MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE FERHAT ABBAS - SETIF -
UFAS (ALGERIE)

MEMOIRE

Présenté devant la Faculté des Sciences de l'Ingénieur,
Département d'Informatique
Pour l'obtention du diplôme de

MAGISTER

Option : Matériel et Logiciel

Par
Riad AOUABED

THEME

Indexation automatique de documents scientifiques dans une bibliothèque électronique

Soutenu le : 27/06/2009 Devant la commission d'examen

Dr A.BOUKERRAM	Maitre de Conférence	UFA, SETIF	Président
Dr A. Hamdi-Cherif	Maître de Conférences	Qassim University,KSA	Rapporteur
Dr A.MOUSSAOUI	Maitre de Conférence	UFA, SETIF	Examinateur
Dr M.TOUAHRIA	Maitre de Conférence	UFA, SETIF	Examinateur

DEDICACE

***A** ma mère*

***A** mon père*

***A** ma femme*

***A** mes chers fils **Med Sallah** &*
Bassem Abedrrahmene

***A** mes frères*

***A** mes sœurs*

***A** tous les membres de ma famille.*

Remerciements

Tout d'abord et avant tout, louanges à Allah Seigneur des Mondes, sans Lequel rien ne s'accomplit dans ce monde de l'éphémère.

Je remercie l'Institut d'Informatique, enseignants et administratifs qui m'ont aidé tout au long de mon périple à l'Université.

J'atteste une reconnaissance indéfectible envers mon encadreur Mr. Aboubekeur Hamdi-Cherif. Malgré la distance géographique qui nous a séparé dès le début de ce travail, il a su par son savoir-faire, sa disponibilité en tout instant, ses conseils éclairés, son support moral, me guider dans les chemins difficiles de la connaissance, vers la réalisation de cette thèse qui, je le souhaite sera à la hauteur des espérances investies.

J'aimerais remercier chaleureusement toute l'équipe YECER INFORMATIQUE, qui m'a accueilli en son sein depuis mon Ingéniorat, pour l'ambiance amicale et le travail de groupe qui règne entre ses membres, pour les discussions et le soutien qu'elle m'a apporté tout au long de ce travail.

Je remercie enfin les membres de jury, pour avoir accepté de juger ce travail.
Enfin, à tous mes camarades de promotion un merci amical.

Table des matières

Chapitre 2 Indexation classique des citations scientifiques

Chapitre 3 Indexation automatique des citations – Bibliothèque Citeseer-

Chapitre 4 Modélisation de la bibliothèque électronique UFASeer

Chapitre 5 UFASeer Réalisation et discussions des résultats

Liste des figures

Résumé

Se situant entre l'informatique et les sciences bibliothécaires, l'indexation automatique est un domaine qui utilise des méthodes logicielles pour établir un index pour un ensemble de documents afin d'en faciliter l'accès ultérieur. Un index est une liste de descripteurs à chacun desquels sont associés une liste des documents, ou passages de documents, auxquels ce descripteur renvoie. Un index très simple à établir automatiquement est la liste ordonnée de tous les mots apparaissant dans les documents avec la localisation exacte de chacune de leurs occurrences. Evidemment, un tel index est volumineux et inexploitable. L'indexation de citation est une méthode pour organiser le contenu d'une collection de documents de manière à surmonter les imperfections des méthodes d'indexation existantes. L'automatisation de cette méthode facilite la tâche du chercheur. Elle élimine la majeure partie de l'effort manuel demandé pour construire un index des documents scientifiques publiés sur le Web. Elle permet ainsi aux chercheurs de suivre les dernières publications et les plus appropriées. Ce travail présente une architecture originale d'une bibliothèque qui utilise l'indexation de citation comme méthode d'indexation. UML (Unified Modeling Language) est utilisé comme langage de modélisation, pour raison de lisibilité et maintenance.

Le travail présenté étudie les volets suivants :

1. Indexation classique telle que celle utilisée par *Scientific Citation Index.*
2. Fonctionnalités de la bibliothèque *CiteSeer* utilisant l'indexation de citation.
3. Indexation dans *CiteSeer* et son *Autonomous Citation Index (ACI).*
4. Développement de *UFASeer*, une nouvelle bibliothèque archivant et indexant les documents scientifiques par indexation de citation.
5. Présentation de la majorité des vues produites par UML.

Mots clés: Indexation de citation, recherche documentaire, bibliothèque électronique, UML, extraction des informations, Fouille de patrons (Template mining).

Automatic indexing of scientific documents in a digital library

Abstract

Situated at the crossroads of information technology and librarianship, automatic indexing is an area which uses software methods to establish an index to a set of documents for facilitating further access to these documents and to their content. An index is a list of descriptors associated with a list of documents, or passages of documents, to which this descriptor refers. A very easy index to establish is a representation of a list of ordered words appearing in a given document with the exact location of each of these words' occurrences. Obviously, such an index is large and unmanageable. Citation indexing is a method for organizing the contents of a collection of documents so as to overcome the shortcomings of existing methods of indexing. The automation of this method facilitates the researcher's task and eliminates much of the manual effort required to build an index of scientific papers published on the Web. It thus allows researchers to follow the latest and most appropriate publications in a given discipline. This paper presents an original architecture of an electronic library that uses the citation indexing as a method. UML (Unified Modeling Language) is adopted as a modeling language, for readability and maintenance purposes.

The present work considers the following tracks:

1. Classical indexing such as that used by *Scientific Citation Index (SCI).*
2. Features of *CiteSeer*, a system that provides citation and other searching of scientific literature, primarily in the fields of computer and information sciences.
3. Indexing in *CiteSeer* and its *Autonomous Citation Index (ACI).*
4. *UFASeer* development; a new library archiving and indexing documents by citations.
5. Majority of UML views presentation.

Keywords: Citation index, Information retrieval, Digital library, UML, Information extraction, Template mining.

?????T????T?? ? ??????? ??? ???? ??? ?? ?? ??µ T??

? ???

???? ????Td??? ???T???????? ? ? ???????????? ????? ?? ???? ?? ???????T?? ??????? ? ? ? ????T? ? ? ??? ???? ?? ???? ??dT? ???? ??? ? ?????? ??a ? ???? ???? ???? ??? ?s??? ??? ????? ? ?????? ?? ? ??? ? T??? ??? ?????? ? ????? ? ? ?? ? ??? ?? ? ????? ?? ? ?T? ?? ???? ?? ?? ???? ???? ??? ???d?? a ?????? ? ??? ? ? ???? ?? ? ?? ? ?? ??? ?? ??????? ? ? ? ??? ????? ???? ?? ? ???? ? ? T??? ???? ? ???T? ??? ?? ???S ? ?? ?? ?a ? ????

?? ????? ??? ? ???? ??? ? ??t ??? ???S ??????????? ?? ???? ???? G???T?????? ????? ????? ?????? ????? ??????? ?? ? ?? ??t ??? ???? ? C???T? ??? ? ???? ???????? ? ???? ????C??? ???? ?? ? ???? ???? ??????? ?? ??????? ? ???d?? ?? ???? ???????t ??? ??? ?????? ??? G?? ?????????? ?? ??????? ?? ??????? ?µ ?? a ?????? ? ??t ?? ?????? ??? ???? ???? ?? ????? ? ??? ???? ????????T ? ?? ? ??S ???? ????? ???? ???? ? ???T ? ???? ?? ??????????? ? ????? ??? G???T????? ?? ?????? Td ?????? T?????T??????? ??????????????? ??? ???? ?? ?????? ??T ???????G?S ??????????? ? UML ??? ????µ ????? ? ???? ?? ???? ??????µ ?G????????d?? ??

???T??? ? ???? ?? ??????d?? ???????????????a ??? ? ? ?*SCI* ? ? ???Td? T?? ?T ? ?? ???T??? ????? ? ??????????????? ?? G??T?? ???? ??? ??? ? ? Citeseer ???T???? ?? ?? ? ????????? ? ? ?? ???? ?? ? ?C??? ? ? ??????? ?? G??T???????? ??? ? ???? Citeseer ? ?? ???????????s ? ? ?????? dS ????????? ?? ???Td?????T?? ?T?? ? ?*UFASeer* ???t ? ???????G?S ?????????? ?s ? ? ?S ?????????? ???p ?? ? ??

??????? ????? ? T??? ?????T??? ? ????????? ????µ ? T?? ? ?? ?????? ??? ??? ***?????S ?T????? ????????***

? ????? ?? ? ???T??? ? ???????? ?? ? T? ? ????????G?S ????

Introduction

L'ère de l'Internet a développé l'accès à l'information et sa diffusion. Ce moment historique a ouvert des nouvelles possibilités dans pratiquement tous les secteurs d'activité principalement celle les bibliothèques électroniques, l'éducation, le gouvernement, le commerce, la santé, et le divertissement, pour ne citer que quelques activités [Lawr 98].

La quantité d'informations publiées et disponibles sur le Web augmente rapidement. Cette information est vaste, diversifiée et grandissante car tout le monde peut à son tour y mettre ses propres informations. Ceci transforme le Web en une bibliothèque numérique colossale [Gils 98]. Cette augmentation pose les difficultés de mise à jour pour les chercheurs. Afin de relever ce défi, des méthodes d'accès rapide j l'information sont constamment développées.

De nombreux outils sont désormais proposés aux utilisateurs du Web parmi lesquels nous trouvons ceux relatifs j l'extraction de l'information, principalement des moteurs de recherche et des dictionnaires informatisés. Malgré le fait que les moteurs de recherche offrent la possibilité d'accès à l'information stockée sur le Web en quelques fractions de secondes, il n'en demeure pas moins que les résultats fournis peuvent être très vagues, incomplets et sans pertinence. Dans le meilleur des cas, les moteurs de recherche, même s'ils offrent des réponses aux requêtes des utilisateurs ne fournissent pas pour autant des réponses précises ou valides. D'autre part, les dictionnaires informatisés présentent des outils de recherche d'information valides et précises dans des domaines particuliers, mais n'offrent pas à l'utilisateur la possibilité de l'enrichir. De ce point de vue-la, les méthodes existantes de gestion de l'information sur le Web demeure chaotique et nécessite donc des améliorations [Lawr 99].

Afin de faciliter la tâche d'accès à l'information, des techniques documentaires sont apparues soit pour organiser le savoir comme les systèmes de classement, les thesaurus, et les listes d'autorités, soit pour la description du contenu comme l'analyse du contenu et l'indexation [Lawr 98]. C'est ainsi que le documentaliste bibliothécaire a eu longtemps la fonction essentielle d'intermédiaire entre

l'information, contenue dans le document, et l'utilisateur. Petit à petit, le rôle du documentaliste a changé avec l'apparition de l'outil informatique et les bases de données bibliographiques en ligne. Ensuite, et avec le développement des supports de stockage et la migration du document du support papier au support électronique, la multiplication des documents et des versions d'un même document commence à poser un problème sérieux d'accès. Par exemple, quand on soumet une requête au Web, on obtient la plupart du temps des milliers et des milliers de réponses, parfois l'équivalent du contenu de toute une bibliothèque en document papier, qui peuvent renfermer ou pas l'information pertinente ; ce qui impose un système de filtrage pour faciliter l'accès à cette information utile [Lawr 98].

Nous ne portons un intérêt particulier qu'aux publications scientifiques. Ces derniers sont identifiés par des apostilles et références aux autres articles ou livres scientifiques. Cette représentation de l'information rend les textes scientifiques différents de l'histoire d'un journaliste ou d'un roman. Les références présentées dans les documents scientifiques fournissent des liens entre la littérature qui peut être employée pour un certain nombre de buts comprenant la recherche, l'évaluation, et l'analyse documentaire [Boll 99].

L'indexation de citation est une méthode pour organiser le contenu d'une collection de documents, permettant d'évaluer l'influence et la signification intellectuelle des recherches publiées dans le temps. Elle fournit une capacité unique d'indiquer exactement quand et où un document ou un auteur particulier a été cité ainsi que l'évolution de ces citations dans le temps [Kurt 99]. L'indexation de citation peut renvoyer aux individus, aux établissements, et aux pays en termes de citations de publications enregistrées soit individuelles ou collectives. A notre connaissance, aucune autre méthodologie ne permet une telle identification qui précise le poids d'influence des individus et des collectivités sur le développement scientifique et technologique.

Dans ce mémoire, nous proposons une architecture de bibliothèque qui utilise l'indexation automatique des citations pour aider les chercheurs j effectuer leur travail de façon automatique.

Ce travail est divisé en cinq chapitres.

Le ***Chapitre 1*** décrit l'aspect théorique de la recherche sur le Web, et présente les documents électroniques et leurs différents formats, les outils de recherche disponibles sur le Web tels que les moteurs de recherche, les annuaires et les méta-moteurs ainsi que les différentes méthodes d'indexations utilisées.

Le ***Chapitre 2*** donne un aperçu sur l'aspect de l'indexation des citations classique, sa définition, son apparition et ses avantages.

Le Chapitre 3 présente la bibliothèque électronique ***Citeseer*** qui utilise l'indexation automatique des citations, avec ces différents composants : la recherche, l'indexation et l'interface Web, ainsi que les algorithmes lips j l'extraction des informations et l'indexation des documents.

Le ***Chapitre 4*** décrit la modélisation de la bibliothèque électronique *UFASeer*. L'approche y est utilisée.

Le ***Chapitre 5*** présente l'extraction automatique des informations à partir d'un document électronique afin de les classer dans une base de données.

Le mémoire se termine par une conclusion et des perspectives de développement.

Chapitre I

L'accès aux Documents Scientifiques

1.1 Introduction :

L'explosion documentaire a engendré des problèmes de repérage et d'accès à l'information. Des techniques documentaires sont apparues pour résoudre ce problème. En effet, tout savoir doit être organisé. Ce qui signifie le recours aux méthodes de classement, de thesaurus, entre autres. Le contenu doit être décrit avec précision soit par analyse du contenu, soit par indexation, ou autre [Lawr 98].

L'apparition des nouvelles technologies et plus particulièrement l'explosion d'Internet offrent au grand public un accès à une information à la fois hétérogène et illimitée. L'utilisateur se trouve dans la position d'un manipulateur d'une énorme masse d'informations textuelles. Une problématique nouvelle a donc vu le jour ces dernières années : comment donner les moyens à chaque usager d'Internet d'accéder à une information précise?. [Boll 99].

Dans ce chapitre nous essayons de répondre à cette question en parcourant les différentes méthodes disponibles. Nous avons trois parties à explorer. La première partie présente les documents scientifiques et leurs différents formats électroniques. La deuxième partie présente les différentes méthodes d'accès j ces documents. La troisième partie explique les différentes méthodes d'indexation avec leurs avantages et inconvénients.

1.2 Les documents Scientifiques

1.2.1 Les documents électroniques.

I.2.1.1 Définition

Un document électronique est défini par l'ISO42 comme "Un document existant sous forme électronique à manipuler avec des outils informatiques" [Buck 97].

Sur le Web on peut trouver plusieurs types de fichiers consultables sur écran et/ou récupérables (par téléchargement) sur un support électronique (disque, disquette,...) ensuite sur support papier. En effet, en plus des fichiers sous format HTML (qui est hypertextuel), on trouve des fichiers PDF, des fichiers PS (PostScript), des fichiers texte (.txt) des fichiers Word (.doc, .Rtf…)...Tous ces types de fichiers sont des documents électroniques.

I.2.1.2 Les formes des documents électroniques:

D'une manière générale, un document électronique se présente sous deux formes quel que soit l'outil de traitement de ce dernier :

- ✍ Forme binaire qui est le plus compact.
- ✍ Forme textuelle qui permet un échange plus aisé des documents entre les applications.

On trouve généralement 4 formats de documents électroniques :

- ✍ Documents numérisés (exp. format TIFF)
- ✍ Documents formatés (exp. PostScript, PDF)
- ✍ Documents textuels (exp. RTF)
- ✍ Documents structurés (exp. SGML, XML)

Le Tableau 1 présente les différents formats de documents électroniques avec le type d'information représenté ainsi que l'usage et le traitement possible de chaque type de format [Bonh 97].

Tableau 1.1 : Les formats de documents électroniques

Format du document	Information représentée	Traitements possibles
Numérisé	Matrice de pixels	Impression, affichage, OCR, analyse d'image.
Formaté	Suite de caractères, balise de formatage	Edition, formatage, recherche simple
Structuré	Contenu, balise de typage, structure hiérarchique, attributs	Formatage multiple, indexation, recherche contextuelle, hypertexte

En naviguant sur le Web, nous pouvons rencontrer plusieurs types de formats de documents. Les principaux langages de formatage de documents existants sont :

1. L'*ASCII* (American Standard Code for Information Interchange) : utilisé dans les premiers documents électroniques. Il est assez limité dans la mesure où il ne peut représenter que du texte linéaire sans typographie ni mise en page.
2. Le langage *SGML* (Standard Generalized Markup Language) est un standard de balisage de la structure logique des documents se basant sur la DTD (Document Type Définition). Ce langage permet de coder la structure logique des documents ayant des contenus composites : texte, graphique, image, mais aussi des structures telles que celles des formules mathématiques, des formules chimiques et des tableaux.

3. Le langage *HTML* (HyperText Markup Language) : utilisé pour le balisage des documents multimédia manipulés par le Web. C'est une application de SGML avec une DTD (Document Type Définition) particulière. En termes simplifiés, on peut dire que HTML tire ses origines de SGML. Il s'agit d'un type particulier d'annotations qui correspond à une collection de styles reconnaissables par les navigateurs. Il est utilisé sur le Web depuis 1990.
4. Le langage *XML* (eXtensible Markup Language) est un langage de description de structures de documents recommandé par le W3C (World Wide Web Consortium). Il est plus riche que l'actuel format HTML dans la mesure où il se base sur la structure, le contenu et la sémantique indépendamment de la mise en page. Comme *SGML*, il se base sur des DTD tout en offrant la possibilité d'être utilisé dans un environnement réparti. Ce standard est en plein développement. Différents groupes au sein du W3C travaillent sur plusieurs chantiers connexes.
5. Le format *PDF* (Portable Document Format) de la société Adobe qui a aussi développé le format PostScript. Il permet de supporter à la fois la structure et la forme du document. Son objectif est de permettre l'échange de documents formatés contenant à la fois l'aspect visuel et la structure du document interprétables pour l'affichage et l'impression sur une large gamme de plates-formes.
6. Le format *PS* (Post Script) : tout comme le format *PDF*, il est développé par la société Adobe. Le PostScript est un langage de description de page utilisé par certaines imprimantes. Il est utilisé pour produire une documentation de bonne qualité. Certaines communautés scientifiques utilisent fréquemment ce format. Il s'agit surtout des physiciens, mathématiciens, informaticiens et autres.
7. Autres formats d'échange : d'autres formats sont plus ou moins utilisés dans la diffusion et l'échange des données sur Internet. Parmi œs formats, on trouve le format *RTF* (Rich Text Format), *TEX* et LaTex, largement utilisés dans le milieu des mathématiciens, chimistes, physiciens et informaticiens car ils offrent des grandes facilités pour représenter les formules mathématiques, et d'autres en utilisation ou encore en développement.

Avec toutes ces catégories de langages de balisages et de normes, certaines disciplines restent à l'écart et rencontrent des problèmes sérieux pour permettre à leur documentation d'être indexée par les moteurs de recherche et par conséquent restent inaccessible aux chercheurs.

I.2.1.3 Avantages et inconvénients des documents électroniques:

Par rapport aux documents papiers, un document électronique présente des avantages en relation avec le support de stockage, les délais de publication, le mode de présentation et d'accès (hypertextualité, ...), etc…

Du côté des supports de stockage, un document électronique occupe des espaces nettement moindres que le support papier. Sur un simple DVD, on peut sauvegarder plusieurs millions de pages voir des encyclopédies entières.

Pour les délais de publication, et par rapport à ceux des documents papiers, on remarque que les publications électroniques sont généralement publiées dans des délais assez courts.

Un autre avantage du document électronique est la fonction d'archivage et de récupération de l'information. En effet, le support électronique permet, d'une part le stockage de volumes importants d'informations sur des espaces minimes avec plusieurs modes de présentation à l'utilisateur, et d'autre part la recherche de l'information avec toutes ses facilités sur les documents électroniques alors que cette recherche sur support papier reste toujours liée aux outils de référence (index, bibliographies, bases de données...). [Caro 95].

A ces avantages s'ajoute ceux de l'usage des revues électroniques :

- Une publication rapide et continue (2 jours après l'acceptation de l'article);
- Accès immédiat ;
- Possibilités de recherche optimales
- Outils de navigation hypermédiatiques permettant aux usagers de butiner dans les différentes parties d'un article : sommaire, paragraphes, références, figures, tableaux.
- Possibilité pour l'usager d'établir des profils de recherche pour obtenir les articles d'intérêt.

Toutefois plusieurs problèmes restent à résoudre pour les documents électroniques. Certains aspects juridiques d'un document électronique comme le copyright, les droits d'auteurs, la valeur légale, restent à préciser. Du point de vue de l'usage, les problèmes de standards des formats de fichiers, le problème d'ergonomie, de lecture sur écran, sont aussi à améliorer. En effet l'objectif d'un document électronique n'est pas uniquement la réduction de l'espace de stockage.

I.2.2 Les documents Scientifiques

I.2.2.1 Définition

Un « document scientifique » est tout objet utilisé ou produit par les scientifiques dans l'exercice de leurs fonctions de recherche pour consultation,

étude, expérimentation ou preuve. De ce fait tout objet utilisé par un chercheur, indépendamment de son support ou de sa forme ?article d'une revue scientifique, communication dans un colloque, livre, norme ou brevet - est un document scientifique.

Sur le Web, le document scientifique électronique a commencé à prendre une place importante. Toutefois, il doit offrir des avantages propres. Selon Chartron [Char 97], le document scientifique électronique accessible par Internet devrait permettre un accès:

- Plus rapide aux informations,
- Démultiplié pour les étudiants et les chercheurs,
- Fédérant plusieurs types de ressources : bibliographie, textes d'articles, données de mesure.

I.2.2.2 L'usage des documents scientifiques

L'usage des documents scientifiques diffère d'une discipline à une autre. Certaines disciplines se basent sur des sources plutôt que sur d'autres. En règle générale, les sciences humaines et sociales ont tendance à utiliser plutôt les livres que les articles de périodiques. A l'opposé, les sciences exactes ou encore les sciences expérimentales se basent sur l'article dans des revues scientifiques dites de prestige. D'ailleurs, l'évaluation et la validation des travaux dans ces sciences passent souvent par la publication d'articles dits primaires dans des revues internationales avec comité de lecture.

I.2.2.3 Les publications scientifiques électronique :

Une publication scientifique est un rapport écrit et publié décrivant les résultats d'une recherche originale. Ce rapport doit être présenté selon un code professionnel qui résulte de l'éthique scientifique, de l'expérience d'édition et de la tradition.

On peut trouver sur support électronique plusieurs types de publications scientifiques allant du sommaire des revues jusqu'au texte intégral des articles et même des livres entiers. On trouve aussi des dictionnaires sur support électronique ou encore des encyclopédies entières en version électronique.

On peut distinguer deux types de publications scientifiques:

- *Les publications scientifiques sous format électronique*, où on peut intégrer les fichiers RTF, PDF, Post Script et même les revues électroniques qui se contentent

d'un simple balisage des documents papiers. Ces publications peuvent rassembler aussi bien des articles scientifiques, des communications dans des séminaires et colloques. Ces publications sont généralement conçues pour être consultées et exploitées sur support papier. L'objectif principal de leur mise en ligne est de faciliter leur diffusion par des transferts beaucoup plus rapides que les méthodes classiques de diffusion.

- *Les publications scientifiques électroniques proprement dites* : ce sont les documents scientifiques publiés sur un support électronique, en ligne ou non. On classe sous cette catégorie tous les documents scientifiques diffusés sur CD/DVD, Internet ou autres. Ils sont généralement sous des formats plus ou moins adaptés à une exploitation et consultation sur ordinateur. On pense surtout aux formats SGML, HTML et XML et leurs dérivées. Toutefois, les documents renfermant uniquement des références, avec ou sans résumé, ne sont pas considérés comme électroniques.

1.2 La recherche des documents sur le Web

1.2.1 Le Web

Le World Wide Web est un service Internet qui a été mis en place en 1989 par des chercheurs du CERN, en Suisse. Il se base sur le protocole HTTP (HyperText Transmission Protocol) qui permet d'établir une communication entre un serveur Web et un logiciel client appelé navigateur. Le Web est devenu assez rapidement le service principal d'Internet. La plupart des autres protocoles comme le FTP, le Telnet, la messagerie, sont accessibles par les navigateurs Web facilitant ainsi l'accès aux différentes ressources documentaires réparties j travers le monde.
Une personne ne peut contrôler ni quantifier le volume d'information circulant et accessible aujourd'hui sur le Web, il est par contre possible d'avoir une idée sur la nature de cette information. On trouve ainsi sur le Web deux types d'informations à savoir :

1. L'information structurée représentée par les bases de données de touts types (Medline, Pascal,…), les catalogues des bibliothèques (les OPACs, les catalogues collectifs comme le CCFR,…)…
2. L'information brute représentée par les différents types de documents (revues, magasines, livres,….), les sites personnels et/ou des institutions, …

De ce fait, on peut considérer le Web comme un gigantesque système d'information puisqu'il permet d'accéder à un nombre illimité de systèmes d'informations de tout genre et dans tous les domaines.

1.2.2 Les outils de recherche et d'accès à l'information électronique

Devant ce développement du Web et avec l'augmentation continue des documents en ligne, l'accès à l'information électronique pose de plus en plus des problèmes.

Dans ce qui suit nous allons passer en revue les principaux outils de recherche et d'accès à l'information électronique tout en décrivant l'utilisation de ces outils dans la recherche des documents scientifiques.

1.2.2.1 Les moteurs de recherche

Il s'agit d'outils de recherche d'information automatiques assez puissants permettant de formuler une requête avec des mots clés. Le principe de fonctionnement se base sur trois composantes essentielles à savoir :

- *Un robot* : généralement assez puissant qui visite régulièrement des dizaines de millions de documents afin de stocker le contenu.
- *Un système d'indexation* : qui permet d'analyser et indexer l'information stockée par le robot afin de la rendre accessible à l'utilisateur.
- *Une interface d'interrogation* : qui permette à l'utilisateur de rechercher l'information indexée par le système. Selon le moteur, elle offre soit une recherche simple, soit une recherche experte ou encore les deux à la fois.

Les avantages de ces moteurs sont essentiellement la rapidité d'indexation et d'analyse des sites, l'exhaustivité relative, la finesse de leur analyse (indexer uniquement le titre, les balises métas des pages Web, le document entier,…).

I.2.2.2 Les annuaires

Appelés encore répertoires de recherche ou index. Ces outils sont entretenus par des personnes qui évaluent les sites et les classent par thème. Chaque thème est divisé en rubriques et en sous-rubriques de plus en plus précises.

Ces outils sont très appréciés quand il s'agit d'un domaine qu'on appréhende mal ou qui est très vaste. A l'opposé, ils donnent rarement de bons résultats pour une recherche pointue. Ces répertoires n'adoptent pas forcément ou uniquement un classement thématique. On trouve des répertoires géographiques qui permettent de rechercher une information selon une démarche géographique et d'autres sites appelés "sites carrefour" qui recensent les sites clés d'une discipline ou d'une spécialité.

I.2.2.3 Les méta-moteurs

Face à la prolifération de ces outils de recherche, une nouvelle catégorie s'est développée qui se déclare comme une race hybride entre les deux catégories

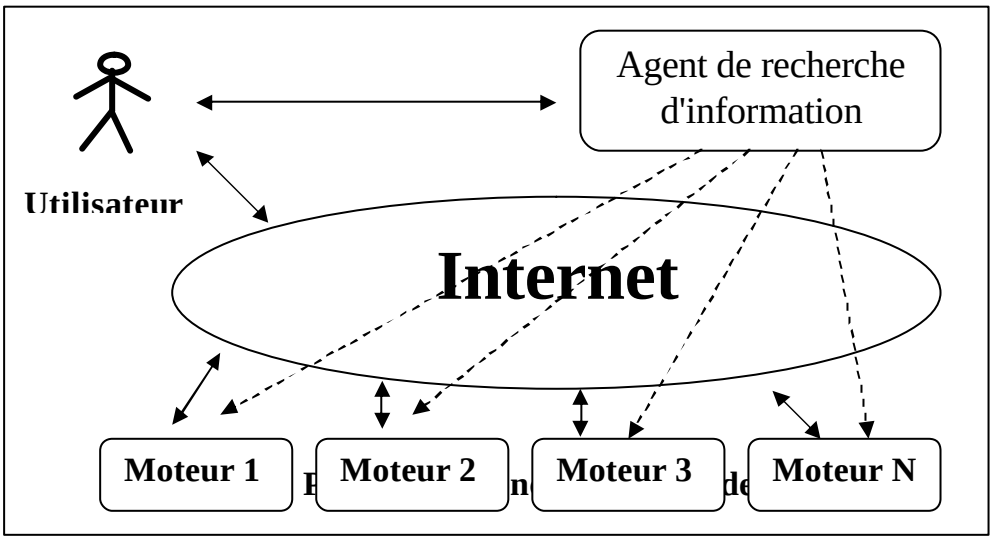
précédentes. Il s'agit des ***méta-moteurs***. Ce sont des agents intelligents permettent d'interroger simultanément plusieurs outils de recherche (annuaires et moteurs de recherche).

Exemples: *MetaCrawler* (http://www.metacrawler.com) et *ProFusion* (http://designlab.ukans.edu/profusion), *Inquirus* (http://www.Inquirus.com).

On peut définir un méta-moteur comme un "logiciel permettant de lancer une requête dans plusieurs moteurs de recherche simultanément" [Reve 98]. On peut considérer qu'ils constituent la première génération des agents dits "intelligents".

Selon la fonction accomplie, ces agents peuvent appartenir à un ou plusieurs types :

- *Les agents de recherche d'information :* ces agents sont généralement capables de rechercher de l'information sur plusieurs moteurs de recherche ou répertoires. En plus de cette fonction de recherche, ils sont capables de rapatrier les réponses en local et de les traiter : classement, indexation, fusion des résultats, élimination des doublons, etc…
- *Les agents sectoriels :* ils fonctionnent sur le même principe que les précédents tout en étant spécialisés dans un domaine précis.
- *Les agents pour la veille :* ces agents ont comme rôle d'assurer la veille scientifique ou technologique sur un sujet précis ou un produit.



1.3 L’indexation des documents

L’indexation consiste à identifier dans un document certains éléments significatifs qui serviront de clé pour retrouver ce document au sein d’une collection. Ces éléments comprennent le nom de l’auteur, le titre de l’ouvrage, le nom de l’éditeur, la date de publication et l’intitulé du sujet traité.

Les recherches en indexation se sont développées vers la fin des années 60 en particulier avec les travaux de H.P Luhn, à partir des méthodes KWIC (KeyWord In Context) qu'il a élaboré. Il proposa les principes d'une indexation automatique sur un modèle statistique capable d'extraire les mots-clés.

