forme de mots clés. Le système se charge de traduire (analyser) ces mots clés en une requête de langage de base de données ou une autre forme interne utilisable par le système. Les systèmes SMART, SPIRIT, OKAPI *Recherche* et MercureO2 sont interrogeables en langage naturel [2].

3. **Interrogation en langage graphique :**

Une interface d'aide à la formulation de la requête est proposée à l'utilisateur. En effet, une vue d'ensemble de la base d'information et en particulier une vue de termes représentant le contenu sémantique des documents, est donnée a l'utilisateur pour l'assister à formuler sa requête. Dans PROTEUS [2], l'interface d'aide à la formulation de requête propose un gestionnaire de thesaurus. Ce dernier est représente par un graphe, les nœuds étant les termes du thesaurus et les liens étant les relations sémantiques entre ces termes. L'utilisateur peut identifier le type de relation qu'il souhaite utiliser et sélectionner un terme. Le projet NEURODOC [2] est plus adapte a l'utilisation d'un thesaurus volumineux. NEURODOC offre à l'utilisateur un tableau de bord ou chaque nœud possède un nom et résume le sous-ensemble de mots et de documents fortement lies.

I.5.2.4 La représentation des documents et des requêtes (indexation ou analyse)

La représentation des documents et des requêtes est supportée par un ensemble de règles et notations permettant la traduction d'une requête ou d'un document d'une description brute vers une description structurée. Ce processus de conversion est appelé Indexation.

L'indexation est une opération permettant d'extraire d'un document ou d'une requête une représentation paramétrée qui couvre au mieux son contenu sémantique.

Le résultat de l'indexation constitue le descripteur du document ou de requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leur degré de représentativité du contenu sémantique de l'unité qu'ils décrivent. Les descripteurs des documents (mots, groupe de mots) forment le langage d'indexation. L'indexation, est une étape primordiale dans la recherche d'information. De sa qualité dépend en partie la qualité des réponses du système. Conscients de son importance, et soucieux de bien la réaliser, les développeurs des SRI ont proposé plusieurs manières de procéder. Les principales sont l'indexation manuelle et l'indexation automatique.

Elles sont définies comme suit :

– Indexation manuelle : dans le cas de l'indexation manuelle, chaque document

est analysé par un spécialiste du domaine ou par un expert documentaliste.

En fonction de ses connaissances, Cet expert détermine, les mots clés qui lui semblent les plus significatifs pour représenter le document. L'indexation humaine est une activité fondée sur le jugement d'un être humain. Elle se caractérise par sa profondeur, sa cohérence (ce qui est fondamental pour la cohérence du fond et des fichiers) et sa qualité (exhaustivité - spécificité). Elle est cependant trop dépendante de l'état des connaissances des indexeurs. Cela induit à la subjectivité de ses résultats. Elle nécessite la lecture de l'intégralité des documents. Son application est de ce fait inadaptée à des collections de taille importante. Les collections TREC[1] constituent un exemple significatif.

Elles contiennent des millions de documents extraits d'Internet (le web).

L'indexation automatique permet de pallier à ce problème.

– Indexation automatique : l'indexation automatique reconnait des chaines de caractères constitutifs de mots non vides. Elle détecte automatiquement les termes les plus représentatifs du contenu du document. Ce type d'indexation est actuellement la méthode la plus répandue.

Elle comprend deux étapes fondamentales : l'identification des termes d'indexation et l'évaluation de leurs poids.

L'identification des termes d'indexation consiste à analyser le texte du document mot a mot. Son objectif est d'en extraire les mots vides qui ne jouent qu'un rôle syntaxique. Ces mots sont identifies puis élimines grâce a un anti dictionnaire (Stoplist en Anglais). Les mots apparaissant trop souvent n'ont aucun intérêt. Ils sont également éliminés. Seuls les mots significatifs représentant les concepts du document sont retenus.

Afin d'augmenter la qualité de la recherche, la pondération des termes extraits est primordiale. Pour mettre en évidence les diverses contributions d'un terme dans la représentation d'un document un poids lui est attribue.

I.5.2.5 L'appariement requête-document

Le processus d'appariement requête-document est le noyau d'un système de recherche d'information. Il permet d'associer à chaque document une valeur de pertinence vis à vis d'une requête. Les documents ayant une pertinence positive sont sélectionnes.

La mesure de pertinence est calculée à partir d'une fonction de similarité, notée *RSV* (*Q,d*)(*Retrieval Satus Value*), *Q* étant une requête et *d* un document.

Elle tient compte des poids des termes détermines en fonction d'analyses statiques et probabilistes. Notons que ce processus est étroitement lie aux représentations des documents et des requêtes.

En effet, si l'opération d'indexation est la même dans la plupart des modèles de recherche d'information, ces derniers différents souvent par rapport aux fonctions utilisées pour la mesure des poids et pour l'appariement requête-document.

I.5.2.6 La notion de 'besoin' dans la recherche d'information

La notion de 'besoin d'information' est centrale dans le domaine de la recherche d'information puisque elle est définie comme une interaction entre « un individu qui a besoin d'information » et « un document qui contient ou non la réponse à ce besoin » [2].

L'utilisateur doit donc formuler une requête, c'est-à-dire exprimer son besoin en information sous forme de descripteurs ou mots clés plus au moins liés, dont la relation est exprimée par la présence d'opérateurs entre eux. La requête peut s'effectuer sur l'ensemble des mots du texte, ou dans certaines zones précises du document, lorsque l'information est indexée et structurée selon différents champs (titre, auteur, ...).

I.6 Evaluation des performances des systèmes de recherche d'information

L'évaluation des systèmes de recherche d'information constitue une étape importante dans l'élaboration d'un modèle de recherche d'information. En effet, elle permet de caractériser le modèle et de fournir des éléments de comparaison entre modèles.

D'une façon générale, tout système de recherche d'information présente deux objectifs:

– retrouver tous les documents pertinents,

– rejeter tous les documents non pertinents.

Et cela pour répondre aux besoins de l'utilisateur.

Ces deux objectifs sont évalués par les mesures de précision et de rappel définis ci-dessous.

Nous allons définir également les mesures à *x* documents et d'autres mesures de performance.

I.6.1 La notion de pertinence

Pour être en mesure d'offrir aux utilisateurs les informations répondant le mieux à leurs besoins, tout système de recherche d'information s'appuie sur un modèle de calcul de pertinence qui, pour chaque requête, calcul le score de pertinence de chaque donnée (document). Celles qui auront le meilleur score de pertinence seront présentées à l'utilisateur.

Cette approche permet d'évaluer ce qu'on nomme la pertinence système, c'est-à-dire la pertinence que les systèmes de recherche d'information calculent. Or, La notion de pertinence est très complexe, elle est évaluée par les systèmes de recherche d'information et également liée au jugement des utilisateurs.

On distingue classiquement deux types de pertinence : la **pertinence utilisateur**, qui est le jugement apporté par l'utilisateur sur le document, en fonction de son besoin d'information, et la **pertinence système**, qui correspond à la valeur de correspondance entre le document et la requête, calculée par les systèmes. La satisfaction de l'utilisateur est liée à la correspondance entre ces deux pertinences.

Un étudiant en droit qui doit étudier un cas précis et qui dispose du corpus de toute la jurisprudence du droit français et ne disposant que d'un accès chronologique ou thématique aux documents, va chercher à identifier dans son besoin en information les critères qui peuvent cerner soit la période pendant laquelle des actes de jurisprudences qui lui sont pertinents ont pu être émis, soit la thématique traité dans sa requête. D'autres critères vont certainement intervenir dans l'estimation de la pertinence d'un document. Certains documents ne seront pas utiles, car déjà connus, d'autres peuvent être éliminés puisque ils demanderaient trop de travail pour être utilisés.

Cet exemple donne une idée sur la grande diversité des facteurs qui interviennent lorsqu'un utilisateur évalue la pertinence d'un document.

Il existe une distance plus ou moins grande entre les résultats d'un système de recherche d'information et les jugements de pertinence de l'utilisateur. L'utilisation d'un système de recherche d'information est plus généralement conçue comme un processus itératif visant à améliorer progressivement l'adéquation entre pertinence système et pertinence utilisateur. Pour ce faire, une nouvelle fonction est très fréquemment ajoutée au schéma fonctionnel classique : le bouclage de pertinence (relevance feedback). Une fois un premier ensemble de documents retrouvés, l'utilisateur peut émettre des jugements de pertinence sur ces documents, jugements qui sont pris en compte pour définir une nouvelle requête (reformulation de la requête).

I.6.2 Les mesures de Précision/Rappel

Les mesures de précision/rappel sont obtenues en partitionnant l'ensemble des documents restitués par le SRI en deux catégories : les documents pertinents et les documents non pertinents. Ces deux catégories se définissent comme suit:

– Taux de précision : La précision mesure la capacité du système à rejeter tous les documents non pertinents à une requête. Il est donné par le rapport entre l'ensemble des documents sélectionnes pertinents et l'ensemble des documents sélectionnés.

– Taux de rappel : Le rappel mesure la capacité du système à retrouver tous les documents pertinents répondants à une requête. Il est donné par le rapport entre les documents retrouves pertinents et l'ensemble des documents pertinents de la base.

Les taux de précision et de rappel sont donnés par les formulations suivantes :

$$\text{Précision} = \frac{R+}{M}$$

$$\text{Rappel} = \frac{R+}{R}$$

Ou:

R: le nombre total de documents pertinents dans la collection

M : le nombre de documents sélectionnés

R+ : le nombre de documents pertinents sélectionnés

La figure 2 illustre la précision et le rappel d'une requête d'une façon générale.

Toutefois, seule une partie des documents restitués par le système est examinée par l'utilisateur. Dans ce cas, la paire des mesures (taux de rappel, taux de précision) est calculée à chaque point de rappel (document pertinent restitue). Il s'agit de considérer la liste ordonnée des documents évalues, de calculer pour chaque document sélectionné la précision et le rappel, puis exprimer en fonction des valeurs trouvées la précision en fonction du rappel. Avec ces valeurs, on trace une courbe représentant la précision en fonction du rappel.

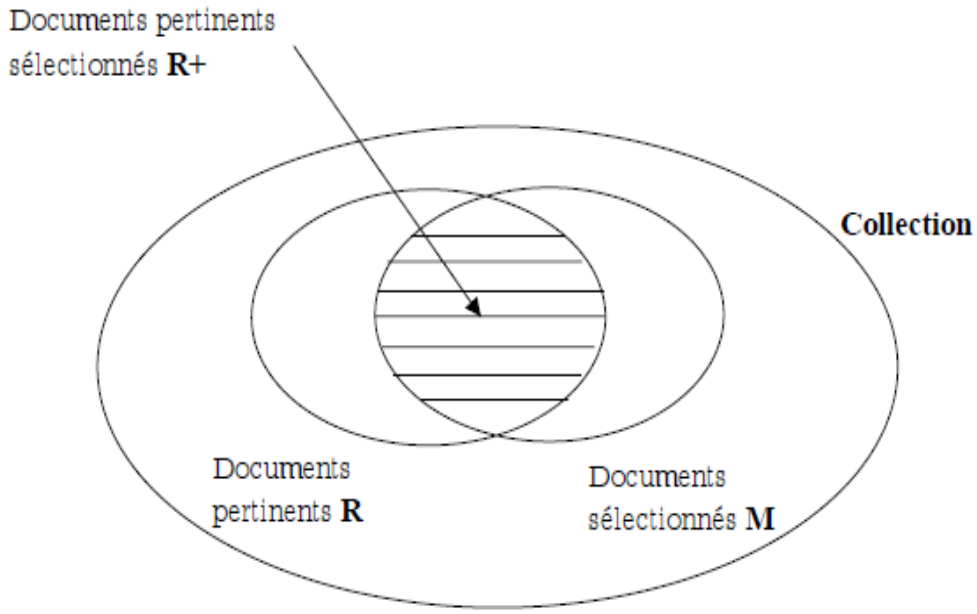


Figure. 2 – *Exemple de rappel et de précision pour une requête*

Exemple :

Le tableau 1 présente une liste de documents triés par ordre décroissant de pertinence, et les mesures de précision et de rappel engendrées. La figure 3 illustre la courbe rappel-précision. On calcule la précision pour chacune des valeurs de rappel 0.1, 0.2 . . . 1.0 par interpolation linéaire. Cette méthode d'évaluation est très significative. La précision mesurée indépendamment du rappel et inversement sont par contre peu significatives. En effet, un

système même peu performant a de très fortes chances d'attribuer la plus grande valeur de pertinence à un document pertinent s'il sélectionne seulement ce document.

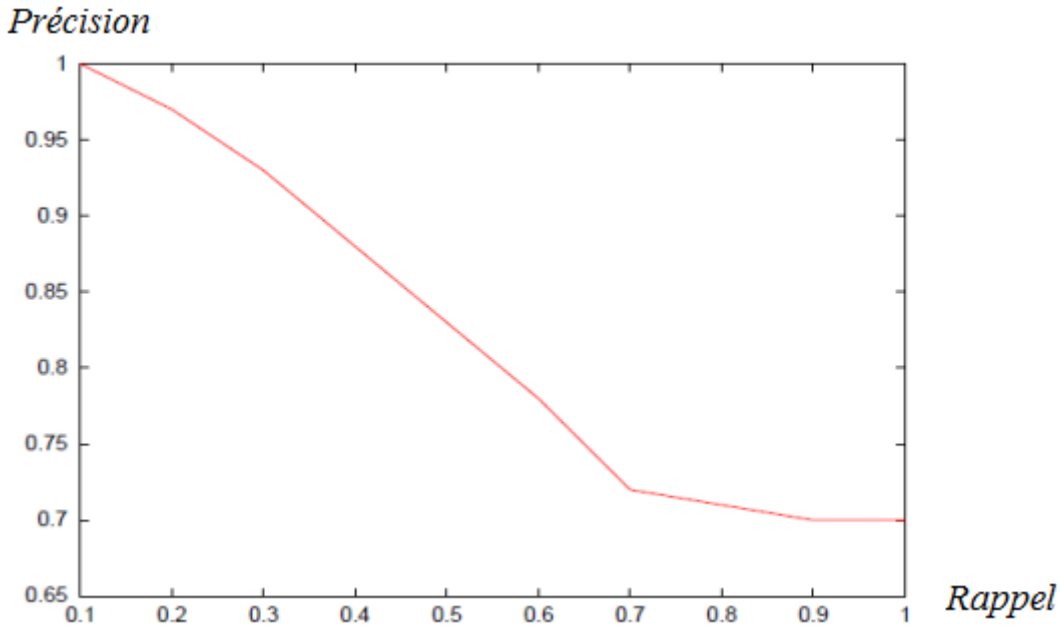


Figure. 3 – *Courbe rappel-précision*

document	score	[pertinent]	précision	rappel
d1	9.45	*	1	0.14
d2	9.34		0.5	0.14
d3	8.72	*	0.67	0.29
d4	8.66	*	0.75	0.43
d5	7.93	*	0.8	0.57
d6	6.53		0.67	0.57
d7	6.41	*	0.71	0.71
d8	4.87		0.62	0.71
d9	4.31	*	0.67	0.86
d10	4.09	*	0.7	1
d11	3.82		0.64	1
d12	2.21		0.58	1
d13	2.18		0.54	1
d14	1.64		0.5	1
d15	0.02		0.47	1

Tableau. 1 – *Exemple de valeurs rappel-précision*

La précision vaudra $\frac{1}{1} = 1$, mais le rappel sera très faible ($\frac{1}{R}$). S'il sélectionne tous les documents, le rappel vaudra $\frac{R}{R} = 1$ mais la précision sera très faible ($\frac{R}{S+R}$). Par conséquent, les mesures de précision et de rappel évoluent inversement. En d'autres termes, la courbe de précision en fonction du rappel est décroissante. La combinaison des mesures de précision et de rappel est un critère d'évaluation très significatif.

Plus cette courbe est élevée, plus le système est performant.

Les mesures à *x* documents et la précision moyenne

Deux mesures communément utilisées, sont la précision à *x* documents (*x* = 5,10,15,20,*etc.*) et la précision moyenne.

– La précision à *x* documents est souvent reliée à ce que l'on appelle la précision exacte ou la R-précision. La précision exacte est la précision au point ou la précision vaut le rappel. Si la requête admet *n* documents pertinents, la précision exacte est la précision calculée a partir des *n* premiers documents de la liste ordonnée des documents restitues.

rappel	précision
0.10	1.00
0.20	0.97
0.30	0.93
0.40	0.88
0.50	0.83
0.60	0.78
0.70	0.72
0.80	0.71
0.90	0.70
1.00	0.70

Tableau. 2 – Valeurs utilisés pour la courbe rappel-précision

– La précision moyenne est la moyenne des valeurs de précision a chaque document pertinent de la liste ordonnée. Elle tient compte à la fois de la précision et du rappel. Elle est

mesurée comme la moyenne des précisions (non interpolées) calculées pour chaque document pertinent à trouver, au rang de ce document.

Si un document pertinent est retourne à la 10*e* position, la précision pour ce document est la précision à 10 documents. Si un document pertinent n'a pas été trouve par le système, la précision pour ce document est nulle.

I.6.3 Autres mesures de performance

– Le temps de réponse acceptables : un SRI doit pouvoir fournir à l'utilisateur les documents correspondants à sa demande dans des temps très courts.

– La présentation des résultats claire et facilité d'utilisation : capacité du système à comprendre les besoins de l'utilisateur et a mettre en valeur les documents correspondants a ceux-ci. Ceci est lie à l'interface avec l'utilisateur.

– Le nombre total de documents pertinents retournés, ou le rappel à

1000 documents : ces mesures permettant d'évaluer la performance globale du système au final, en fonction ou non du nombre de documents pertinents total.

– Le rang du premier document pertinent : cette mesure a été proposée pour prendre en compte la satisfaction de l'utilisateur qui chercherait un seul document pertinent (comme c'est éventuellement le cas pour les moteurs de recherche sur Internet).

– La longueur de recherche : elle est égale au nombre de documents non pertinents que doit lire l'utilisateur pour avoir un certain nombre *n* de documents pertinents.

D'autres mesures qui combinent les scores de précision et de rappel, appelées aussi les mesures composites d'évaluation ont été définies. Par exemple [2] propose une mesure générale d'efficacité (efficiency), appelée E-mesure, qui est une combinaison de la précision et du rappel, et qui prend en compte un paramètre que l'utilisateur ajuste pour contrôler l'importance qu'il donne a la précision et au rappel.

I.7 Améliorations techniques

De nombreuses études ont porté sur des améliorations possibles de techniques d'indexation et de recherche. Parmi les tentatives les plus marquantes, on retrouve notamment:

- Rétroaction de pertinence (relevance feedback) : Cette technique vise à étendre la portée de la recherche en intégrant les termes issus des documents pertinents, ou des documents en tête de la liste de réponses trouvées automatiquement,

- expansion de requête : Cette technique vise à renforcer l'expression de la requête de l'utilisateur (qui est souvent très courte) par l'intégration des termes reliés (soit en exploitant un thésaurus, soit en utilisant un calcul basé sur des cooccurrences),

- regroupement (clustering) des documents : Il vise à créer une structure entre les documents selon leurs similarités. Cette structure peut aider à la fois la recherche et la présentation des résultats.