La recherche d'information dans les documents a été l'un des thèmes principaux de recherche en informatique jusqu'au début des années 1970. C'est ce que nous montrent les références comme [Maro 60], [Gard 72] et [Salt 72] qui ont mené des recherches approfondies sur la recherche d'information.

C'est à partir des années 70, avec le développement de l'informatique que de nombreux progrès sont réalisés, grâce entre autre à l'utilisation de langages documentaires et l'intégration de connaissances syntaxiques. Les années 80 sont marquées par le développement des analyseurs morphologiques et syntaxiques et leur intégration dans des programmes d'indexation.

I.3.1 Les approches classiques

Afin d'effectuer une recherche d'information efficace et pertinente, il apparaît comme nécessaire de donner une représentation mieux structurée, et si possible normalisée, du contenu des textes. Lors de la recherche, la requête de l'utilisateur exprimée en langage naturel doit être transformée en une représentation structurée et normalisée, qui va permettre d'apparier celle-ci avec la représentation du contenu des documents. Il faut pour cela résoudre les problèmes linguistiques les plus visibles [Pier 00] :

La synonymie : un même concept peut être exprimé par des mots différents.

La polysémie : un même mot peut représenter plusieurs concepts.

Tous les concepts ne sont pas tous exprimables avec des mots simples.

I.3.1.1 L'indexation manuelle avec vocabulaire contrôlé

Les index contrôlés proviennent d'un vocabulaire dont la pertinence est contrôlée. Il constitue un langage documentaire, artificiel qui se définit en prélevant dans le langage naturel les expressions dont il a besoin pour couvrir un certain univers conceptuel [Pier 00]. L'indexation manuelle sur vocabulaire contrôlé a longtemps été le moyen principal de constitution des bases de données

bibliographiques commerciales. Elle consiste à représenter le contenu du document par une liste de groupes nominaux qui expriment les principaux thèmes traités dans le document. Avec un langage ainsi défini, on peut produire une représentation formalisée des documents et des requêtes afin de faciliter leur appariement.

L'indexation manuelle est un exercice très subjectif, étant donné qu'il dépend des connaissances du documentaliste sur le sujet traité dans le texte et par conséquent de la manière dont il va hiérarchiser les thèmes retenus.

Une liste de groupes nominaux représentant les éléments du langage documentaire est dressée, nous pouvons distinguer, les descripteurs, termes qui seront habilités à figurer dans l'index des documents, et les non-descripteurs ou termes interdits qui ne figureront pas dans l'index. Les descripteurs et les non-descripteurs sont liés par des relations sémantiques.

Le graphe constitué de descripteurs et de non-descripteurs liés par des relations sémantiques s'appelle *un thésaurus*. Certaines relations ont pour but de limiter les problèmes d'ambiguïté du langage, alors que les autres ont pour but de suggérer de nouveaux descripteurs à partir de ceux auxquels le documentaliste a pensé.

Le thésaurus s'attaque à deux problèmes différents, celui de *la synonymie* et celui de la *polysémie*.

Le principe du thésaurus est de constituer pour chaque concept la liste des mots qui peuvent l'exprimer. L'un des mots est alors choisi comme descripteur, les autres mots sont des non-descripteurs et leur usage dans l'index est interdit. Dans un thésaurus, il existe deux relations duales : la première relation part du descripteur pour arriver au non-descripteur, la seconde relation part du non-descripteur pour arriver au descripteur.

Concernant la polysémie, le danger est de trouver des documents non pertinents lors d'une interrogation, ce qui est fort possible lorsqu'on utilise des mots ayant plusieurs sens. Une des solutions serait d'utiliser comme descripteur un mot qui ne possède pas d'ambiguïté. Mais cette solution n'est pas toujours possible. En tenant compte du fait que le langage documentaire est un langage artificiel, il est alors utile de choisir comme représentant un mot composé non ambigu ou encore construire un pseudo-mot en qualifiant le mot ambigu par un autre.

Outre l'aide qu'il apporte pour résoudre les ambiguïtés du langage, le thésaurus aide le documentaliste dans l'exhaustivité de la description par des relations sémantiques de "suggestion" de nouveaux termes à mettre dans l'index. Il existe deux types de relations sémantiques de suggestion :

Les relations hiérarchiques (relations termes génériques et relations termes spécifiques).

Par exemple : Europe ---- terme spécifique ✍ France,

Espagne ✍ terme générique --- Europe.

La relation de terme associé. Il peut s'agir de termes que l'on trouve fréquemment associés : relation entre un agent d'une action et l'action elle-même, entre l'action et son objet, des termes co-occurrents.

Si l'indexation manuelle sur vocabulaire contrôlé permet une recherche sur des thèmes assez généraux de manière assez efficace par un personnel formé, ses principaux inconvénients sont la perte importante d'information par rapport au texte intégral et par conséquent la difficulté de répondre à des questions très précises. De plus, l'indexation manuelle engendre un coût financier non négligeable, il représente également une lourde tâche. Il apparaît donc illusoire d'envisager son utilisation pour des gros volumes de documents comme les bases de presse ou les pages Internet. Par ailleurs, le vocabulaire contrôlé est sujet à des variations dans les documents et peut être ambigu. L'ambiguïté des unités lexicales peut être également source de bruit.

I.3.1.2 L'interrogation booléenne de bases de documents indexés avec un vocabulaire contrôlé

Le modèle booléen permet de retrouver les documents dont la représentation correspond exactement à celle de la requête. Les requêtes booléennes consistent à combiner les termes entre eux, à l'aide d'opérateurs booléens. Il est alors possible de créer des recherches plus ou moins complexes. Il existe trois types d'opérateurs booléens : ET (AND), OU (OR) et SAUF (NOT).

L'interrogation booléenne donne pour résultat une partition de la base en deux sous-ensembles.

Le premier sous-ensemble donne accès aux documents dont l'index donne vrai à la fonction booléenne question, et qui de surcroît sont jugés pertinents par le système.

Le deuxième sous-ensemble est celui des documents dont l'index donne faux à la fonction booléenne question et qui sont jugés non pertinents. On appelle les documents de ce deuxième ensemble *« le bruit »*, mais il peut également comprendre des documents pertinents qui ne seront pas visualisables qu'on appelle « *silence* ».

L'interrogation booléenne pose plusieurs problèmes quant à l'accès au contenu des documents. En effet, ce type d'interrogation manque de souplesse. Il ne s'applique bien qu'à des recherches qui ne sont pas trop précises et pour lesquelles les concepts de sélection sont pris en compte par le vocabulaire contrôlé. Ce manque de souplesse est également vrai en vocabulaire libre où, bien que tout le vocabulaire soit pris en compte, les traitements linguistiques d'indexation sont souvent

inexistants ou élémentaires, ce qui rend difficile l'appariement parfait que nécessite le booléen.

De plus, l'interrogation booléenne est inadaptée pour traiter les documents partiellement pertinents et ne permet pas à un utilisateur de poser des questions précises en une seule fois. En effet, un utilisateur doit, pour obtenir une réponse pertinente, réaliser une requête booléenne qui associe plusieurs mots connectés par des opérateurs et ainsi prendre le risque de ne trouver aucune réponse à la requête. Une solution serait de poser des questions simples et de les combiner peu à peu pour trouver des questions optimales, mais ce n'est pas à la portée de tous les utilisateurs. Il est à noter qu'il existe, par ailleurs, des algorithmes qui permettent à partir d'un ensemble de mots, de trouver les meilleures questions booléennes qui ont des réponses non nulles (booléen pondéré).

I.3.1.3 Le texte intégral

C'est dans le milieu des années 70 que sont apparues les premières bases de textes intégraux et notamment dans le domaine juridique. La réponse à une requête n'est alors plus constituée d'une référence mais du texte du document ce qui constitue un vrai progrès.

La principale difficulté est alors de pouvoir établir une interrogation efficace, afin de parvenir à des réponses précises et exploitables. La technique la plus utilisée consiste à prendre comme index chaque chaîne de caractères comprise entre deux blancs, à l'exception des mots vides (ou mots outils). Ces derniers représentent à eux seuls près d'un tiers des mots d'un texte. Toutes les formes sont confrontées à un anti-dictionnaire comprenant les mots grammaticaux (prépositions, déterminants, pronoms) ce qui permet de les exclure de l'index. Un autre moyen est d'éliminer les mots qui reviennent le plus souvent parce que non discriminants. Ce type d'indexation représente beaucoup moins de perte d'information que l'indexation manuelle sur vocabulaire contrôlé qui ne concerne qu'un nombre réduit de mots.

Un des principaux problèmes de l'indexation en texte intégral est qu'elle ne tient pas compte des problèmes linguistiques (synonymie, polysémie, prise en compte des mots composés). Le problème de la synonymie s'est de plus largement aggravé, car on trouve dans le texte intégral les mots fléchis et dérivés, mal orthographiés ou encore pouvant accepter plusieurs variantes orthographiques. Il s'est donc avéré nécessaire de trouver des solutions mais ces dernières engendrent une complexification non négligeable des techniques d'interrogation.

L'utilisation d'opérateurs de troncature s'est trouvée être une solution pour trouver toutes les dérivations d'un même mot. En revanche, leur utilisation notamment sur une racine courte peut engendrer des aberrations (énormités).

Un autre problème engendré par l'accès au texte intégral est la quasi-impossibilité de trouver des mots composés. En effet, les systèmes d'indexation en texte intégral ne disposant pas de connaissances linguistiques, ne possèdent aucune représentation interne et précise de ce qu'est un mot, sinon celle d'une suite de caractères encadrés par des "blancs".

Il est donc très difficile de trouver des expressions figées ou des dates, des chiffres, des acronymes. En réponse à ce problème, il a été créé en plus des opérateurs booléens, des opérateurs de proximité permettant de trouver des mots relativement proches les uns des autres. C'est le cas des guillemets qui permettent de trouver une expression dans son intégralité ou encore l'utilisation d'un opérateur d'adjacence ADJ qui impose que les deux mots soient dans un ordre donné et qu'ils ne soient séparés que par des mots vides.

Enfin, le dernier problème engendré par l'avènement du texte intégral, est que l'on trouve vite confronté à la localisation des informations pertinentes dans un document.

En effet, on est passé d'une recherche de document à une recherche d'information dans des documents. Il est donc apparu comme nécessaire de mettre en évidence les mots de la question dans les documents et de pouvoir passer d'une occurrence à l'autre pour faciliter le repérage des passages pertinents.

I.3.2 Les modèles statistiques

A partir des années 70, les équipes de recherche se sont intéressées à un mode de comparaison question/document non plus booléen mais pondéré (mesuré). Une relation d'ordre de pertinence est établie sur l'ensemble de la base par rapport à la question. Les documents réponses apparaissent donc classés par ordre de pertinence. Ce type de comparaison est bien adapté aux questions longues et précises, et autorise de prendre une portion de texte comme critère de recherche. Ce sont essentiellement les équipes de recherches anglo-saxonnes qui ont travaillé sur ces modèles statistiques afin d'en obtenir une comparaison pondérée.

I.3.2.1 Le modèle vectoriel

Le modèle vectoriel a été crée par Gérard Salton [Salt 89] et concrétisé par le système de recherche SMART. Ce modèle consiste à représenter les documents et la requête dans un espace vectoriel dont les axes sont les termes (concepts ou vocabulaire non vide du texte intégral). Les coordonnées du vecteur dépendent de la

fréquence des termes dans le document ou la requête. Une distance est calculée entre le vecteur représentant chaque document et le vecteur représentant la requête. Chaque axe représente un mot. La dimension sur l'axe dépend en général de la fréquence du mot dans le document et sa répartition dans la base.
Le modèle vectoriel évalue une ressemblance entre la représentation de la requête celle des documents [Salt 75]. Un poids est affecté à chaque terme indexé. Ce poids détermine l'importance relative du terme dans le document. Chaque document est ainsi décrit par un vecteur dont les coordonnées correspondent au terme qu'il contient. De même une requête est traduite en vecteur. La distance entre un vecteur document et un vecteur requête donne une mesure de similarité qui est utilisée pour sélectionner les documents correspondant à une requête. Plus les vecteurs sont proches, plus leur contenu est semblable. Il est alors possible de calculer une distance entre la question et chaque document, mais également entre deux documents. On peut ainsi regrouper les documents par proximité de contenu.

I.3.2.2 Le modèle Latent Semantic Indexing

Le problème de l'accès au texte est essentiellement dû à l'écart entre les termes utilisés dans les questions et les documents. Pour diminuer cet écart, il faut pouvoir inférer, à partir des mots de la question, les mots équivalents contenus dans les documents. Pour cela, on peut se servir des relations sémantiques implicites qui sont représentées par les cooccurrences entre termes dans les documents. L'approche LSI (Latent Semantic Indexing) [Deer 90] consiste à réduire le nombre dimension de l'espace vectoriel tel que l'a défini Gerard Salton en s'appuyant sur le fait que les documents traitant des mêmes sujets ont des vocabulaires proches et sont donc proches dans l'espace vectoriel.

I.3.2.3 Le modèle Booléen pondéré

Cette approche consiste à regrouper les réponses par "intersections conceptuelles identiques" [Pier 00] et à trier les classes de documents obtenues par ordre décroissant de pertinence. La différence fondamentale avec les approches vectorielles ou LSI est que le poids du document ne dépend pas du fait que le mot recherché est courant ou non. Les applications visées par cette approche sont plutôt la recherche d'information dans les textes, alors que l'approche vectorielle consiste à évaluer si un document est globalement pertinent ou non.

L'avantage incontestable est qu'en une seule requête, un algorithme peut déterminer toutes les intersections possibles entre les mots de la question et les documents. Cette combinatoire ne pourra être faite que très difficilement manuellement.

I.3.3 La linguistique et la recherche d'information

La recherche d'information a utilisé à des degrés divers des méthodes faisant appel à des connaissances de nature linguistique. La nécessité d'identifier de la manière la plus précise possible les événements linguistiques des documents fait intervenir dans ce domaine différents sous-domaines de la linguistique. Outre le découpage en phrases et en mots d'un texte, le TAL (Traitement Automatique du Langage) intervient à différents niveaux de l'indexation notamment dans l'analyse morpho-syntaxique et l'analyse sémantique lexicale.

L'intérêt de l'analyse est de faire simuler à une machine le processus de décodage d'un texte tel qu'il peut être fait par un être humain. Pour automatiser une opération, on la décompose en tâches élémentaires, ainsi l'analyse se fait à différents niveaux.

I.3.3.1 L'analyse morphosyntaxique

Les connaissances morphologiques concernent la manière dont les mots sont construits à partir des unités minimales de signification [Boui 01]. Cette analyse permet de ramener à une forme canonique les mots reconnus en séparant les variations grammaticales afin d'identifier les morphèmes, les terminaisons grammaticales, les caractères spéciaux, etc…

L'analyse morphosyntaxique est utilisée pour reconnaître les mots et en proposer une ou plusieurs normalisations. Cette analyse servira à identifier les différentes représentations d'un même concept. Les mots reconnus peuvent être des mots simples ou des expressions idiomatiques et peuvent faire l'objet de correction orthographique de façon automatique. A l'inverse, les mots inconnus lors de l'analyse, sont soit traités par défaut, soit signalés pour une mise à jour du dictionnaire ou une correction. Lors de l'analyse morphologique, chaque mot est étiqueté en fonction de ses propriétés linguistiques. Cette étape est capitale sans elle le reste de l'analyse est impossible car elle permet un comptage des concepts.

L'analyse morphologique permet de passer à l'analyse syntaxique. L'étiquetage syntaxique permet de déterminer les mots vides, inutiles dans l'indexation en distinguant un nom d'une conjonction (le nom "or" et la conjonction "or"). Cette analyse syntaxique facilitera ensuite les problèmes de traduction et de synonymie. Par ailleurs, au niveau de la phrase, cette analyse permet de déterminer la structure syntaxique en vue d'une recherche et d'un résultat plus pertinent.

I.3.3.2 Analyse sémantique lexicale

La sémantique est le point d'articulation entre la forme des textes, la syntaxe et le lexique, et le sens référentiel (les objets et les relations des domaines) ou pragmatique (le contexte d'énonciation ou les connaissances partagées sur le monde). La description sémantique se réalise grâce à des traits sur les éléments lexicaux, des liens entre ses éléments et des règles de construction du sens.

En TAL (Traitement Automatique du Langage), l'analyse sémantique consiste à associer à une séquence des marqueurs linguistiques censés y consigner le sens. La plupart des représentations sémantiques s'appuient sur une analyse syntaxique effectuée préalablement. Le niveau sémantique est encore plus complexe à décrire et à formaliser que les autres analyses. Le sens se compose d'un certain nombre de mots identifiés par l'analyse morphologique et regroupés en structure par l'analyse syntaxique. Ces mots et ces structures constituent autant d'indices pour le calcul du sens.

L'analyse sémantique permet la désambiguïsation des relations sémantiques entre les mots. Ainsi elle rapproche les termes utilisés dans une requête de ceux contenus dans les documents. Plus le domaine sera général, plus l'analyse sera difficile. A l'inverse, dans un lexique restreint, les ambiguïtés seront plus rares et les documents en deviendront plus exploitables. Ainsi, en effectuant une recherche dans un domaine clairement délimité, le résultat sera plus pertinent pour l'utilisateur.

Une analyse purement sémantique n'est pas suffisante, c'est pourquoi, l'analyseur comporte des règles concernant le contexte dans lequel se situe le texte analysé ainsi qu'un certain nombre de règles faisant partie de la syntaxe.

I.4 Conclusion

Avec l'augmentation prodigieuse de l'information scientifique sur Internet et les difficultés présentées par les techniques de recherche courantes, plusieurs voies d'amélioration de l'accès à cette information sur le Web sont mises en œuvre. Par exemple, l'amélioration des méthodes de recherche sur Internet, y compris les agents intelligents de la recherche et les nouvelles techniques d'indexation des documents.

Plusieurs approches d'indexation sont utilisées pour faciliter l'accès aux documents scientifiques. Parmi les approches qui ont prouvé leur efficacité et leur performance, on peut évoquer l'indexation des citations. Cette approche est détaillée dans les chapitres suivants.

Chapitre II

Indexation classique des citations scientifiques

2.1 Introduction

L'indexation de citation est une méthode pour organiser le contenu d'une collection de documents de manière à surmonter les imperfections des autres méthodes d'indexation. L'avantage primaire de l'indexation de citation est qu'elle identifie les relations (rapports) entre les documents qui sont souvent négligés dans les méthodes d'indexation traditionnelle.

Un avantage secondaire important est que la méthode d'indexation de citation est particulièrement bonne convenue à l'utilisation des méthodes d'indexation automatique qui n'exigent pas des indexeurs spécialisés. Ceci aide à rendre les indices de citation plus usuels que les autres indices traditionnels. En outre, les citations, qui sont des descriptions bibliographiques des documents, ne sont pas sensibles à l'obsolescence scientifique et technologique de même que les termes utilisés dans les indices traditionnels [Boll 99].

L'analyse des citations est une technique permettant d'évaluer l'influence et la signification intellectuelle des recherches publiées dans le temps. Elle fournit la capacité unique d'indiquer exactement quand et où un document ou un auteur particulier a été cité la première fois Les historiographes de la science et de technologie peuvent employer l'analyse de citation pour identifier les individus, les établissements, et les pays haut-cités dans le temps en terme de publications enregistrées soit individuelles ou collectives [Kurt 99].

2.2 L'indexation de citation

2.2.1 Les citations et les références bibliographiques

Une citation est les reconnaissances reçus d'un document par un autre. La première raison de l'usage des citations est le respect aux créateurs intellectuels, et représente une identification des publications originales dans lesquelles une idée ou un concept a été décrit [Wout 99].

Il est important de clarifier la distinction terminologique entre "*citation*" et "*référence*" ; si le papier **R** contient une apostille bibliographique qu'il utilise et décrit le papier **C**, alors **R** *contient une référence* à **C**, et **C** *a une citation de* **R**.

Le nombre de références d'un papier est mesuré par le nombre d'articles dans sa bibliographie, alors que le nombre de citations d'un papier est mesuré par le nombre d'articles qui le mentionnent [Pric 86].

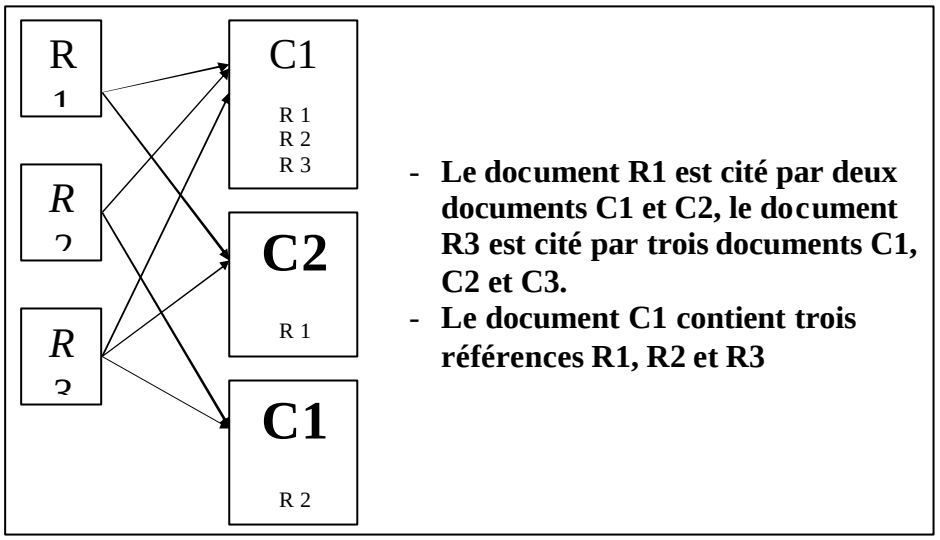


Figure 2.1 Les citations et les références bibliographiques

En résumé, les citations symbolisent l'association conceptuelle des idées scientifiques comme reconnues entre les recherches publiées par différents auteurs [Garf 88]. Par les références qu'ils citent dans leurs papiers, les auteurs font des liaisons explicites entre leurs recherches courantes et le travail antérieur dans les archives de la littérature scientifique.

Ces associations conceptuelles ont été décrites en tant que " *transactions intellectuelles",* reconnaissances formelles *"dette intellectuelle"* à une première source d'information [Garf 88].

2.2.2 L'indexation des citations scientifiques

L'indexation de citation est basée sur un concept simple: Les références d'un auteur aux informations précédemment enregistrées identifient une grande partie des premiers travaux qui sont convenables au sujet de son présent document. Ces références s'appellent généralement les citations.

Un Index de citation est une liste structurée de toutes les citations dans une collection de documents. De telles listes sont habituellement arrangées de sorte que le document cité soit suivi des documents qui le citent [Garf 88].

2.2.2.1 Historique

La première application pratique de l'indexation de citation était *les citations de Shepard,* un outil de référence légal qui a été en service depuis 1873. *Les citations de Shepard* doit son existence au fait que la loi américaine, comme la loi anglaise, fonctionne sous la Théorie du *Stare Decisis* [Wout 99].

Stare Decisis signifie que toutes les cours de justice doivent suivre leurs propres précédents aussi bien que ceux établis par des cours plus élevées. Les précédents sont les décisions antérieures prononcées.
Pour vérifier un cas sous la théorie de *Stare Decisis***,** un avocat doit baser son argument sur des décisions précédentes concernant un point similaire *vis-à-vis* de la loi. Avant de présenter la décision précédente comme argument, l'avocat doit s'assurer que la décision n'a pas été annulée (rejetée, outrepassée), ou n'est pas limitée d'une manière quelconque. *Les citations de Shepard* permettent à l'avocat de faire ceci avec un minimum de dérangement.

Un cas légal est toujours mentionné par un code qui comprend le volume et le numéro de page du document dans lequel le cas est rapporté. Une fois qu'un cas est rapporté d'une manière permanente, son code de référence devient fixe pour tout le temps.

Exemple:

301.U.S.356 est une référence au cas rapporté dans le volume 301 des rapports de cour suprême des Etats-Unis à la page 356. Des statuts sont également mentionnés d'une façon semblable. Ainsi, ch16Sec24NJRS se rapporte au chapitre 16, la section 24 du règlement corrigé de New Jersey.

Tirant avantage de ce système de codage, *Frank Shepard* a inventé une liste qui montre chaque cas dans lequel une décision rapportée est citée dans un cas ultérieur. La liste montre également quels statuts et journaux citent la décision originale.

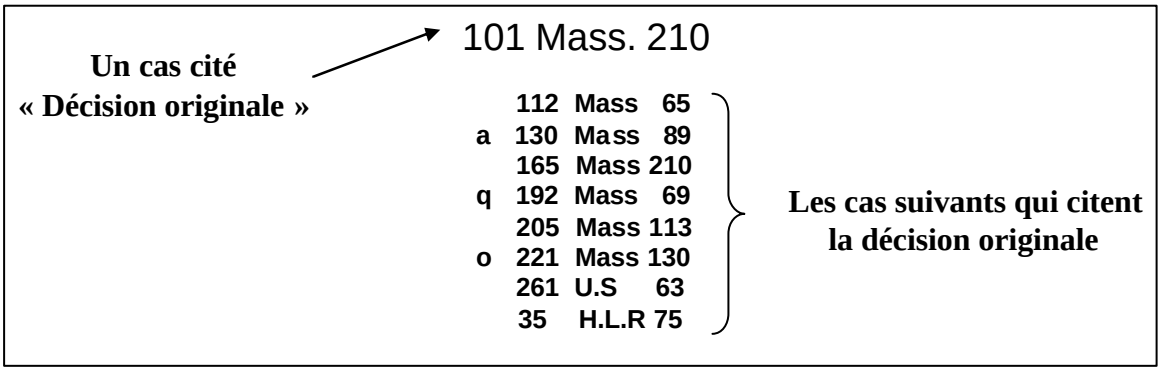


Figure 2.2 Les citations de *Shepard's* (Un cas cité et les citations ultérieurs)

La Figure 2,2 représente une liste de citations de *Shepard* pour un cas imaginaire. Le cas cité est *101Mass.210,* les articles énumérés au-dessous sont les citations. Les lettres qui précédent les codes des cas prouvent la décision originale dans les citations, (a) affirmé puis (q) remis en cause et (o) finalement rejetée (outrepassé).

Pour utiliser les citations de *Shepard*, un avocat doit d'abord localiser une *décision précédente* associée (reliée) à son cas courant. Il fait ceci en consultant un *sommaire*, un *index*, ou une *encyclopédie* qui lui fournira le nombre de cas pour n'importe quelle décision donnée. L'avocat recherche alors le nombre de cas dans les citations de *Shepard* et trouve tous les cas suivants qui citent cette décision.
À partir de cette information, il peut déterminer si la décision originale a été affirmée ou modifiée de quelque façon. Ainsi, dans l'exemple donné sur la Figure 2.2, la décision originale ne pourrait pas être employée comme précédent parce qu'elle a été plus tard rejetée (outrepassée).

2.2.2.2 Les problèmes des index traditionnels

Après la 2ème guerre mondiale, les utilisateurs de la littérature scientifique et technologique trouvaient de plus en plus des difficultés j trouver l'information convenable à leur propre travail. Plusieurs facteurs ont causé cette situation. L'un de ces facteurs était que la taille de la littérature se développait très rapidement
Pendant que le volume croissant des informations scientifiques commençait à encombrer le nombre limité des indexeurs qui pourraient être économiquement soutenus, il y avait un retard de six mois à plusieurs années avant que les papiers soient classifiés.
Ceci a eu comme conséquence que de plus en plus les scientifiques passent le temps reproduisant inutilement les travaux existants.
Un autre facteur dans les problèmes de la littérature scientifiques était le besoin croissant d'échanger l'information entre les disciplines scientifiques. La majorité d'index des matières a couvert seulement un champ ou une discipline.
Par exemple, un chimiste choisissant les matériaux appropriés pour les implants chirurgicaux ou les organes internes artificiels pourrait trouver l'information utile dans les journaux chimiques, les journaux médicaux, ou dans les journaux techniques.
Il y avait également d'autres problèmes. Les termes de l'indexation utilisés dans les index des matières sont souvent *ambigus* et se prêtent à différentes interprétations, particulièrement quand l'utilisateur n'est pas entièrement familier avec les détails d'un *système d'indexation* particulier.
Les index de matière rencontrent également le problème *d'assigner des termes* pour les nouveaux concepts, particulièrement dans les domaines qui se développent rapidement comme la biochimie. En effet, dans la plupart de cas, l'accord de ce qui est le terme approprié pour un concept ne peut être atteint jusqu'à une certaine heure

après que le papier original présentant le concept ait été indexé par un terme inadéquat.

Puisque les indexeurs possèdent différentes capacités intellectuelles et connaissances techniques, deux indexeurs emploieront souvent différents *mots clés*, ou *des titres*, ou des *termes de sujet* quand ils indexent le même document.

Ainsi, il n'est pas étonnant de trouver les documents relatifs classifiés sous les titres des matières entièrement différentes sans indiquer au chercheur comment ceci s'est produit.

Par exemple, un papier important 1963 sur le sujet "les variations saisonnières de naissance"[Tlet 63] est indexé sous le titre de matière de la *"périodicité"* dans l'édition 1964 de *l'index Medicus.* Il est peu probable que n'importe qui recherchant l'information sur des *« variations saisonnières de naissance »* penserait regarder sous le terme *"périodicité"* puisque c'est tout à fait un concept différent que "*la variation saisonnière*".

Ces types de problèmes ont fait clairement apparaître le besoin de système qui fournirait un index unifié de la littérature scientifique, exempt des difficultés sémantiques, et ne dépendant pas de la connaissance soumise par les sélecteurs [Wout 99].

2.2.2.3 Le début des index de citations pour la Science.

La tradition scientifique exige que quand un scientifique ou un technologue honorable publie un article, il doit se référer à des articles qui se relient à son thème. Ces références sont censées identifier les premiers chercheurs dont les concepts, méthodes ou appareils, ont inspiré ou ont été employés par l'auteur dans le développement de son propre article.

Quelques raisons spécifiques pour l'usage des citations sont :

1. Respecter les créateurs (inventeurs).
2. Donner la confiance pour les travaux relatifs.
3. Identification des méthodologies, des équipements, …
5. Corriger son propre travail.
6. Corriger les travaux des autres.
7. Critiquer les travaux précédents.
8. Justifier ses travaux.
9. Alerter les chercheurs du prochain travail.
10. Fournir le chemin du travail *mal disséminé*, *mal indexé*, ou *non cité*.
11. Authentifier les données et indexer les constantes physiques, ...

12. Identifier les publications originales dans lesquelles une idée ou un concept a été décrite.

Au début des années 50, la disponibilité de ce système intégré pour lier les articles scientifiques a commencé à susciter l'attention comme base possible d'un système d'indexation pour la littérature scientifique.

2.2.2.4 Exemple le Projet d'indexation : La Bibliothèque médicale galloise

C'est un projet apparu en 1952, sou la direction de Dr. Chauncey Leake, il a comme objectif de trouver la meilleur méthode pour indexer la littérature médicale dans la bibliothèque médicale de Johns Hopkins.

GARFIELD

Eugène Garfield, un des investigateurs du projet gallois. Garfield s'est rendu compte que presque chaque phrase dans une revue d'article révisé est soutenue par une citation à des travaux précédents. Ainsi, l'article révisé peut vraiment être considéré comme une série de rapports d'indexation.