I.8 Conclusion

Dans ce premier chapitre, nous nous sommes essentiellement intéressés à l'étude des systèmes de recherche d'information, d'une façon générale, ainsi qu'une présentation des éléments constituant l'architecture de tels systèmes.

La finalité de chaque système de recherche d'information est de satisfaire les besoins des utilisateurs. Ces derniers sont préoccupés par un seul problème : celui de pouvoir récupérer tous les documents dont ils ont besoin d'une façon rapide et efficace. Et pour cela différentes techniques ont été mis en point, dont la plus efficace : L'indexation sémantique latente.

CHAPITRE 2

Indexation sémantique latente

« Latent semantic indexing »

2.1 Introduction

Le monde devient de plus en plus digitalisé avec une expansion massive en volume de données qui sont accessibles en ligne sous diverses formes. En outre, en raison de la révolution Internet, n'importe qui peut accéder à des millions de pages sur le Web. En 1998, des chercheurs ont estimé qu'il y avait environ 300 millions de pages Web sur Internet [24] [25] et maintenant, cette estimation est passée à 1000 milliards.

De même, Deux tendances dominent la recherche d'information et l'analyse dans l'entreprise d'aujourd'hui: le volume d'informations a considérablement augmenté, et la valeur de cette information est en croissance tout aussi rapide. Les entreprises modernes doivent faire face à des téraoctets de texte, comme le courriel, qui jouent souvent un rôle important. Même les petites et moyennes entreprises sont face au volume croissant de textes qui nécessitent un accès rapide et une analyse significative.

D'où le besoin clair de fournir de nouvelles approches qui augmentent les procédés d'extraction de données et c'est l'une des raisons principales derrière l'intérêt continu pour les systèmes de recherche d'informations (IR). En particulier, dans le corps de recherche des indexations sémantique latente (LSI).

Ce chapitre est organisé comme suit :

On commence par une vue d'ensemble du modèle de l'espace de vecteur (VSM), ensuite on présentera le travail existant et relatif pour les algorithmes de LSI sur la décomposition de prétraitement et de matrice, puis un choix de recherche sur l'utilisation des systèmes de LSI dans différents domaines d'application est fourni, et on finira avec un résumé des désavantages de cette technique et une conclusion.

Approches de traitement sémantique

Les efforts pour intégrer des informations sémantiques dans les systèmes de traitement de texte remontent à près d'un demi-siècle. Au fil des années, les concepteurs ont suivi différentes approches pour intégrer un certain degré de traitement sémantique dans leurs systèmes de récupération de l'information:

» Les structures auxiliaires

» L'indexation sémantique latente

Les structures auxiliaires

Les vocabulaires contrôlés, ou les structures auxiliaires, tels que les dictionnaires et les thésaurus, permettent des termes plus larges, des termes plus précis, et les termes connexes doivent être intégrés dans les queries. Les vocabulaires contrôlés sont une façon de surmonter certaines des contraintes les plus sévères de requêtes de mots clés booléenne qui ont des significations semblables (synonymie), et des mots qui ont plus d'un sens (polysémie). La synonymie et la polysémie sont souvent la cause de l'inadéquation dans le vocabulaire utilisé par les auteurs des documents et des utilisateurs de systèmes de récupération de texte.

Au fil des années, d'autres structures auxiliaires d'intérêt général, tels que le grand ensemble de synonymes de Wordnet, ont été construits. La tendance la plus récente a été de créer des modèles de données qui représentent des ensembles de concepts dans un domaine (ontologies), qui peuvent intégrer les relations entre les termes.

Les vocabulaires contrôlés peuvent contribuer à l'efficacité et l'exhaustivité de la recherche d'informations et opérations liées à l'analyse textuelle. Mais cette approche pour le traitement sémantique fonctionne mieux lorsque les sujets sont étroitement définis et la terminologie est normalisée. Néanmoins il n'est pas bien adapté aux besoins de la plupart des entreprises modernes et des volumes croissants de données non structurées qui contiennent des milliers de termes uniques couvrant un nombre illimité de sujets.

Certains autres inconvénients de l'utilisation de structures auxiliaires:

- Établir une structure auxiliaire exige beaucoup de moyens humains et de surveillance
- La langue évolue rapidement, nécessitant la mise à jour constante des vocabulaires contrôlés
- Les vocabulaires contrôlés peuvent souvent représenter la vision du monde de leurs créateurs, en introduisant une source potentielle d'asymétrie conceptuelle

- Les vocabulaires contrôlés capturent une vision du monde à un moment donné. Ils peuvent être difficiles à modifier en tant que concepts changeant dans un sujet précis.

Latent Semantic Indexing (indexation sémantique latente)
Latent Semantic Indexing est une méthode de recherche d'information statistique qui est capable de récupérer le texte basé sur les concepts qu'il contient, non seulement par correspondance des mots clés spécifiques. D'abord appliqué au texte à Bell Labs dans les années 1980, il a été appelé LSI en raison de sa capacité à corréler les termes sémantiquement liés dans une collection de texte.

LSI utilise une matrice terme-document appelée TDM pour identifier l'apparition des termes dans un ensemble de documents, en se basant sur la fréquence d'apparition des termes dans les différents documents pour refléter le fait que certains termes sont plus importants que d'autres dans un corps de texte, puis effectue une décomposition en valeur singulière (SVD) sur la matrice pour déterminer les modèles dans les relations entre les termes et concepts utilisés dans les documents.

LSI utilise une technique de transformation mathématique pour réduire le nombre de dimensions représentant la matrice termes -document pour la rendre utilisable et efficace. Une conséquence du traitement LSI est la création d'associations entre des termes qui apparaissent dans des contextes similaires. En conséquence, les requêtes sur un ensemble de documents qui ont subi LSI renvoi des résultats qui sont conceptuellement similaires dans un sens à la requête même si elles ne partagent pas un mot ou des mots spécifiques à la requête.

Les avantages théoriques de LSI ont été soigneusement testés et sont soutenus par des résultats expérimentaux. La tâche de catégorisation de documents en fonction de leurs similitudes conceptuelles par exemple, a démontré la supériorité de LSI sur les autres approches pour extraire des informations sémantiques à partir de documents.

2.2 Introduction à VSM (Vector Space Model)

Dans cette section on présente une brève illustration du mécanisme le plus fondamental de l'algèbre linéaire. VSM, un modèle présenté par G. Salton, est une technologie d'IR qui est basée sur le concept d'un espace de vecteur.

Les termes, les documents et les requêtes sont représentés comme des vecteurs dans un espace de vecteurs. Dans ce modèle, la base de données est représentée comme une matrice de documents et de termes (TDM), tous les documents dans la base de données sont stockés dans les colonnes de la matrice et tous les termes sont stockés dans les lignes de la matrice.

Un document est représenté par un vecteur d = (d1 ; d2 ; : : : : Dn) où chaque élément di est un nombre indiquant le degré d'importance (nombre d'apparition) d'un terme Ti (le modèle de VSM est fondé sur l'hypothèse que la signification d'un document peut être connu a partir des termes contenus dans le document). En d'autres termes chaque document est représenté comme vecteur dans un espace de vecteur de dimension n. De même, la requête de l'utilisateur peut être représentée comme vecteur q [28] [29] [30] dans cet espace.

L'utilisation d'une telle technique se fonde sur un modèle mathématique fondamental. Dans ce modèle, les documents sont représentés comme des ensembles de termes qui peuvent être pesés et manœuvrés. Ainsi on peut comparer la représentation de la requête à la représentation de chaque document dans l'espace de vecteur. Des documents de la base de données seront sélectionnés comme résultat de la recherche par l'intermédiaire des opérations de vecteur simples. Plus de détails au sujet des outils utilisés pour mesurer la similitude entre les termes, les documents et la requête sont fournis dans le chapitre 3.

VSM a été développé pour éliminer plusieurs problèmes liés aux techniques utilisant des mots-clés traditionnels, spécialement la synonymie et polysémie comme décrit en chapitre 1. La fonction de recherche pour ce modèle est basée sur la signification sémantique et conceptuelle des documents elle fournit un mécanisme pour comparer les termes dans une requête aux termes dans un document, aussi bien que la comparaison entre les documents dans la base de données. Avoir tous les composants d'IR dans le même espace de vecteur, et le calcul des similitudes entre elles permettent d'avoir le résultat désiré. Ceci signifie que des résultats qui sont conceptuellement plus appropriés aux utilisateurs peuvent être retournés automatiquement aux utilisateurs [28].

Une représentation de l'espace de vecteur des documents tridimensionnels est montrée dans figure.4, où chaque document se compose de trois termes. Comme mentionné ci-dessous, l'exemple tridimensionnel peut être prolongé à n dimensions quand on a n termes différents qui représentent le document.

Puisque la configuration de l'espace de document est en fonction des termes et des poids des termes qui sont assignés aux divers documents de la base de données, on peut se demander si une configuration « optimum » de l'espace de document existe, c.-à-d, une configuration qui produit une performance optimale de recherche. Si rien de spécial n'est connu au sujet des documents à l'étude, on pourrait considérer qu'un espace de documents est idéal quand des documents qui sont conjointement appropriés à la requête d'utilisateur sont groupés ensemble, donc ils seraient proposés conjointement en réponse à la rêquete de l'utilisateur.

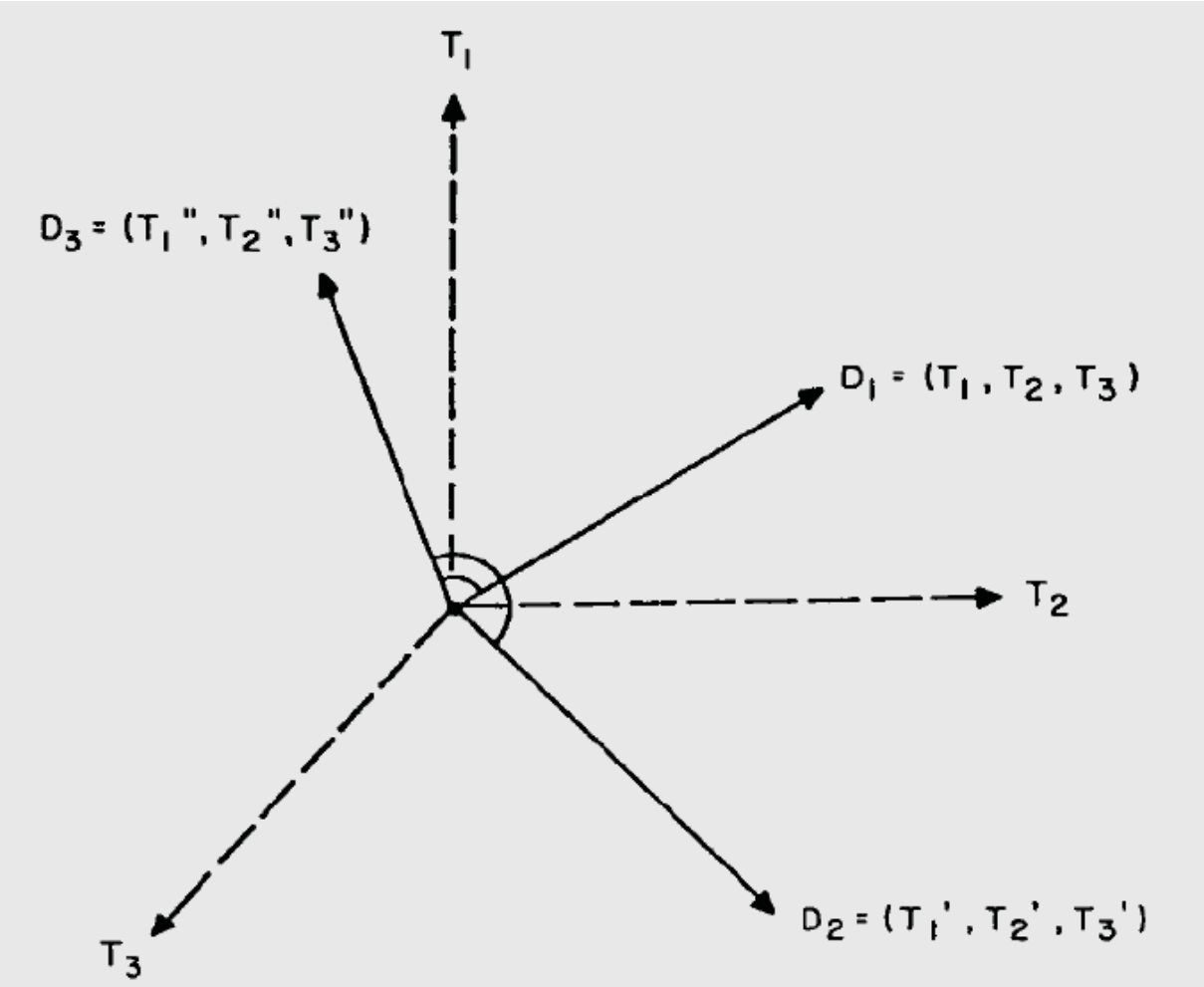


Figure 4 : Représentation de document de l'espace de vecteur [27]

De même les documents éloignés dans l'espace de documents ne seront jamais proposés conjointement en réponse.

Une telle situation est montrée dans figure. 5, où la distance entre deux croix représentant deux documents est inversement liée à la similitude entre les vecteurs correspondants.

La configuration de documents de la figure. 5 peut représenter la meilleure situation, en supposant que les documents pertinents et non pertinents en ce qui concerne les diverses requêtes sont séparables comme le montre la figure. Dans son travail de brevet [27] Salton clarifie cela, aucune manière pratique n'existe pour produire réellement un tel espace, parce qu'il est difficile de produire la configuration optimale en l'absence de la connaissance des détails complets de la recherche pour la base donnée. Dans ces circonstances, le besoin d'avoir recours à LSI se fait sentir, puisque cette technique peut aider en fournissant un tel espace de vecteur riche.

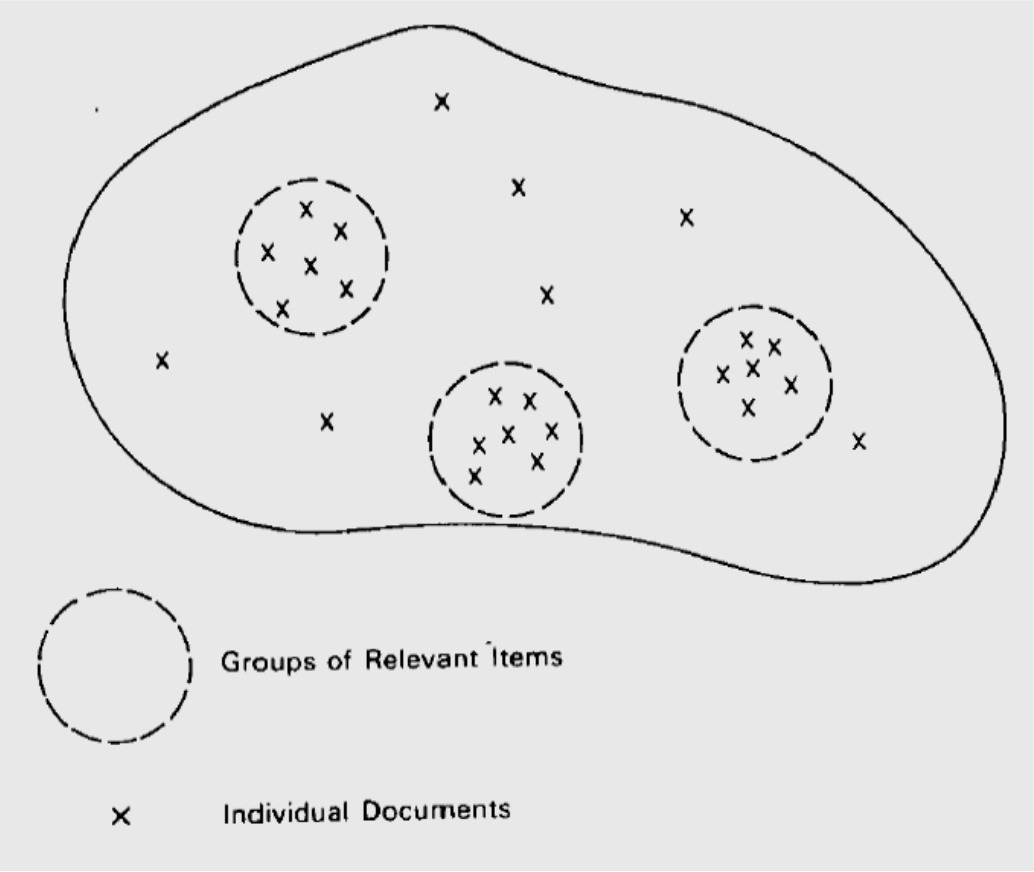


Figure 5 : Représentation idéale de l'espace de document [27]

Comme défini précédemment la LSI est un modèle de l'espace de vecteur qui a recours à la décomposition en valeur singulière (SVD). Cependant, il y a une différence importante entre la LSI et le VSM, à savoir la LSI utilise une approximation de qualité inférieure pour la représentation de l'espace de vecteur de la base de données. C'est-à-dire, la TDM originale est remplacée par une autre matrice qui est assez semblable à la TDM originale mais dont l'espace de colonne est seulement un sous-espace de l'espace de colonne de la matrice originale. L'algorithme de SVD est employé dans la LSI pour réduire l'espace de vecteur, enlever le bruit ou la redondance lexicologique (qui sont illustrés dans la prochaine section) de la TDM, afin d'essayer de résoudre le problème d'inexactitude lié à la synonymie et à la polysémie.

Le LSI fournit clairement un espace de vecteur riche, qui exploite les rapports sémantiques latents entre les limites et les documents. La réduction de l'espace de vecteur a l'effet d'indiquer le rapport sémantique fondamental parmi les documents, parce qu'une grande partie du bruit dans la matrice est enlevé [28].

2.3 Bruit lexicologique

Après la création de la TDM, qui est une matrice bidimensionnelle représentant le nombre de fois un mot-clé apparaît dans chaque document de la base de données, la matrice résultante sera creuse, une grande proportion d'éléments seront des zéros, car chaque mot-clé apparaîtra seulement dans quelques documents. Les zéros dans la matrice représentent le bruit ou la redondance lexicologique dans la matrice à faible densité. Dans le système de LSI, l'algorithme de SVD est employé pour enlever ce bruit lexicologique dans la TDM originale, afin d'établir le rapport sémantique parmi les termes et les documents et d'essayer de surmonter les problèmes d'inexactitude liés au mot-clé traditionnel assortissant des méthodes d'IR.

On pourrait suggérer l'existence de trois (au moins) types de bruit dans le cadre de LSI qui sont :

- Bruit traditionnel par exemple les mots d'arrêt (stoplist) mentionnés en chapitre 1 (a, de, etc.). Ce type de bruit est habituellement traité et enlevé à l'étape de prétraitement de la base de données suivant les indications du chapitre 3.