Le problème est alors comment transformer ces rapports en format cohérent qui serait utilisé comme un index ?

ADAIR

En 1953, le projet gallois a conduit j une conférence dont les publications ont été rapportées dans un journal du Colorado. Cet article a été lu par *William C. Adair*, qui était un ancien vice-président de la société qui a produit *les citations de Shepard*. Adair a écrit au projet gallois et a suggéré de considérer la méthode utilisée par *Shepard* comme technique possible d'indexation.

Après avoir examiné *les citations de Shepard,* Garfield s'est rendu compte que le principe de "citation" pourrait fournir des moyens pour indexer les articles des revues. Ceci pourrait être étendu à la littérature scientifique en général.

Après la fin du projet gallois, Garfield a commencé son travail dans la bibliothèque de la Science à l'Université de Columbia. Pendant cette période, il a continué la correspondance avec Adair et a commencé à écrire un article détaillé sur les index de citation pour la littérature scientifique.

Garfield a proposé qu'Adair écrive un article plus court qui expliquerait, d'une façon générale, l'opération des citations de Shepard, qui est paru en juillet de 1955 [Wout 99].

2.1.2.5 Index de citation de la Génétique.

En 1961, l'institut national de la santé lance un programme coopératif avec l'institut de l'information scientifique (ISI) de Garfield pour préparer un index de citation pour le champ de la génétique.

En plus de la préparation de l'index, le programme devait étudier et vérifier les points suivants au sujet d'index de citation :

1. Devrait-il y avoir un seul index de citation pour toutes les sciences et technologies, et les différentes spécialités.
2. Quelles sont les meilleures possibilités pour classer l'index de citation (auteur, journal, etc.) ?
3. Quelles sont les techniques que l'on peut employer pour regrouper les informations des citations ?

Garfield décide d'entreprendre une approche complète et interdisciplinaire pour préparer un index de citation et puis d'extraire l'index de citation de la génétique à partir de cette base d'information. La base de données interdisciplinaire a été par la suite employée pour produire le premier *index de citation* de la Science qui a été édité en 1963. Le premier SCI *(Scientific Citation Index)* a couvert la littérature de l'année civile de 1961. Il a couvert 613 journaux avec un contenu de 1.4 million de citations.

L'index multidisciplinaire et interdisciplinaire de citations scientifiques est maintenant employé presque par toutes les bibliothèques de chaque université importante aux Etats-Unis. C'est l'index le plus utilisé dans la science et la technologie [Garf 88].

2.3 Description de l'index des citations scientifiques (SCI)

L'index de citation scientifique (SCI) fournit un index pour le contenu de chaque sujet publié. Les journaux sont considérés comme des journaux originaux et les articles qu'ils contiennent s'appellent les articles sources. Tous les journaux sont indexés d'une manière globale pour éliminer le problème de savoir si un article particulier est couvert ou non. Tous les articles originaux et la plupart des autres articles utiles dans un journal sont traités, y compris les éditoriaux, les lettres, et les réunions. Les articles éphémères tels que les annonces et les notices des nouvelles sont négligés [Melv 83].

Avant l'impression de l'index:

- ✍ Chaque thème d'un journal source est *publié* et *étiqueté* pour assurer que toutes les données appropriées seront enregistrées.

- Tous *les titres en langue étrangère* sont traduits en anglais.
- Toutes *les citations* sont traitées, où en pratiques, les citations sont également extraites à partir du texte.
- Une carte perforée séparée est préparée pour chaque *article cité* apparaissant dans un *article source* traité.
- Pour chaque *article source*, un ensemble de cartes perforées contenant *l'auteur (s), le titre, le journal*, etc... est également préparé.
- Chaque carte perforée est vérifiée par comparaison directe avec le journal original.
- Une fois que les cartes perforées sont vérifiées, les données sont transférées à partir des cartes aux bandes magnétiques. Pendant ce processus, l'ordinateur exécute une routine d'unification qui élimine les erreurs de la littérature originale telle que les fautes d'orthographe dans les *noms des auteurs* et *les titres des publications cités*.

2.3.1 Format d'arrangement

SCI se compose de trois indices séparés mais reliés. Ce sont *l'index des citations, l'index des sources,* et *l'index des termes.* Chacun de ces derniers constitue le SCI qui est publié chaque trimestre de l'année, les indices de quatrième trimestre sont incorporés dans la cumulation annuelle pour chaque index.

2.3.1.1 L'index des citations

- Les auteurs cités sont classés par ordre alphabétique.
- Une entrée pour un article cité contient le premier nom des auteurs initiaux, l'année de publication de l'article cité, et le nom de la publication dans laquelle l'article cité est apparu avec son volume et numéro de page.
- Quand il y a plus d'un article cité pour n'importe quel auteur, ceux-ci sont arrangés chronologiquement par année de publication.
- *Les articles source*s qui citent un article particulier de référence sont classés par ordre alphabétique par l'auteur de source immédiatement au dessous de chaque ligne de référence.
- La ligne d'article source contient le nom de l'auteur de citation, titre de la publication dans laquelle la citation est apparue, l'année de publication, volume, et la page.
- Il y a également un symbole codé qui indique si l'article de citation est un article, un résumé, un éditorial…
- Dans l'index de citation, seulement le premier auteur est montré pour les articles cités et les citations.

✍ Une section séparée de l'index de citation employée pour les articles inconnus (aucun auteur personnel indiqué pour le travail cité). Ces articles sont classés par ordre alphabétique par les titres des publications citées.

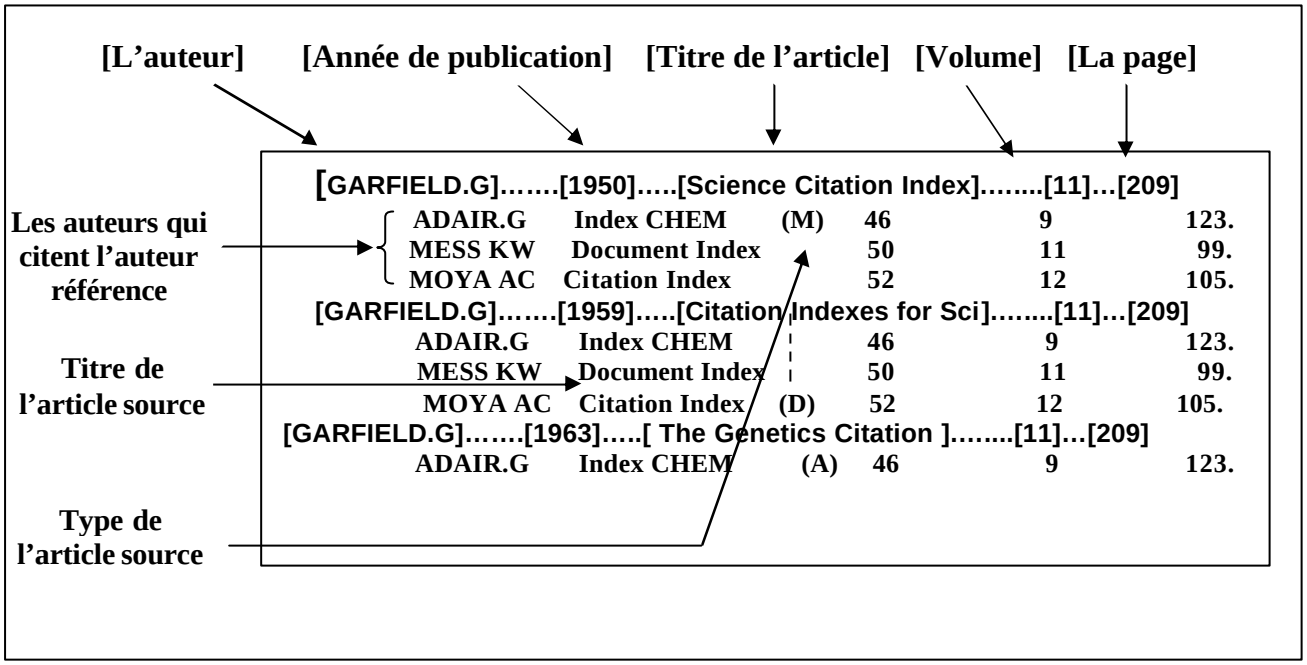


Figure 2.3 : Exemple d'index des citations de SCI

Les lettres alphabétiques indiquant le type de publication :

- (A) Résumés des articles publiés
- (C) Corrections, errata, etc…
- (D) Discussions, articles de conférence
- (E) Éditoriaux,
- (I) Articles au sujet des individus (hommages, nécrologes,…etc.)
- (L) Lettres, communication, etc.
- (M) Résumés des réunions.
- (N) Notes techniques.
- (Q) Bibliographie pour SCI fournie après publication primaire, par l'auteur source.
- (R) Revues et bibliographies.
- Blancs : Articles, rapports, exposés techniques, etc...

2.3.1.2 Index des sources

Pour chaque citation:

- ✍ Les auteurs sont classés par ordre alphabétique.
- ✍ Les entrées fournissent tous *les co-auteurs*, le *titre complet de citation, le titre du journal, le volume, origine, la page, l'année, le type d'article* (revue, lettre, correction, etc.), et *le nombre de références dans la bibliographie de l'article de source*.

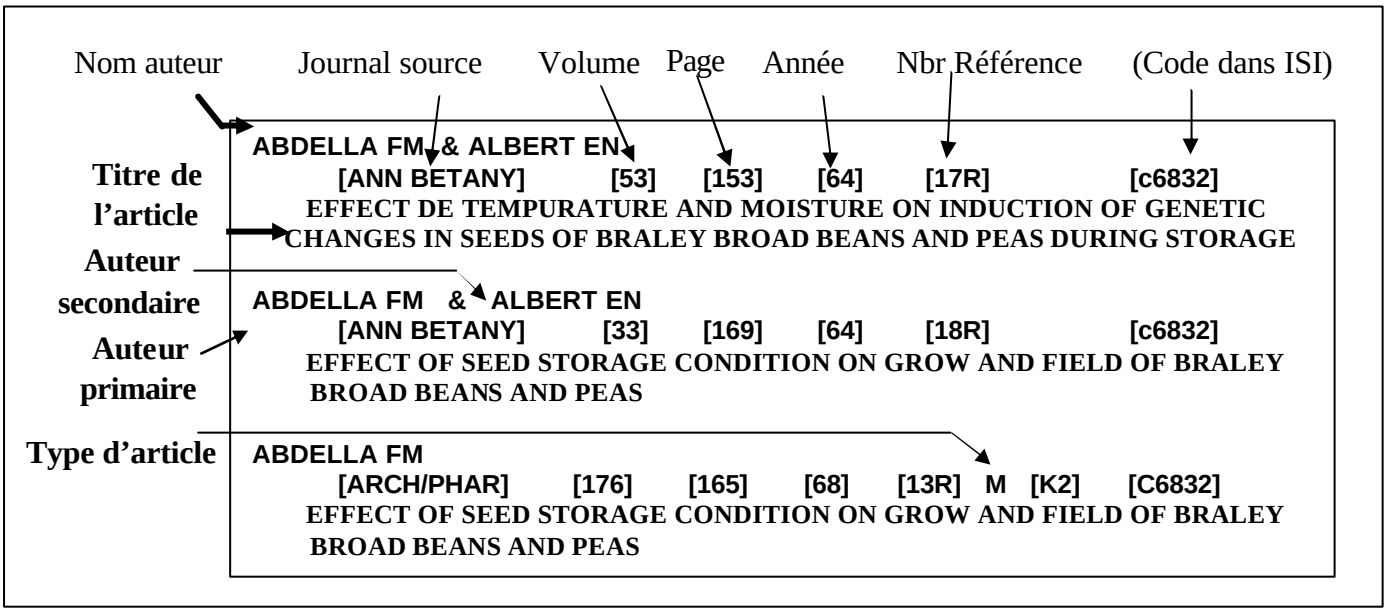


Figure 2.4: Exemple d'index des sources de SCI

2.3.1.3 Permuterm Subject Index (PSI)

Signifie l'index des termes permutés. En PSI**,** les termes "permutés" sont employés dans leur sens mathématique correct. Ils sont extrais à partir de *l'index des Mots Clés dans le Contexte* (KWIC). Cet index donne les mots clés dans le titre d'un article et fait la permutation entre eux.

Pour produire *PSI***,** un programme est utilisé pour permuter tous les mots significatifs dans chaque *titre* et *sous-titre* de chaque article inclus dans *l'index des sources*. Toutes les paires possibles des termes sont formées. Avec ce système, chaque mot significatif prend est considéré comme un terme primaire aussi bien que comme *co-terme* [Melv 83].

La Figure 2.5 ci dessous montre les entrées d'indexation quand la technique de permutation des termes est employée.

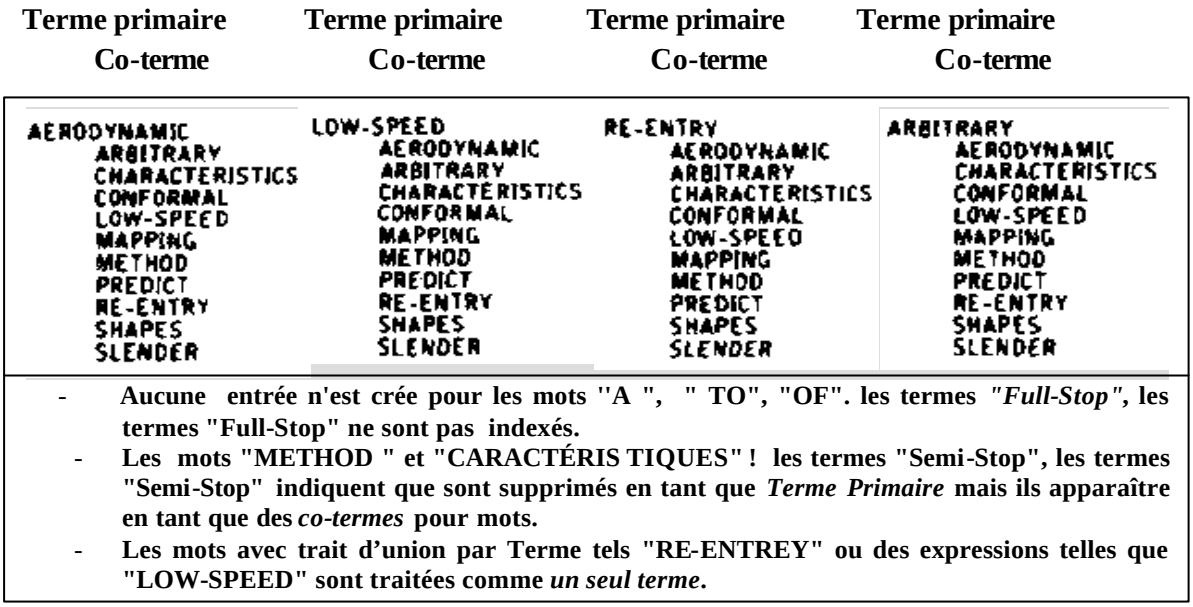


Figure 2.5 : Exemple d'index qui résulte quand on applique la technique des termes permutés sur le titre d'un article

- PSI est classé par ordre alphabétique par rapport au terme primaire. Tous les co-termes se combinant avec un terme primaire particulier sont découpés et énumérés dans l'ordre alphabétique au-dessous du terme primaire.
- Les co-termes commençant par des nombres apparaissent à la fin de la liste.
- Chaque co-terme est associé au nom de l'auteur dont l'article contient ce co-terme et son *terme primaire* associé**.**
- Pour les entrées inconnues, le titre du journal est donné au lieu du nom de l'auteur.

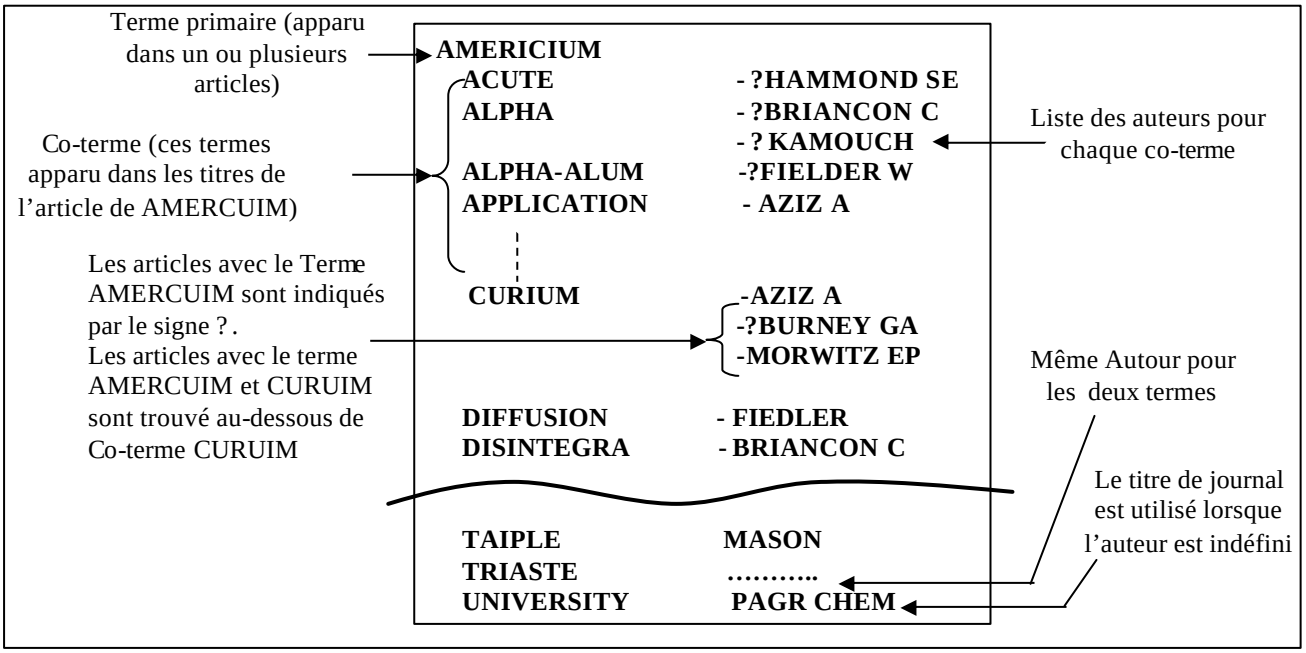


Figure 2.6 Exemple d'index des termes permutés

2.4 Techniques de bases de la recherche avec SCI

L'utilisation de *l'index de citation* implique les étapes suivantes :

- Le chercheur commence par le nom d'un auteur qu'il a identifié.
- Il accède alors à *l'index de citation* et il cherche le nom de cet auteur. Une fois que le nom de l'auteur est localisé, le chercheur peut voir les articles qui l'ont cité.
- Le chercheur note alors l'auteur, le journal, le volume, et la page de chaque article de citation.
- Le chercheur alors se tourne vers l'index de source et cherche le nom de l'auteur de citation. A cette entrée, il trouvera les données bibliographiques complètes de citation comprenant le titre complet et tous les co-auteurs.
- Le chercheur peut maintenant examiner les titres des articles de source et choisir les articles qui semblent très probablement être appropriés à sa matière.
- Il peut alors obtenir les journaux contenant les articles d'intérêt de la bibliothèque.

La recherche peut être facilement augmentée afin d'établir une bibliographie plus étendue pour une requête particulière. Par exemple, une fois qu'il trouve un certain nombre d'articles de source, le chercheur peut employer les bibliographies d'une ou plusieurs de ces derniers pour fournir les noms d'autres auteurs pour les rechercher dans *l'index de citation.* Ce processus s'appelle "cycle"**.**

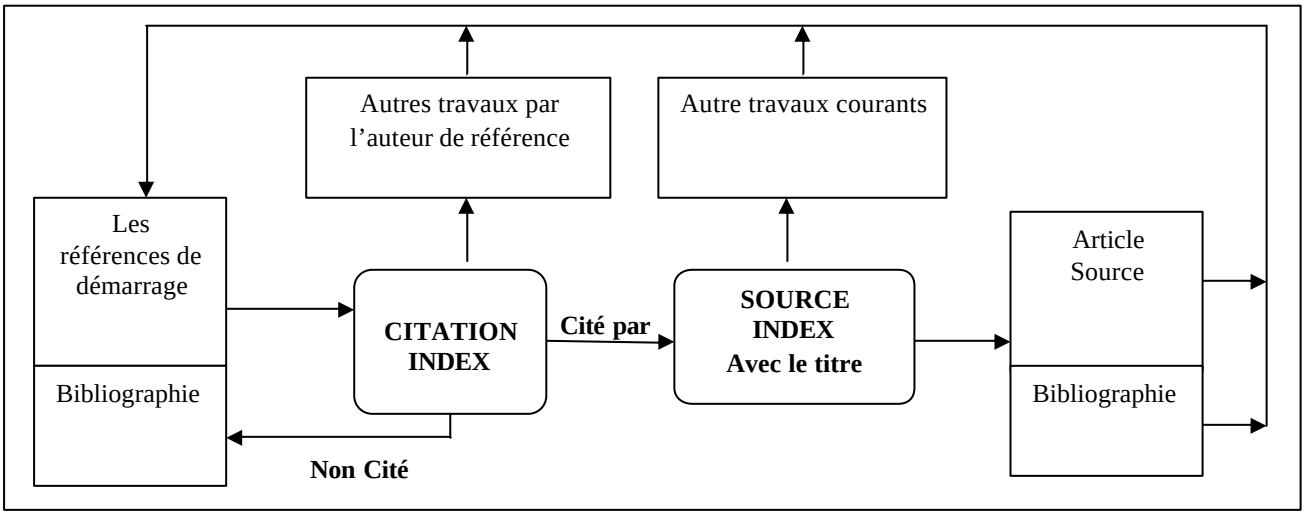


Figure 2.7 L'utilisation de "Cycling" dans la recherche par SCI

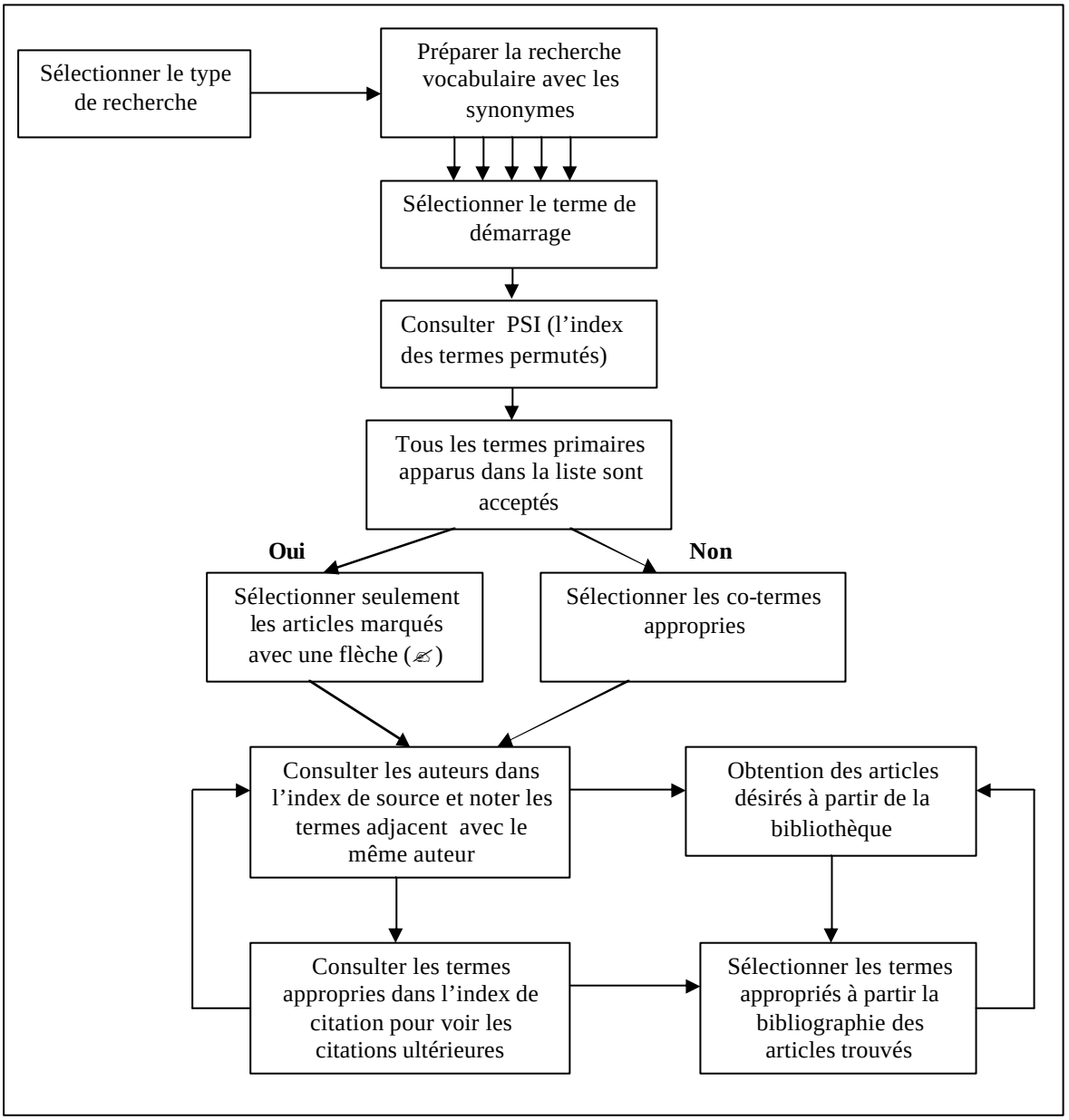


Figure 2.8 L'utilisation de PSI dans la recherche par SCI

2

La meilleure manière d'évaluer l'indexation de citation est d'examiner comment *l'index de citation de la Science (SCI)* répond aux insuffisances précédemment discutées avec les indices traditionnels. Ces derniers peuvent généralement être résumés comme suit :

1. Incapacité de traiter le volume croissant de la littérature scientifique sur une base convenable.
2. Capacité limitée pour lier les disciplines pour rassembler les informations relatives.
3. Difficultés sémantiques en préparation et utilisation des indices.

2.5.1 Globalité et opportunité

Pour obtenir la couverture complète de la littérature, SCI se base sur la loi de Bradford [Brad 69]. En général, cette loi déclare qu'un petit pourcentage des journaux explique un grand pourcentage des articles significatifs dans n'importe quel domaine indiqué de la science. À l'appui de cette loi, une étude des abrégés sur la Physique par Keenan et Atherton montre que 50% des articles sont pris seulement de 19 journaux. En outre, les études sur *Index Chemistry* prouvent que 100 journaux expliquent 98% de tous les nouveaux articles en chimie synthétique [Wout 99].

D'autres analyses ont indiqué que cette concentration d'information dans quelques journaux, non seulement pour des disciplines particulières, mais dans la littérature scientifique en général. Professeur Derek J. indique qu'environ 1000 journaux contiennent 80% *de tous les* articles scientifiques [Wout 99]. L'Institut Scientifique d'Information (ISI) que les 1000 journaux les plus fortement cités expliquent 90% de la littérature *significative*.

Ces résultats mènent au raisonnement que si les 2200 journaux couverts par *le SCI* sont correctement choisis, la majeure partie de la littérature scientifique importante du monde sera indexée malgré l'existence de 30 à 50 mille journaux dans le monde.

L'éditeur du *SCI* utilise plusieurs méthodes pour assurer que les journaux couverts sont significatifs.

- ✍ L'utilisation des experts en différentes disciplines pour la sélection des journaux.
- ✍ Des analyses à grande échelle de citation sont faites pour voir les journaux le plus fréquemment cités.

Ces informations sont particulièrement utiles en déterminant les journaux qui sont plus utilisés dans les domaines naissants de la science.

2.5.2 Possibilité de recherche multidisciplinaire

L'index de citation a gagné par la construction des liens intégrés entre les documents fournis par les citations des auteurs en listant l'ensemble de tous les articles avec les citations communes. Avec cette capacité unique de grouper un ensemble d'articles qui sont apparemment indépendants et qui sont très importants pour les chercheurs. Par exemple, la chimie de l'eau est convenable à l'océanographie, mais elle est également convenable à un vaste choix d'autres problèmes dans la biologie, la physique, la chimie, et d'autres domaines appliqués.

Avec SCI, un article courant cite un article précédent donné, il sera indexé sous l'article cité. Cela ne change rien si l'article de citation apparaissait dans un journal de physique, un journal de chimie, un journal de technologie, ou n'importe quel

autre type de journal. Par conséquent, un chercheur employant le SCI peut identifier un groupe d'articles liés à sa matière, mais qui ont été édités dans une variété de journaux.

2.5.3 Résolution des problèmes sémantiques

L'index de citation résout les problèmes sémantiques dans les indices traditionnels en employant les signes conventionnels de citation plutôt que les mots pour décrire le contenu d'un document.

Il est plutôt difficile pour que la plupart des personnes comprennent ce concept d'abord. Par conséquent, l'exemple plutôt prolongé suivant est présenté comme aide à la compréhension [Wout 99].

En 1963, professeur J. Lederberg a publié un papier dans *Nature* sous le titre *"Molecular Biology, Eugenics and Euphenics*". Dans cet article, il a établi le mot *"euphenics"* comme synonyme pour le concept de *"engineering human development*". Aussi longtemps que cet article était le seul dans la littérature sur *l'euphenics*, il y avait une équivalence entre le mot *"euphenics"* et les citations qui ont identifié le document dans lequel il est apparu la première fois. Le mot *"euphenics"* et la citation *"Lederberg J., 63, Nature 198, 428"* étaient essentiellement des symboles équivalents pour le sujet discuté dans le papier de Lederberg.

Si d'autres auteurs emploient le terme *"euphenics"* dans un autre journal. D'habitude, les auteurs suivants accorderont l'honneur à Lederberg comme le créateur du terme en citant son papier original.

En conséquence, dans un système d'indexation de citation, les nouveaux papiers seraient automatiquement groupés ensemble sous la citation "Lederberg J., 63, Nature 198, 428". *Dans un système d'indexation de mot*, les papiers suivants seraient groupés ensemble sous le terme *"euphenics"*. Les deux méthodes atteindraient le même objectif pour rendre l'information sur l'*"euphenics"* accessible. Dans un système, *le mot* est le terme d'indexation ; dans l'autre, *la citation* est le terme d'indexation.

Une fois qu'on comprend comment une citation peut servir de terme d'indexation, il n'est pas difficile de montrer pourquoi les citations sont fréquemment meilleures que des mots dans ce rôle.

- En portant l'exemple de Lederberg plus loin, si un autre auteur parle de *"engineering human development"* mais ne mentionne pas le mot *"euphenics"*. Aussi longtemps que l'auteur cite le papier original par Lederberg (qui est fortement probable), le nouveau papier sera indexé sous le papier de Lederberg dans un index de citation. La chance est très mince, cependant, que n'importe quel Indexeur par

mot compare *“engineering human development”* avec *"euphenics"*. Ainsi, le papier qui ne mentionne pas spécifiquement *"euphenics"* a une probabilité élevée d'être indexé sous d'autres termes.