- Bruit produit par la structure pauvre de la base de données ou du modèle de requête employé. Plus de détails sur les conséquences d'une telle structure sont donnés dans le chapitre 3. Les descriptions plus longues de document augmentent le nombre de mots-clés et la distribution des valeurs différentes de zéro dans la TDM, qui aide alternativement en améliorant la signification sémantique parmi les documents dans la base de données. D'une autre part, les descriptions plus courtes de document représentent la structure pauvre de la base de données qui ne soutient pas la technique de recherche par LSI, à mesure que la redondance dans la TDM augmente pour de tels types de structure.
- Nouveaux types de bruit produit par des *spammers* ou d'autres essayant d'éviter le filtre des systèmes sur des annonces.

2.4 Algorithmes de LSI

Comme mentionné dans l'introduction, une grande partie du travail existant sur le LSI s'est concentrée sur les étapes de prétraitement effectuées sur les bases de données, et sur les algorithmes de décomposition utilisés pour l'approximation de la TDM. Cette partie du chapitre présente ces étapes.

2.4.1 Prétraitement

Dans cette section, les secteurs noyau dans l'étape de prétraitement à savoir : l'identification et l'élimination des mots d'arrêt, steamming algorithm, pondération de termes, contrôle de pertinence et l'entretien de la base de données (mise à jour) sont présentés.

- **StopWords (mot d'arrêt):** La recherche dans ce secteur a été centrée sur ce qui constitue les mots-clés dans une base de données, qui décrivent la base de données et sont employés comme références aux titres de documents. La règle d'analyse employée par la plupart des chercheurs [3] [6] a exigé que les mots-clés apparaissent dans plus d'un document mais n'apparaît pas dans tous les documents. Des termes qui sont présents dans seulement un document, ou bien dans tous les documents devraient être supprimés car elles ont peu ou pas de capacité d'améliorer la signification sémantique parmi les documents de la TDM. Le but de ce travail a été d'extraire les termes qui ont une signification et d'enlever la ponctuation, les adjectifs, et les mots qui sont considéré n'avoir aucune signification, par exemple « et », « ou », « dans ».

donc tous les termes se produisant dans plus d'un document et qui ne font pas partie de la liste des mots d'arrêt (stopwords) seront inclus. La liste de mots d'arrêt construite par Fox [20] a été largement acceptée comme norme pour identifier les mots non-significatifs qui peuvent être éliminés d'une liste de mots-clés. Le processus d'exclure de tels mots à haute fréquence et sans signification est connu comme stoplisting [1].

- **Stemming Algorithm** : la première publication sur le stemming algorithm fût en 1968 [31]. Tandis que celui qui en a le plus parlé était Porter (1980) [32]. Une quantité de travail considérable a été consacrée à produire des algorithmes de provenance efficaces pour IR [33] [34] [35]. Un algorithme de provenance « décompose » des mots en tiges, par exemple les mots-clés « voyage », « voyageur », « voyager », peuvent tout « être décomposés » en tige « voyage ». Cette tige peut alors être employée comme un mot-clé plutôt que devoir stocké les trois mots-clés séparément. En conséquence, la provenance réduit le stockage exigé pour tenir des mots-clés en réduisant le nombre de mots-clés à tenir [1]. L'idée principale derrière la provenance est que les utilisateurs recherchant des informations sur le mot « recherche » seront également intéressés par les articles sur : recherché, recherchant, ainsi de suite [26]. Cependant, l'utilité de la provenance pour améliorer la qualité de recherche a toujours été remise en cause dans la communauté de la recherche [6] [26]. Car dans beaucoup de cas elle pourrait mener à une information non pertinente qui cause une récupération pauvre, ne correspondant pas à la requête et l'utilisateur peut être considérablement ennuyé [26] [34] [35]. Par conséquent, beaucoup de chercheurs ont évité l'utilisation du Stemming Algorithm dans leur travail sur LSI [3] [22] [17], particulièrement depuis que les mémoires sont disponibles de nos jours et qui possèdent suffisamment de volume pour enlever ces soucis de stockage.

- **Pondération de terme :** Comme avec le Stemming Algorithm, il existe un corps de travail considérable sur la pondération de terme [3] [36] [22] [37] [38]. La pondération de terme est l'une des méthodes communes pour améliorer la performance de la recherche. Elle consiste à donner des poids aux différents termes de TDM. Dans la pratique, des poids locaux et globaux sont appliqués pour augmenter, ou diminuter, l'importance des termes dans les documents de la TDM. Les poids globaux reflètent l'importance d'un terme dans tous les documents d'une base de données. les poids locaux reflètent quant à eux l'importance d'un terme dans un document donné.

Quelques chercheurs ne tiennent compte d'aucune pondération de terme dans la TDM et emploient un modèle non pondéré simple de TDM dans leur recherche [6].

Dans le VSM classique proposé par Salton et d'autres [27] le poids des termes dans le document correspondant dans la TDM est le produit des poids locaux et globaux. Le vecteur de poids pour le document d est : Vd = $[w_{1,d}\ ;\ w_{2,d}; \ldots\ldots;\ w_{n,d}]^T$

Où :

$w_{n,d}=tf_t.log\frac{|D|}{|\{t\in d\}|}$

- tf_t est la fréquence du terme t dans le document d (un paramètre local).
- $log\frac{|D|}{|\{t\in d\}|}$ est la fréquence inverse de document (paramètre global). $|D|$ est le nombre total de document dans la base de données ; $|\{t \in d\}|$ est le nombre de document contenant le terme t.

Dans un VSM simple les poids de terme n'incluent pas le paramètre global. Les poids utilisés dans la TDM sont juste les occurrences de terme (paramètre local) :

$$w_{n,d} = tf_t$$

✓ Contrôle de pertinence : Ce processus peut être identifié comme procédé commandé ou automatique pour la reformulation de la requête [39]. Souvent, les utilisateurs ne rechercheront pas tous les documents appropriés à la première tentative, ceci est dû à la requête pauvre en information, qui n'exprime pas exactement ce que les utilisateurs recherchent. La recherche peut être améliorée en rassemblant la rétroaction d'utilisateur sur la pertinence des documents renvoyés. Fondamentalement le processus est comme suit :

- Après que des résultats préliminaires de recherche soient présentés, permettre à l'utilisateur de fournir la rétroaction sur la pertinence des documents recherchés.
- Utiliser cette information de rétroaction pour reformuler la requête.
- Produire de nouveaux résultants bases sur la reformulation de la requête.

La requête reformulée sur la base de la rétroaction est générée comme suit :

- Expansion ou reformulation de la requête : Ajouter les nouveaux termes des documents pertinents à la requête.
- Repondération des termes : Augmenter le poids des termes des documents pertinents et diminuer le poids des termes des documents non pertinents [1] [40].

Dans d'autres systèmes de LSI une méthode différente pour la reformulation de requête est adoptée. L'information d'utilisateur pour le contrôle de pertinence est employée pour formuler une nouvelle requête en ajoutant les vecteurs des documents appropriés au vecteur de la requête, qui peut être regardé comme « somme » des documents appropriés de la première requête [3]. Ou une autre méthode qui consiste à soustraire les vecteurs des documents non pertinents du vecteur de la requête.

Figure. 6 (obtenue à partir d'un site Web de cours d'IR à l'Université du Texas à Austin [40]) dépeint le procédé de contrôle de pertinence dans les systèmes d'IR. Le travail décrit dedans examine et évalue l'utilisation de cette technique dans des systèmes de recherche. Le contrôle de pertinence fournit beaucoup d'avantages à la recherche. Le succès de ces méthodes est que beaucoup de mots (ceux des documents appropriés) enrichie la requête initiale qui est habituellement tout à fait appauvrie [3]. Mais il est important de noter que, dans les grandes collections de document, le jugement des documents à chaque requête induit un coût élevé de calcul.

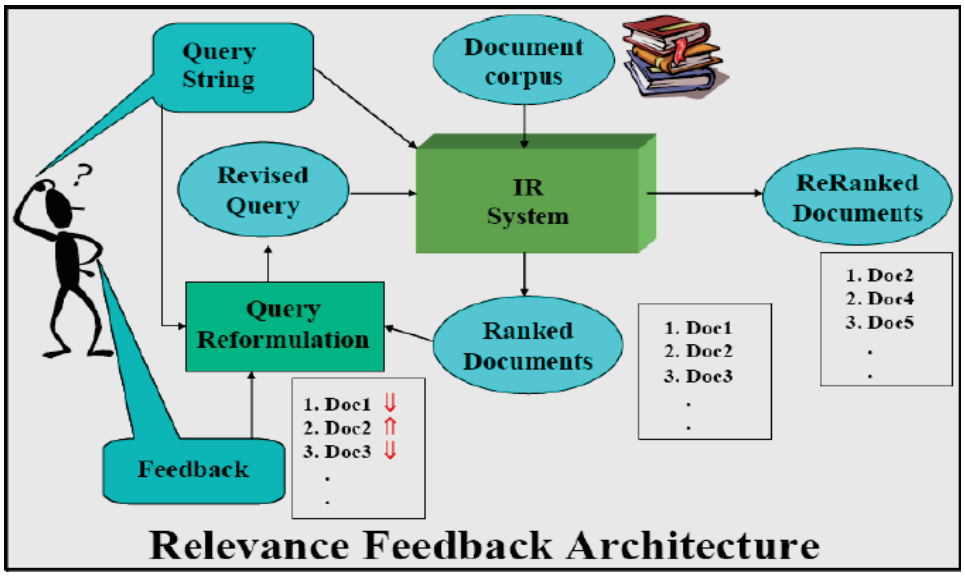


Figure 6 : Représentation de contrôle de pertinence [40]

- ✓ Mise à jour : Il est probable que les bases de données doivent être modifiées. L'information est continuellement ajoutée ou enlevée. Dans un système de LSI, l'approche standard qui consiste à faire des additions (de nouveaux termes ou documents) est de réappliquer la SVD sur la nouvelle TDM, cette approche comporte de nombreux calculs, particulièrement pour de grandes bases de données. Dans l'effort d'éviter ce coût de calcul élevé, d'autres techniques ont été considérées, par exemple folding-in et SVD updating. Celles-ci sont discutées dans [3] [6] [4] et [41]. folding-in n'est pas cher en termes informatiques mais elle a comme conséquence une représentation imprécise de la base de données. On lui recommande que des documents soient pliés de temps en temps. SVD updating est plus chère en termes informatiques mais elle a l'avantage significatif de préserver la représentation de la base de données [1].

2.4.2 Décomposition de Matrice

La recherche s'est déplacée au delà des fondations du procédé de LSI, et un grand nombre de recherches a été effectué dans le but d'accélérer le procédé de LSI. Beaucoup d'articles ont décrit les composants du système de LSI en détail en présentant l'essai empirique afin d'améliorer l'arrangement du LSI pour IR [42] [17]. Les questions clés dans l'exécution de LSI ont été identifiées et discutées par Telcordia [42].

Dans ce travail séminal, les points suivants ont été soulevés:

- ✓ Les questions de mise en œuvre de la LSI dans la pratique ont été discutées en décrivant les composants fonctionnels de la LSI. En particulier, les problèmes d'évolutivité dans les différentes composantes du système ont été abordés.

- ✓ Le travail s'est concentré sur le calcul couteux des tâches informatiques dans le système de la LSI, en donnant des suggestions pour son exécution afin de réduire le coût de calcul. Les issues proposées d'exécution peuvent simplement être récapitulées à un certain nombre de points : prêtant une attention particulière à l'exécution de la requête en se concentrant sur le temps de réponse de la recherche (le temps qu'il prend pour répondre à une requête), présentant des techniques pratiques d'exécution pour réduire des frais généraux de recherche en fournissant différentes méthodes pour améliorer la vitesse de recherche. Tandis que le travail fournit une illustration

intensive importante pour la LSI, il a quelques inconvénients comme énuméré dans les sections suivantes.

Le travail vise à répondre aux questions essentielles sur les performances de la LSI en exécutant un certain nombre de tests empiriques qui traitent de nombreux problèmes dans la LSI, par exemple le choix d'un rang optimal pour la SVD et la distinction entre un document pertinent et non-pertinent. La recherche fournit une bonne compréhension du processus de LSI, cependant, ces résultats ont été quelque peu insatisfaisants et les questions importantes mises en évidence dans le travail restent des questions ouvertes ou sans réponses dans la RI.

Par exemple, le seuil à utiliser en vue d'identifier les documents pertinents dépend de l'application.

Plusieurs algorithmes de décomposition alternative à la SVD ont été proposés, y compris la factorisation QR [3] [1] [43] et semi décomposition discrète de la matrice (SDD) [44]. Décomposition QR consiste à identifier et à ignorer les dépendances dans les colonnes de TDM qui n'apportent pas de nouvelles idées pour les documents dans la base de données. Bien que le QR soit plus simple que la SVD, elle n'a pas été utilisée dans les méthodes IR, car l'algorithme de SVD est plus puissant en termes de volume de résultats trouvés [1].

La décomposition SDD [45], développé pour être utilisé dans la compression d'image, a une fonction de base similaire à la SVD dans son approximation de la matrice. Cependant, dans la décomposition SDD les *m* vecteurs et les *n* vecteurs (où *m* et *n* représentent le nombre de termes et de documents, respectivement) sont limitées aux entiers 1, 0, -1 [44]. Tel que revendiqué, l'algorithme SDD renvoie les mêmes résultats que l'algorithme de SVD et présente les avantages de l'utilisation de moins d'un vingtième de l'entreposage et seulement la moitié du temps de la requête. Cependant, l'algorithme SDD a l'inconvénient de prendre cinq fois plus de temps à calculer que la SVD.

Dans les opérateurs unitaires sur l'espace de document [46], Hoenkamp montre que la décomposition sous-jacente LSI est un exemple d'un opérateur unitaire. Hoenkamp propose l'utilisation de la transformée de Haar (HWT) comme une alternative pour son coût de calcul nettement moindre. Cet axe de recherche a montré des résultats prometteurs. En outre, la notion de représentation de la TDM comme une image au niveau de gris, comme illustré sur la Figure. 7 (La base de données Cochrane contient des titres d'études médicales [47]) a été postulée. Dans ce modèle, les points blancs dans l'image (valeurs non nulles) représentent les

mots-clés dans les ensembles de documents. En outre, il a fait savoir que l'aide de la HWT pour supprimer le bruit d'une image est équivalente à l'aide de la HWT pour supprimer le bruit lexical de la TDM. Cependant, c'est un travail théorique qui doit être prouvé dans la pratique.

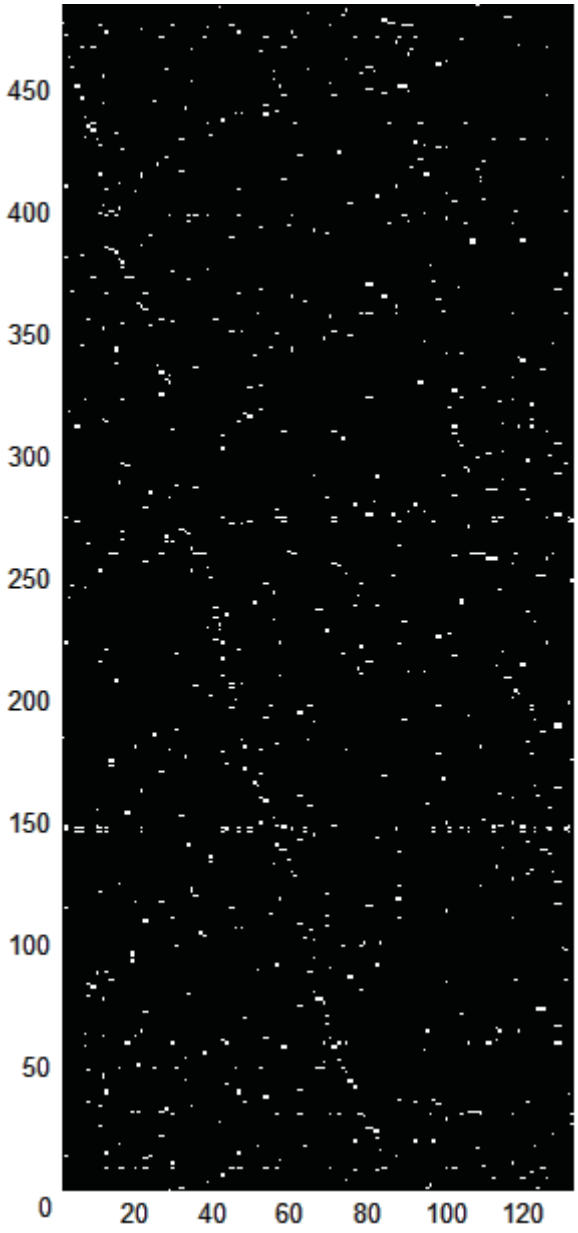


Figure 7 - TDM Cochrane représentée comme une image en niveaux de gris

2.5 Application de la LSI :

La LSI est évidemment employé dans la recherche d'information, mais la richesse d'autres applications qui emploient la LSI démontre combien cette technique est efficace, et que le travail aura un impact plus large dans le futur. Cette section présente une vue d'ensemble des diverses applications dans différents secteurs qui emploient les techniques de la LSI. Peut-être les applications les plus étonnantes de la recherche de LSI ont été dans les domaines autres que l'IR.

La SVD a été employée avec des algorithmes de filigranage pour résoudre le problème de la protection de copyright des documents multimédia [55] et comme technique robuste et stable dans des applications à traitement d'images. La réduction du rang a été employée dans la cryptographie [56] et dans le traitement d'images [7] [57] [58]. De plus, comme le montre notre travail, la LSI est la méthode révolutionnaire dans l'IR, Les résultats expérimentaux prouvent que la LSI, avec les configurations textuelles et visuelles, a la capacité d'identifier le concept sémantique fondamental des documents d'une base de données, ayant pour résultat l'amélioration de l'exécution de récupération.

D'autres chercheurs ont employé la LSI dans le domaine de la récupération d'image [8] [9] [59] [60]. Traditionnellement les images sont stockées dans de grandes bases de données. On suggère l'utilisation du LSI pour extraire les images requises, Les techniques de récupération d'image accèdent habituellement aux images en se basant sur leurs contenus, c'est plus efficace que de rechercher manuellement dans les grandes bases de données. Pour les travaux récents, une matrice visuelle de mot-clés image est créée, alors la LSI est employée pour découvrir le rapport sémantique entre les mots-clés et les images visuelles, afin d'améliorer le procédé de récupération.

Le concept du LSI a été également employé dans les systèmes de récupération audio et vidéo [10] [61] [11] [62] [63]. Dans les applications audio, le jet audio est converti en jet de texte par un système de reconnaissance de la parole. Alors le texte de chaque partie du discours est représenté dans un vecteur de document qui est la somme des mots que contient la parole. On peut aussi utiliser la structure sémantique appropriée à base de données qui peut être obtenue par LSI, afin de réduire l'effet du bruit produit des erreurs de reconnaissance de la parole.