- Considérer la même situation du point de vue de l'utilisateur d'un index. Si le chercheur est au courant du terme *"euphenics"*, les indices du mot lui permettront de trouver l'article de Lederberg et les articles qui mentionnent spécifiquement le terme *"euphenics"* sont ignorés.

Avec le SCI, il suffit que le chercheur sache seulement que Lederberg avait publié sur sujet. Il cherche simplement le nom de Lederberg afin de trouver le papier original et tous les papiers qui le citent. C'est particulièrement utile à un chercheur qui n'est pas au courant du langage des différentes disciplines.

2.5.4 Possibilités uniques des index de citation

L’index de citation résout plusieurs difficultés inhérentes aux indices traditionnels. De plus, il peut accomplir certaines taches inaccessibles aux autres indices. Le plus important de ces possibilités est probablement la capacité d'apporter le chercheur en avant dans le temps à partir d'une référence connue. Le SCI est établi de sorte que quand n'importe quel article origine est apparu, il soit indexé dans l'index de citation aussi longtemps qu'il est cité au moins une fois dans l’année actuelle dans un des journaux.

Dès que le chercheur localise les premiers "articles cités", il est avancé aux articles qui citent actuellement l'article original. Ceci pourrait établir un lien de cinquante ans ou de plus, ou il peut prendre le chercheur en avant dans les incréments aussi petits qu'une année.

En utilisant ces capacités des indices de citation, il est nécessaire de répondre aux questions des chercheurs comme celles-ci :

- ✍ Ce concept de base a-t-il été appliqué ailleurs ?
- ✍ Cette théorie a-t-elle été confirmée ?
- ✍ Cette méthode a-t-elle été améliorée ?
- ✍ Y a-t-il une nouvelle synthèse pour ce vieux composé ?
- ✍ Y a-t-il eu des errata ou des notes de correction éditées pour cet article ?

En outre, n'importe quel scientifique peut légitimement souhaiter déterminer si son propre travail a été appliqué ou critiqué par d'autres.

Les indices de citation facilitent ce type de rétroaction (FeedBack) dans le cycle de communication. Une autre utilisation des indices de citation est qu’ils permettent d’identifier rapidement les scientifiques travaillant actuellement dans les branches spéciales de la science pour but de correspondance ou sélection personnelle.

En conclusion, Une remarque devrait être faite au sujet de la capacité exceptionnelle des indices de citation d'être un outil pour évaluer les pratiques littéraires et la structure de la littérature scientifique.
L'utilisation des données de citation et les réseaux des articles reliés servent pour tracer l'histoire d'un sujet. Les comptes de citations peuvent également être employés pour déterminer la durée d'une recherche ou l'intérêt d'un article ou d'une matière donnée.
Plusieurs études ont été faites sur l'indexation de citation. Barlup [Bard 69] décrit une étude sur l'efficacité de la recherche par SCI (la pertinence des résultats). Cette recherche a été appliquée sur une gamme de sujets médicaux. Une équipe de médecins ont été employée pour évaluer la pertinence. Les résultats obtenus montrent que 72% des articles de citation localisés sont *"strictement liés aux contenu de l'article cité"*, 22% sont *"indirectement relié"* et environ 5% pourrait être considéré *bruit*.
Dans ce cas, le SCI a produit une efficacité élevée de récupération. En général, SCI peut fournir pleins d'avantages avec l'amélioration du système lui-même aussi bien que la conception, recherche et développement de nouvelles applications.

2.5.5 Evaluation des auteurs

Bien que le SCI ait été à l'origine conçu pour être un outil de recherche pour l'usage dans la bibliothèque et la science de l'information, il y a des indications qu'elles auront des applications importantes comme un outil pour évaluer le personnel scientifique. En employant la base de données de SCI, il est possible de compter le nombre de citations d'un auteur donné. Bien qu'il y ait des exceptions, les auteurs fréquemment cités sont habituellement ceux qui ont effectué le travail le plus important dans un domaine donné [Garf 88].
Par exemple, en employant la base de données de SCI, il était possible d'énumérer les cinquante auteurs les plus cités dans l'année 1967. Les deux prix Nobel, Derek 1969 H. R. Barton et de Murray GelI-Mann sont apparus sur la liste. La production d'une liste de cinquante qui contient deux gagnants de prix Nobel est compilée par une méthode purement mécanique qui n'a pas exigé la lecture des travaux de ces hommes.
La capacité de l'index de citation de mesurer *l'impact du travail* d'un scientifique a des conséquences économiques pratiques. Les administrateurs de la recherche pourraient utiliser un outil tel qu'une aide dans le personnel scientifique d'évaluation ou en recruter de nouvelles personnes. Les officiers de diverses bases pourraient l'employer en attribuant des prix, des concessions, des camaraderies, et d'autres formes d'aide de recherches.

2.6 Conclusion

Quand les indices de citation pour la littérature scientifique ont été présentés la première fois, ils ont été considérés comme suppléments aux méthodes d'indexations traditionnelles. Cependant le temps a indiqué clairement que les indices des citations qui sont complets et opportuns sont autorisés à être considérés comme un outil performant d'indexation dans les bibliothèques électroniques de la science de l'information. De plus, les indices de citation jouent maintenant les rôles évaluatifs, analytiques, et prédictifs importants qui n'ont été jamais imaginés avec les indices traditionnels. L'index de citation a approuvé une simplicité de son automatisation. Le chapitre suivant traite un exemple de bibliothèque électronique, en l'occurrence *Citeseer*, qui utilise l'index de citation automatisé pour indexer ses documents.

Chapitre III

Indexation automatique des citations

– Bibliothèque Citeseer-

3.1 Introduction

Il est important pour les bases de données documentaires d'incorporer un service d'indexation de citation afin de faciliter les taches des usagers, comme la recherche la consultation et la production des statistiques, comme les indices traditionnels de citation. Par exemple, Cameron [Came 97]. A proposé une base de données bibliographiques de citation liant tous les travaux scientifiques publiés ".
L'institut de recherche NEC a réalisé une bibliothèque numérique des publications scientifiques qui crée un index de citation de façon automatique et autonome sans aucun effort additionnel de la part des auteurs ou des établissements, et sans aide manuelle [Gils 98].
Ce chapitre décrit l'architecture de la bibliothèque *Citeseer*, et présente ses différents composants ainsi que les différentes étapes du processus d'indexation automatique.

3.2 La bibliothèque électronique Citeseer

3.2.1 Les laboratoires Américains NEC

Le laboratoire Américain NEC Inc. (NEC Labs America, http://www.nec-labs.com/), est une filiale de NEC USA situé au Princeton à New-Jersey. Il a été établi en 2002. Il est reconnu comme étant l'un des innovateurs dans les domaines des logiciels d'Internet, les systèmes robustes et la gestion des réseaux à large bande, Ce laboratoire a mis au point plusieurs solutions technologiques largement utilisés par les entreprises, les fournisseurs de service et les utilisateurs individuels dans le monde entier.

3.2.2 La bibliothèque électronique Citeseer

Citeseer est une bibliothèque numérique de littérature scientifique qui vise à améliorer la diffusion et la rétroaction de la littérature scientifique. C'est l'un des agents Web qui ont été construits pour aider l'utilisateur à trouver facilement les documents scientifiques intéressants et appropriés en créant une "vue" adaptée aux besoins de chaque d'utilisateur ou groupe d'utilisateurs [Lawr 98]. Citeseer fournit des algorithmes et des techniques portables qui peuvent être déployés dans d'autres bibliothèques numériques pour indexer les fichiers PDF (Portable Document Format) et Postscript sur le Web en utilisant l'indexation automatique des citations.

3.2.3 Indexation autonome des citations (ACI)

Un système d'indexation autonome des citations **(**ACI) est un système qui peut automatiquement créer un index de citations des documents dans un format électronique. Un tel système doit, de façon autonome, localiser des articles, extraire les citations, identifier les citations d'un même article dans différents formats, et identifier le contexte des citations dans le corps des articles [Isaa 05].

Citeseer est un prototype d'une bibliothèque électronique qui accomplit avec succès ces tâches avec ponctualité suffisante. *Citeseer* télécharge les papiers (articles, journal, document…) sous format PDF et Post Script, les convertit en texte, analyse les documents pour extraire les citations et stocke ces informations dans une base de données. Citeseer offre deux méthodes de recherche, à savoir la recherche par mot-clé et la recherche par les liens des citations.

3.3 Architecture de Citeseer

La bibliothèque électronique *Citeseer* est composée de trois parties :

1) Un outil pour localiser et acquérir automatiquement les documents.
2) Un analyseur des documents et un créateur de base de données.
3) Une interface pour exploiter la base de données par la recherche par mot-clé ou par la navigation par les liens de citation.

La figure ci-dessous montre un diagramme de cette architecture.

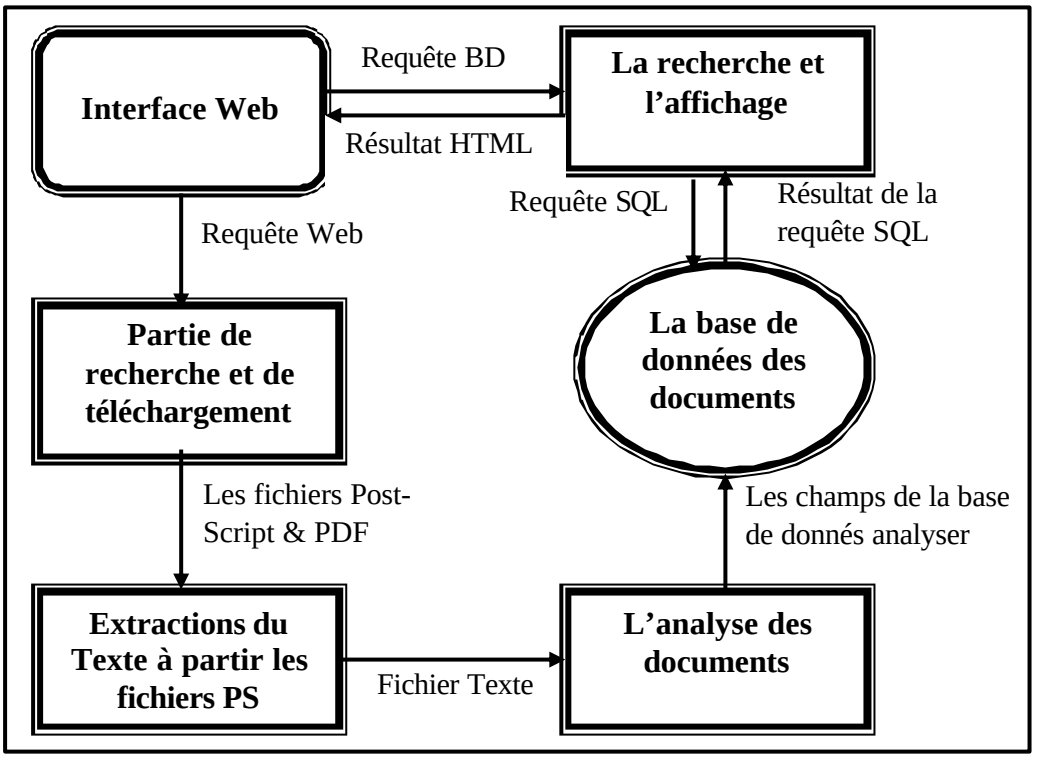


Figure 3.1 : L'architecture de la bibliothèque Citeseer

L'utilisation de *Citeseer* est relativement simple et directe. Quand un utilisateur veut explorer un nouveau thème, une nouvelle instance de C*iteseer* est crée pour cette matière particulière.

3.3.1 Acquisition des Documents

La première étape est la recherche des pages Web qui sont susceptibles de contenir des liens avec les documents pertinents. Lorsque L'utilisateur lance une recherche, en utilisant un ensemble de mots clés ; *Citeseer* utilise les moteurs de recherche (par exemple Google, AltaVista, HotBot..) et les heuristiques (par exemple recherche des pages qui contiennent également les mots "publications", "post-script" ...). Afin de localiser et télécharger les dossiers Post-Script identifiés par l'extension "Ps", le "PS.Z", ou "PS.GZ. *Citeseer* consulte les listes de publipostage (mailing lists) et les annonces dans les nouveaux groupes de message *Usenet* (news groups) pour trouver les articles ou pour permettre de se relier directement aux éditeurs [Alia 06].

Une fois que les chercheurs adoptent le système ACI, ils peuvent être automatiquement informés des nouvelles publications. Dans le cas des revues qui prennent en charge typiquement l'accès aux papiers en ligne, la seule méthode pour indexer ces articles est de faire des contrats avec les éditeurs eux-mêmes.

3.3.2 L'analyse des Documents

Citeseer utilise des programmes d'analyse pour extraire les composantes sémantiques des documents et sauvegarde ces derniers dans une base de données. *Citeseer* sauvegarde le répertoire de téléchargement et commence le processus d'analyse sur les documents dès qu'ils deviennent disponibles.

3.3.2.1 Les étapes d'analyse des documents

1) La première étape dans l'analyse des documents est l'extraction du texte brut à partir du fichier PS (conversion des fichiers PS en texte). *Citeseer* emploie le programme *PreScript* de la Bibliothèque Numérique de la Nouvelle-Zélande (*http://www.nzdl.org/technology/prescript.htm*), pour extraire le code ASCII en utilisant des informations à partir du formatage original du texte Post-Script.

2) La deuxième étape dans l'analyse des documents est l'utilisation des heuristiques pour identifier les champs suivants dans un document [Isaa 06]:

- *URL* : l'adresse URL du document téléchargé.
- *En-tête* (Header) : contient les informations au début du document comme *le titre*, *l'auteur*, *l'établissement*, *l'Année* et les autres informations qui viennent avant le texte intégral du document.
- *Résumé* (Abstract) **:** le texte du résumé, s'il existe, est extrait.
- *Introduction*: les 300 premiers mots de la section d'introduction, si elle existe. sont extraits
- *Citations*: La liste des références utilisées par le document est extraite et analysée.
- *Fréquence des mots:* les fréquences de touts les mots dans le document sont extraits et enregistrées, excepté les mots des citations et les mots d'arrêt. Les mots sont lemmatisés, c'est-à-dire ramenés à leurs racines, en utilisant l'algorithme de Porter [Port 80].

Une fois que l'ensemble de références bibliographiques est identifié, les différentes citations sont extraites [Hui 03].

- Chaque citation est analysée en utilisant des heuristiques pour extraire les champs suivants : *Titre, Auteur, Année de Publication, journal, Nombre de pages, et Etiquette de citation.*
- L'étiquette de citation est une information dans la citation qui est employée pour citer cette citation dans le corps du document (par exemple, **[Giles97]**, [**Mar82**]). Les étiquettes des citations sont employées pour trouver les endroits

dans le corps du document où les citations sont apparues. Ceci permet d'extraire le contexte des citations pendant la navigation dans la base de données.

✍ La fréquence des mots de chaque citation est également enregistrée après lemmatisation.

✍ Les heuristiques utilisées pour analyser les citations sont construites selon la philosophie *"invariants first*". Par exemple l'étiquette d'une citation existe toujours au début d'une citation, l'année de la publication existe presque dans toute citation comme un numéro de quatre chiffres commençant par les chiffres **"19"** ou **"20"**.

Exemple, les informations sur l'auteur précèdent toujours le Titre, et l'Editeur suit presque toujours le Titre.

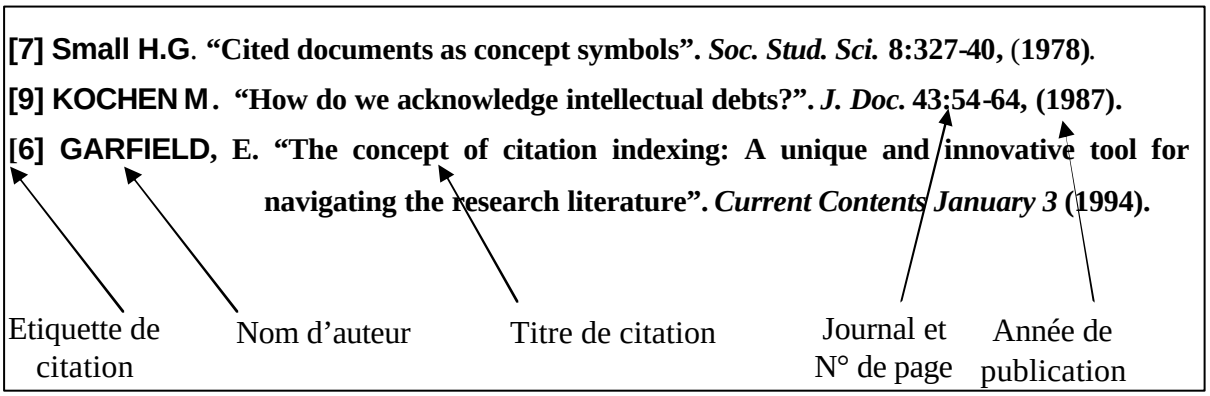


Figure 3.2 : Exemple des citations d'un document

Enfin la base de données résultante contient les tables suivantes :

- *Document* : Contient le titre, l'auteur, les parties de texte du document, l'*URL du document*, et un seul *Identificateur d'article* (***UAID***).
- *Les mots de document* : Contient les fréquences des mots d'informations dans le corps du document référencés sur le tableau du document.
- *Citation :* Contient le texte des citations utilisées par un document ainsi que les informations sur les champs analysés. Chaque enregistrement dans cette table a un seul Identificateur de citation (***UCID***) et un champ pour ***UAID*** correspondant.
- *Mots de citation* **:** Contient les mots fréquents dans la citation.
- *Groupe de citations et le Poids de Groupe***:** Contient le nombre de groupes, et le poids des informations lorsqu'en regroupe les citations identiques dans différentes formes. Ces informations sont employées pour la recherche automatique de documents similaires.

3.3.2.2 Difficulté dans l'analyse des documents

L'analyse des citations en langage naturel présente plusieurs difficultés [Isaa 06], qui obligent l'utilisation des heuristiques pour extraire certains sous champs. Cependant, le tableau ci-dessous montre des statistiques d'analyse sur un exemple d'une base de données *Citeseer* qui a été créé à partir de 5093 documents liés aux " réseaux de neurones ".

Tableau 3.1 : Les résultats d'analyse des citations dans un exemple de base de données *Citeseer* [Isaa 06]

Base de données	Réseau de neurone
Documents analysés	5093
Citations trouvées	89614
Titres identifiés	71908/89614 = 80.2 %
Auteurs identifiés	73539/89614 = 82.1 %
Numéro des pages identifiés	39595/89614 = 44.2 %

Nous pouvons voir que la probabilité de détection est de 80% pour les titres et les auteurs et 40% pour les numéros des pages. Cette diminution est due au fait que beaucoup de citations ne contiennent pas des numéros des pages. Les résultats trouvés impliquent l'utilisation des nouvelles techniques d'étude et d'heuristiques adéquates afin d'extraire les champs additionnels des citations.

3.3.4 La navigation dans la base de données

Le troisième composant de *Citeseer* est le navigateur (browser). *Citeseer* prend les requêtes des utilisateurs dans une syntaxe appropriée, les traitent puis renvoie la réponse sous forme HTML [Alia 06]. L'interface Web *Citeseer* fournit plusieurs possibilités qui permettent à l'utilisateur d'interroger et de naviguer facilement dans la base de données des documents. Le premier accès à la base de données de publication doit être une recherche par *mot-clé*, après n'importe quelle réponse d'une requête (non vide), l'utilisateur peut passer au parcours des citations. L'exemple suivant explique les différents cas d'interrogation de la bibliothèque électronique *Citeseer*.

Exemple

Une base de données de *Citeseer* a été crée en utilisant les mots-clés initiaux *"réseaux de neurone"* pour la démonstration des objectifs. Supposons que l'utilisateur veut trouver tous les documents cités conjointement écrits par **Giles** et **Chen**. L'exemple de la requête : *Citation : +Giles +Chen* renvoie toutes les citations qui contiennent les mots "Giles" et "Chen".

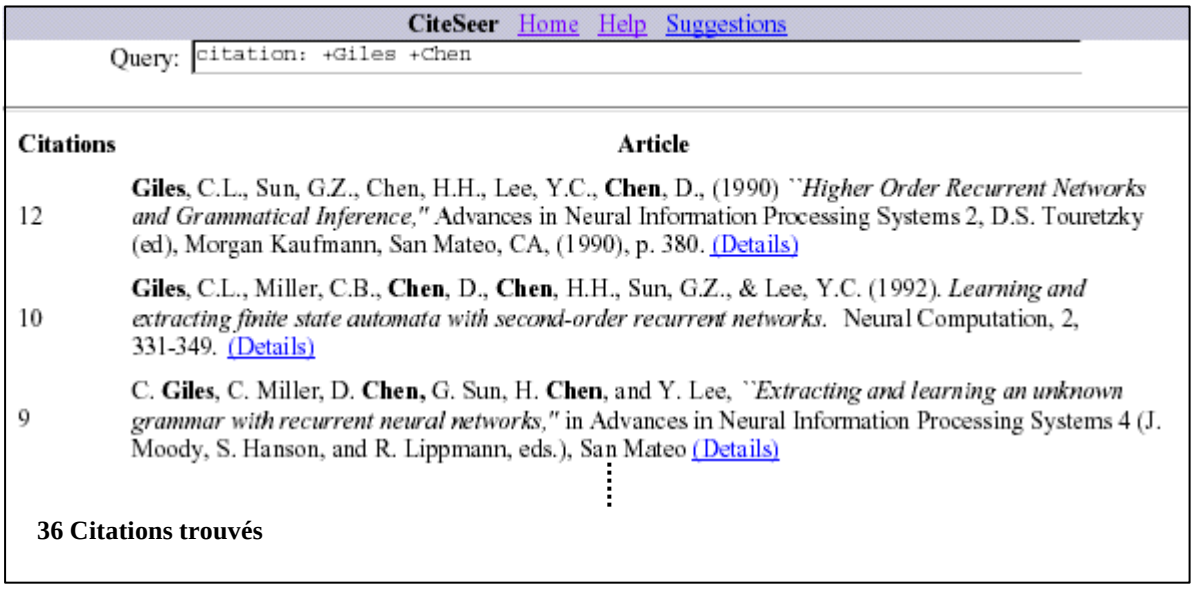


Figure 3.3 : La recherche des citations par mot clé dans Citeseer

Si l'utilisateur veut connaître quels sont les documents qui citent le troisième article de la figure 3.3 *"extracting and learning an unknown grammar with recurrent neural networks*", il aura le resultat de la Figure suivante.

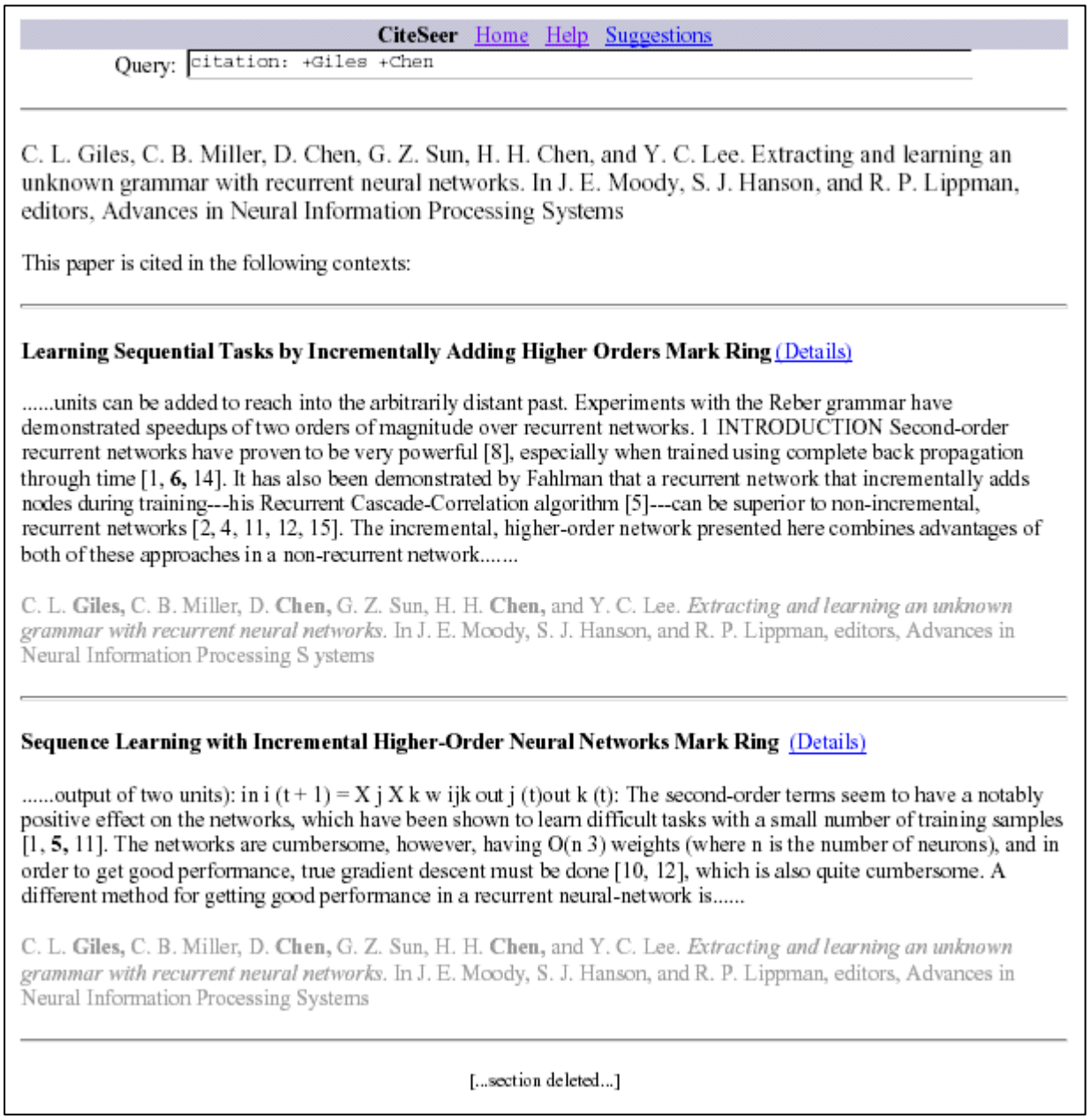


Figure 3.4 : Les informations détaillées des citations dans Citeseer

Toutes les citations ont une même forme (i .e . Un lien à la citation, et une partie de contexte de la citation). L'utilisateur peut accéder aux détails d'une citation en choisissant le lien vers le document cité. La figure suivante montre le format du résultat retourné.

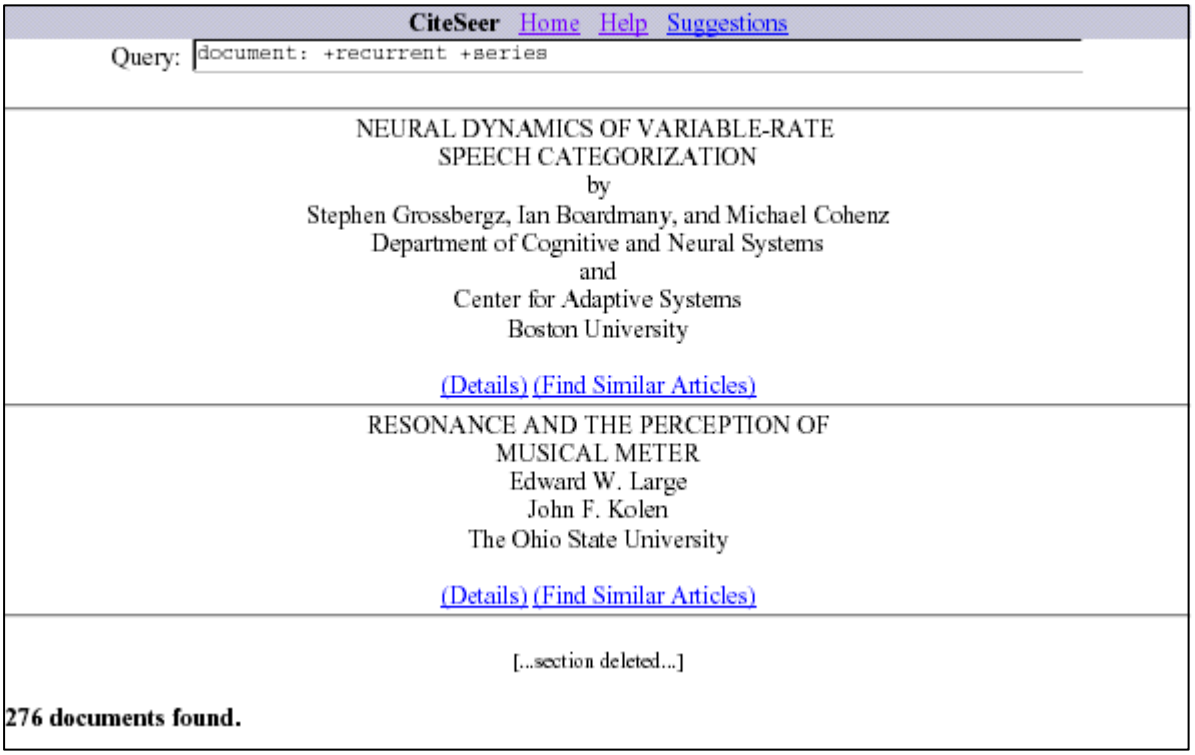
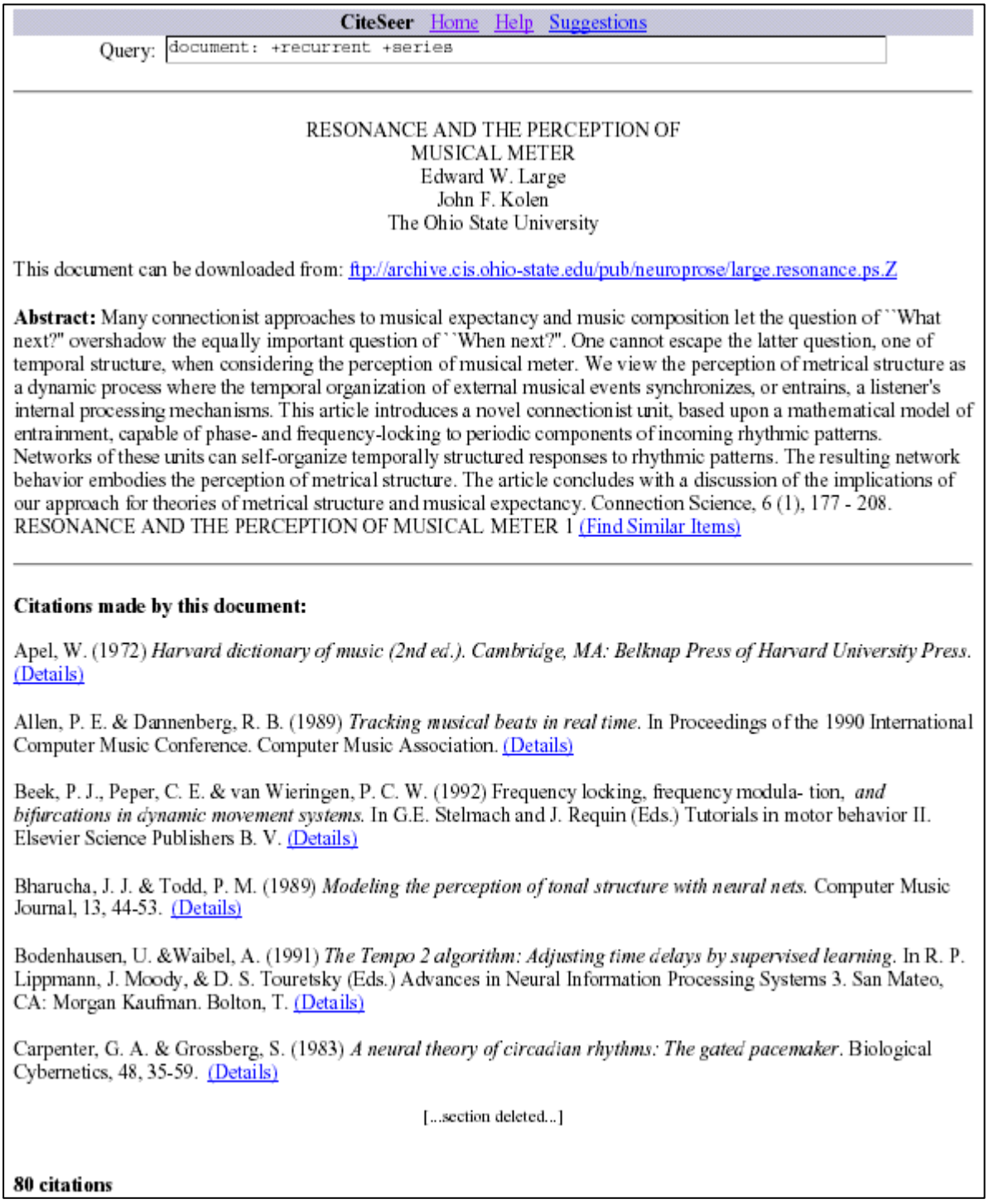


Figure 3.5 : La recherche par mot clé sur les documents dans *Citeseer*

La figure ci-dessus présente la première page du résultat d'une recherche par mot-clé sur les documents eux-mêmes, *document : +recurrent +series*. Ici l'en-tête est fourni pour les documents qui contiennent les mots-clés dans leurs corps. Le lien donne les détails d'un document particulier. La première page des détails du deuxième article de la figure précédente est montrée par la Figure 3.6. *L'en-tête*, *le résumé*, *l'URL,* et *la liste des références* sont affichés.