Dans les applications vidéo, les séquences vidéo sont décomposées en contenu visuel (représentant les séquences vidéo) et mots (décrivant le contenu visuel) pour former une matrice. Alors la LSI est employée pour déterminer les rapports entre les mots et le contenu visuel selon les Co-occurrences des mots dans le contenu dans la matrice. Comme technique, la LSI peut modeler le contenu visuel, réduire le bruit et augmenter l'information de Co-occurrence.

LSI n'est pas spécifique à la langue anglaise, il n'utilise pas la syntaxe ou la sémantique étendue de l'anglais mais relève plutôt des mots dans la base de données. À cet égard LSI peut être appliquée à n'importe quelle langue. Les principes de LSI ont été appliqués à la recherche inter-langues. Avec l'explosion de l'Internet et les réseaux distribués, il ya de nombreuses collections de documents qui existent dans plusieurs langues.

Dans la recherche inter-langue, en saisissant une requête dans une langue, la LSI peut être utilisée pour retourner les documents dans une autre. Ce qui est requis pour les applications inter-langues, un espace commun dans lequel les mots de plusieurs langues sont représentés [3]. Dans [12], une méthode de recherche documentaire automatisée totalement inter-langues, dans laquelle aucune traduction de la requête n'est requise, est décrite. Les requêtes dans une langue peuvent récupérer des documents dans d'autres langues. Ceci est accompli par une méthode qui construit automatiquement un espace multi-langue sémantique en utilisant la LSI. L'analyse sémantique latente inter-langue a été utilisée pour développer une représentation de faible dimension constituée de mots et de documents dans plusieurs langues.

Inconvénients des travaux existants :

Comme on peut le voir dans les sections précédentes, il reste encore beaucoup de possibilités pour davantage de recherche dans l'amélioration de la performance du système de LSI. Les principales limites du travail existant peuvent être identifiées comme suit:

- Ces dernières années, le volume de recherches sur les étapes de prétraitement effectuée sur les bases de données, devient faible en comparaison à la recherche sur les autres phases, la plupart des travaux existants sur l'étape de prétraitement ont été largement accepté par la plupart des recherches. En outre, des outils suggérés pour l'amélioration de la recherche ont été proposés. Seulement, sur certaine bases de

données, ils réalisent de petites améliorations sur les résultats de la recherche, alors que dans d'autres bases de données ils rendront la recherche pauvre. De plus, la technique de pertinence entraîne un coût de calcul élevé avec les grandes bases de données. En outre, certains chercheurs sont enclins à tester leurs méthodes sans utiliser d'outils.

- L'amélioration des résultats pour nombreuses approches du LSI, en collaboration avec d'autres techniques, a été négligeable, et il ya de nombreux inconvénients et faiblesses qui peuvent être identifiés. En Telcordia LSI Engine, le travail n'a pas abordé la précision et le rappel de LSI comme des mesures standard pour l'efficacité de LSI, et utilise seulement le temps de réponse des requêtes. Un tel système métrique n'est pas suffisant pour les questions de mesure du rendement.

- Certains travaux peuvent être considérés comme simplement un test empirique pour LSI, offrant une bonne perception des phases du système, ainsi que d'essayer de répondre à de nombreuses questions sans réponses pour LSI telles que la détermination de la meilleure valeur de k. Le résultat, comme l'indiquent les chercheurs, n'était pas satisfaisant. Tous les algorithmes de décomposition alternative à SVD qui ont été suggérées ont échoué, et la norme SVD basée LSI reste le moyen le plus efficace de chercher en termes de nombre de documents retournés.

- En termes d'applications de LSI, peu de lacunes peuvent être soulevées, car le succès de la technique de LSI dépend du milieu et des objectifs qu'on veut atteindre. Toutefois, dans les applications d'apprentissage, il est clair que LSI manque d'importantes capacités cognitives que les humains possèdent et utilisent pour appliquer les connaissances expérimentés.

- La plupart des recherches pour améliorer la performance de LSI ont été portées sur la complexité de l'étape de décomposition. très peu de travaux étudient l'amélioration de la précision des documents retournés.

Basé sur les limitations des travaux existants, et les principaux objectifs présentés dans le chapitre 1, le travail exposé dans cette thèse peut être résumée comme suit:

- Le système LSI, comme indiqué ci-dessus, peut former un domaine fertile de la recherche, nous allons évaluer empiriquement l'effet qu'engendre le changement du paramètre le plus important, le nombre de dimensions k extraite par SVD, sur la

performance de LSI. Comment déterminer les dimensions optimales de la SVD, afin de répondre à une question critique. En outre, l'utilisation d'un outil d'analyse SNR et mesure du bruit dans la TDM.

- L'importance de la structure de la base, comme un facteur clé dans l'amélioration des résultats de LSI, a été démontrée par le volume considérable de recherches menées sur cette question. Des recherches sont menées pour présenter et décrire la structure la plus efficace pour la base de données qui aideront à la modélisation du bruit lexicale et son enlèvement de la TDM, afin d'améliorer la recherche dans le système LSI.
- LSI, comme indiquée ci-dessus, a été utilisée en tandem avec d'autres techniques pour obtenir de meilleures performances dans les résultats. Une approche couramment utilisée dans le traitement de l'image est de combiner différentes techniques afin d'améliorer la réduction du bruit. La comparaison de la TDM à une image en niveaux de gris invite un traitement similaire. Une approche hybride, efficace et nouvelle de LSI pour une utilisation efficace en RI basée sur l'utilisation de techniques de traitement d'image en tandem avec les éléments existants seront présentées.

2.6 Conclusion

Ce chapitre résume les travaux existants sur le domaine de la recherche d'information, en particulier sur la LSI. Une brève introduction à l'architecture originale de VSM (l'origine du système LSI) a été présentée. Des algorithmes de décomposition existants, utilisés pour le rapprochement TDM ont également été mentionnés et examinés.

Un aperçu des différentes applications dans différents domaines ayant utilisé les techniques de LSI a été donné.

En outre, les limites du travail en vigueur ont été fixées. C'est le but du travail de recherche présenté dans ce projet de fin d'étude pour répondre aux limites présentées dans la section précédente grâce à une approche hybride d'analyse nouvelle. Une nouvelle méthodologie, basée sur l'utilisation de techniques de traitement d'image, sera étudiée dans le chapitre suivant, afin de faciliter l'analyse de grandes bases de données.

Chapitre III

Les ondelettes de Haar

3.1 Introduction

Comme cela a été mentionné précédemment, La LSI prend en charge de nombreux avantages sur les techniques traditionnelles de ciblage des mots clés. Toutefois, ces avantages ont un coût de calcul élevé. En dépit de l'augmentation des vitesses allant jusqu'à cent fois plus grâce à l'algorithme original du LSI [1], actuellement, même les meilleurs systèmes sont trop lents pour de nombreuses applications.

Dans Unitary Operators on the Document Space [4], Hoenkamp affirme que la propriété fondamentale de la SVD est son caractère unitaire, sa capacité à représenter l'espace document dans un nouvel espace de telle sorte que les documents connexes restent ensemble et les documents sans rapport éloignés, et cela avec un bruit généré très faible.

Il postule encore l'idée de la matrice de terme document comme une image en niveaux de gris, l'équivalence de l'utilisation de la décomposition de Haar pour supprimer le bruit lexical et en utilisant la décomposition de Haar pour supprimer le bruit d'une image.

Ce chapitre vise à étudier les ondelettes de Haar pour comprendre sa fusion avec la SVD et l'amélioration que nous pourrons avoir en combinant les deux techniques.

3.2 La Transformée en Ondelettes

Les deux approches mathématiques présentées dans les précédemment sont adaptées aux processus stationnaires. De nouvelles méthodes élaborées et mises au point ces dernières années, unifient et généralisent les idées et les pratiques développées précédemment et permettent d'analyser des signaux non-stationnaires. La Transformée en ondelettes fait partie de ces nouvelles méthodes, son principe est de décrire l'évolution temporelle d'un signal à différentes échelles de temps.

La théorie des ondelettes est apparue au début des années 1990 [44], elle touche de nombreux domaines des mathématiques, notamment le traitement du signal et des images [45], [42].

Cette section présente un rapide aperçu des fondements théoriques des Ondelettes, pour aller plus loin sur cette théorie du traitement du signal à l'aide des Ondelettes, le lecteur pourra se reporter au livre de Mallat [43].

Malgré une origine aux nombreuses racines, on attribue le point de départ de l'utilisation des ondelettes au géophysicien Jean Morlet, qui envisageait de les utiliser pour l'analyse de sismogrammes utilisés dans la recherche de pétrole sous terre.

Dans la transformation par Ondelettes, comme dans l'analyse de Fourier, en cherche à transformer un signal quelconque en une série de nombres que l'on pourra ensuite utiliser pour reconstruire au mieux le signal d'origine. Cependant dans la transformation par Ondelettes, on utilise plusieurs niveaux de résolution pour examiner le signal et faire ressortir les différentes variations.

L'analyse multi résolution donne un ensemble de signaux d'approximation et de détails d'un signal de départ en suivant une approche fin-à-grossier. On obtient une décomposition multi-échelle du signal de départ en séparant à chaque niveau de résolution les basses fréquences (approximation) et les hautes fréquences (détails) du signal.

Cette approche à un sens quand le signal à des composantes à haute fréquence pour des courtes durées et des composantes de basse fréquence pour de longues durées. Pour accomplir une telle tache une Ondelette sera employée au lieu d'une fonction de fenêtrage, la Transformée en ondelettes est capable de fournir les informations de temps et de fréquence simultanément et donc une représentation temps – fréquence du signal.

3.2.1 Définition

La Transformée en ondelettes est une représentation multi-résolutions, qui exprime les variations d'un signal à différentes résolutions. Une Ondelette ψ est une fonction oscillante, comme les fonctions sinus et cosinus, mais localisée. Cela se traduit par le fait qu'elle est intégrable de valeur 0.

$$\int_{-\infty}^{+\infty} \psi(x)dx = 0$$

Le signal est étudie aux échelles ½, ¼,...,2^j, avec $j \in Z$ et $j \leq -1$

3.2.2 L'Ondelette de Haar

L'Ondelette de Haar est l'Ondelette dont le support est le plus petit, cela implique que sa transformée du signal nécessitera le minimum d'espace de stockage.

1. Cas d'une dimension :

Soit *h* la fonction, dite de base de Haar, définie sur $\Re$ par :

$$h(x) = \begin{cases} 1 \text{ si } 0 < x < \frac{1}{2} \\ \\ 0 \text{ sinon} \end{cases}$$

Avec une période de Haar unitaire.

Supposons que nous avons un signal définit sur l'intervalle [0,1]. Pour avoir une approximation discrète du signal nous allons calculer ses valeurs dans deux points, quatre points, huit points et ainsi de suite ; le diviser en deux fonctions, de 0 à 1 / 2 et de 1 / 2 à1, puis en quatre fonctions, de 0 à 1 / 4, de 1 / 4 à 1 / 2, de 1 / 2 à 3 / 4, et de 3 / 4 à 1 etc.…

On obtient différentes résolutions et pour chacune on peut avoir une représentation dans l'espace des fonctions à l'aide d'un système de fonctions de base, nommées fonctions de base multi-résolutions ou multi-échelles.

Les Ondelettes sont des fonctions de base multi - échelles qui assurent le passage cohérent entre les différentes résolutions, la décomposition et la reconstitution de la fonction représentée. Si on utilise les Ondelettes comme système de fonctions de base, à chaque niveau on dispose des approximations (moyennes) de la fonction initiale et des informations de détails.

3.2.3 Exemple de calcul

La transformée de Haar de la fonction f(x) = $[y_1\ y_2\ y_3 \dots \dots y_n]$ génère :

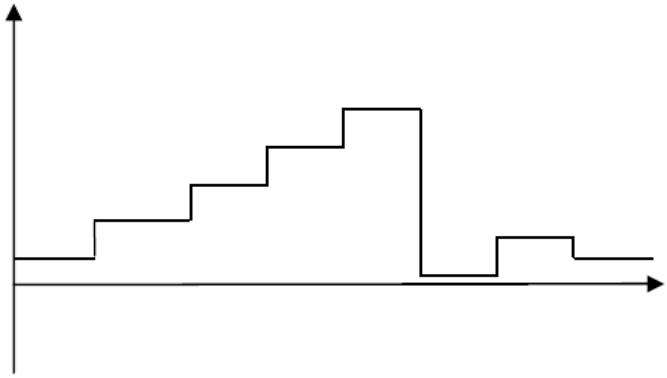
- Des approximations $[a_1\ a_2\ a_3 \dots \dots a_{n/2}]$ qui sont les moyennes des valeurs initiales de la fonction prisent deux par deux $a_1 = (y_1 + y_2)/2$ …

- Coefficients de détail ou les différences$[d_1\ d_2\ d_3 \dots \dots d_{n/2}]$,

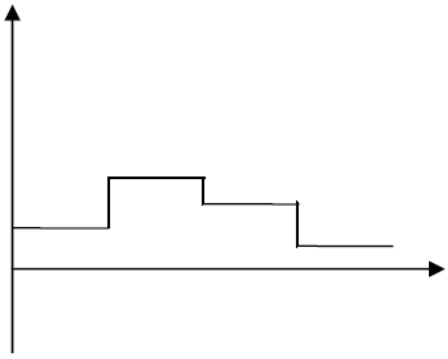
Avec : $d_1 = y_1 - a_1$; $d_2 = y_3 - a_2$

Considérons un signal monodimensionnel composé de quatre échantillons

S = [2 4 8 12 14 0 2 1]



Pour calculer sa transformée de Haar, moyennons d'abord les paires de valeurs voisines pour obtenir [3 10 7 1.5]



Afin de récupérer le signal initial nous devons également enregistrer d'autres valeurs représentant la perte d'information.

[-1 -2 7 0.5]

Le signal peut donc être représenté par sa résolution inférieure et le signal de détail.

En appliquant ce procédé, récursivement sur le signal on aboutit à sa transformée de Haar, à la fin signal est représenté par un seul coefficient de moyenne du signal et l'ensemble de coefficients des signaux de détails successifs.

Résolution	Moyenne	Détails
8	[2 4 8 12 14 0 2 1]	
4	[3 10 7 1.5]	[-1 -2 7 0.5]
2	[6.5 4.25]	[-3.5 2.75]
1	[5.375]	[1.125]

Tableau 3. Transformée de Haar du signal S

Observons la transformée de Haar ainsi obtenue, en plus du coefficient de moyenne du signal, les coefficients de détails expriment les variations du signal aux différentes résolutions. A une même échelle, plus le coefficient est grand en valeur absolue, plus ces variations sont importantes. Le signal original sera présenté par :

[5.375 1.125 -3.5 2.75 -1 -2 7 0.5]

2. Cas de deux dimensions :

Il ya un certain nombre de façons d'appliquer la décomposition de Haar à une structure à deux dimensions (matrice). La méthode utilisée dans le présent document est la méthode standard qui consiste à appliquer en premier lieu la décomposition de Haar à toutes les lignes de la matrice puis d'appliquer la décomposition à toutes les colonnes de la matrice résultante.

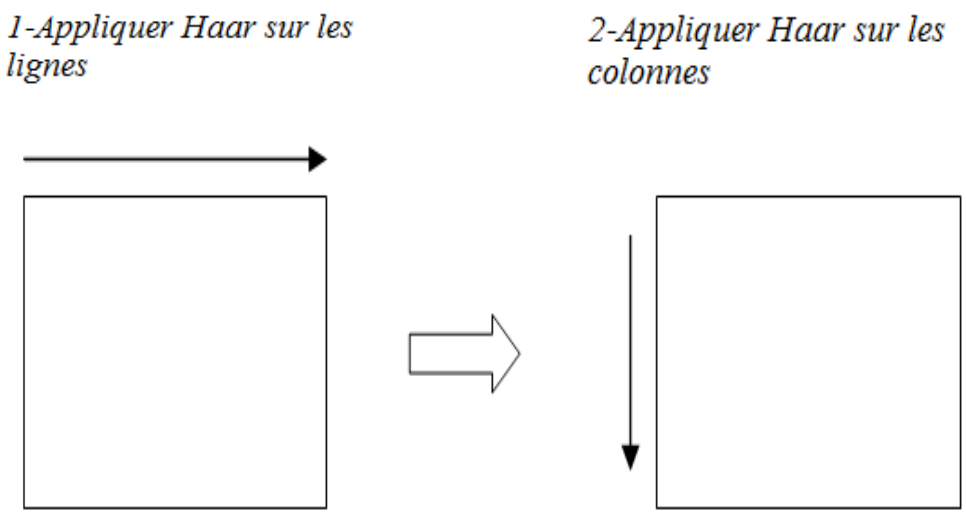


Figure 8 : Décomposition de Haar d'une matrice

3.2.4 Le débruitage

Le bruit est une erreur aléatoire d'une variable mesurée. Il existe plusieurs raisons possibles pour générer des données bruitées, tel que les erreurs de mesures pendant l'acquisition des données, les erreurs humaines, ou les erreurs de machines lors de la saisie des données. On peut définir, donc, le débruitage comme le processus d'identification des données optimales parmi les données bruitées disponibles.

La réduction du bruit à l'aide de Haar peut être réalisée par une variété de systèmes de seuillage. Dans ce document la méthode du seuil dur sera utilisée. Cela mettra toutes les valeurs en dessous du seuil à zéro. Par exemple, si le vecteur (1, 10, 4, 6) possède un seuil de 5 dur, le résultat est (0, 10, 0, 6).

Analyse :

Figure 9: Une image et la décomposition de premier niveau de Haar de l'image.

Comment peut-on expliquer cela? Si nous revenons à l'analogie Hoenkamp de la matrice terme document comme une image en niveaux de gris, nous pouvons faire la lumière sur le processus. En traitement d'images, la décomposition de Haar peut être utilisée pour révéler la structure d'une image; les différents niveaux de résolution montreront les différentes caractéristiques de l'image: la structure de pointe, détails de fond etc. Ceci est illustré dans la figure3.1.

Si on considère la matrice termes documents comme une image, alors les mêmes règles doivent s'appliquer.

Une autre considération est la valeur des pixels dans une image en gris, comparée à l'image de la matrice de termes documents. Une image de 8 bits en échelle de gris aura des valeurs comprises entre 0 et 255, et est susceptible d'avoir une bonne répartition des valeurs

dans cette plage. Dans une matrice terme document, la plupart des valeurs sont 0 ou 1. La réduction du bruit à l'aide de Haar exploite la redondance en remplaçant les valeurs proches de zéro par zéro après l'application de la décomposition de Haar. Toutefois, avant même la décomposition de Haar, la plupart des valeurs dans une matrice de termes documents sont déjà proches de zéro, d'où le besoin de bien choisir le seuil pour une bonne réduction du bruit sans pour autant causer du dommage à la matrice.

Après cette brève étude des ondelettes de Haar, on constate que l'application de cette technique est en mesure d'améliorer le processus de LSI. Tant que la décomposition de Haar [9] est utilisée dans le domaine de la réduction du bruit de l'image. L'approche combinée SVD-Haar pourrait potentiellement avoir un effet positif en comparaison d'application de la SVD seule. Une étude complémentaire est donc proposée.

3.3 Étude proposée :

L'étude vise à examiner l'effet de la décomposition de Haar comme une étape post-traitement sur le processus standard de LSI.

Aperçu du processus de révision :

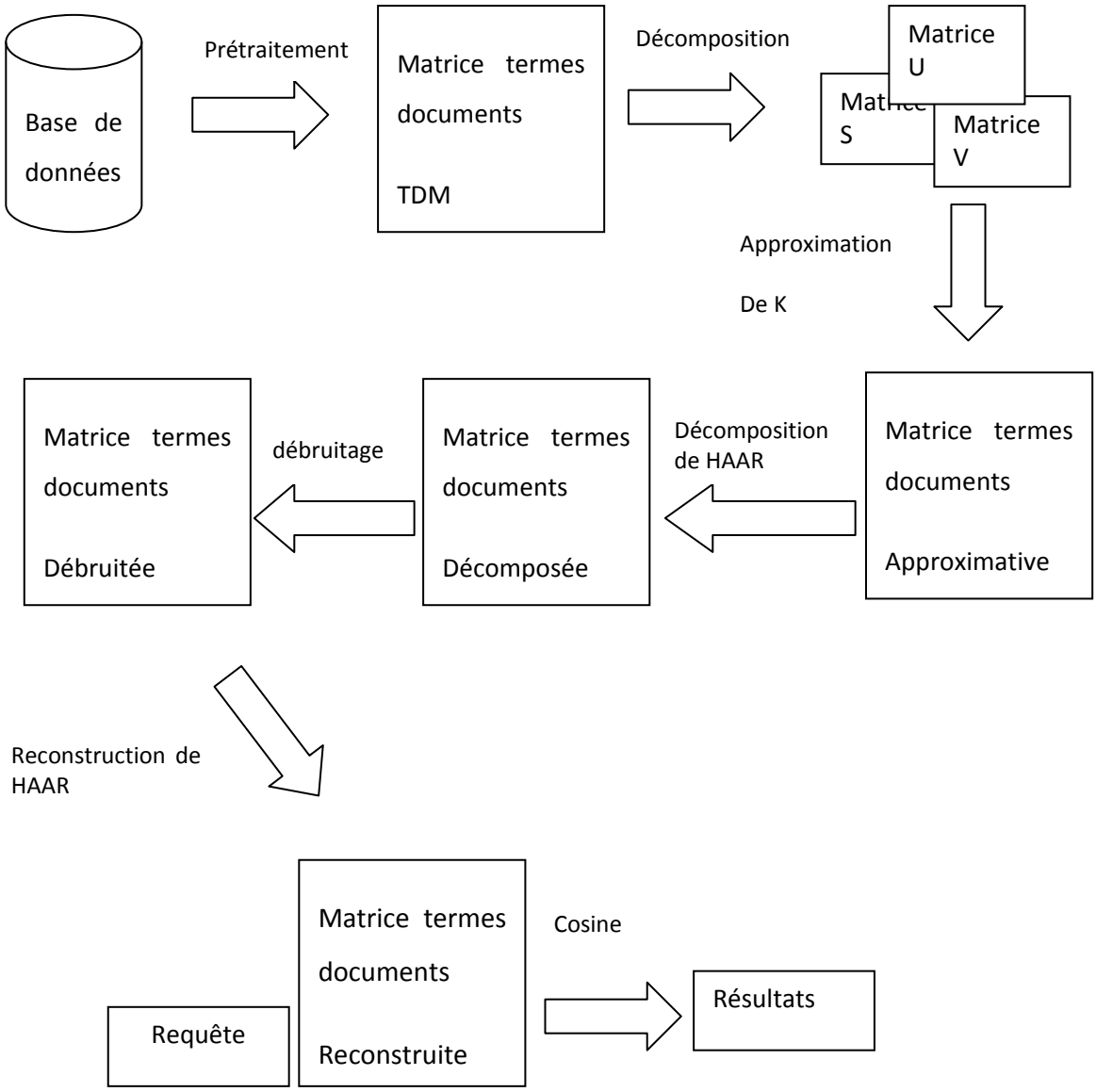


Figure 10 Processus révisé

3.4 Conclusion :

Une approche couramment utilisée dans le traitement de l'image est de combiner différentes techniques afin d'améliorer la réduction du bruit, Le but de la recherche présentée dans ce chapitre est de développer une nouvelle approche de l'indexation sémantique latente (LSI) dans le cadre de recherche de documents texte basé sur des techniques de traitement d'image, ceci grâce à la combinaison SVD-HAAR.

L'étude et l'analyse des résultats dus à cette fusion sont présentées dans le chapitre suivant.

Chapitre 4

Étude expérimentale et analyse des résultats

4.1 Introduction

Dans ce chapitre, une nouvelle approche pour l'analyse lexicale du bruit dans le processus de l'indexation sémantique latente (LSI) est présentée. Cette approche, basée sur l'utilisation d'outils de traitement d'image, est considérée comme une nouvelle philosophie pour la mesure et l'analyse de LSI en recherche d'information (RI). Une étude sur la catégorisation de textes est proposée, dans laquelle, différentes caractéristiques des bases de données peuvent être utilisées pour améliorer la recherche.

Afin de fournir une base claire pour expliquer les étapes impliquées dans le modèle proposé, les étapes de mise en œuvre d'un système standard de LSI sont examinées en premier dans la section 4.2. Un aperçu de l'architecture et l'application des algorithmes classiques de décomposition de matrice est fournie, ainsi qu'une explication des méthodes utilisées pour générer les résultats et les outils de mesure utilisés pour l'évaluation des performances des systèmes. Les sections 4.3 et 4.4 présentent la nouvelle approche pour l'analyse lexicale. A la section 4.5 on présente brièvement l'interface graphique du projet. La conclusion de ce chapitre est donnée dans la section 4.6.

4.2 Les composants du système LSI

Dans cette section, tous les éléments associés à la phase de traitement du système LSI sont clarifiées. Traditionnellement LSI est mis en œuvre en plusieurs étapes [3] [6] [16]. La première étape consiste à prétraiter les données de documents. Ce processus comprend la suppression de tous les termes de ponctuation et les «stop words» comme *the, as, and*, etc, c'est à dire ceux qui n'ont pas de sens sémantique distinctif dans un document.

L'étape suivante consiste à construire une matrice terme document (TDM), qui représente la relation entre les documents et les mots qu'ils contiennent dans la base de données. Un algorithme de matrice convenable de décomposition est alors utilisé pour décomposer la TDM, afin d'éliminer le bruit dans la matrice, en réduisant la dimensionnalité de la TDM.

L'algorithme de décomposition initial proposé par Berry [3] [18] et al, et de loin le plus largement utilisé, est la décomposition en valeurs singulières (SVD) [4] [19] [20]. La décomposition est utilisée pour éliminer, ou au pire de réduire le bruit (représenté par rareté) de la matrice. Elle fonctionne en réduisant la dimensionnalité des TDM facilitant ainsi la détermination de la relation sémantique entre les termes et les documents dans le système. Un avantage supplémentaire est que cette approche favorise l'élimination de la polysémie et de synonymie. Wiemer-Hastings montre que le pouvoir de LSI vient principalement de l'algorithme SVD [21]. Choisir un paramètre optimal de réduction de dimensionnalité (k) reste insaisissable. Traditionnellement, le k optimal est choisi par l'exécution d'un ensemble de requêtes avec des ensembles connus de documents pertinents pour plusieurs valeurs de k [22]. La valeur de k qui retourne les meilleurs résultats est choisie comme le k optimal pour chaque collection. Enfin, l'ensemble de documents est comparé avec la requête et les documents qui sont les plus proches de la requête de l'utilisateur sont retournés.

4.2.1 Description de la base de données

Dans cette recherche, la base de données contient un ensemble de titres de documents sur lequel la recherche est effectuée. Cette section décrit la structure et le contenu des bases de données utilisées dans ce travail.

Structure de la base :

La base de données utilisée est un tableau dans Microsoft Access. Le tableau est sous la forme: ID, Titre; l'onglet « Titre» contient les titre des documents à partir duquel les mots-clés sont générés. L'onglet «ID» permet de référencier les documents.

Contenu de la base de données :

Les documents utilisés dans les expériences sont organisés comme un ensemble de deux bases de données. La base de données *Mémos* très petite, elle est souvent utilisée comme un exemple de travail dans de nombreux articles traitant la LSI [3] [16]. Une telle structure s'est avérée utile pour tracer les grands principes de LSI. Nous avons inclus cette base de données dans notre étude pour fournir une référence de base.

La base de données Cochrane est une petite base de 135 documents contenant les titres d'études médicales dans l'administration du médicament qui est un autre système de test couramment utilisé dans la littérature LSI. Il peut être trouvé sur le site de Cochrane [47].

4.2.2 Description de prétraitement de documents

La table des documents de la base de données doit être convertie en une TDM. Avant que cela puisse être réalisé, un prétraitement doit être effectué sur l'ensemble des documents. Comme mentionné précédemment, la ponctuation et les mots sans signification (stop list) doivent être supprimés, et les mots-clés nécessaires à la construction doivent être extraits pour la comparaison avec la requête de l'utilisateur [3]. Initialement tout le texte de la base de données est extrait, et réunis pour former une grande collection de termes qui apparaissent dans chacun des documents. Cette liste est ensuite traitée par:

- Suppression des caractères de ponctuation "0123456789.,;:() [] etc:«, ceux-ci ne contribuent pas à la signification des termes dans les documents.
- Surpression des «mots vides». Ce sont des mots de la stop-list qui n'ont aucun sens pour la recherche et donc ne représentent en rien la structure sémantique des documents. Des exemples de «mots vides» sont : les, probablement, toutefois, etc.
- La prochaine étape de prétraitement consiste à retirer les mots en double dans la liste des mots clés. Parce que tous les mots ont été extraits de chaque entrée dans la table, beaucoup de mots apparaissent plus d'une fois et doivent être enlevés. Ce résultat est obtenu en triant tous les termes dans la liste des mots-clés par ordre alphabétique, donc tous les mots répétés sont adjacents dans la liste. Une fonction récursive peut être utilisé pour comparer chaque paire adjacente de mots dans la liste des mots clés, et si elles sont les mêmes, le terme dupliqué est supprimé.
- Enfin, une liste qui comprend tous les mots-clés de l'ensemble des documents est obtenue, avec une liste de mots-clés dans chaque document.

Exemple base de données Mémos

Pour une bonne illustration de l'étape de prétraitement, on utilise la base de données *mémos* comme exemple [6]. Les titres de la base *mémos* sont présentés dans le tableau 4:

1	human computer interface for ABC computer applications
2	a survey of user opinion of computer system response time
3	the EPS user interface management system
4	system and human system engineering testing of EPS
5	relation of user perceived response time to error measurement
6	the generation of random, binary and ordered trees
7	the intersection of paths in trees
8	graph minors IV: widths of trees and well-quasi ordering
9	graph minors: a survey

Tableau 4: Ensemble des documents de la base de données Memo [6]

Matrice terme document

Une fois le prétraitement terminé, la TDM est construite à partir d'une liste de termes qui caractérisent la structure de tous les documents et la liste des mots clés pour chaque document qui a été généré à l'étape précédente. Chaque rangée de la matrice est attribuée à un terme, et chaque colonne de la matrice est attribuée à un document. La valeur qui apparaît dans la position (i, j) de la matrice est le nombre de fois que le mot-clé attribué à la ligne i apparaît dans le document attribué à la colonne j. La plupart des valeurs dans la matrice sont nulles, seulement sous-ensemble de mots clés apparaît dans un document donné. Il est intéressant de voir la relation des termes dans les documents.

La TDM généré pour l'exemple Mémos est présentée au tableau 5.

	B1	B2	B3	B4	B5	B6	B7	B8	B9
Computer	2	1	0	0	0	0	0	0	0
Eps	0	0	1	1	0	0	0	0	0
Graph	0	0	0	0	0	0	0	1	1
Human	1	0	0	1	0	0	0	0	0
Interface	1	0	1	0	0	0	0	0	0
Minors	0	0	0	0	0	0	0	1	1
Response	0	1	0	0	1	0	0	0	0
Survey	0	1	0	0	0	0	0	0	1
System	0	1	1	2	0	0	0	0	0
Time	0	1	0	0	1	0	0	0	0
Trees	0	0	0	0	0	1	1	1	0
User	0	1	1	0	1	0	0	0	0

Tableau 5: TDM pour l'exemple Mémos [6]

Chaque colonne de la base de données peut être considérée comme un vecteur décrivant le document qu'elle représente, chaque ligne peut être considérée comme un vecteur décrivant le terme qu'elle représente. Les documents sont décrits en termes de mots clés qui les composent, et les mots clés sont exprimés en termes des documents dans lesquels ils apparaissent. Il est sans doute une grande partie de la redondance (bruit lexicales) dans ce processus, comme illustré par la faible densité de la matrice. Le processus LSI vise à éliminer cette redondance en décomposant la TDM en utilisant l'algorithme SVD.

4.2.4 Vecteur requête

Pour pouvoir mener la recherche, les requêtes doivent également être représentées sous forme vectorielle. Ceci est réalisé par le même procédé qui est utilisé pour convertir les documents en colonnes dans la TDM. Les mots-clés sont extraits de la requête, et si un mot-clé apparaît également dans le document figurant alors le nombre de fois qu'elle apparaît dans la requête est enregistré en utilisant le même format que l'un des vecteurs de documents dans la TDM.

4.2.5 Implémentations des algorithmes de décomposition matricielle

Comme mentionné précédemment, maintenant que la TDM a été généré, la décomposition de matrice peut être effectuée afin de générer un lien sémantique entres les termes et les documents de la TDM, Dans notre étude on utilisera l'algorithme de SVD.

Décomposition en valeurs singulières

Appelée aussi Singular Value Decomposition (SVD), cette technique consiste à projeter la matrice dans un espace de dimension plus faible où les descripteurs considérés ne sont plus de simples termes. Avec cette méthode, les termes apparaissant ensemble sont projetés sur la même dimension. Cette représentation est censée résoudre partiellement le problème de synonymie et de polysémie. Elle permet de trouver des documents pertinents pour une requête même s'ils ne partagent aucun mot avec cette requête. Grâce à une analyse statistique de grands corpus, le sens de chaque mot est caractérisé par un vecteur dans un espace de grandes dimensions, la proximité entre deux vecteurs correspondant à la proximité de sens de ces mots. Cette analyse statistique consiste à construire une matrice d'occurrences qui sera réduite afin de faire ressortir les relations sémantiques *«latentes »* entre mots ou entre textes.

En effet, deux mots peuvent être considérés sémantiquement proches s'ils sont utilisés dans des contextes similaires. Le contexte d'un mot est ici défini comme l'ensemble des mots qui apparaissent conjointement avec lui. Cette notion de co-occurrence est évidemment statistique, la méthode fonctionne si un nombre suffisant de textes est utilisé.

Cette approche permet donc de représenter les termes de la collection suivant la structure sémantique latente.

Le LSI utilise une matrice *X* (terme-document) qui est composée des vecteurs de termes et de documents. Elle utilise la technique de décomposition à valeur singulière afin d'approximer la matrice terme-document par des combinaisons linéaires et permet donc de créer un nouvel espace vectoriel :

$$X_{t*d} = U_{0_{t*m}} * S_{0_{m*m}} * V'_{0_{m*d}}$$

Où

– $U_{0_{t*m}}$ est la matrice orthogonale des vecteurs singuliers de gauche,

– $V_{0_{m*d}}$ est la matrice contenant les colonnes orthogonales des vecteurs singuliers de droite,

– $V'_{0_{m*d}}$ est la transposée de la matrice $V_{0_{m*d}}$,

– $S_{0_{m*m}}$ est la matrice diagonale (triée) des valeurs singulières.

– *t* est le nombre de lignes dans *X*, *d* est le nombre de colonnes dans *X* et *m* est le rang de *X* tel que ($m \leq min(t, d)$).

Il est prouvé qu'il existe une seule décomposition de cette manière.

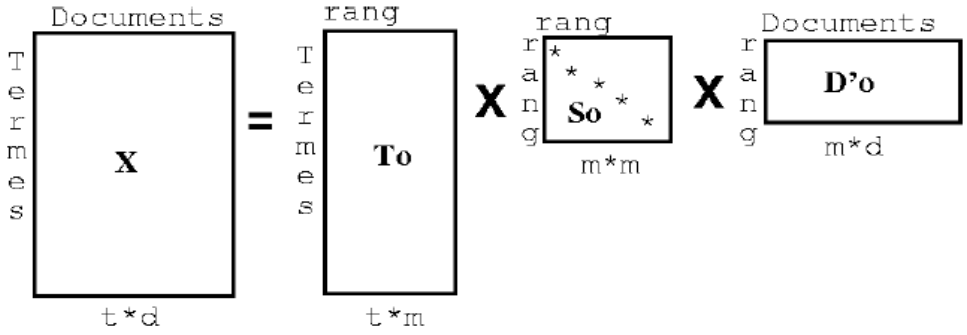


Figure. 11 – Représentation de la décomposition en valeurs singulières de la matrice X

Cette matrice est par la suite réduite par la matrice *Xh* contenant les plus grandes valeurs singulières *k* ($k \leq m$).

$$Xh_{k*d} = U_{0_{t*k}} * S_{0_{k*k}} * V'_{0_{k*d}}$$

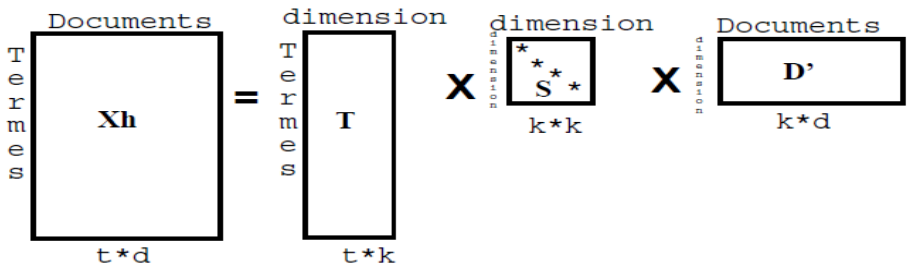


Figure. 12 – Réduction de la SVD de la matrice X

Xh ne garde que les *k* premières valeurs et permet donc de représenter les documents dans un espace de dimension *k*.

L'espace sémantique étant construit, la proximité sémantique entre deux mots est déterminée par le cosinus de leur angle.

Les documents qui n'ont pas servis à la phase de la création de la TDM sont ajoutés à cet espace réduit en approximant leur position suivant le vecteur contenant le vocabulaire qui le caractérise. Ce qui suppose que l'espace LSI créé au départ caractérise bien les dimensions importantes de similarité pour pouvoir approximer un nouveau terme ou un nouveau document dans la collection. Ce genre d'approche suppose que l'échantillon utilisé pour la mise en œuvre de la TDM est réellement représentatif de la collection de documents.