CiteSeer Home Help Suggestions

Query: document: +recurrent +series

RESONANCE AND THE PERCEPTION OF
MUSICAL METER
Edward W. Large
John F. Kolen
The Ohio State University

This document can be downloaded from: ftp://archive.cis.ohio-state.edu/pub/neuroprose/large.resonance.ps.Z

Abstract: Many connectionist approaches to musical expectancy and music composition let the question of ``What next?'' overshadow the equally important question of ``When next?''. One cannot escape the latter question, one of temporal structure, when considering the perception of musical meter. We view the perception of metrical structure as a dynamic process where the temporal organization of external musical events synchronizes, or entrains, a listener's internal processing mechanisms. This article introduces a novel connectionist unit, based upon a mathematical model of entrainment, capable of phase- and frequency-locking to periodic components of incoming rhythmic patterns. Networks of these units can self-organize temporally structured responses to rhythmic patterns. The resulting network behavior embodies the perception of metrical structure. The article concludes with a discussion of the implications of our approach for theories of metrical structure and musical expectancy. Connection Science, 6 (1), 177 - 208.
RESONANCE AND THE PERCEPTION OF MUSICAL METER 1 (Find Similar Items)

Citations made by this document:

Apel, W. (1972) *Harvard dictionary of music (2nd ed.). Cambridge, MA: Belknap Press of Harvard University Press.* (Details)

Allen, P. E. & Dannenberg, R. B. (1989) *Tracking musical beats in real time.* In Proceedings of the 1990 International Computer Music Conference. Computer Music Association. (Details)

Beek, P. J., Peper, C. E. & van Wieringen, P. C. W. (1992) Frequency locking, frequency modula- tion, *and bifurcations in dynamic movement systems.* In G.E. Stelmach and J. Requin (Eds.) Tutorials in motor behavior II. Elsevier Science Publishers B. V. (Details)

Bharucha, J. J. & Todd, P. M. (1989) *Modeling the perception of tonal structure with neural nets.* Computer Music Journal, 13, 44-53. (Details)

Bodenhausen, U. &Waibel, A. (1991) *The Tempo 2 algorithm: Adjusting time delays by supervised learning.* In R. P. Lippmann, J. Moody, & D. S. Touretsky (Eds.) Advances in Neural Information Processing Systems 3. San Mateo, CA: Morgan Kaufman. Bolton, T. (Details)

Carpenter, G. A. & Grossberg, S. (1983) *A neural theory of circadian rhythms: The gated pacemaker.* Biological Cybernetics, 48, 35-59. (Details)

[...section deleted...]

80 citations

Figure 3.6 : Les informations détaillées d'un document dans Citeseer

Une fois une première recherche par mot-clé faite, l'utilisateur peut passer au parcours de la base de données en employant les liens des citations- documents. L'utilisateur peut trouver les documents cités par une publication particulière ainsi que les documents qui citent cette dernière.

3.4 Mesures de la distance sémantique

Plusieurs types d'algorithmes existent pour mesurer la distance entre deux documents texte. Dans ce qui suit, nous décrivons les trois types de modèles les plus utilisés :

1) *La distance de chaîne* ou la *distance de texte* : considère la distance entre deux symboles comme étant le nombre de différence en caractères entre les deux. Par exemple *Levenshtein* distance [Leve 65] définit la différence entre deux chaînes de caractères comme étant le nombre d'insertions, de suppressions ou de substitutions des lettres nécessaires pour transformer une chaîne à une autre. Un autre algorithme plus sophistiqué a été proposé par est *LikeIt* [Yian 97] utilise un algorithme qui essaye de «construire le poids optimal d'assortiment des lettres et des mots ».

2) Un autre type de mesure de distance entre les documents textuel est basé sur les statistiques et exploite les fréquences des mots dans l'ensemble des documents. La méthode la plus utilisé est connue sous le nom (Term Frequency Inverse Document Frequency TFIDF) [Salt 73].

On considère un dictionnaire de tous les mots dans les corps des documents. Dans un document ***d***, la fréquence de chaque racine de mot (***s***) est ($\boldsymbol{f_{ds}}$), le nombre de documents ayant la racine ***s*** est $\boldsymbol{n_s}$, et la fréquence la plus élevée s'appelle $\boldsymbol{f_{dmax}}$.

Dans l'algorithme de TFIDF [Salt 97] le poids d'un mot $\mathbf{W_{ds}}$ est calculé par :

$$Wds ? \frac{(0.5 ? 0.5 \frac{fds}{fd\max}) (\log \frac{ND}{ns})}{\sqrt{? _{j?d} ((0.5 ? 0.5 \frac{fdj}{fd\max})^{2} (\log \frac{ND}{nj})^{2})}}$$

Où ***le ND*** est le nombre de tous les documents. Afin de trouver la distance entre deux documents, un produit scalaire simple des deux vecteurs des mots pour ces deux documents est calculé. Les mots très communs, parfois appelés les mots d'arrêt, tels que : *a, de…* etc. sont ignorés afin de réduire la taille du dictionnaire des mots. Des heuristiques sont utilisées pour lemmatiser les mots afin de ne garder dans le dictionnaire que les racines de ces derniers; Par exemple "walking", "walk", "walked" proviennent du mot "Walk", L'heuristique de lemmatisation la plus utilisée est celle de « stemming heuristique » proposée par Porter [Port 80].

3) La mesure de distance basée sur la sémantique : dans ce type de méthodes la structure du document est utilisée pour extraire des connaissances qui seront utilisées pour comparer les documents. Par exemple, les sous champs tels que *le*

titre, l'année de publication et *l'auteur* peuvent être employés pour comparer les citations des différents documents, pour créer les indices de citation [Isaa 05].

Citeseer met en application la mesure de la distance sémantique dans deux applications:

- Citeseer utilise la fréquence de mot et la distance du texte pour grouper les différentes formes de la même citation.
- Citeseer utilise la fréquence des informations de citation pour trouver des documents relatifs à l'intérêt de l'utilisateur.

3.4.1 Le regroupement des citations identiques

Parmi les difficultés rencontrées dans l'automatisation de l'indexation des citations sont la variation de la syntaxe des citations. Les citations d'un même document peuvent être présentées dans différents formats et les champs des citations contiennent d'habitude des erreurs, et il peut être difficile de déterminer les sous champs d'une citation de façon automatique, par exemple les virgules et les points d'arrêt peuvent signifier des rôles différents [Isaa 06].
La figure ci-dessous montre des exemples de citations du même extraites à partir de trois différentes publications sur les réseaux de neurones.

[7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J.Stone. "Classification and Regression Trees". Wadsworth, Pacific Grove, California, 1984.

[6] L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and Regression Trees". Wadsworth and Brooks, 1984.

[1] L. Breiman et al. "Classification and Regression Trees" Wadsworth, 1984.

Figure 3.7: Exemple de différentes formes des citations

Nous présentant dans ce qui suit les méthodes et les algorithmes utilisé par Citeseer pour regrouper les citations identiques (**ICG**).

1) La normalisation

Pour améliorer les résultats, *Citeseer* utilise les normalisations suivantes :

a) Convertir toutes les lettres en lettres minuscules,

b) Conversion des traits d'union en espaces,

c) Enlèvement des étiquettes de citation, par exemple [3], [valse 92] au début de la citation,

d) Expansion des abréviations communes, par exemple *Conf = Conférence*, *P = Procédure.*

e) Enlèvement des mots et des caractères étrangers, par exemple "en cours d'impression", "admis pour la publication".

Les algorithmes suivant sont utilisés par *Citeseer* pour identifier et regrouper les formes variables des citations du même papier.

2) Algorithme de base (Baseline)

La Figure 3.8 montre l'algorithme de base que *Citeseer* utilise pour la comparaison entre les citations.

```
- Pour chaque Citation A
      Si A est non pas encore groupé alors
                  - créer un nouveau groupe G avec cette citation
                        Pour chaque citation non groupé B faire
                  Ajouter B à G si le nombre de mots assortis est plus grand que X%
                  de la langue de la plus courte citation.
            FinPour
      FinSi
FinPour
```

Figure 3.8: Algorithme de *Baseline*

3) Algorithme d'assortiment des mots

La Figure 3.9 présente le deuxième algorithme basé sur *l'assortissant des mots* que *Citeseer* utilise pour le regroupement des citations.

```
Classer les citations par ordre décroissant de longueur
Pour chaque citation C faire
      Chercher le groupe G avec le plus grand nombre de mots assortis
      Si le taux = Nbr mots non-assortis/ Nbr des mots assortis < seuil minimum
            Alor ajouter C au groupe G
                  Sinon créer un nouveau groupe pour la citation (identifié par cette
citation).
      FinSi
FinPour
Classer les groupes, pour que chaque deux groupes successifs contiennent le plus grand
nombre des mots communs.
```

Figure 3.9: Algorithme d'assortiment des mots

Cet algorithme classe les citations selon leurs longueurs, et utilise la première citation dans un groupe comme un identificateur de groupe dans les futures comparaisons. Le classement initial par longueur permet de trouver la citation la plus longue et la plus complète dans un groupe des citations comme par exemple la plus longue citation peut donner tous les auteurs. L'implémentation de cet algorithme est basée sur une table de Hashage ou chaque entrée contient une liste de groupes contenant un mot donné,
L'algorithme est donné comme suit:

```
- Classer les citations par ordre décroissant de longueur
Pour chaque citation C faire
        Pour chaque mot dans la citation
            Chercher la table de groupes qui contient ce mot à partir de la table de
              Hashage.
        FinPour
    FinPour
- Chercher le groupe G le plus commun dans la table des groupes pour chaque mot.
- Si le taux = Nbr mots non-assortis/ Nbr des mots assortis  <  seuil minimum
        Alors  Ajouter C  au groupe G
        Sinon créer un nouveau groupe pour cette citation, en ajoutant le numéro
              de ce dernier à la table pour chaque mot.
        FinSi
```

Figure 3.9b: Algorithme d'assortiment des mots

Cette méthode peut être améliorée en divisant les données, par exemple selon l'année de la publication ou le premier auteur.

5) Algorithme d'assortiment des mots et des expressions

L'algorithme est semblable à l'algorithme précédent, avec l'addition des informations d'expression. L'algorithme considère les expressions comme étant l'ensemble de chaque deux mots successifs dans chaque section de la citation contenant trois mots successifs ou plus (ces sections sont limitées par un. (Point) ou **,** (Virgule) ou le début/fin de la citation), L'algorithme ainsi ajoute l'ordre du mot pour les sous-champs de la citation contenant trois mots ou plus, comme par exemple le titre, mais il est insensible à l'ordonnancement des sous-champs.

Classer les citations par ordre décroissant de longueur
Pour chaque citation **C**
Chercher le groupe **G** avec le grand Nbr des mots assortis
Soit le **Taux a** = Nbr des mots non-assortis/ Nbr des mots assortis.
Soit le **Taux b** = Nbr des phrase non-assortis/ Nbr des phrases assorties où une phrase est un ensemble de deux mots successifs dans chaque section contient 3 mots ou plus (la section est limitée par**. ,** ou le début et la fin de citation)
Si (**a** < **Seuil1**) ou (**a** < **seuil2** et **b** < **seuil3**)
Alors ajouter **C** au groupe **G**
Sinon créer un nouveau groupe pour cette citation.
FinPour
Classer les groupes, pour que chaque deux groupes successifs contiennent le plus grand nombre des mots communs.

Figure 3.10: Algorithme d'assortiment des mots et d'expressions

L'implémentation est semblable à l'algorithme précédent et est basée sur une table de Hashage.

6) Algorithme LikeIt la distance de chaîne de caractères

Comme cité précédemment, *LikeIt* est une forme sophistiquée de la distance de chaîne présentée par Yianilos [Yian 97]. *Citeseer* emploie *LikeIt* comme le montre la figure 3.11. Il faut noter que *LikeIt* calcule une mesure de similitude contrairement à la mesure de distance, et que ses valeurs de similitude sont normalisées par la longueur de la chaîne de caractères).

Classer les citations par ordre décroissant de longueur
Pour chaque citation **C**
Chercher le groupe **G** avec le plus grand valeur **d= LikeIt(c,c)/ LikeIt(c,g)**
Si d > seuil **alors** ajouter **C** dans le groupe **G**
Sinon créer un nouveau groupe pour cette citation
FinSi
FinPour
Classer les groupes, pour que chaque deux groupes successifs contiennent le plus grand nombre des mots communs.

Figure 3.11: Algorithme basé sur la mesure LikeIt

Où *LikeIt (c,g)* est une valeur qui représente le nombre des insertions, des suppressions ou de substitutions de lettres pour transformer la citation c en g.

7) Algorithme de Sous-champ

Cet algorithme est basé sur l'extraction automatique des sous-champs à partir des citations. Cette opération est parfois difficile et nécessite l'utilisation d'heuristiques complexes, par exemple :

- Les virgules peuvent être employées pour séparer les champs, mais sont également employées pour séparer les listes d'auteurs, et peuvent être utilisées dans les titres.
- Les points peuvent être employés pour séparer les sous-champs ou pour les abréviations.
- Parfois il n'y a aucune ponctuation entre les champs.

Citeseer utilise cet algorithme sur les sous-champs titre et auteur. (*Les références bibliographiques).*

Exemple : le titre est le premier sous-champ contenant trois mots ou plus. L'heuristique réelle est plus complexe.

- ✍ Les citations sont groupées ensembles si le titre et l'auteur principal sont les mêmes.
- ✍ Plusieurs normalisations sont employées pour rendre l'algorithme moins sensible, par exemple à l'utilisation des tirets et des espaces.
- ✍ Plusieurs hypothèses sont employées pour extraire l'auteur principal. Par exemple l'algorithme est moins sensible si un ou tous les auteurs sont listés, ou si les prénoms d'auteur sont employés.

8) Résultats

Nous avons présenté quelques d'algorithmes utilisés par *Citeseer* pour identifier et regrouper les différentes formes des citations du même papier. L'exactitude et l'efficacité de tous les algorithmes ont été quantitativement comparées sur un certain nombre d'ensembles de données.

Une étude de Citeseer [Lawr 99] sur 1947 documents téléchargés sous forme Post-Script. Les documents sont converti en texte par *Prescript*, 39166 citations sont extraites à partir de ces documents, un échantillon de quatre groupes des citations qui contiennent les mots : "reinforcement", "constraint", "face" et "reasoning" sont extraites. Ces ensembles ont contenu 406, 295, 349, et 514 citations respectivement

Nous avons manuellement groupé les citations dans chaque ensemble afin d'avoir constitué des groupes corrects. Pour chaque algorithme, nous avons comparé les groupes automatisés aux groupes corrects. Les erreurs sont présentées sur le tableau suivant :

Tableau 3.2 Comparaison entre les différents algorithmes de regroupement des citations [Lawr 99]

	Constraint	**Face**	**Reasoning**	**Average**
Nombre des citations	295	349	514	
Baseline Simple	12%	6%	13%	10.3%
Assortiment des mots	9%	4%	8%	7%
Assortiment des mots et des expressions	**6%**	**3%**	**7%**	**5.3%**
LikeIt	13%	12%	17%	14.3%
Algorithme de Sous-champ	12%	9%	16%	12.3%

Les résultats obtenu montrent que:

1. L'algorithme basé sur *l'assortiment de mot* et *des expressions* s'est avéré plus performant que le les autres algorithmes. Il est suffisamment précis et peut être facilement intégré dans un système automatisé d'indexation de citation.
2. L'algorithme basé sur la distance de *chaîne de caractère* est le moins performant de tous ces algorithmes.
3. L'algorithme basé sur l'extraction des *sous-champs* des citations produit une exactitude relativement faible.

3.4.2 La recherche des Documents Similaires

Dans une base de données de documents, un utilisateur peut trouver les documents similaires manuellement en employant les dispositifs sémantiques tels que l'auteur, le groupe de recherche, ou l'année de publication pour le document. *Citeseer* utilise un mécanisme pour la recherche automatique des documents similaires basés sur les mesures de distance entre les dispositifs sémantiques extraits à partir de ces documents.

1) LikeIt

Citeseer emploie la distance de chaîne de caractères de *LikeIt* [YIAN 93] pour mesurer la distance de texte entre les en-têtes des documents dans une base de données. *LikeIt* compare les sous-chaînes avec les chaînes plus grandes. Les champs liés aux *auteurs*, *établissements*, et *mots communs dans le titre* tendent à réduire la distance de *LikeIt* entre les en-têtes.

2) CCIDF "Common Citation X Inverse Document Frequency"

Citeseer utilise les citations communes pour déterminer si les documents téléchargés dans la base de données sont le plus profondément liés à un document sélectionné par l'utilisateur. Cette mesure est CCIDF **(**Common Citation *X* Inverse

Document Frequency) est semblable a TFIDF pour exprimer les poids orientés des mots [SAL 97]. L'algorithme pour calculer CCIDF de tous les documents dans la base de données par rapport à un document d'intérêt ***A*** et pour choisir les meilleurs documents ***M*** est comme suit :

✍ Employer l'algorithme qui regroupe les citations identiques (ICG) sur la base de données entière des documents pour obtenir le nombre (***Ci***) qui représente le nombre de fois où chaque document ***i*** est cité.

✍ Prendre l'inverse de ces fréquences comme poids pour cette citation (***Wi = 1/Ci***) et stocker ces valeurs dans la base de données. Cette étape doit seulement être exécutée une fois que la base de données a été construite, et est réutilisée pour des questions postérieures.

✍ Déterminer la liste des citations et leurs poids associés pour le document ***A*** et interroger la base de données pour trouver l'ensemble de ***n*** documents **{Bj} : *j=1… n*** qui partagent au moins une citation avec ***A***.

✍ Pour chaque ***j = 1..n***, déterminer la similitude ***Rj*** du document comme la somme des poids des citations partagées avec ***A***

$$Rj? \sum_{(i? Ai)? (i? Bi)} Wi$$

✍ Assortir les valeurs de ***Rj*** et renvoyer les documents ***Bj*** avec les valeurs les plus élevées ***Rj***.

Figure 3.12: CCIDF

Comme pour l'utilisation de TFIDF, CCIDF suppose que si une citation très peu commune est partagée par deux documents, ceci devrait être pesé plus fortement qu'une citation citée par un grand nombre de documents. Actuellement, bien que nous n'ayons pas testé des mesures d'exécution formelles sur CCIDF, cette méthode pourrait donner des résultats meilleurs que celle basée sur le vecteur de mots ou celle basée sur *LikeIt***.**

4.5 Conclusion

On a étudié dans ce chapitre un modèle de bibliothèque électronique appelé *Citeseer* (**www. Citeseer.ist.psu.edu**) qui utilise l'indexation automatique des citations pour aider les chercheurs j localiser et télécharger des documents pertinents. *Citeseer* utilise les moteurs de recherche et des heuristiques pour localiser et télécharger les documents à partir d'un ensemble des mots-clés spécifié par l'utilisateur. Les documents sont analysés afin d'extraire les dispositifs d'informations tels que : le titre, l'auteur, le résumé et les citations individuellement identifiées. Ces informations sont placées dans une base de données.

L'interface Web de *Citeseer* peut être employée pour trouver les pièces appropriées dans la base de données en utilisant la recherche par *mot-clé* ou par le *parcours des liens* entre les documents *(les références bibliographiques),* constitués par les citations, les liens vers *les citations* et *les documents cités*. En plus, on peut trouver les documents qui sont similaires au document recherché en utilisant l'analyse du texte intégral ou l'analyse des citations communes.

Les citations des articles peuvent être présentées dans différents formats, Pour cela il est important pour un système automatique d'indexation de citation de pouvoir détecter ces différents formats. *Citeseer* propose un ensemble de techniques et d'algorithmes pour identifier et grouper les formes variables des citations au même article. L'exactitude et l'efficacité de tous les algorithmes présentés ont été quantitativement comparées en utilisant le taux d'erreurs.

Dans le chapitre suivant nous proposant l'architecture d'une bibliothèque électronique qui utilise l'indexation automatique des citations pour indexer ses documents. La modélisation est faite en utilisant l'approche objet représentée par le langage de modélisation **UML** « Unified Modeling Language ».

Chapitre IV

Modélisation de la bibliothèque électronique -UFASeer-

4.1 Introduction

Dans ce chapitre nous proposons la modélisation objet d'une bibliothèque électronique qui utilise l'indexation des citations en utilisant Le langage de modélisation unifié UML -Unified Modeling Language. Ce dernier présente une liste de meilleures pratiques d'ingénierie qui ont prouvé leurs succès dans la modélisation d'un grand nombre de système complexes. Nous présentons ensuite les différents diagrammes UML statiques et dynamiques, afin de faciliter la tâche pour réaliser ce type des bibliothèques. Cette bibliothèque permet de chercher les documents par mots clés, et de naviguer entre les documents en utilisant les citations comme liens implicites pour finalement afficher les documents appropriés au besoin du chercheur.

4.2 Bibliothèque UFASeer

La bibliothèque *UFASeer* est une bibliothèque électronique conçue et réalisée à l'université Ferhat Abbas à Sétif (UFAS). Cette bibliothèque utilise l'indexation des citations pour indexer les thèses Magister et les thèses doctorales comme première étapes et par suite d'autres supports scientifiques.

Comme *Citeseer*, *UFASeer* propose la recherche par mots clés comme première étape, une fois le chercheur trouve un document qui le satisfait, il commence la recherche par consultation des citations en avant et en arrière pour trouver les documents appropriés. *UFASeer* utilise les moteurs de recherche et les méta-moteurs pour chercher et télécharger les documents demandés par les usagers et les documents qui ne sont pas encore téléchargés (exemple : les citations d'un document), pour remplir sa base de données. L'opération de recherche est permanente tant qu'il y a des citations non téléchargées.

4.3 Modélisation UML de la bibliothèque UFASeer

4.3.1 Le diagramme des cas d'utilisation

Les cas d'utilisation sont employés pour représenter le modèle conceptuel en permettant la structuration des besoins des utilisateurs et les objectifs du système cible et cela en se limitant aux préoccupations réelles des utilisateurs tout en restant loin de toute solution d'implantation. Le diagramme des cas d'utilisation (Figure 4.1) représente l'ensemble des cas d'utilisation de la bibliothèque UFA*Seer*.

La bibliothèque est divisée en deux parties : La partie utilisateur (chercheur) et la partie module UFA*Seer*.

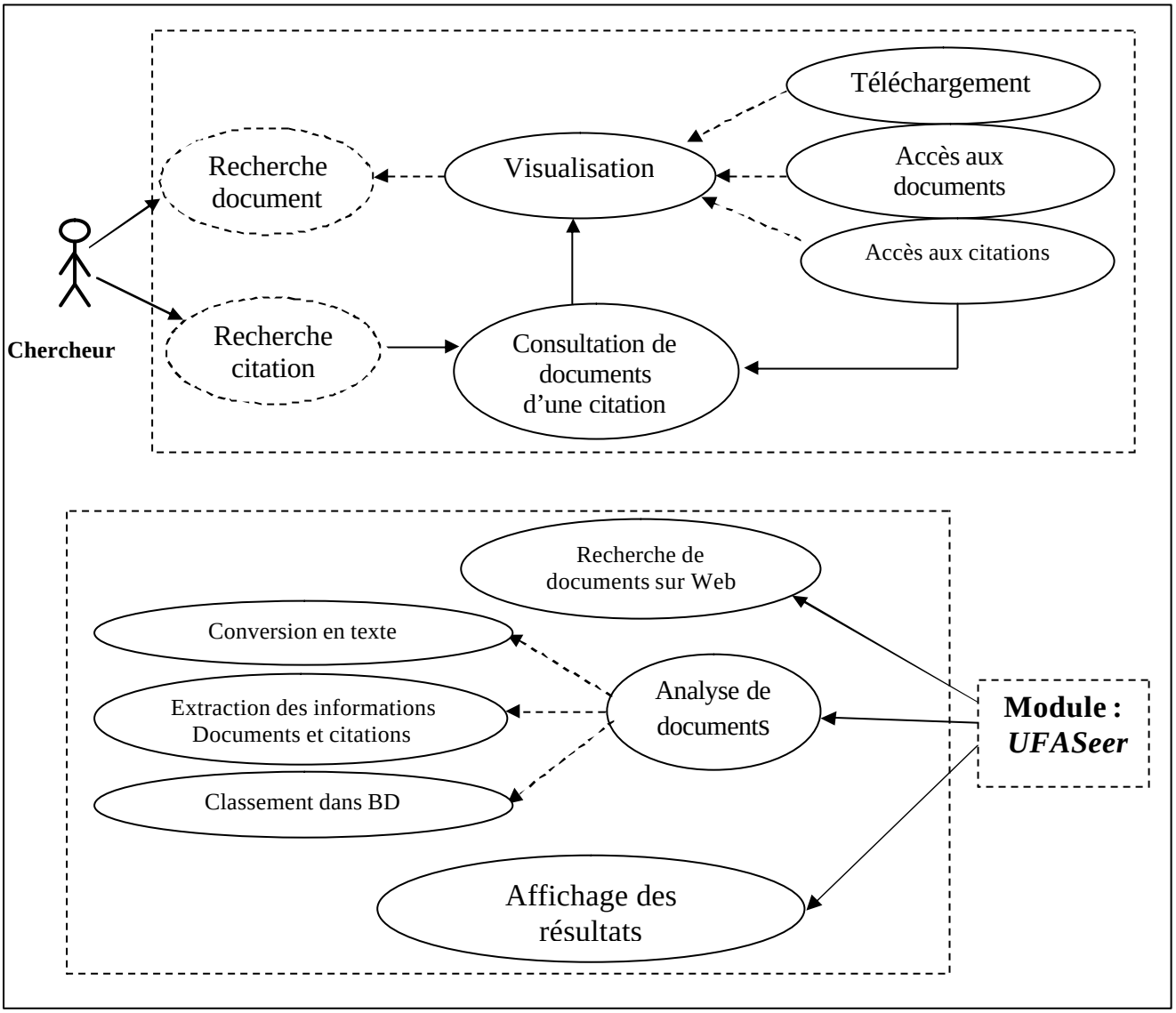


Figure 4.1 : Diagramme de cas d'utilisations de UFA*Seer*

a) Partie Utilisateur

- Le chercheur peut lancer la recherche sur les documents ou les citations par mots clés.
- Une fois qu'un résultat s'affiche, il peut sélectionner un document et afficher ses informations. Par suite, il peut télécharger ce document ou accéder à ses documents similaires ou parcourir leurs citations en avant et en arrière.
- Si le chercheur lance la recherche par citation, il peut consulter tous les documents où cette citation apparaît.

b) Partie Module UFASeer

- Une fois le système reçoit une requête, il fait la recherche sur la base de données interne pour afficher les résultats appropriés à l'utilisateur.
- Le système offre la possibilité de naviguer entre les documents, par la consultation des citations en avant et en arrière.
- **UFA*Seer*** fait la recherche sur le Web, pour alimenter la base de données, il télécharge les documents appropriés à la demande de l'utilisateur ou les citations qui ne sont pas encore téléchargées.
- Une fois les documents sont localisés, l'agent ***UFASeer*** commence l'opération d'indexation automatique, il extrait les informations des documents (l'entête) et les citations pour les sauvegarder dans la base de données afin d'être accessibles la prochaine fois.

4.3.2 Description des scénarios

Les scénarios sont des versions simplifiées des diagrammes de séquence et mettent en évidence les interactions qui existent entre le système ***(UFASeer)*** et les acteurs de son contexte. Ils jouent alors le rôle d'illustrateurs du cas d'utilisation dont ils sont une instance.