Le paramètre *k* est important à définir car une réduction à un espace de trop grande dimension ne ferait pas suffisamment émerger les liaisons sémantiques entre mots, et un trop petit nombre de dimensions conduirait à une trop grande perte d'informations. Le nombre adéquat de dimensions ne peut pas être actuellement déterminé théoriquement ; seuls des tests empiriques peuvent situer cette valeur qui varie d'une base à une autre.

De plus, les valeurs de la matrice après réduction ne sont pas interprétables (par les êtres humains).

4.2.6 Méthodologie des métriques

Cette section explique les méthodologies utilisées pour générer les résultats. Chaque colonne de la TDM représente un document mis en forme vectorielle comme le montre le tableau 6. Cela est également vrai pour la TDM approchée.

La requête est un vecteur ligne construite de telle sorte que sa transposée peut être considéré comme équivalent à un vecteur contenant seulement les mots qui apparaissent dans la requête. En effet, la requête est un pseudo-document. Par exemple la requête (0 1 0 1) est un vecteur ligne de 4 dimensions.

0	0	1	0	0
1	0	1	0	1
0	1	1	0	0
0	1	0	1	1

Tableau 6: Chaque colonne représente un document

Chaque vecteur document dans la TDM approchée peut alors être comparé à la requête en calculant le cosinus entre eux. Le cosinus est calculé à partir de l'équation suivante:

$$cos\,\theta = \frac{a_j^T q}{\|a_j^T\| \|q\|}$$

Où

- a_j^T est le vecteur transposé du j^{eme} vecteur de document dans la matrice *a*,
- *q* est le vecteur requête,
- $\|a_j^T\|$ est le module de a_j^T,
- ||q || est le module de *q*.

Le module est équivalent à la norme euclidienne:

$\|q\| = \sqrt{q_1^2 + q_2^2 + q_3^2 \ldots + q_{n-1}^2 + q_n^2}$.

Une valeur du cosinus de 1 signifie que les deux vecteurs existent exactement le même espace dimensionnel. En dessous de cette valeur les vecteurs deviennent de moins en moins similaires. Afin de déterminer les documents qui sont suffisamment semblables pour être renvoyé en réponse à la requête d'un utilisateur, un seuil de 0,5 est fixé par la plupart des chercheurs dans ce domaine. Le temps de calcul est également un facteur important lorsque l'on considère les performances d'un algorithme.

4.2.7 Métriques utilisés

Afin de montrer clairement les mesures, des graphiques illustrant les résultats de différentes recherches pour chaque algorithme ont été tracés, représentant le nombre total de documents retournés et le nombre de documents pertinents retournés. Ou des graphiques formés d'une seule ligne pour chaque algorithme dans le cas ou le nombre total de documents retournés est égal au nombre de documents pertinents retournés. Les performances de chaque algorithme dépendent des paramètres entrés par l'utilisateur, par exemple le rang (k) pour la SVD, et la valeur seuil pour le traitement de l'image. A travers une gamme de valeurs de différents rangs (k), on détermine une valeur optimale, c'est à dire une valeur de k qui renvoit les meilleurs résultats à l'utilisateur.

4.3 Analyse du bruit lexical et des mesures en recherche d'information intelligente

Dans cette section, on présente une nouvelle approche pour l'analyse de la TDM en utilisant des techniques de traitement d'image.

Il est à noter que la visualisation de la TDM comme une image permet d'examiner et analyser plus facilement les grands ensembles de données [72]. La distribution sur le TDM, qui peut être remarquée facilement sur l'image visualisée, dépend de la structure, le contenu et la taille de la base de données comme il sera montré à la section suivante. Ces facteurs constituent une base pour l'étude et la compréhension du processus de LSI.

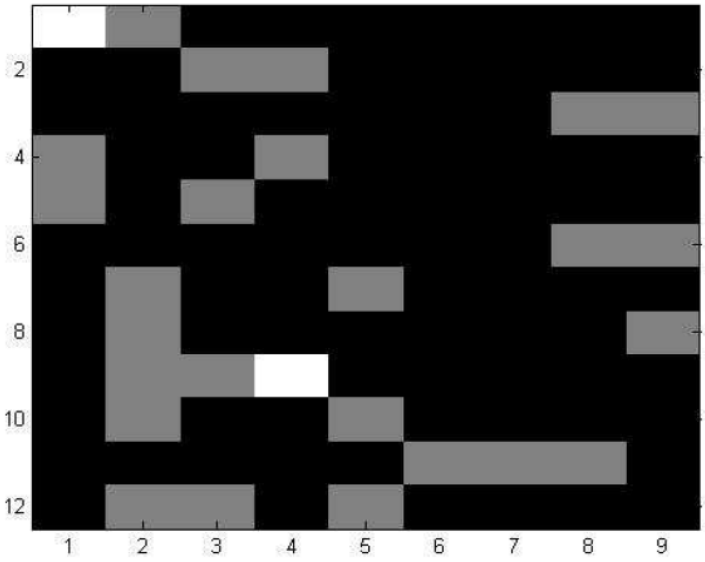


Figure 13: TDM comme une image de la base de données Mémos

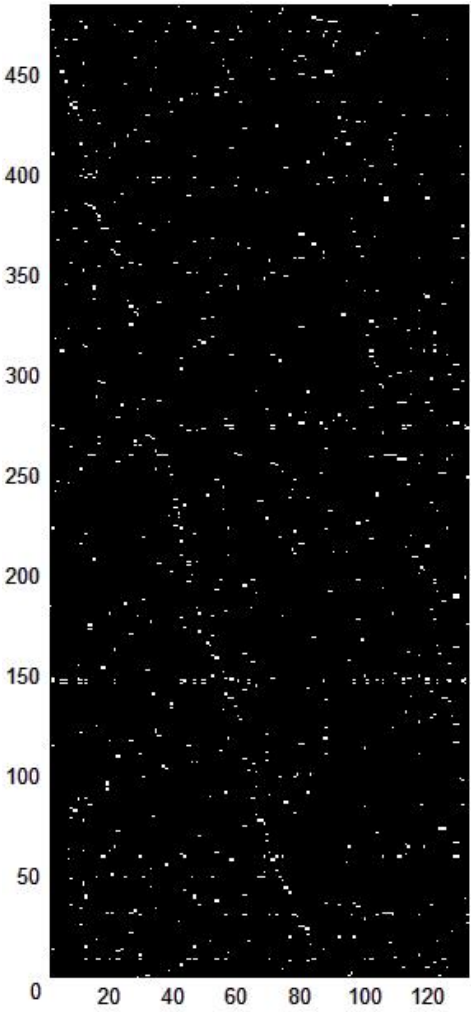


Figure 14: TDM comme une image de la base de données Cochrane

4.3.1 Méthodologie proposée pour la mesure de bruit lexicale

Dans cette section, une nouvelle méthodologie de mesure du bruit lexicale est présentée.

Premièrement, la TDM est générée puis représentée comme une image en niveaux de gris.

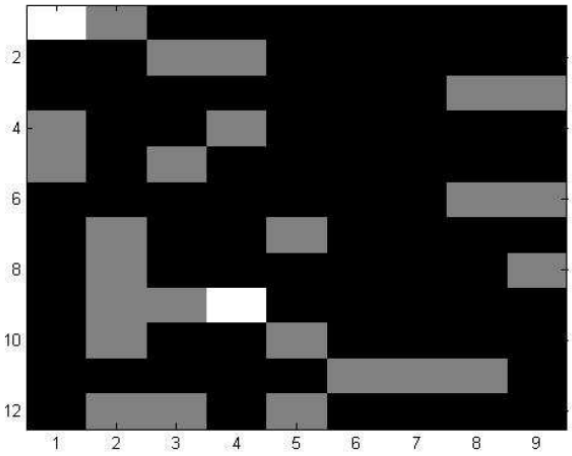


Figure 13: TDM comme une image de la base de données Mémos

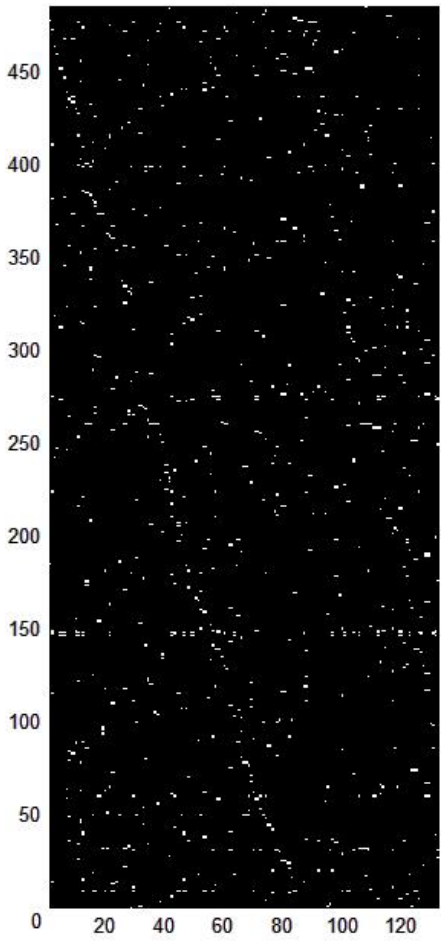


Figure 14: TDM comme une image de la base de données Cochrane

La décomposition SVD est ensuite appliquée pour une gamme de valeurs de *k*, en reconstruisant les matrices approchées, obtient les résultats illustrés dans les figures 3.12 à 3.23.

Le choix de la valeur de *k* est une étape très importante, car cette valeur à un effet majeur sur la structure de la TDM qui peut être clairement remarqué sur les images.

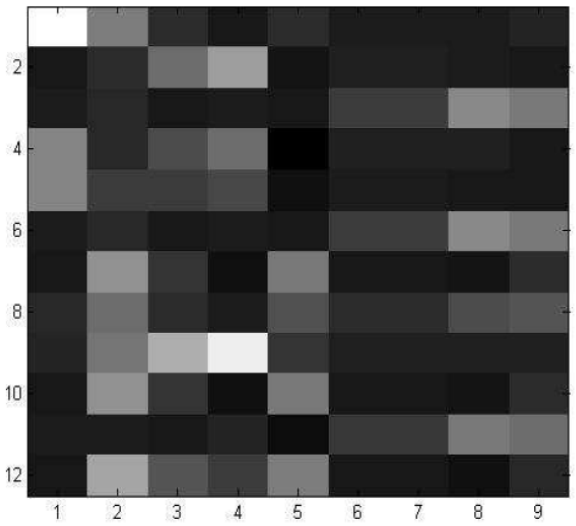


Figure 15: Image TDM après SVD avec k = 4 de base de données Mémos

Figure. 15 montre l'image de l'approximation TDM à *k*=4 de la base de données des mémos. En examinant la TDM approchée, il est clair que la distribution des valeurs non-nulles a été améliorée et la création d'un nouvel espace a été engendré. Dans la Figure. 16 l'image de l'approximation TDM à *k*=1, semble avoir complètement détruit l'information apportée par la TDM. Une valeur très faible de k provoque l'élimination de l'information utile et de ce fait la destruction de l'approximation de la TDM.

D'autre part dans la Figure. 17, à k=8, aucun changement ne peut être détecté dans le TDM. 8 dimensions sont conservées et une seule dimension a été retirée de la matrice diagonale des valeurs propre de la TDM, ce qui a un effet mineur sur la matrice originale.

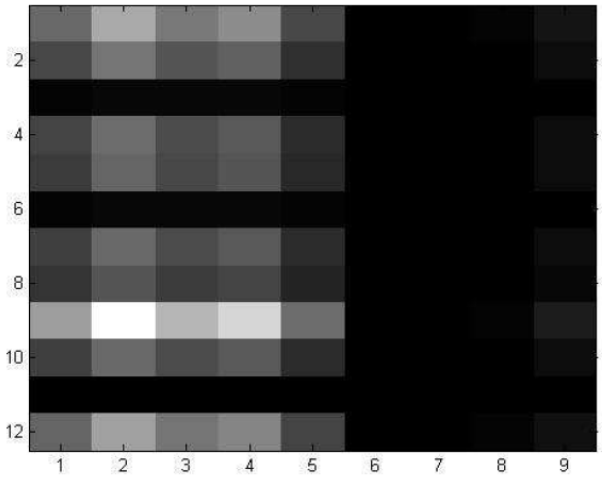


Figure 16: Image TDM après SVD avec k = 1 de base de données Mémos

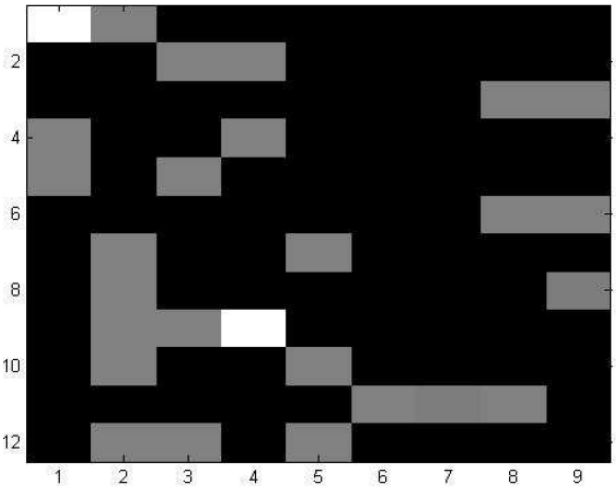


Figure 17: Image TDM après SVD avec k = 8 de base de données Mémos

Figure. 19 montre à nouveau une bonne structure de TDM à *k*=80 de la base de données Cochrane. La propagation des valeurs non-nulles est meilleure en comparaison avec la TDM originale.
Comme le montre la Figure. 18, la valeur k=1 supprime la plupart des informations contenues dans le TDM.

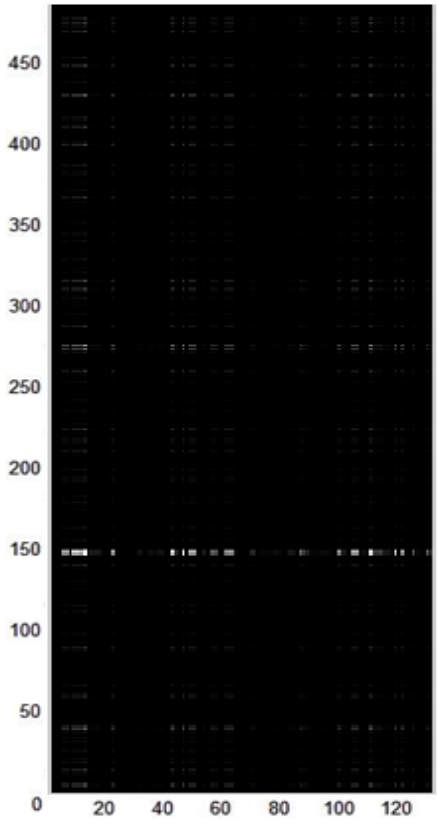


Figure 18: Image TDM après SVD avec k = 1 pour base de données Cochrane

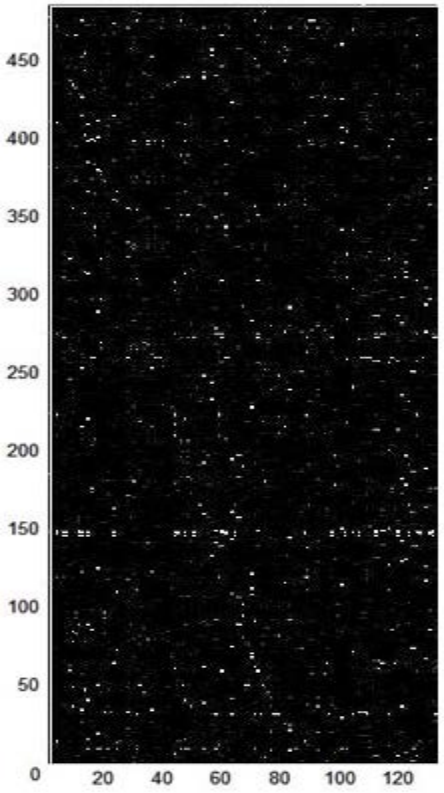


Figure 19: Image TDM après SVD avec k = 80 pour base de données Cochrane

4.4 Approche empirique

Dans cette section l'attention est accordée à l'identification du k optimal pour la base de Cochrane (les valeurs de réduction du rang qui sont utilisés dans l'algorithme de SVD). L'objectif est de déterminer la meilleure structure pour une base de données qui mènera à de meilleurs résultats de recherche de LSI, pour ensuite appliquer le débruitage de HAAR mentionné dans le chapitre précédents afin d'optimiser le système de recherche.

La recherche LSI est effectuée par une requête à différentes valeurs de k. les figures x à z présentent le nombre de documents pertinents et documents retournés pour chaque requête, et les résultats montrent que la meilleure valeur de k où le plus de documents pertinents sont retournés est k=80.