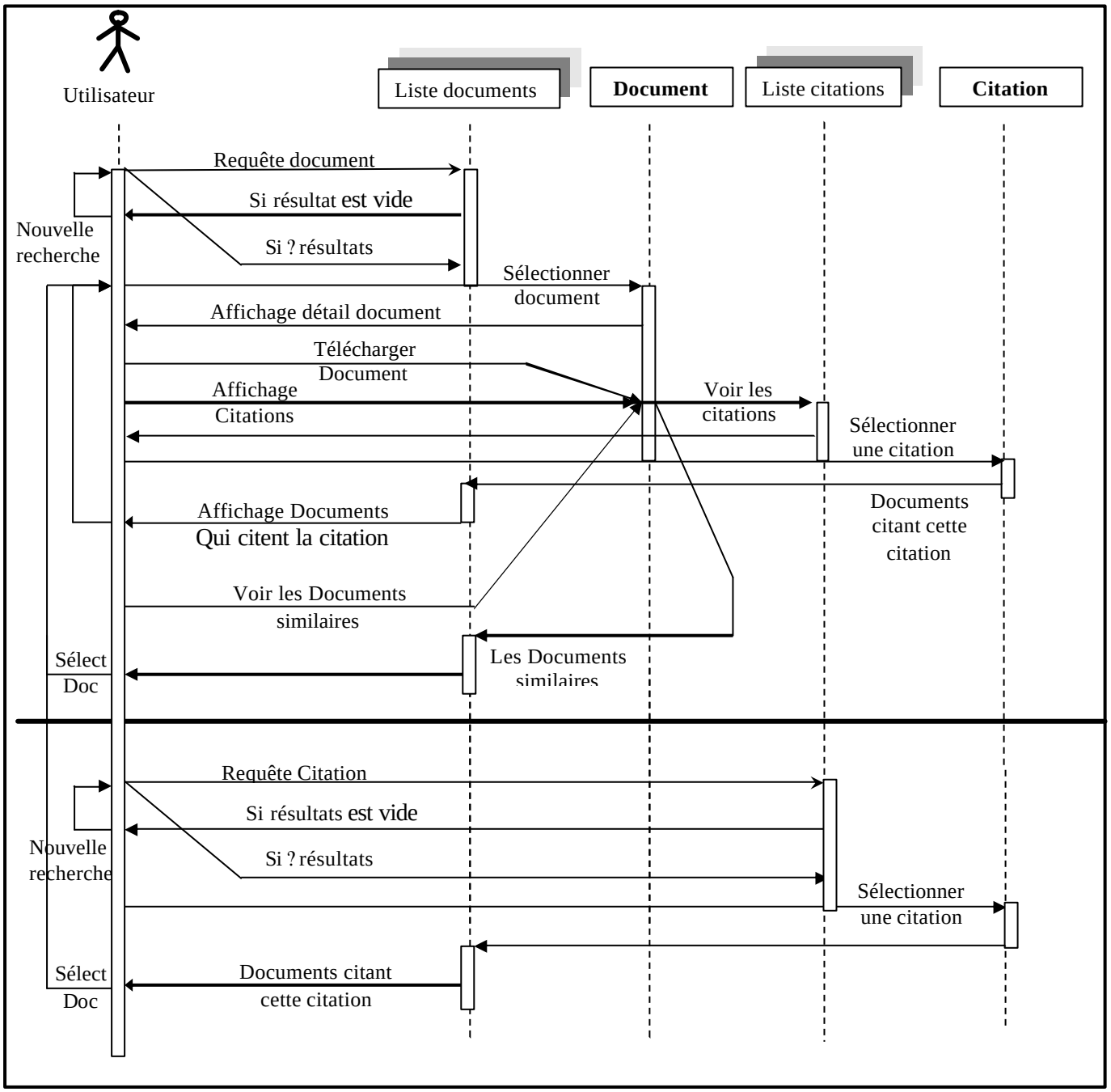


Figure 4.2 : Exemple de scénarios parcours des documents dans UFA*Seer*

4.3.3 Diagramme de classes

Le digramme de classes est un élément essentiel dans la modélisation objet, c'est une collection d'éléments de modélisation statique (classes, paquetages…), qui montre la structure d'un modèle. Il fait abstraction des aspects dynamiques et temporels.

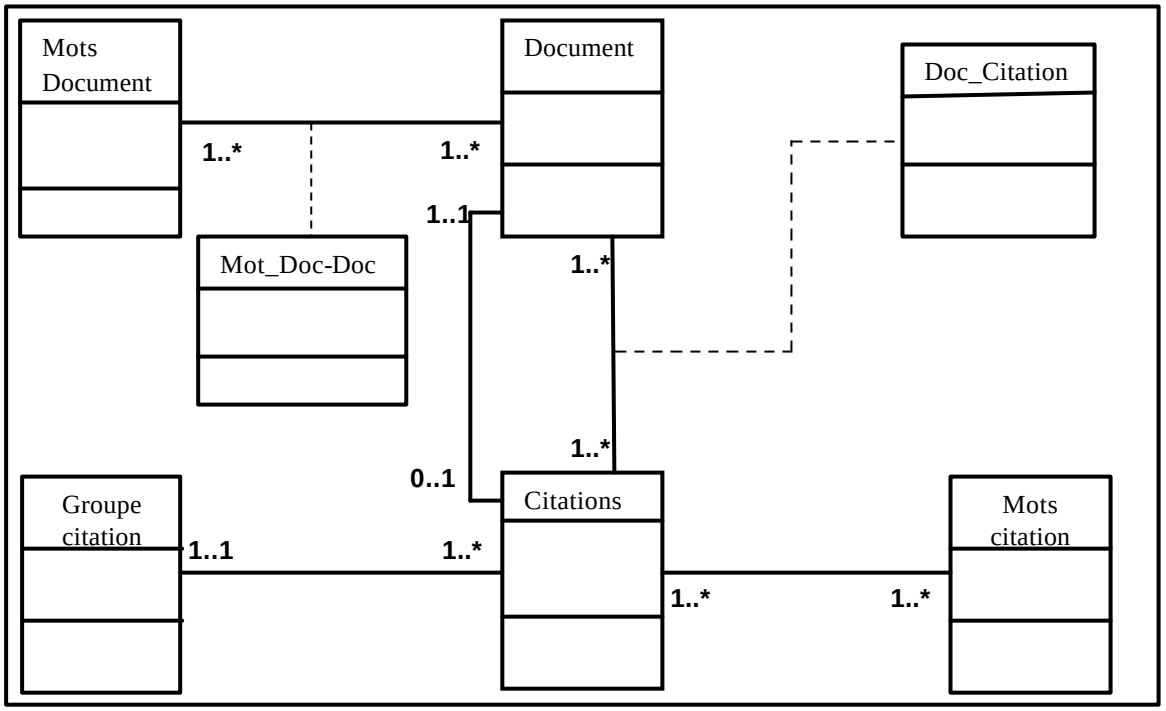


Figure 4.3 : Diagramme de Classes de la bibliothèque UFA*Seer*

4.3.4 Diagramme d'objet

Ce type de diagramme UML montre des objets (instances de classes dans un état particulier) et des liens (relations sémantiques) entre ces objets.

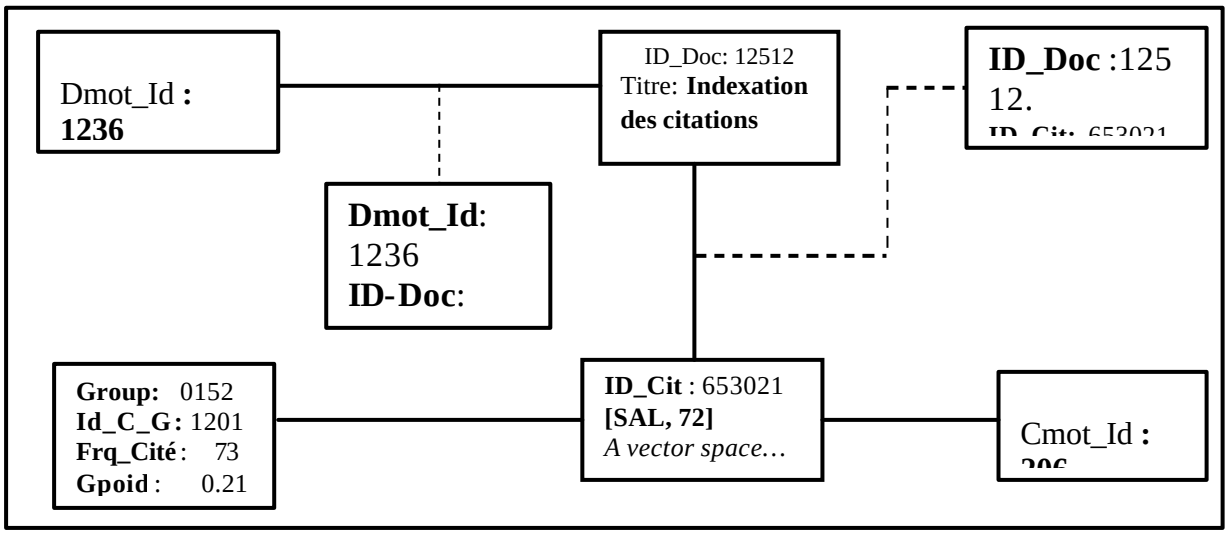


Figure 4.4 : Diagramme d'objets de la bibliothèque UFA*Seer*

4.3.5 Description détaillée des classes

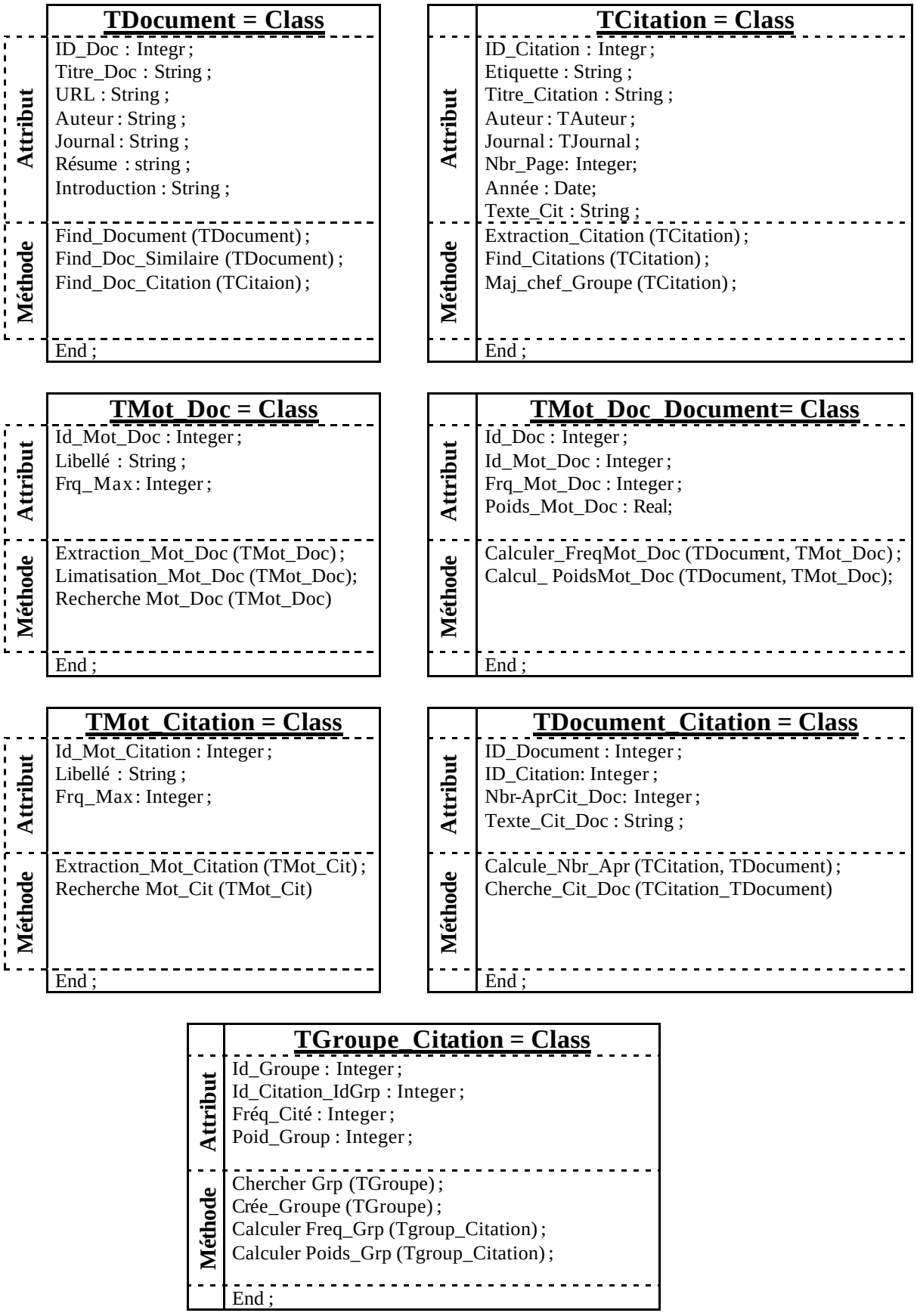


Figure 4.5 : Description des classes de la bibliothèque UFA*Seer*

4.3.6 Les diagrammes des séquences

Les diagrammes de séquence est une vue dynamique d'UML très importante permet de représenter des collaborations entre objets selon un point de vue temporel, on y met l'accent sur la chronologie des envois des messages.

La bibliothèque électronique UFASeer donne la possibilité de recherche soit par document, expliqué par le diagramme de séquence de la Figure 4.7, ou la recherche par citation Figure 4.6. Le diagramme de séquence dans la Figure 4.7 explique le parcours entre les documents et les citations, une fois l'utilisateur trouve un document ou une citation qui l'intéresse il passera au parcours de citations en avant et en arrière.

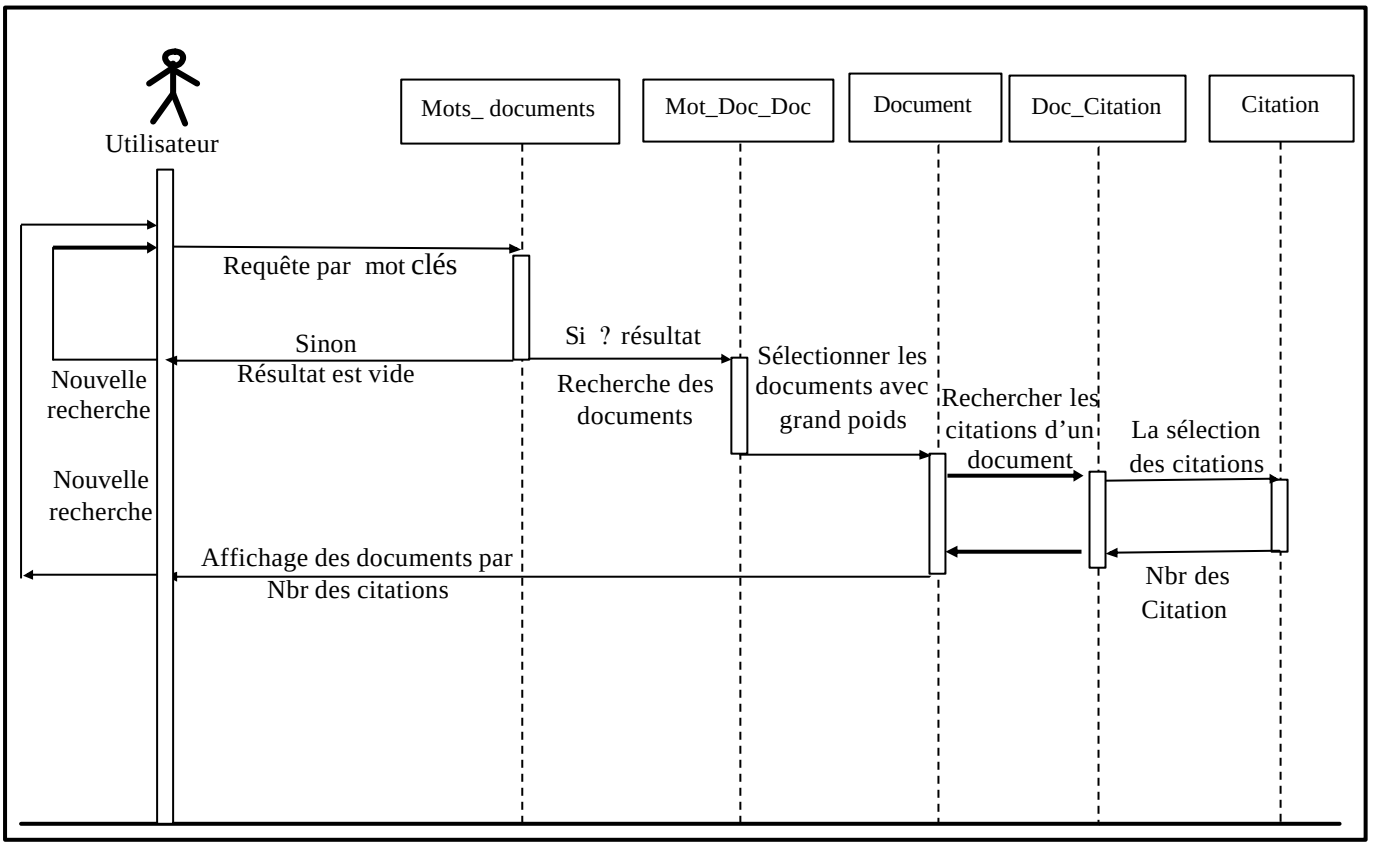


Figure 4.6 : Diagramme de séquence La recherche d'un Document dans UFA*Seer*

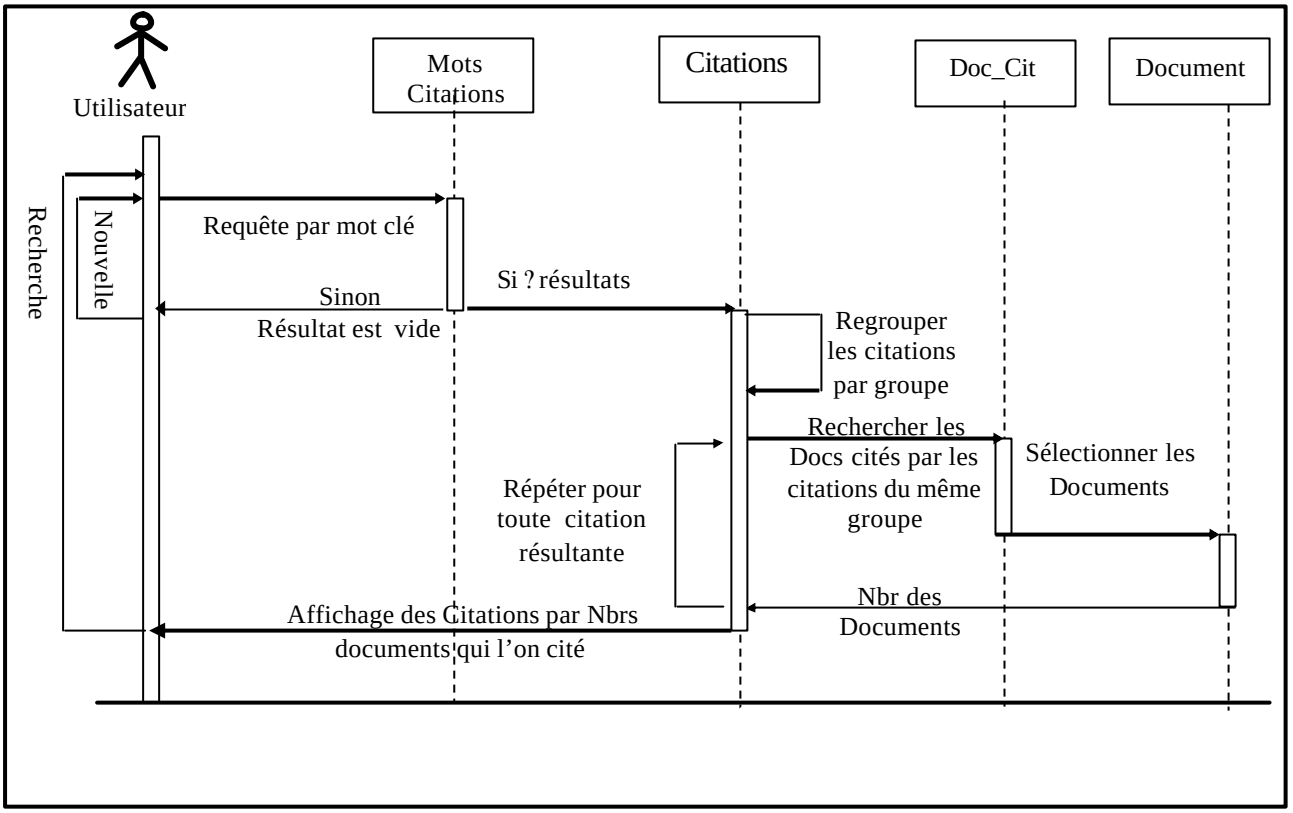


Figure 4.7 : Diagramme de séquence (La recherche d'une citation dans UFA*Seer)*

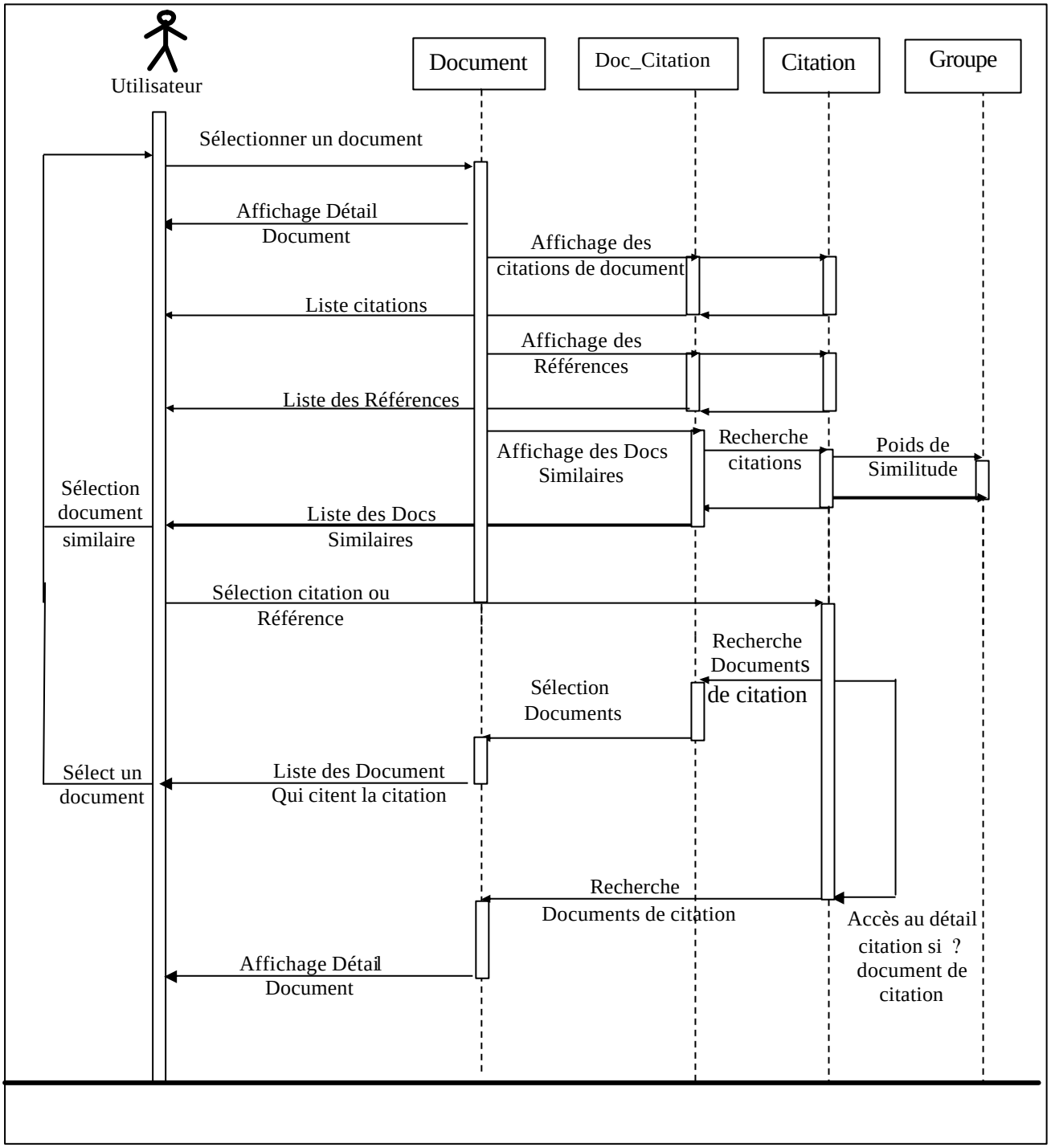


Figure 4.8 : Diagramme de séquences
Affichage détaillé du Document et le parcours des citations

Les deux diagrammes de séquences ci-dessous expliquent les séquences d'indexation d'un document en détaillant l'extraction des informations des documents (l'entête), extraction des mots document (Figure 4.10) et l'extraction des citations (Figure 4.9).

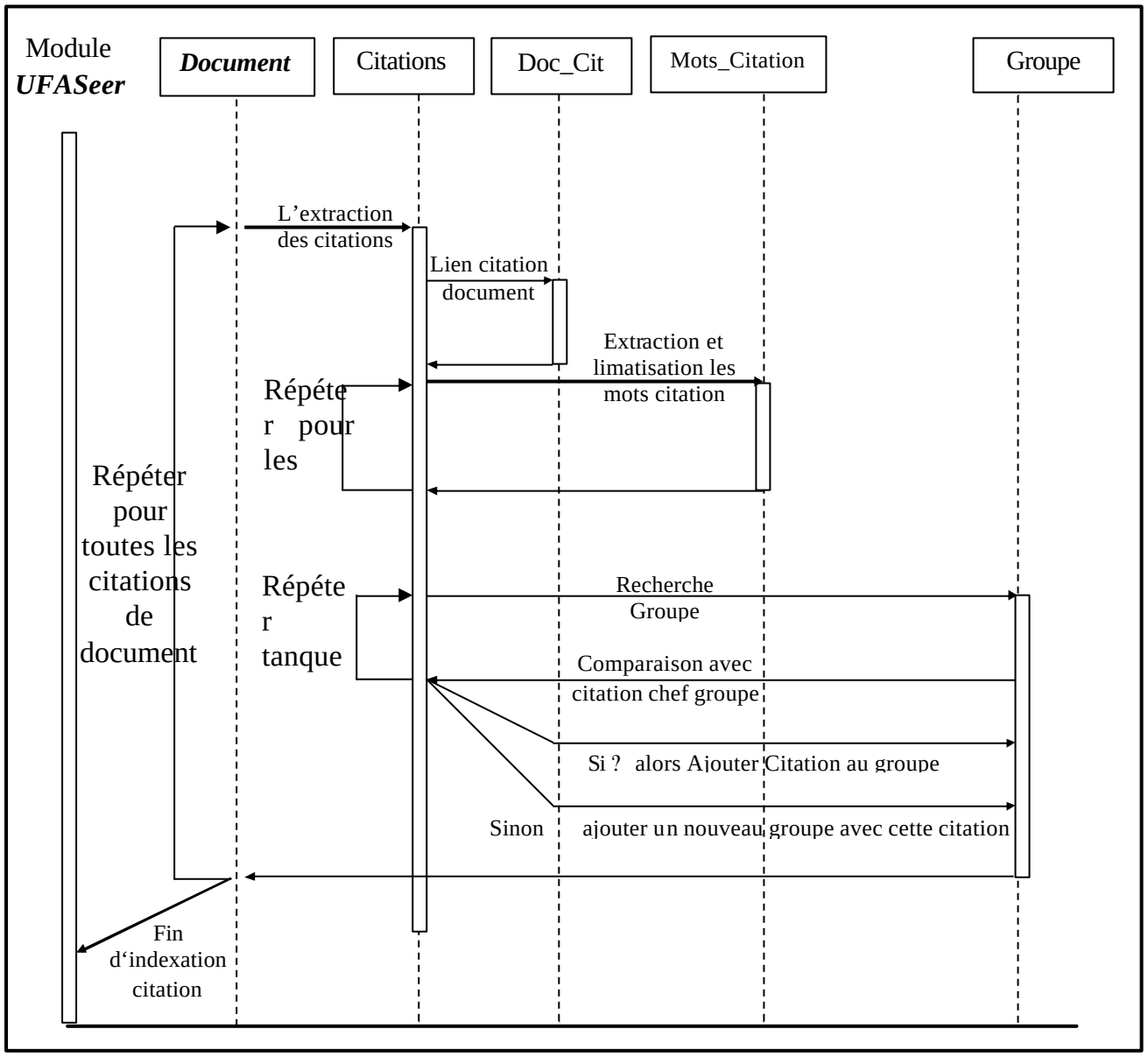


Figure 4.9 : Diagramme de séquence (Indexation des citations d'un document)

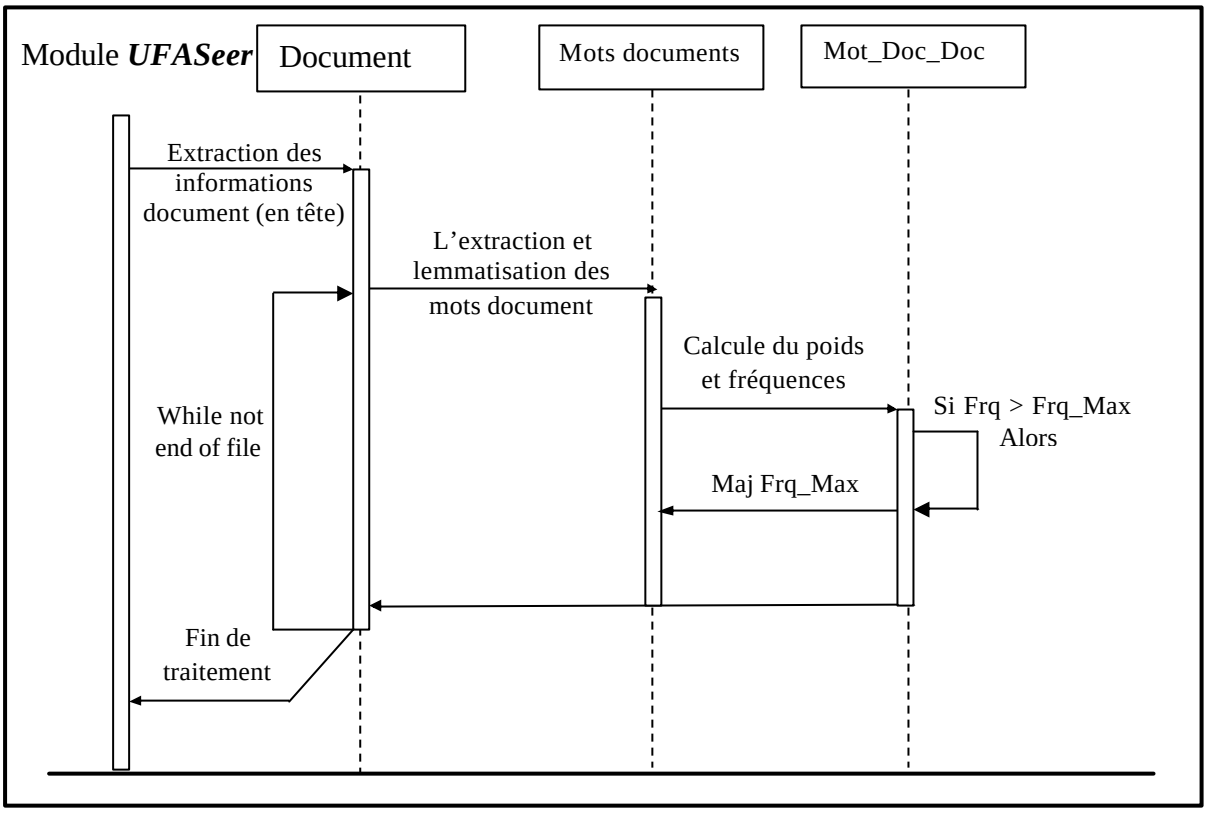


Figure 4.10 : Diagramme de séquences Indexation Document (Extraction de l'en-tête et Mots Document)

4.3.7 Diagrammes états transitions

Ce digramme sert à représenter des automates d'états finis, sous forme de graphes d'états, reliés par des arcs orientés qui décrivent les transitions. Il permet de décrire les changements d'états d'un objet ou d'un composant, en réponse aux interactions avec d'autres objets/composants ou avec des acteurs.

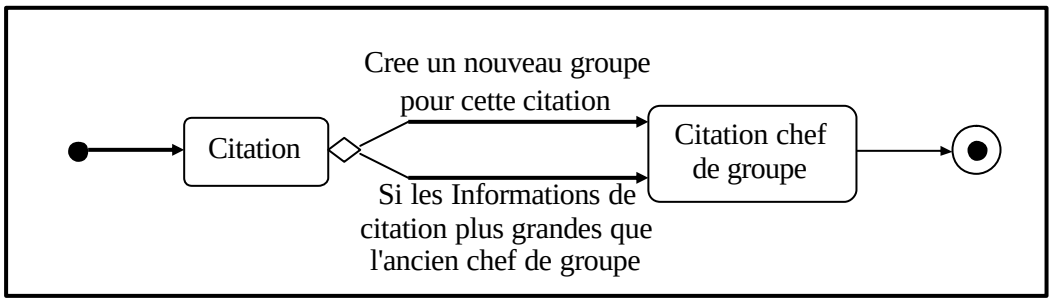


Figure 4.11 Diagramme états transitions (Citation –-> citation chef de groupe)

4.3.8 Diagrammes d'activités

UML permet de représenter graphiquement le comportement d'une méthode ou le déroulement d'un cas d'utilisation à l'aide de diagramme d'activités (une variante des diagrammes d'états-transitions).

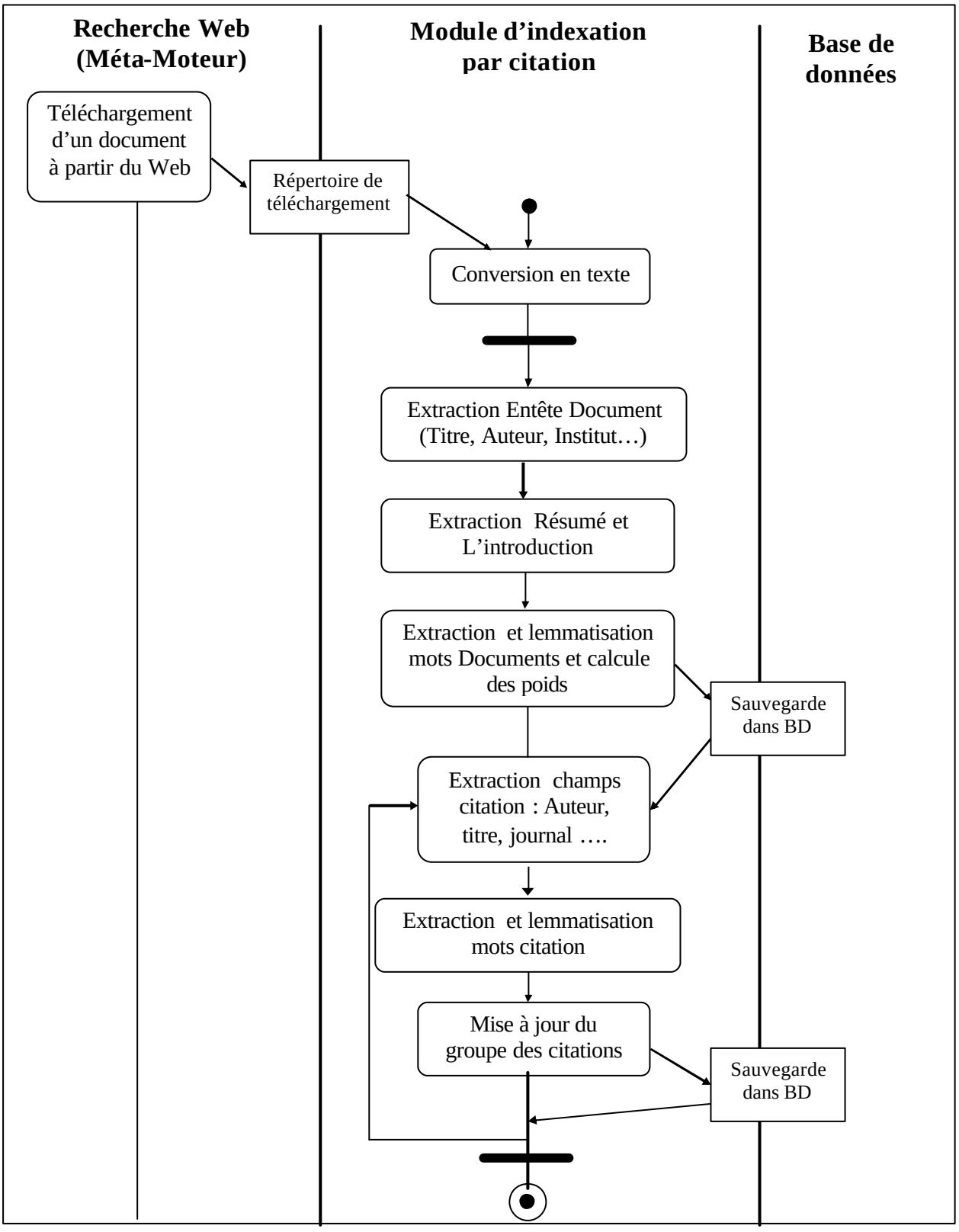


Figure 4.12 : Diagramme D'activités de l'indexation automatique dans UFASeer

2

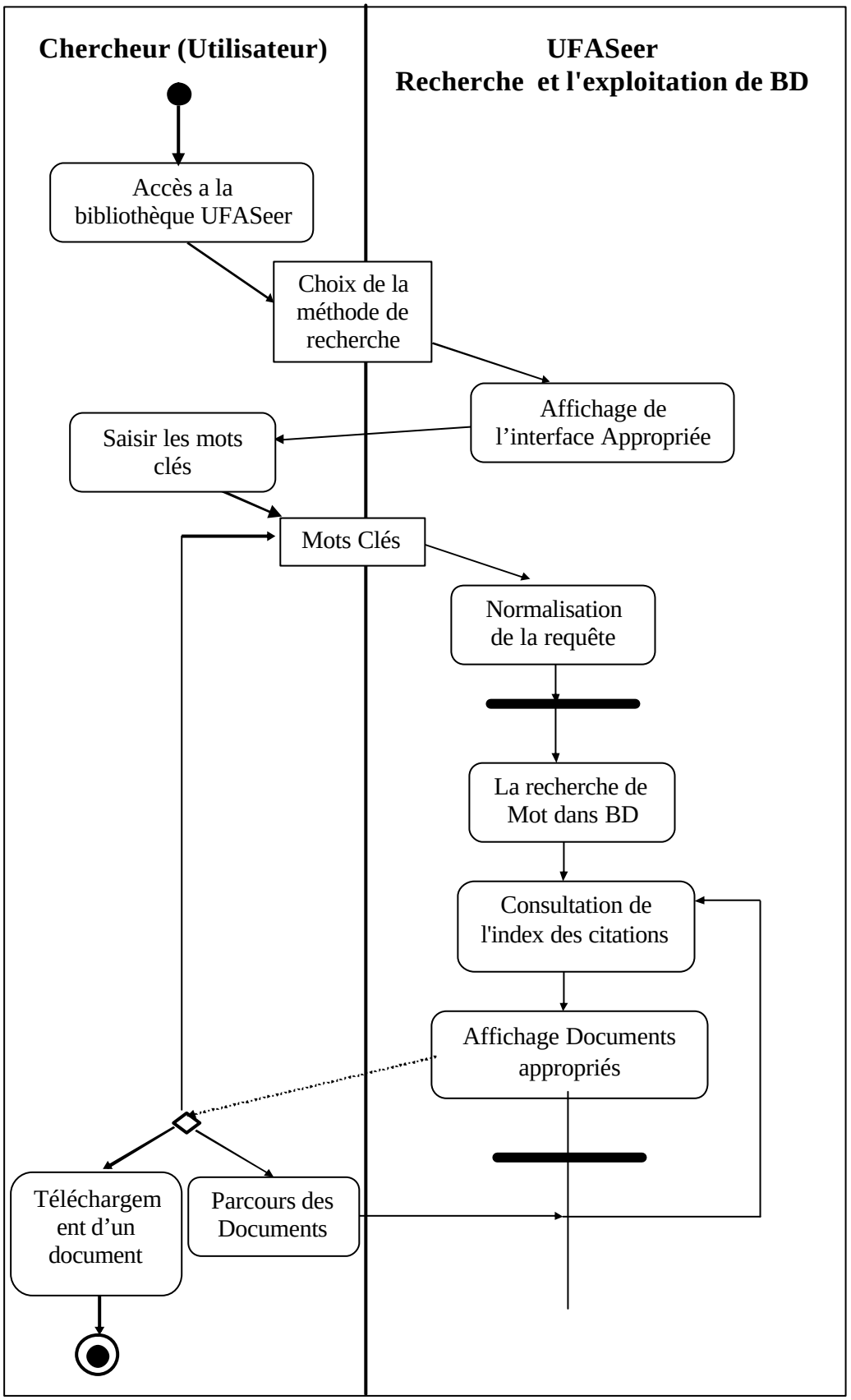


Figure 4.13 Diagramme d'activités de la recherche dans UFASeer

4.3.9 Diagrammes de composants

Le diagramme de composant met en évidence la structure logicielle du système, c'est-à-dire les dépendances entre types de composants. En d'autre terme les diagrammes de composants permettent d'écrire l'architecture physique et statique d'une application en terme de modules : fichiers sources, librairies, exécutables, entre autres. Ils montrent la mise en œuvre physique des modèles de la vue logique avec l'environnement de développement. Les dépendances entre composants permettent notamment d'identifier les contraintes de compilation et de mettre en évidence la réutilisation de composants. Ces diagrammes font apparaître les relations entre les composants comme le couplage entre modules, et parfois l'interface des types de composants peut être mise en évidence.

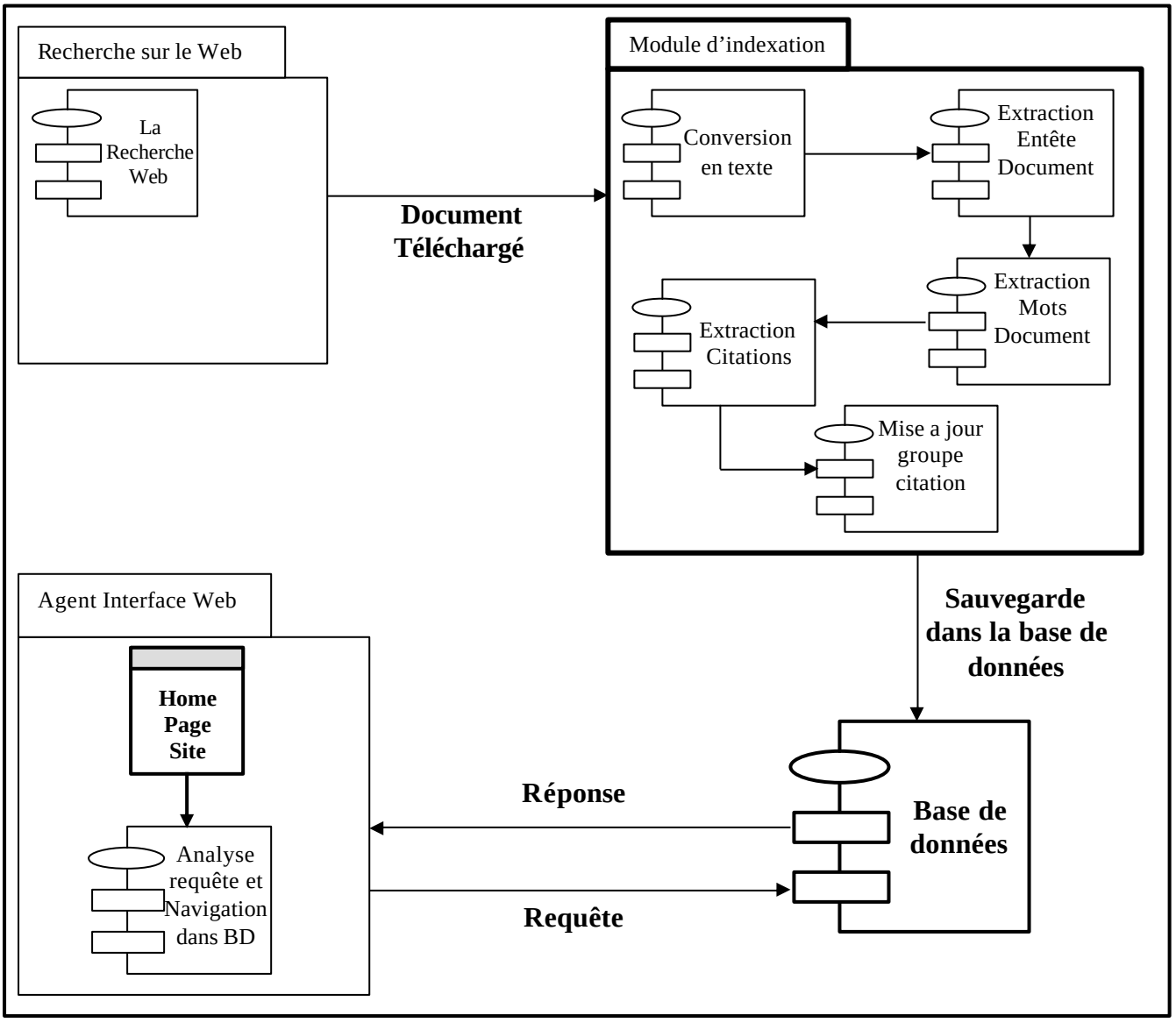


Figure 4.14 : Diagramme des composants (Bibliothèque Electronique UFA*Seer*)

4.3.10 Diagrammes de déploiement

Le diagramme de déploiement met en évidence la répartition des éléments de calculs (processus, composants, objets) sur les unités matérielles (les nœuds). Il est un graphe dont les sommets (nœuds) sont des ressources de calcul (processeurs ou périphériques) et les arcs (non orientés) sont les supports de communication (réseau).

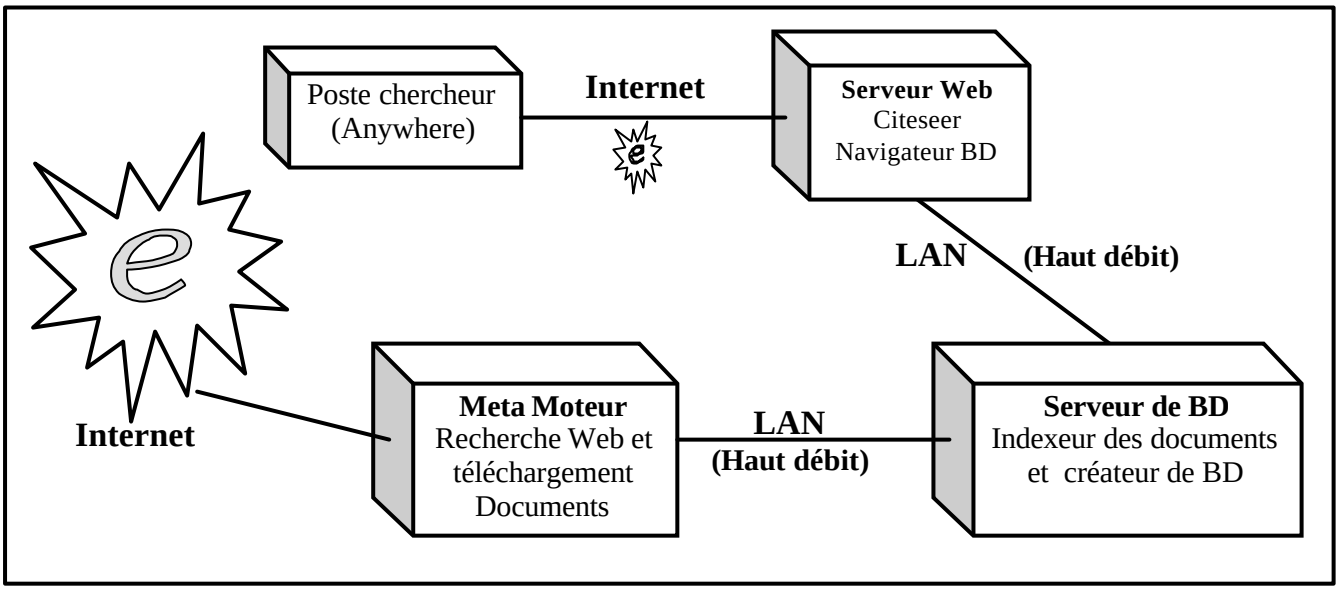


Figure 4.15 : Diagramme de déploiement (Bibliothèque Electronique UFA*Seer*)

4.4 Conclusion

Dans ce chapitre nous avons présenté l'architecture générale et détaillée d'une nouvelle bibliothèque électronique ***UFASeer*** qui utilise l'indexation des citations comme méthode parce que cette dernière a prouvé son succès dans l'indexation des documents scientifiques. Nous avons adopté UML comme langage de modélisation, afin d'en permettre la lisibilité et la maintenance. L'architecture de ***UFASeer*** est mise à la disposition des développeurs de bibliothèques électroniques, comme un outil supplémentaire efficace dans l'indexation des documents scientifiques.

Chapitre V

UFASeer Réalisation et discussions des résultats

5.1 Introduction

Notre objectif dans ce chapitre est de présenter le développement d'un prototype de bibliothèque électronique qui utilise l'indexation automatique des citations. Nous proposons une approche basée sur le *modèle de fouille de texte (Template mining)* pour extraire l'information à partir des articles publiés dans les journaux scientifiques de sorte que nous puissions établir automatiquement une base de données des citations. Nous la généralisons pour inclure les mémoires de Magistèrs et les thèses doctorales. Nous adoptons une représentation appelée *UFASCI* pour extraire automatiquement les informations des articles. Les résultats expérimentaux prouvent que, en employant *UFASCI*, nous pouvons extraire l'en-tête de l'article (Titre, Auteur et affiliation….) et le détail des références (L'auteur, le titre, le journal, le volume, le nombre (issue), l'année, et l'information de page à partir de différents modèles de référence avec un degré élevé d'exactitude.

5.2 Extraction automatique des informations et Template Mining

5.2.1 L'extraction automatique des informations

L'extraction automatique des informations (AIE), représente l'extraction spécifiée à partir du texte en langage naturel [Isaa 06]. En d'autres termes, l'AIE peut être vu comme une activité qui construit automatiquement une source d'informations structurées (une base de données) à partir d'un texte non structuré. Cette base de données peut être employée pour un certain nombre de buts : pour créer une base de données de citation, pour la production d'états, pour l'utilisation de *Data-Mining*, entres autres. La plupart de travail dans l'AIE a émergé de la recherche dans les systèmes basés sur les règles dans l'informatique linguistique et le traitement de langage naturel (NLP). L'exploitation de modèle est une technique particulière utilisée dans l'AIE, qui peut être employée pour extraire des données directement à partir du texte s'il est écrit selon un modèle bien connu [Laws96]. Quand le texte suit un modèle donné, le système extrait les données selon des instructions liées à ce modèle.

5.2.2 L'analyse de citation avec Template Mining

La technique d'exploitation de modèle peut être employée pour construire automatiquement des bases de données des citations qui peuvent contenir une information semblable aux bases de données d'ISI (*Institute for Scientifique Information* - http://scientific.thomsonreuters.com/isi/), telles que les informations

sur le nom du journal d'un article de citation, nom d'auteur, adresse d'auteur, titre, résumé, mots-clés et les détails des articles cités (références bibliographiques).

5.2.3 Modèle de Fouille de texte de (Template mining)

L'objectif de cette étude est d'extraire des informations de citation (les citations et les références citées) à partir des documents numériques, afin de créer les modèles appropriés. La dérivation des modèles est le résultat d'une analyse complète des articles choisis en identifiant la liste de modèles liés à ces articles. Cela suppose que ces modèles représentent une collection des journaux. L'expérience prouve que les modèles de citation sont suivis strictement par les journaux. Et en fait, chaque journal imprimé a un modèle pour les citations, et les auteurs sont requis de suivre ce modèle. Cependant, de tels modèles sont des caractéristiques liées au journal, et ne sont pas communes à tous les journaux électroniques.

5.3 Le modèle général d'un article de citation

Pour chaque article, l'information sur les éléments suivants doit être disponibles : **Informations sur le journal (**Nom, volume, numéro, date de publication, ISSN**…), titre, nom d'auteur, adresse d'auteur, email d'auteur, mots-clés, résumé**, **Introduction, références**. La figure 5.1 montre l'organigramme *du modèle d'article de citation*.

Plusieurs articles de citation ont été examinés pour extraire les descriptions des différents modèles et pour identifier les données à extraire. Et également comme conséquence de l'identification, on obtient les *'signes d'information (Tokens)'* qui décrivent les modèles comme dans le tableau 5.1.

Tableau 5.1 Les tokens dans le modèle d'un article

L'UNITÉ D'INFORMATION	LES TOKENS	PONCTUATIONS
Journal	(une ligne ou. de plusieurs lignes) ' vol ', 'volume', 'issue ', 'no.', 'No.','d, d(d) , 'ISSN', 'pp', d-d;	',' ; '.' ; l'espace; fin de ligne; interligne ;
Nom d'auteur	(une ligne ou plusieurs lignes) a+ +a, c.+c.+ +a, a+ +c.+ +a	','; '&'; 'and'; Fin de ligne; interligne ;
Adresse d'auteur	(une ligne ou plusieurs lignes) 'university' ', 'department', 'school', 'city name, country name,	',' ; fin de ligne; interligne;
Email	'@'	Fin de ligne
Mots-clés	'keywords' , 'key words'	',' ; ';' ; Espace
Abstrait	paragraphes après ' Abstract' ou 'Type of Article'; paragraphes avant 'Introduction', 'Acknowledgements', 'Background', 'Contents'; paragraphes entre 'Abstract' et 'Introduction'	Interligne
Acknowledgement (Remerciements)	'Acknowledgements', 'Acknowledgment'	
Références	'References', 'References and further reading', 'Formal publications cited', 'Notes', 'Bibliography', 'ADDITIONAL READINGS'; 'Works cited'; 'Background reading'; 'A brief bibliography'; 'Notes and references';	

Remarque: d représente un nombre numérique, **a** représente une chaîne, **c** représente un caractère. Le texte qui apparaissent avec le **' '** indique que le texte est constant. + indique: juste après.)

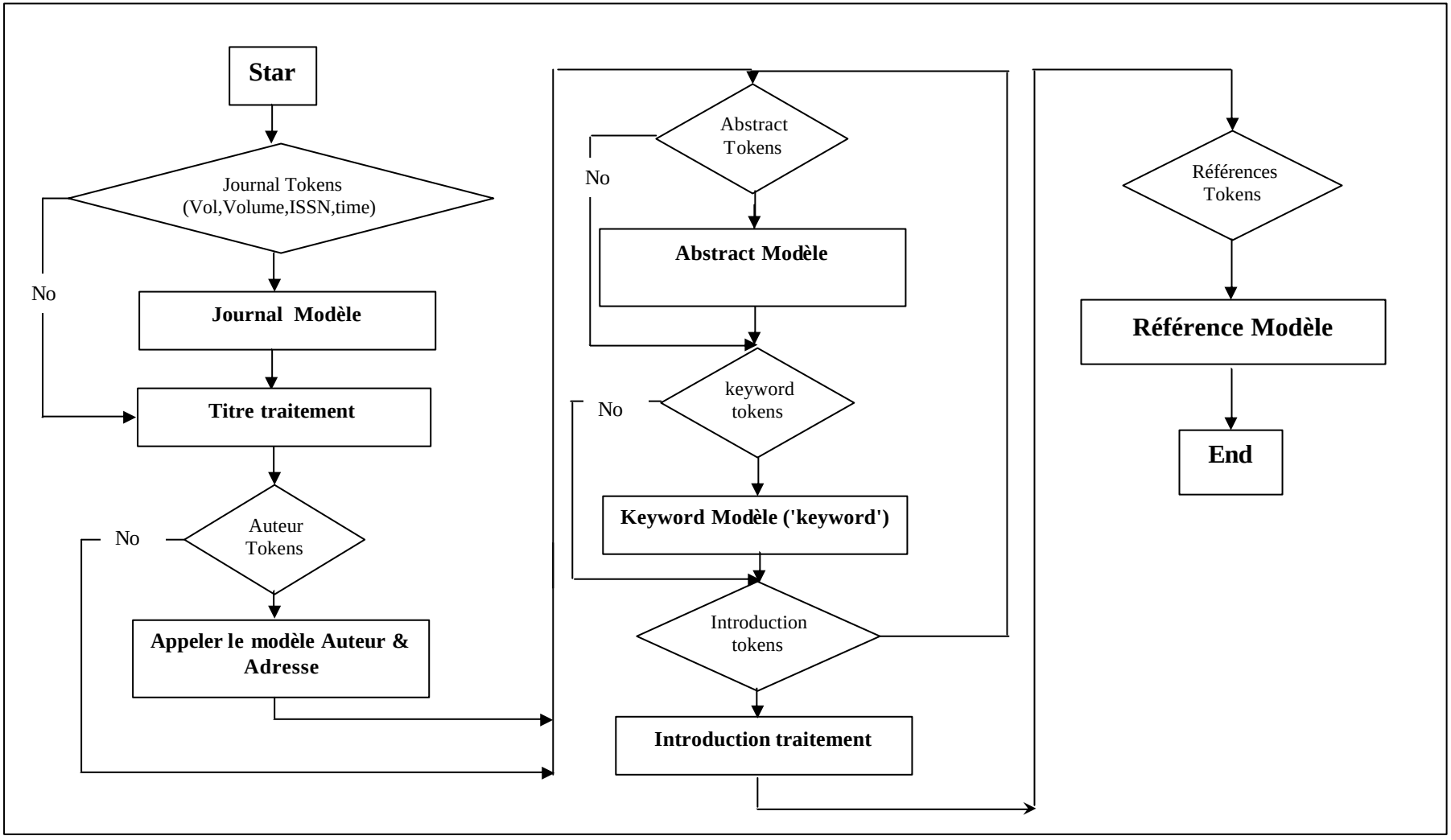


Figure5.1 Organigramme en-tête article

Digital Libraries 98 - Third ACM Conference on Digital Libraries, Edited by I. Witten, R. Akscyn, F. Shipman III, ACM Press, New York, pp. 89–98, 1998.

CiteSeer: An Automatic Citation Indexing System

C. Lee Giles, Kurt D. Bollacker, Steve Lawrence
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
{giles,kurt,lawrence}@research.nj.nec.com

ABSTRACT

We present *CiteSeer*: an autonomous citation indexing system which indexes academic literature in electronic format (e.g. Postscript files on the Web). CiteSeer understands how to parse citations, identify citations to the same paper in different formats, and identify the context of citations in the body of articles. CiteSeer provides most of the advantages of traditional (manually constructed) citation indexes (e.g. the ISI citation indexes), including: literature retrieval by following citation links (e.g. by providing a list of papers that cite a given paper), the evaluation and ranking of papers, authors, journals, etc. based on the number of citations, and the identification of research trends. CiteSeer has many advantages over traditional citation indexes, including the ability to create more up-to-date databases which are not limited to a preselected set of journals or restricted by journal publication delays, completely autonomous operation with a corresponding reduction in cost, and powerful interactive browsing of the literature using the context of citations. Given a particular paper of interest, CiteSeer can display the context of how the paper is cited in subsequent publications. This context may contain a brief summary of the paper, another author's response to the paper, or subsequent work which builds upon the original article. CiteSeer allows the location of papers by keyword search or by citation links. Papers related to a given paper can be located using common citation information or word vector similarity. CiteSeer will soon be available for public use.

KEYWORDS: citation indexing, citation context, literature search, bibliometrics.

INTRODUCTION

A citation index [6] indexes the links between articles that researchers make when they cite other articles. Citation indexes are very useful for a number of purposes, including the advantages of traditional (manually constructed) citation indexes (e.g. the ISI citation indexes [10]), including: literature retrieval by following citation links (e.g. by providing a list of papers that cite a given paper), the evaluation and ranking of papers, authors, journals, etc. based on the number of citations, and the identification of research trends. CiteSeer has many advantages over traditional citation indexes, including a more up-to-date database which is not limited to a preselected set of journals or restricted by journal publication delays, completely autonomous operation with a corresponding reduction in cost, and powerful interactive browsing of the literature using the context of citations.

CITATION INDEXING

References contained in academic articles are used to give credit to previous work in the literature and provide a link between the "citing" and "cited" articles. A citation index [6] indexes the citations that an article makes, linking the articles with the cited works. Citation indexes were originally designed mainly for information retrieval [7]. The citation links allow navigating the literature in unique ways. Papers can be located independent of language, and words in the title, keywords or document. A citation index allows navigation backward in time (the list of cited articles) and forward in time (which subsequent articles cite the current article?) Citation indexes can be used in many ways, e.g. a) citations can help to find other publications which may be of interest, b) the context of citations in citing publications may be helpful in judging the important contributions of a cited paper and the usefulness of a paper for a given query [7, 14], c) a citation index allows finding out where and how often a particular article is cited in the literature, thus providing an indication of the importance of the article, and d) a citation index can provide detailed analyses of research trends and identify emerging areas of science [8].

REFERENCES

1. Eytan Adar and Jeremy Hylton. On-the-fly hyperlink creation for page images. In *Proceedings of Digital Libraries '95 -The Second Annual Conference on the Theory and Practice of Digital Libraries*, June 1995.

2. T. A. Brooks. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37:34–36, 1986.

3. Robert D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, 2(4), 1997.

16. Gerard Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372, 1973.

17. H. G. Small. Cited documents as concept symbols. *Social Studies of Science*, 8(327–340), 1978.

18. Peter Yianilos. The LikeIt intelligent string comparison facility. Technical Report 97-093, NEC Research Institute, 1997.

19. Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the 4th ACM-SIAM Symposium on Discrete Algorithms*, pages 311–321, 1993.

Figure 5. Les éléments d'informations dans un article scientifique

Les formats des éléments d'informations peuvent changer de manière significative. Par exemple, la position des auteurs et leurs adresses dans les articles la figure 5.2 représente les différents cas. On rencontre le même problème pour les informations sur le journal, les mots clés et le résumé.