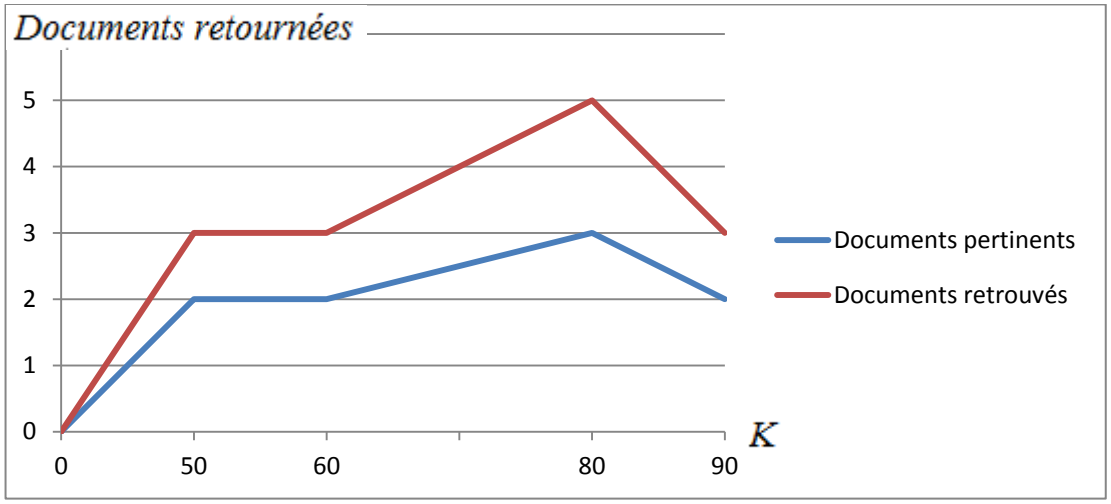


Figure 20 : Rechercher «Intervention treating» pour différentes valeurs de k

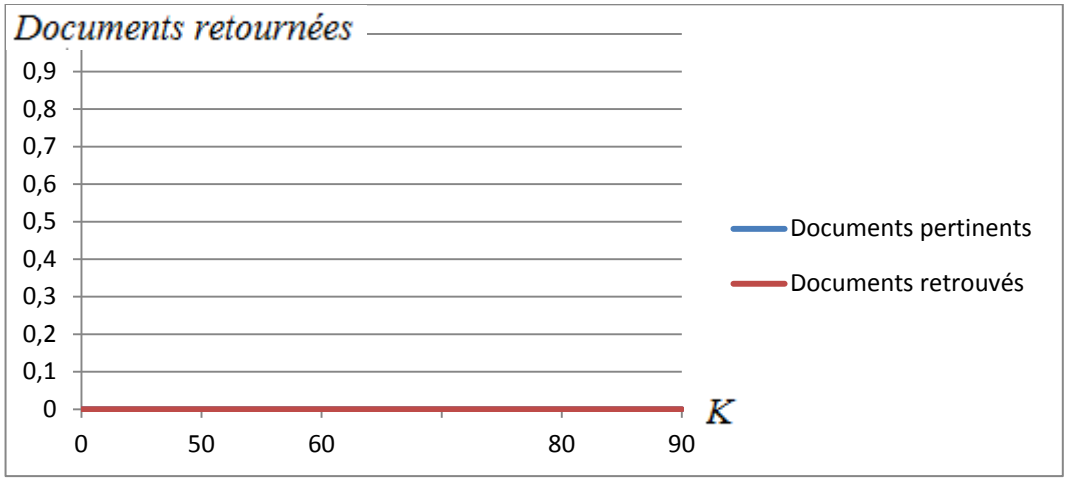


Figure 21 : Rechercher «Immunoglobulin» pour différentes valeurs de k

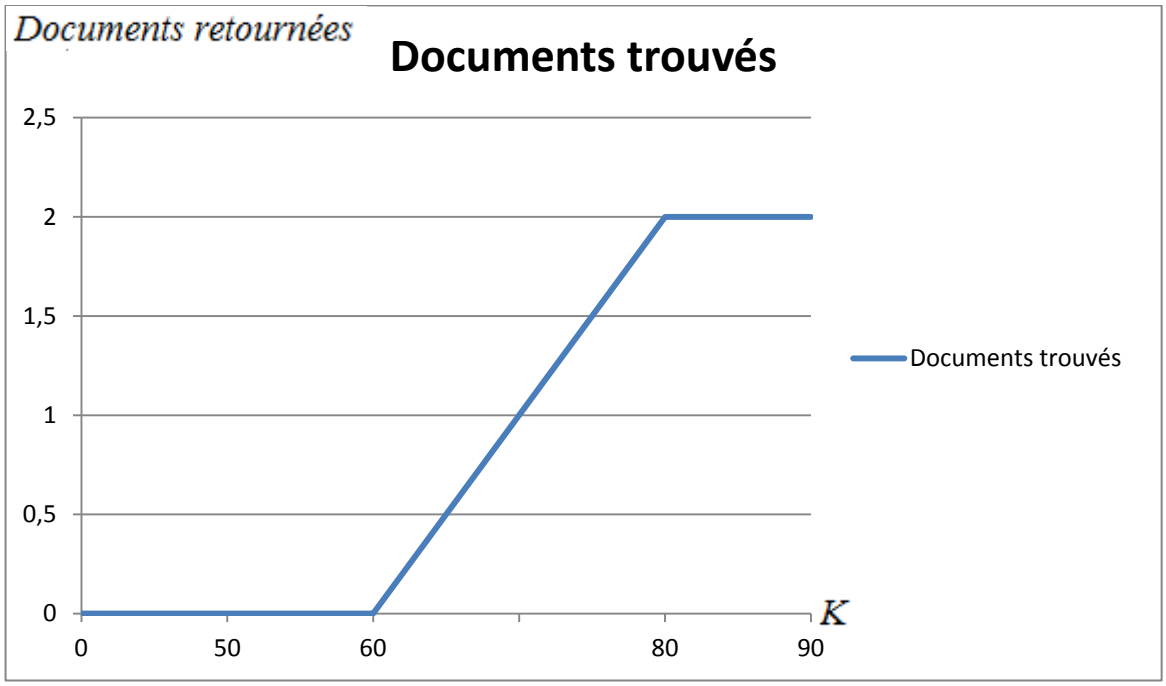


Figure 22 : Rechercher «Acupuncture» pour différentes valeurs de k

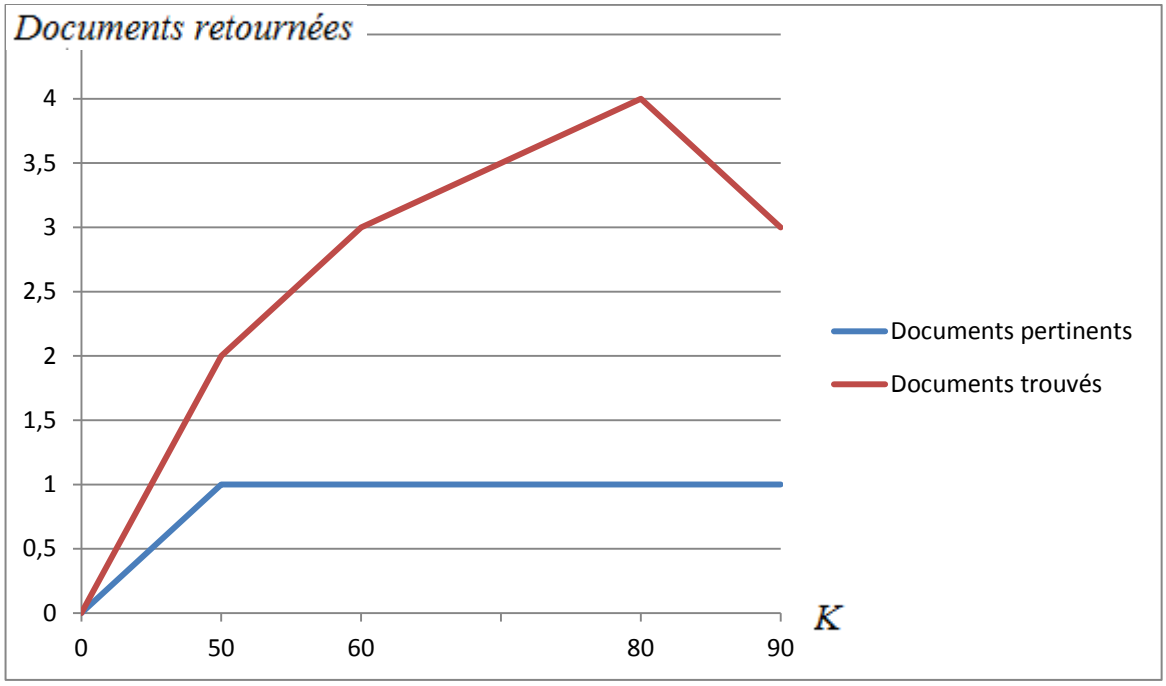


Figure 23 : Rechercher «Acupuncture asthma» pour différentes valeurs de k

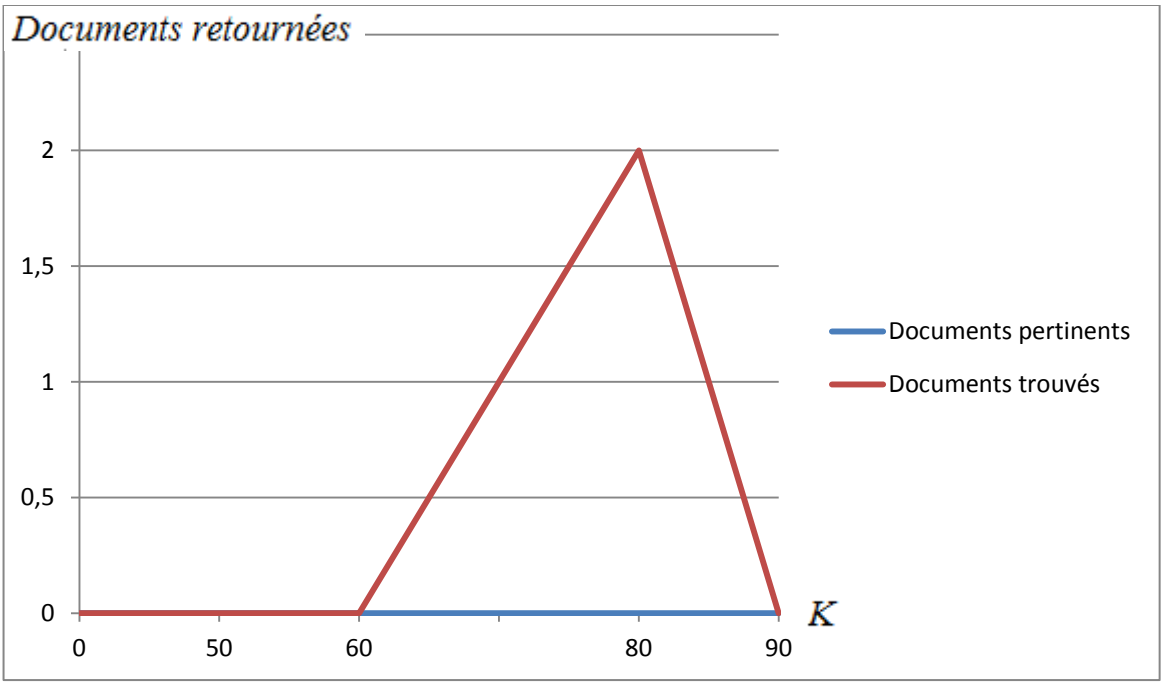


Figure 24 : Rechercher «Treatment effects» pour différentes valeurs de k

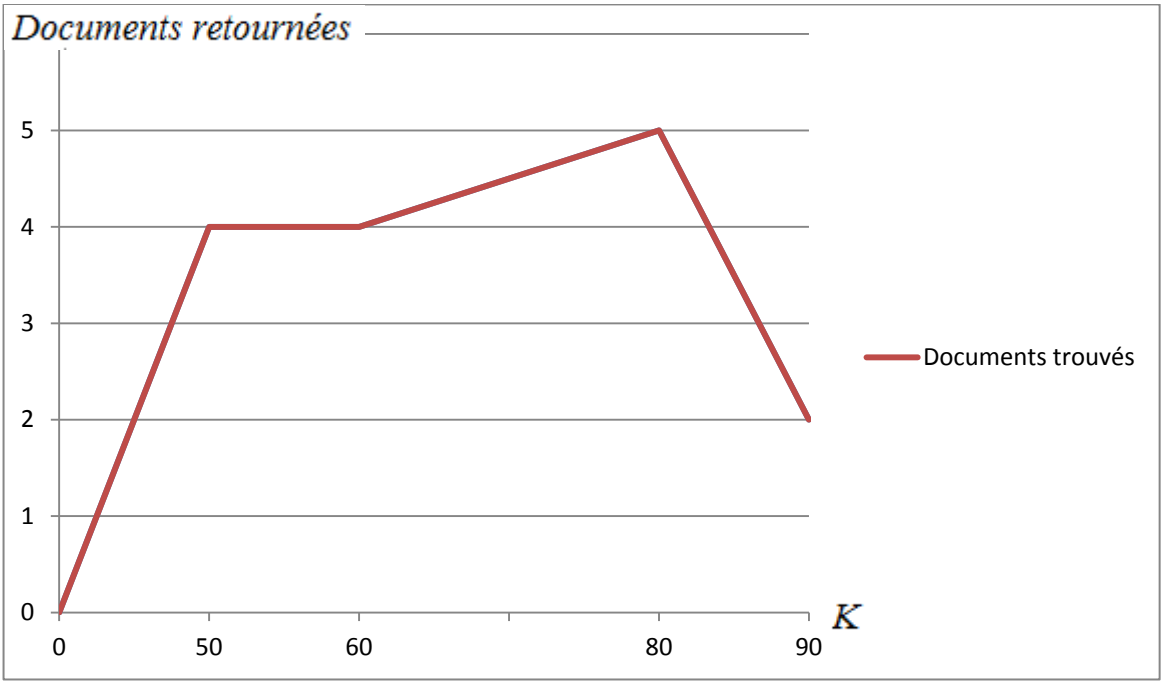


Figure 25 : Rechercher «Therapy» pour différentes valeurs de k

La performance des algorithmes, tels que décrites dans les sections précédentes, est déterminée en examinant le nombre global de documents retournés par la requête, et le nombre de documents retournés qui sont pertinentes à la requête de l'utilisateur. Une combinaison de la SVD avec un k optimal et le débruitage de HAAR montre clairement l'efficacité de cette fusion comme le montrent les figurent 26 à 29.

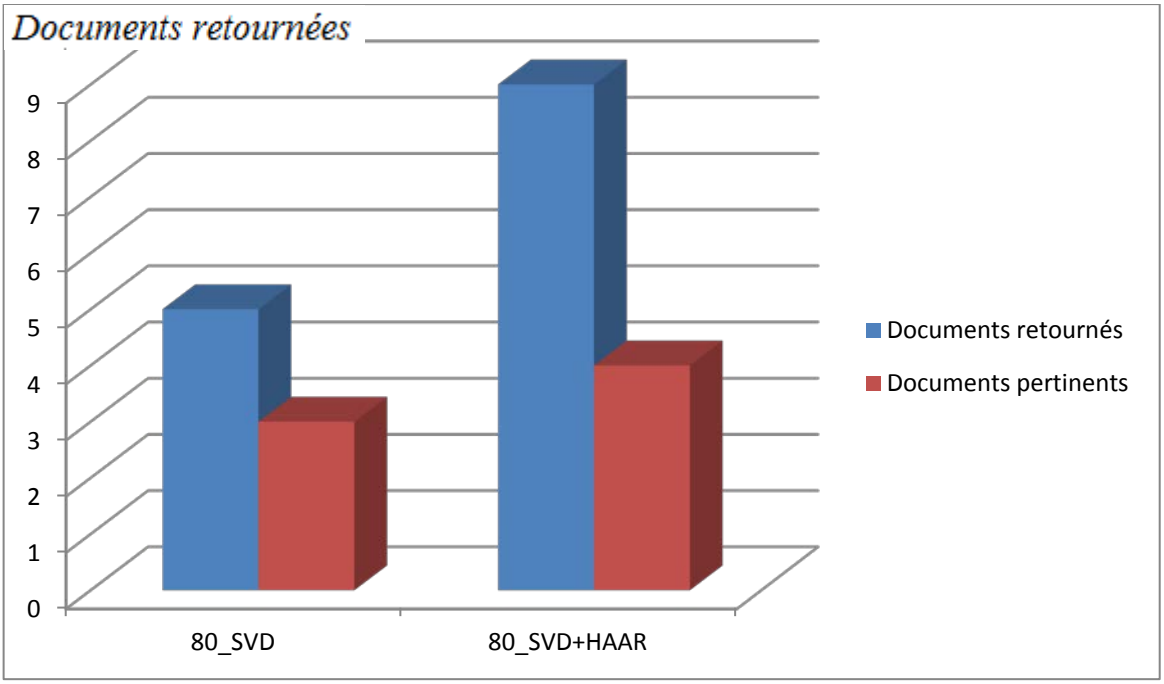


Figure 26 : Rechercher «Intervention treating» en utilisant la 80_SVD et la 80_SVD+HAAR

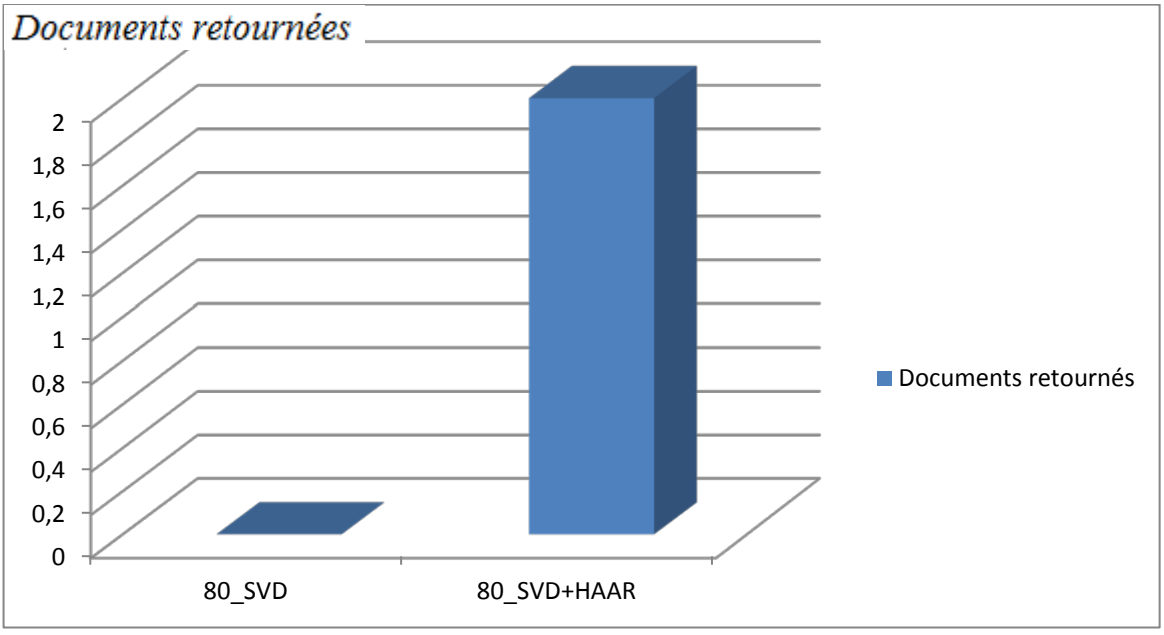


Figure 27 : Rechercher «Immunoglobulin» en utilisant la 80_SVD et la 80_SVD+HAAR

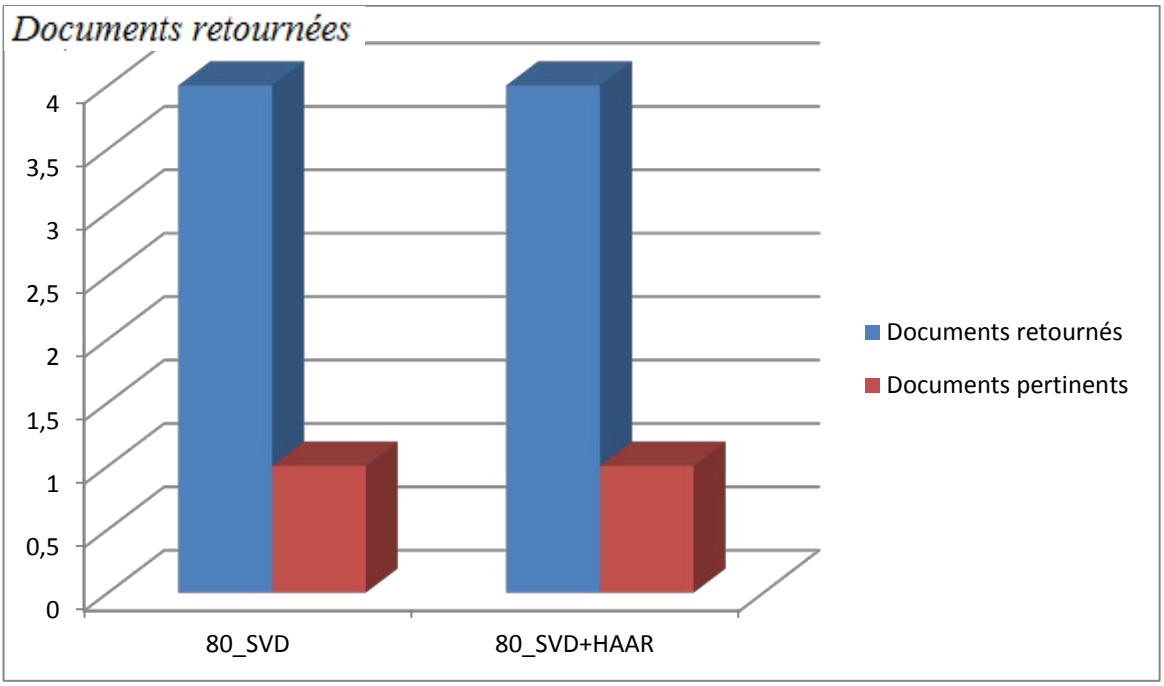


Figure 28:Rechercher «Acupuncture asthma» en utilisant la 80_SVD et la 80_SVD+HAAR

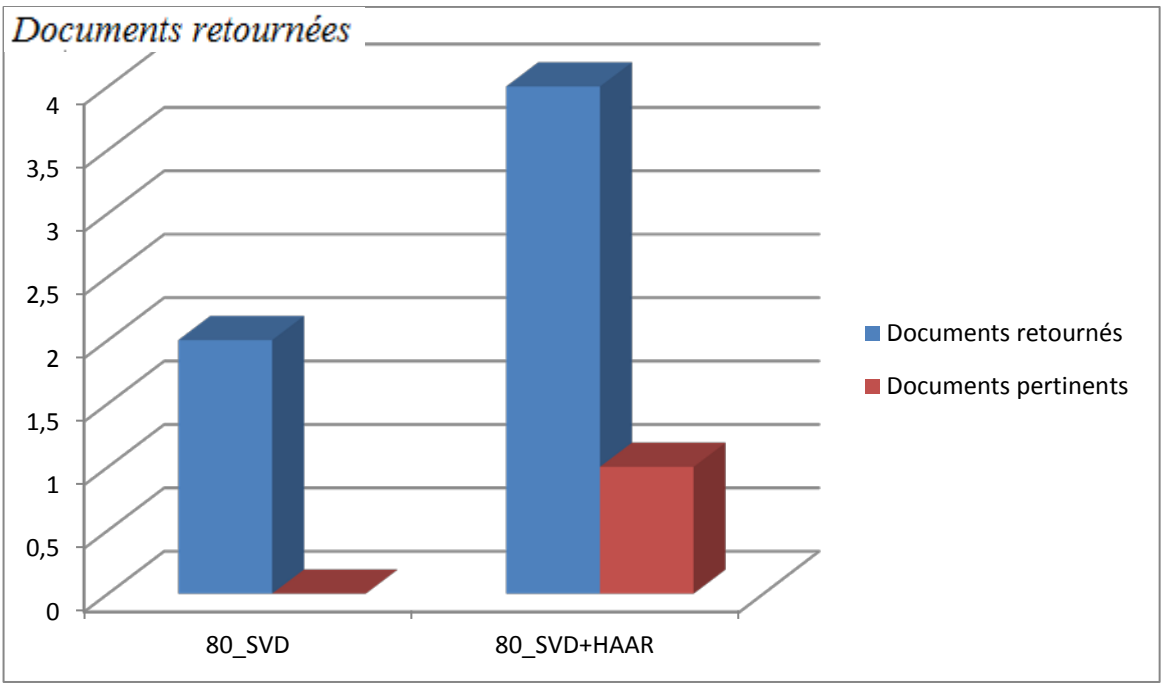


Figure 29 : Rechercher «Treatment effects» en utilisant la 80_SVD et la 80_SVD+HAAR

4.5 Interface graphique

Pour une meilleure illustration des différentes étapes de l'étude, une interface graphique a été construite. Elle permet de présenter de manière ludique les différents traitements appliqués dans la recherche, et de fournir à l'utilisateur une prise en main aisée afin de tester les différents modes de recherches.

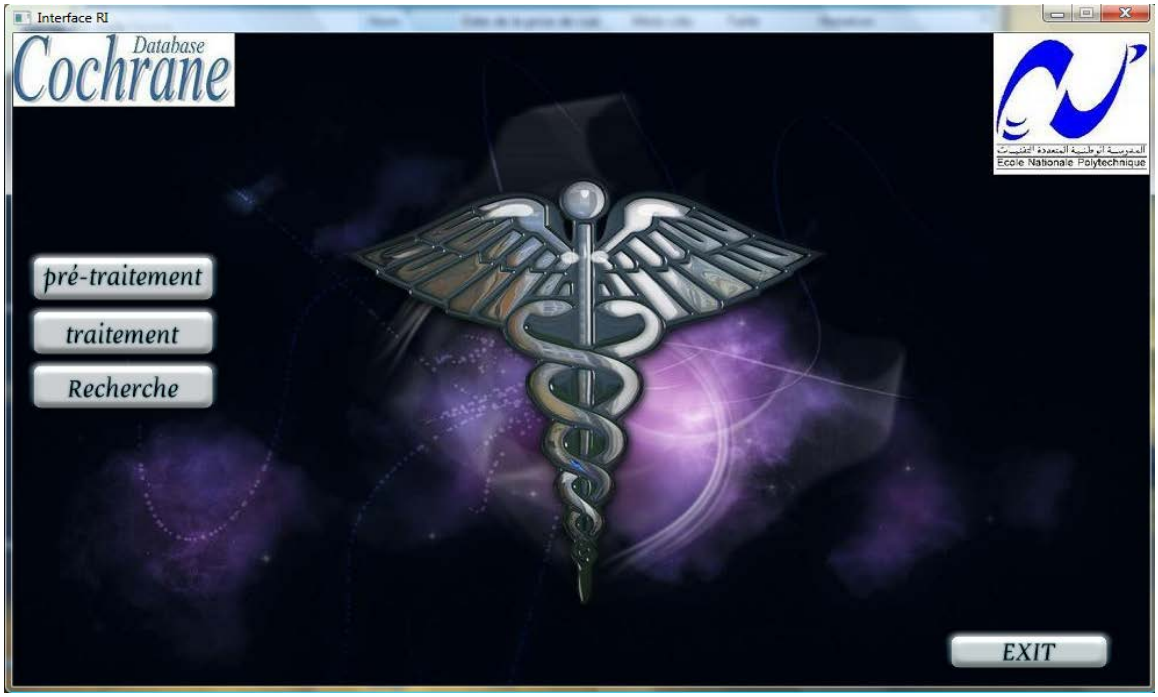


Figure 30: page d'accueil de l'interface graphique

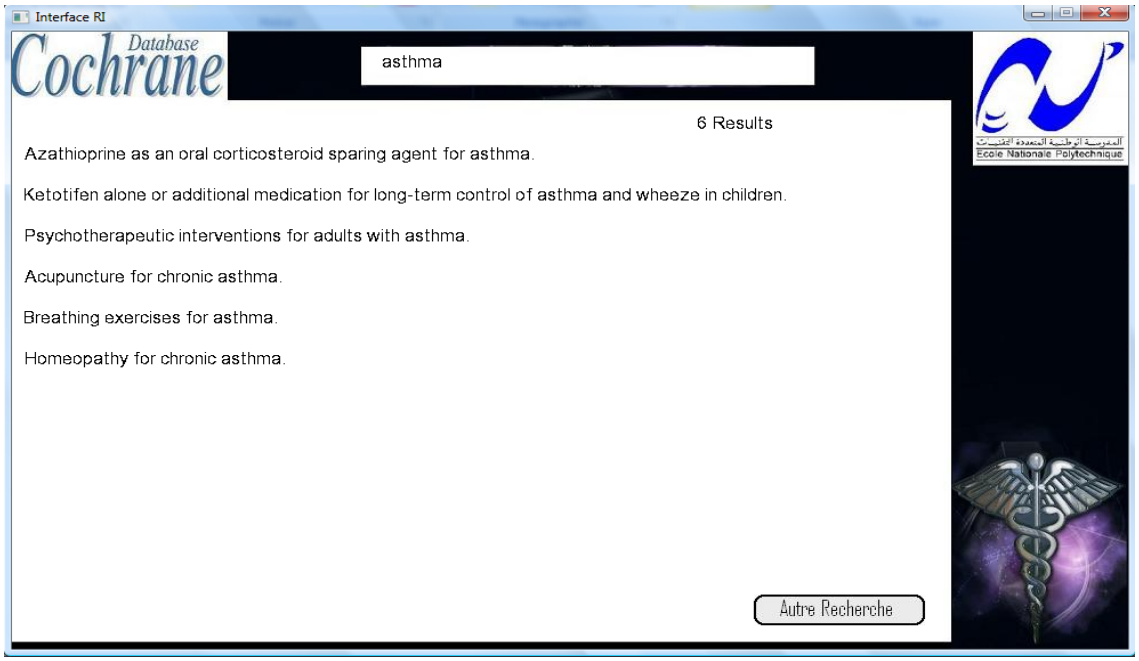


Figure 31: Exemple de recherche dans l'interface graphique

4.6 Conclusion

Une nouvelle approche hybride a été présentée dans ce chapitre pour une utilisation efficace en RI. Les résultats de l'étude pour les différentes approches montrent que, en appliquant HARR comme une étape de post-traitement après la SVD dans le processus de LSI donne de bons résultats. Il est bénéfique de noter que l'action précise de l'étape de traitement dépend de la valeur de k utilisées pour la SVD et la valeur seuil utilisée dans la transformation.

Conclusion générale

Ce travail s'intéresse à la modélisation de l'information textuelle pour l'analyse et la recherche d'information, en se portant essentiellement sur les points suivants :

- Le mode de représentation des documents dans un corpus

- le mode de représentation des requêtes exprimées par un utilisateur, pour la recherche d'information,

- la comparaison, à l'aide de la modélisation que nous avons définie, entre un document et une requête ou entre plusieurs documents.

Les principaux modèles de représentation de l'information : modèle booléen, booléen pondéré, modèle vectoriel, etc. ont été conçus, pour la plupart il y a une trentaine d'années. La grande majorité des recherches actuelles se fonde sur ces modèles pour améliorer les résultats des SRI.

On s'intéresse à l'indexation sémantique latente, dont ses 3 phases qui sont : le prétraitement, le traitement et la requête. On débute avec une étude de la SVD en localisant le k optimal qui donne les meilleurs résultats.

On se propose d'améliorer les performances de cette technique et cela en supprimant le bruit généré par la SVD en utilisant le débruitage de Haar. Des tests montrent clairement la l'efficacité de cette approche hybride.

Une interface graphique a été construite. Elle permet de présenter de manière ludique les différents traitements appliqués dans la recherche.

Bibliographie

[1] Mooers, C.N. "Application of Random Codes to the Gathering of Statistical Information", MIT, Thèse de Master, 1948.

[2] Mizzaro, S. "How many relevance's in information retrieval?", Italie : Departement of Mathematics and Computer Science, University of Udine, 1998.

[3] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public web search engines," Proceedings of the 7th International Conference on World Wide Web 7, Brisbane, Australia, pp. 379-388, 1998.

[4] S. Lawrence and C. Giles, "Searching the world wide web," Science, vol. 280, pp. 98-100, 1998.

[5] T. A. Letsche and M. W. Berry, "Large-scale information retrieval with latent semantic indexing," Information Sciences: International Journal, vol. 100, pp. 105 - 137, 1997.

[6] Z. Wang, S. Wong, and Y. Yao, "An analysis of vector space models based on computational geometry," Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 152 - 160, 1992.

[7] V. V. Raghavan and S. K. M. Wong, "A critical analysis of vector space model for information retrieval," Journal of the American Society for Information Science, vol. 37, pp. 279-287, 1986.

[8] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, p. 613620, 1975.

[9] M. Berry, S. Dumais, and G. OBrien, "Using linear algebra for intelligent information retrieval," SIAM Review, vol. 37, pp. 573 - 595, 1995.

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

[11] C. Fox, "Lexical analysis and stoplists. in information retrieval - data structures & algorithm," Prentice-Hall, pp. 102-130, 1992.

[12] M. W. Berry, Z. Drmavc, and E. R. Jessup, "Matrices, vector spaces, and information retrieval," SIAM Review, vol. 41, pp. 335 - 362, 1999.

[13] J. Lovins, "Development of a stemming algorithm," Mechanical Translation and Computational Linguistics, vol. 11, p. 2231, 1968.

[14] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, pp. 130-137, 1980.

[15] W. B. Frakes, \Stemming algorithms. in information retrieval - data structures & algorithm," Prentice-Hall, pp. 131 - 160, 1992.

[16] D. Hull, "Stemming algorithms - a case study for detailed evaluation," Journal of the American Society for Information Science, vol. 47, pp. 70 - 84, 1996.

[17] M. Fuller and J. Zobel, "Conflation-based comparison of stemming algorithms," Proceeding of the 3rd Australian Document Computing Symposium, pp. 8-13, 1998.

[18] A. Singhal, "Modern information retrieval: A brief overview," IEEE Data Engineering Bulletin, vol. 24, pp. 35-43, 2001.

Bibliographie

[19] A. Kontostathis, \Essential dimensions of latent semantic indexing (lsi)," Proceedings of the 40th Hawaii International Conference on System Sciences - 2007, pp. 73 - 73, 2007

[20] E. R. Jessup and J. H. Martin, "Taking a new look at the latent semantic analysis approach to information retrieval," Computational Information Retrieval, pp. 121 - 144, 2001.

[21] S. Dumais, "Improving the retrieval of information from external sources," Behavior Research Methods, Instruments and Computers, vol. 23, pp. 229-236, 1991.

[22] S. Robertson and S. Walker, \Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 232241, 1994.

[23] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 2129, 1996.

[24] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," Journal of the American Society for Information Science, vol. 41, pp. 288-297, 1990.

[25] "Query operations (relevance feedback / query expansion)," PowerPoint Presentation in Information Retrieval and Web Search Course, University of Texas at Austin URL: www.cs.utexas.edu/ mooney/ir-course/, 2008.

[26] G. OBrien, "Information management tools for updating an svd-encoded indexing scheme," Master's thesis, University of Tennessee, Knoxville, TN, 1994.

[27] H. Zha and H. Simon, "On updating problems in latent semantic indexing," SIAM Journal on Scientific Computing, vol. 21, pp. 782 - 791, 1999.

[28] C. Chen, N. Stoffel, M. Post, C. Basu, D. Basu, and C. Behrens, "Telcordia lsi engine: Implementation and scalability issues applied," Proceedings of the International Workshop on Research Issues in Data Engineering (RIDE), pp. 51-58, 2001.

[29] S. Richards and A. Lovely, "Matrices, vector spaces and information retrieval," Student Project in Linear Algebra, College of the Redwoods, 2002.

Bibliographie

[30] T. Kolda and D. O'Leary, "A semi-discrete matrix decomposition for latent semantic indexing in information retrieval," ACM Transactions on Information Systems, vol. 16, p. 322346, 1998.

[31] D. P. Oleary and S. Peleg, "Digital image compression by outer product expansion," IEEE Transactions on Communications, vol. 31, pp. 441- 444, 1983.

[32] E. Hoenkamp, "Unitary operators on the document space source," Journal of the American Society for Information Science and Technology, vol. 54, pp. 314 - 320, 2003.

[33] Cochrane, "Url: http://www.cochrane.org," 2005.

[34] R. Liu and T. Tan, "An svd-based watermarking scheme for protecting rightful ownership," IEEE Transactions On Multimedia, vol. 4, pp. 121-128, 2002.

[35] C. Moler and D. Morrison, "Singular value analysis of cryptograms," The American Mathematical Monthly, vol. 90, pp. 78-87, 1983.

[36] R. Zhao and W. I. Grosky, "Narrowing the semantic gap-improved text-based web document retrieval using visual features," IEEE Transactions On Multi-media, vol. 4, pp. 189 - 200, 2002.

[37] H. Andrews and C. Patterson, "Outer product expansions and their uses in digital image processing," The American Mathematical Monthly, vol. 82, pp. 1-13, 1975.

[38] "Singular value decomposition (svd) image coding," IEEE Transactions on Communications, vol. 24, p. 425432, 1976.

[39] H. Ito and H. Koshimizu, "Keyword and face image retrieval based on latent semantic indexing," IEEE lntemational Conference on Systems, Man and Cy-bernetics, vol. 1, pp. 358 - 363, 2004.

[40] M. M. Rahman, B. C. Desai, and P. Bhattacharya, "Visual keyword-based image retrieval using latent semantic indexing, correlation-enhanced similarity matching and query expansion in inverted index," Proceeding of the 10th In-ternational Database Engineering and Applications Symposium, pp. 201-208, 2006.

[41] S. Sclaroff, M. L. Cascia, S. Sethi, and L. Taycher, "Unifying textual and visual cues for content-based image retrieval on the world wide web," Computer Vision and Image Understanding, vol. 75, pp. 86-98, 1999.

[42] R. Zhao and W. I. Grosky, "From features to semantics: Some preliminary results," International Conference on Multimedia and Expo., vol. 2, pp. 679-682, 2000.

[43] M. Kurimo, "Indexing audio documents by using latent semantic analysis and som," In: Oja, E., Kaski, S. (Eds.), Kohonen Maps. Elsevier, Amsterdam, p. 363374, 1999.

[44] M. Kurimo, "Thematic indexing of spoken documents by using self-organizing maps," Speech Communication, vol. 38, pp. 29 - 45, 2002.

[45] F. Souvannavong, B. Merialdo, and B. Huet, "Video content modeling with latent semantic analysis," In the 3rd International Workshop on Content-Based Multimedia Indexing, 2003.

[46] F. Souvannavong, B. Merialdo, and B. Huet, "Latent semantic indexing for semantic content detection of video shots," IEEE International Conference on Multimedia and Expo., vol. 3, pp. 1783- 1786, 2004.

[47] "Latent semantic analysis for an effective region-based video shot retrieval system," Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 243 - 250, 2004.

[48] M. Littman, S. Dumais, and T. Landauer, "Automatic cross-language information retrieval using latent semantic indexing," In SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval, pp. 16-23, 1996.

[49] Using LSI for information filtering: TREC-3 experiments. Dumais, S. T. (1995) D. Harman (Ed.), The Third Text REtrieval Conference (TREC3) National Institute of Standards and Technology Special Publication In press 1995.

[50] An Overview of Latent Semantic Indexing Jason I. Hong SIMS 240 Spring 2000.

[51] Unitary operators on the document space Source (2003) Eduard Hoenkamp Journal of the American Society for Information Science and Technology Volume 54, Issue 4.

[52] Filtering noise from images with wavelet transforms J. B. Weaver, X. Yansun, D. M. Healy, Jr., and L. D. Cromwell Magnetic Resonance in Medicine, vol. 21, pp. 288-95, 1991.