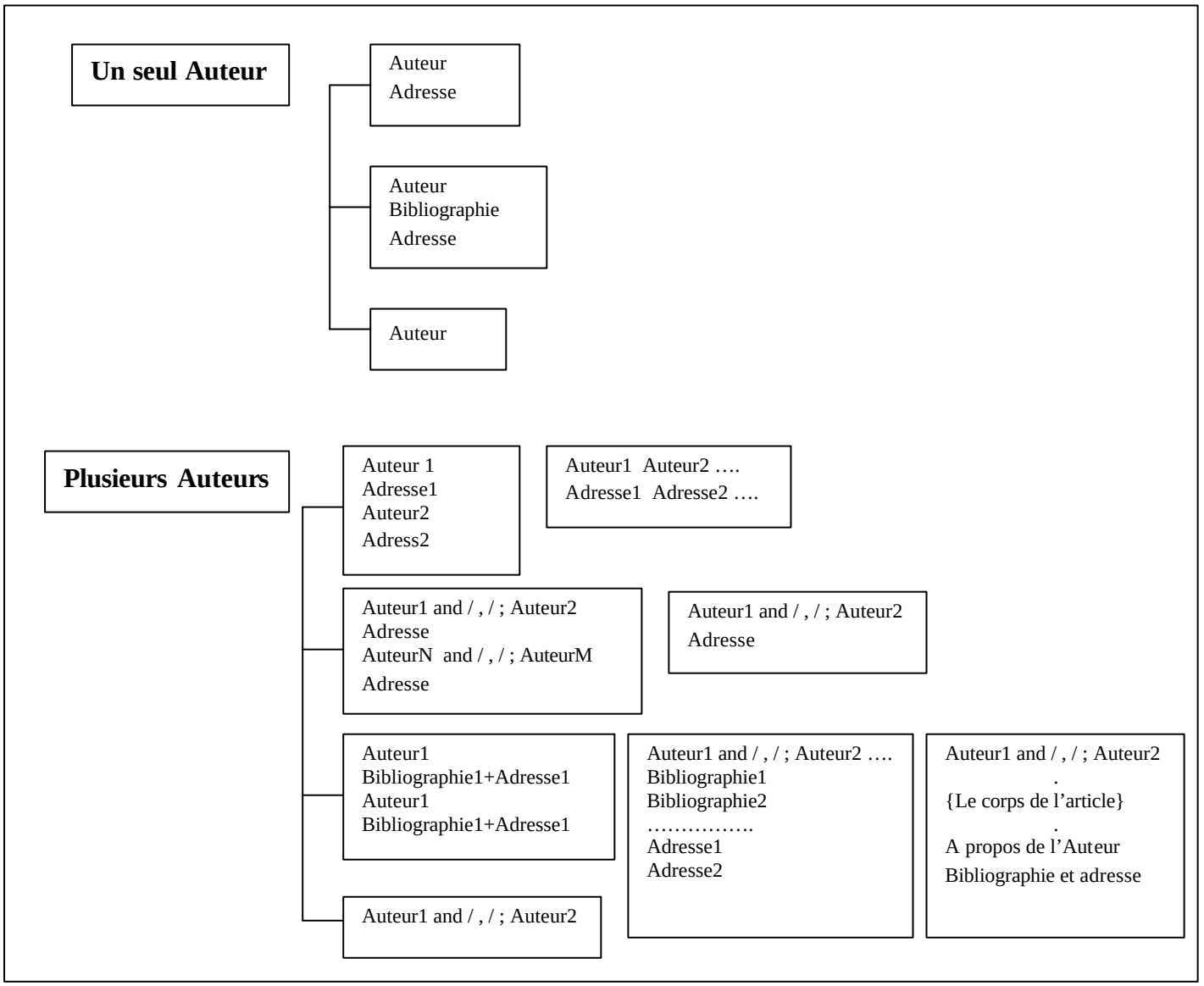


Figure 5.2 La variation de position des auteurs et sont adresses dans les articles scientifiques

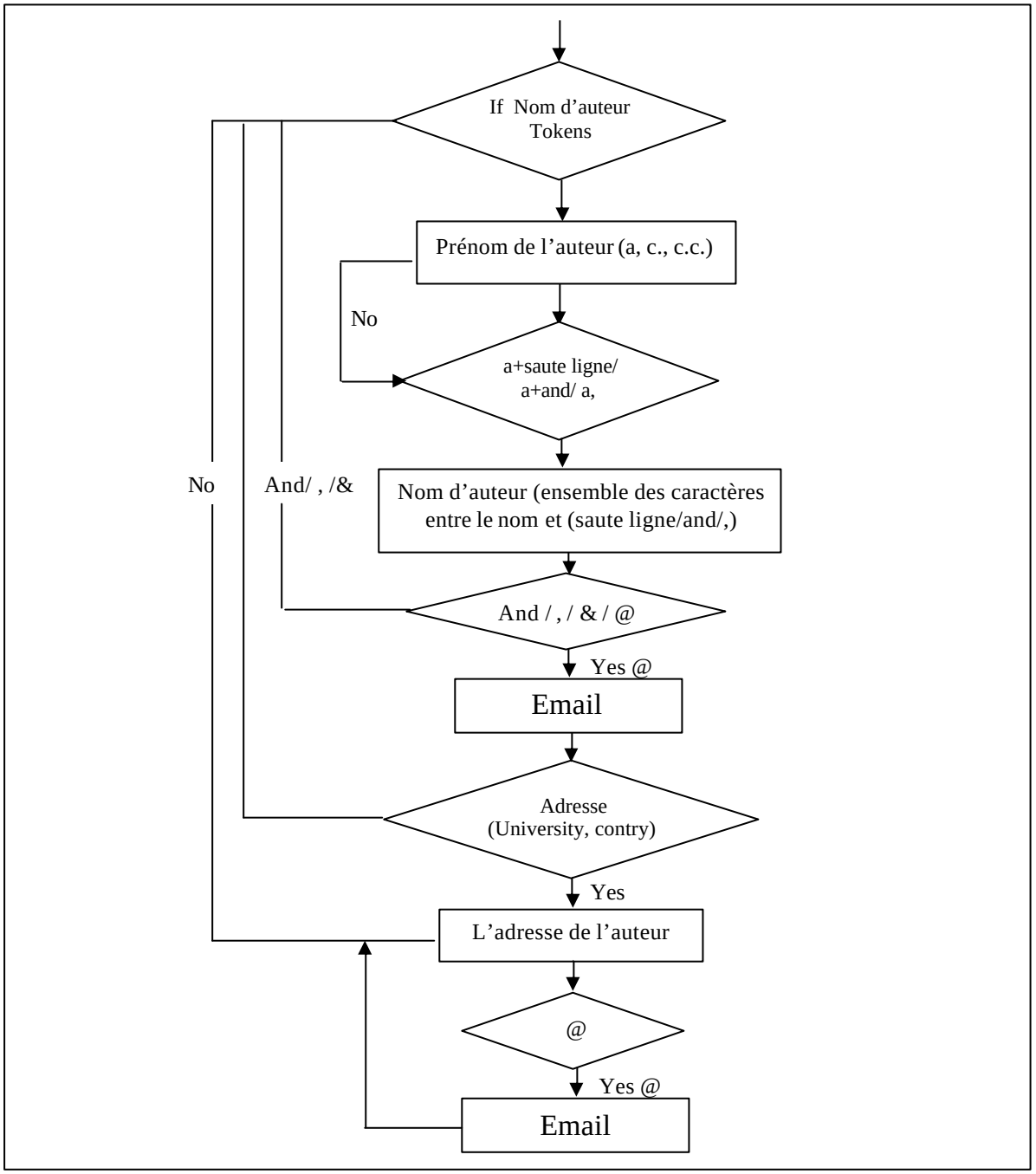


Figure5.4 Organigramme Auteur & adresse

5.4 Le modèle général des citations

L'analyse des différents articles a indiqué qu'il y a plusieurs manières d'identifier les parties des références comme le montre le Tableau 5.2 :

??**Nom d'auteur**. Bien que les noms d'auteurs aient plusieurs variations, ils ont quelques modèles fixes. Dans notre étude, nous avons proposé le modèle ou le nom de famille d'auteur doit être avant le prénom d'auteur.

??**Position de l'année** Il est facile identifier l'année, c'est une ensemble de quatre chiffres entiers consécutifs. Mais il n'est pas aussi facile de trouver sa position dans les références. Il peut apparaître:

- Au début des références (par exemple [Pinker 1994] Steven Pinker, l'instinct de langue, New York : Harper Collins.),.
- Après le nom d'auteur (par exemple Hjortgaard Christensen, F. et Ingwersen, P. (1996), "citation en ligne analyse : une approche méthodologique ", Scientometrics, numéro 1, vol. 37, pp 39-62.).
- A la fin des références (par exemple Mirjana Spasojevic et M. Satyanarayanan. Une étude empirique d'un système de fichiers réparti par secteur. Transactions ACM sur les systèmes informatiques, 14(2) : 200-222, mai 1996.).

Mais nous pouvons employer les quatre chiffres consécutifs (particulièrement il commence par 19, plus tard augmentant à 20) comme marque pour identifier le temps.

?**Titre**. Le titre peut inclure le titre d'article, le titre de livre, le titre de journal, le titre de conférence, le titre de rapport et d'autres. La variété de titres rend leur extraction dans les références difficile est irréalisable. Les titres de journal montrent une variation plus que les titres de livre. Cependant, quelque marque et les signes d'information peuvent nous aider à les identifier, tels que la marque comme le **'journal'**, **'proceddings'**, **'conférence'**, **'rapport'**, **'technical rapport'** , **'in'**, **'(ed.)'**, **'(eds.) '** et leurs abréviations et ainsi de suite.

Tableau 2 Les tokens dans le modèle des citations

L'UNITÉ	LES TOKENS	PONCTUATIONS
Etiquette	[add] ; [dd] ; [d] ;(d) ;	'[‘; ‘]’; ‘(‘; ‘)’
Nom d'auteur	Nom,+ prénom **;** Prénom+ +Nom **;** a,+ +c.+c.**;** a,+ +a+ +c., a+ +a+ +a **;** c.+ +c.+ +a **;** a+ +c.+ +a **;**	**‘,’ ;** Espace, **‘;’ ;** ‘and’**;’**&’**;**
Année	d (ignoré le mois) **;** 1900-1999**,**	‘(‘**;** ‘,’**;** ‘.’**;**
Titre	a+ *a* **;** in **;** in+… +(ed.) **;** in+… +(eds.)**;** “s”**;** s+nom de journal **;** s+ nom de conférence**;**	',' **;** ‘ “ ‘ **;** ‘.’**;** ‘?’**;**
Titre de journal	*'journal*' **;** s+number **;** s+vol. **;** s+d(d) **;**	‘,’**;** ‘.’**;**
Éditeur	L'adresse d'éditeur +:+Nom d'édition **;** le nom d'éditeur +,+ L'adresse d'éditeur**;** nom d'éditeur **;** s+:+s **;** s+, +s **;** s+, **;** pour le nom d'éditeur on inclut le ‘Inc.’**;** pour l'adresse d'éditeur on inclut CC(acronyme des états, tels que le MA, NY..)	‘ : ‘**;** ‘, ‘**;**
Page	d-d **;** d **;** ‘ pp.’**;** ‘ p.’**;** ‘ : ‘**;**	‘ pp.’**;** ‘ p.’**;** ‘ : ‘
Adresse HTTP	‘http’ **;** ‘ftp’ **;** s **;** s **;** ‘Available at’ **;** ‘Available’;	‘http’ **;** ‘ftp’**;**
Journal Volume	‘volume’**;** ‘vol.’**;** d **;** d+(	‘, ‘**;** l'espace **;**
Numéro Issue	‘Issue’ **;** ‘No.’ **;** ‘no.’ **;** (+d+) **;**	‘(‘**;** ‘,’**;**
Page de journal	‘pp.’**;** ‘p.’**;** d-d **;** d **;** ‘:’	‘pp.’**;** ‘p.’**;** ‘: ‘**;**

(**Notes** : **d** représente un nombre numérique, **a** représente une chaîne des caractères, **c** représente un caractère, **s** représente une chaîne des caractères & symboles de ponctuations).

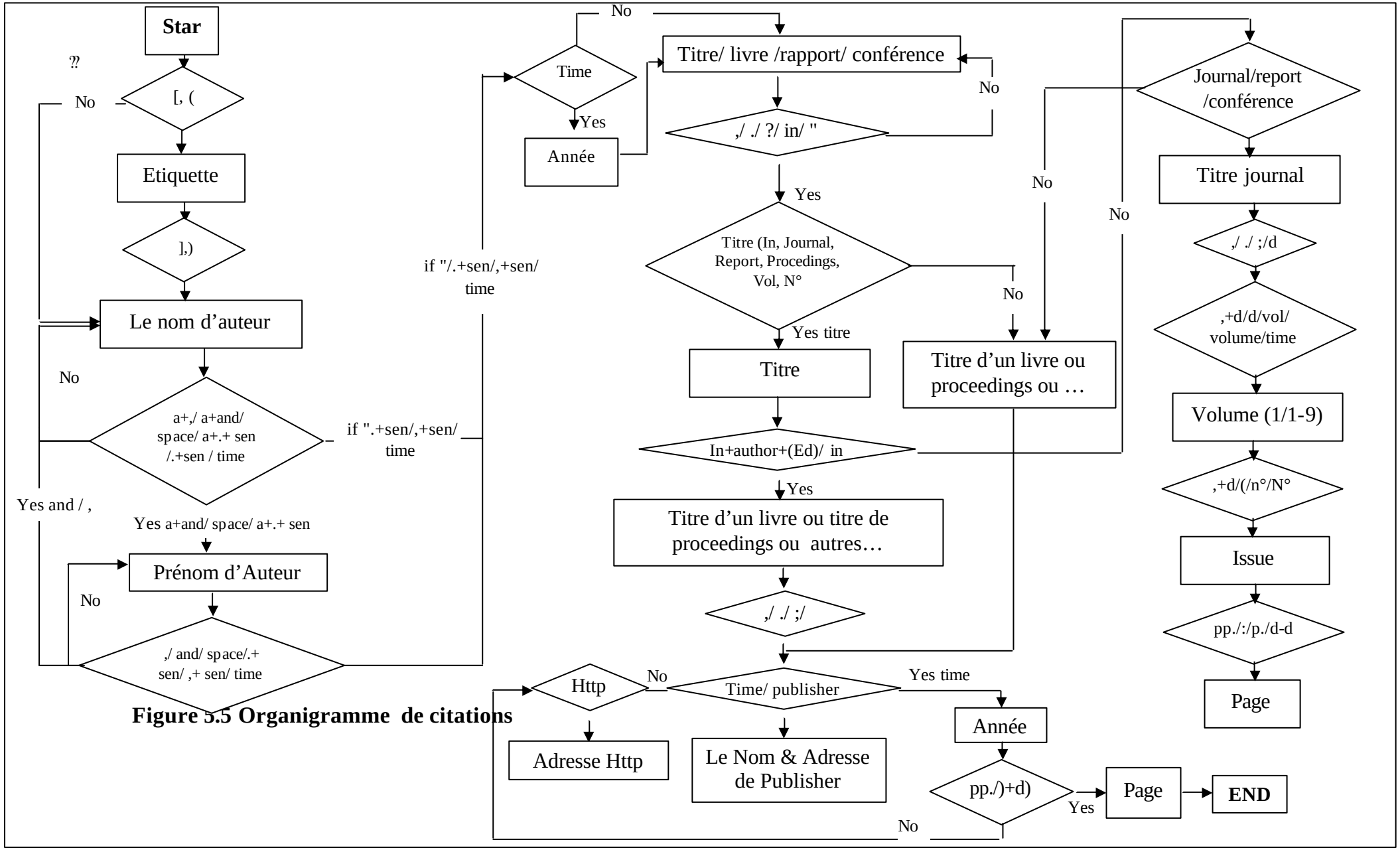


Figure 5.5 Organigramme de citations

Remarque :

Vinkler [Vink 94] a observé que 55% des références des journaux viennent seulement de 10% de titres de journal. Ainsi, une base de données contenant les journaux le plus fréquemment cités peut être employée pour extraire le titre.
Elle peut également nous aider à décider si la chaîne est le nom de journal dans le calibre d'article citation. De la même manière, nous pouvons également remplir la base de données de conférence, base de données des éditeurs, base de données de pays, base de données des universités, la base de données des auteurs, et ainsi de suite, qui peut la faciliter pour distinguer les différents noms.

5.5 Le projet de la bibliothèque *UFASeer*

Notre projet (application) *UFASeer* crée automatiquement la base de données de la bibliothèque électronique en utilisant la technique *Template Mining.*

Le projet UFA*Seer* est représenté par deux composants principaux :

4) Un analyseur des documents et un créateur de base de données (Ufasci).
5) Une interface web (Site UFA*Seer*) pour la navigation dans la base de données qui permet la recherche par mot-clé et la navigation (parcours) par les liens de citation.

La figure ci-dessous montre un diagramme de cette architecture.

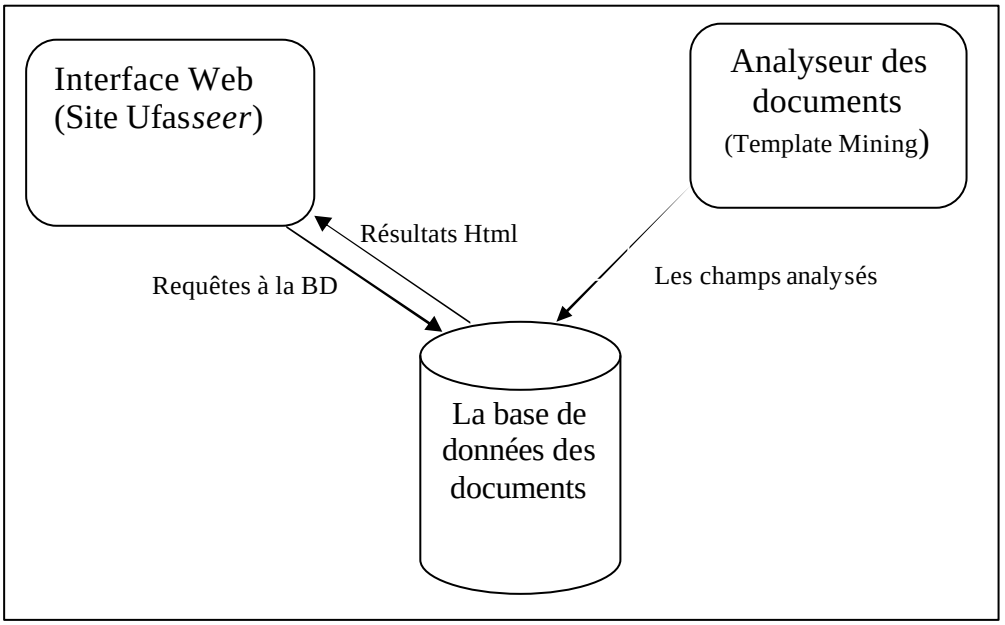


Figure 5.6 : L'architecture de projet UFA*Seer*

5.5.1 L'analyseur des documents

L'analyseur des documents (***Ufasci)*** est un module développé sous environnement DELPHI 7. Il s'exécute d'une manière permanente et automatique avec le démarrage du système d'exploitation Windows. L'utilisateur peut l'arrêter manuellement j la fin des taches. ***Ufasci*** utilise trois répertoires de base, un pour récupérer les fichiers sous format ".txt" afin de les traiter. Après le traitement, les documents correctement traités (sans erreurs) sont classés dans un deuxième répertoire. Le troisième répertoire regroupe les documents non traités par le module ***ufasci***.

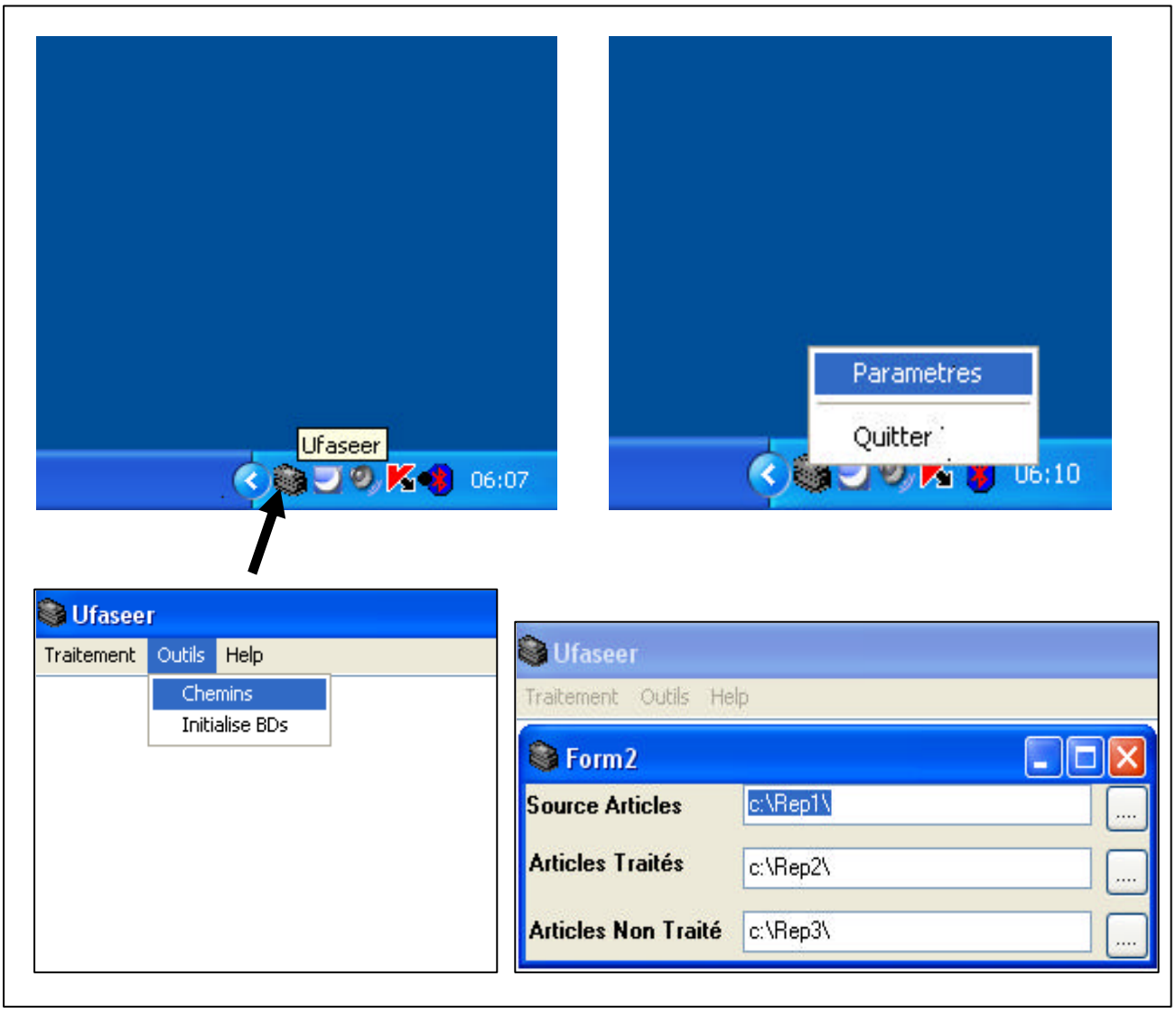


Figure 5.7 : Module *Ufasci* en cour de démarrage avec le paramétrage des répertoires de base

5.5.2 Le principe de fonctionnement de module UFSci

Le module ***UFSci*** teste d'une manière permanente l'existence d'un fichier sous format texte dans le répertoire de base (Rep1), une fois qu'il trouve un fichier, l'opération de traitement commence avec la lecture du fichier et la détection de différents parties contenant les informations importantes, tel que :

- L'entête de fichier regroupe le Titre, les Auteurs et ses adresses….. .
- Le résumé, les mots clés.
- L'introduction.
- La Partie Référence : la partie importante dans notre approche (Indexation par citation).

Voir l'exemple ci-dessous,

The Availability and Persistence of Web References in D-Lib Magazine
Frank McCown, Sheffan Chan, Michael L. Nelson, Johan Bollen
Old Dominion University
Department of Computer Science
Norfolk, VA 23529 USA
{fmccown,chan_s,mln,jbollen}@cs.odu.edu
Abstract. We explore the availability and persistence of URLs cited in articles published in D-Lib Magazine. We extracted 4387 unique URLs referenced in 453 articles published from July 1995 to August 2004. The availability was checked three times a week for 25 weeks from September 2004 to February 2005 .
……..
1 Introduction
D-Lib Magazine plays a pivotal role in the documenting and advancing of trends in the digital library community [3]. Given its importance to the community, appropriate measures have been taken to preserve the primary contents of D-Lib Magazine; it is officially mirrored in six other locations throughout the world. D-Lib Magazine is highly interlinked with other digital libraries and the general web. It is published on-line, and all its articles are HTML formatted, …….
7 Acknowledgements
We would like to thank Bonnie Wilson, editor of D-Lib Magazine, for reviewing our paper.
References
1. Bar-Yossef, Z., Kumar, R., Broder, A. Z., Tomkins, A.: Sic Transit Gloria Telae: Towards an Understanding
of the Web's Decay. In Proceedings of the International WWW Conference. (2004)
2. Berners-Lee, T.: Cool URIs Don't Change. http://www.w3.org/Provider/Style/URI.html
3. Bollen, J., Nelson, M. L., Manepalli, G., Nandigam, G., Manepalli, S.: Trend Analysis of the Digital Library
Community. D-Lib Magazine, Vol. 11 1 (2005) http://www.dlib.org/dlib/january05/bollen/01bollen.html
4. Giles, C. L., Bollacker, K. D., Lawrence, S.: CiteSeer: An Automatic Citation Indexing System. Proceedings
of ACM Digital Libraries. (1998) 89-98
5. Harter, S., Kim, H.: Electronic Journals and Scholarly Communication: A Citation and Reference Study. Research, Vol. 2 1 Paper 9. (1996) http://informationr.net/ir/2-1/paper9a.html

Figure 5.8 : Une partie d'un article sous format texte

Si on prend le fichier texte de la Figure 5.8 ci-dessus, et si on le copie dans le répertoire de base (Rep1) de module ***UFasci***, après le traitement, la base de donnée résultante doit contenir toutes les informations nécessaires (voir les figures ci-dessous).

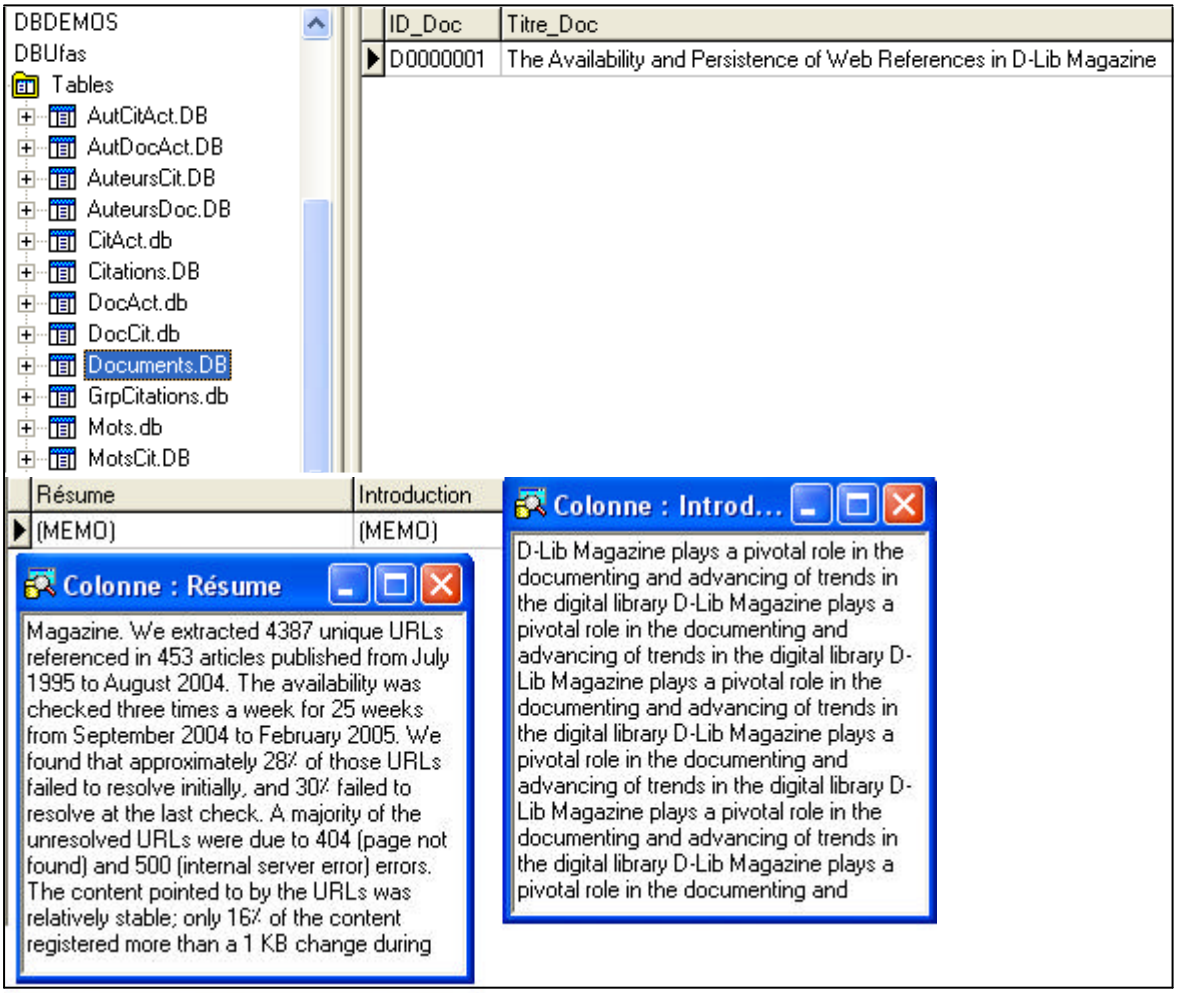


Figure 5.9: La table documents

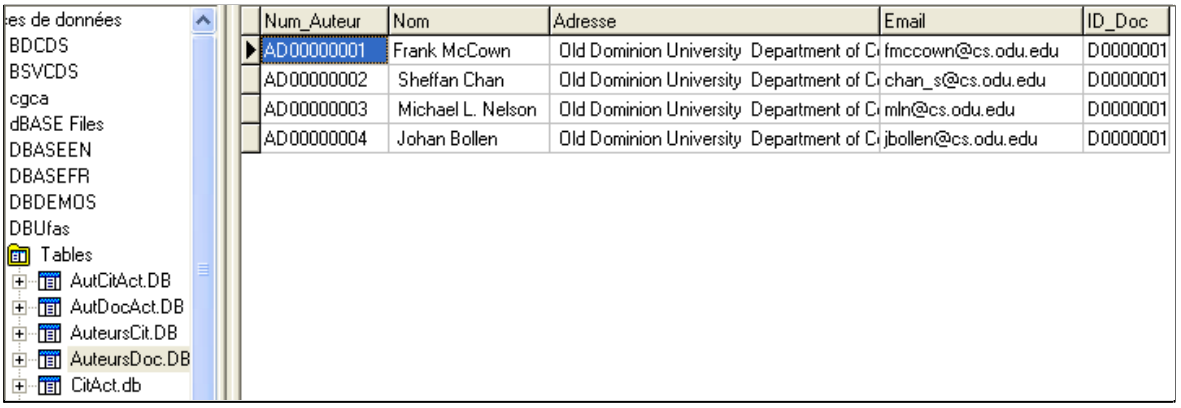
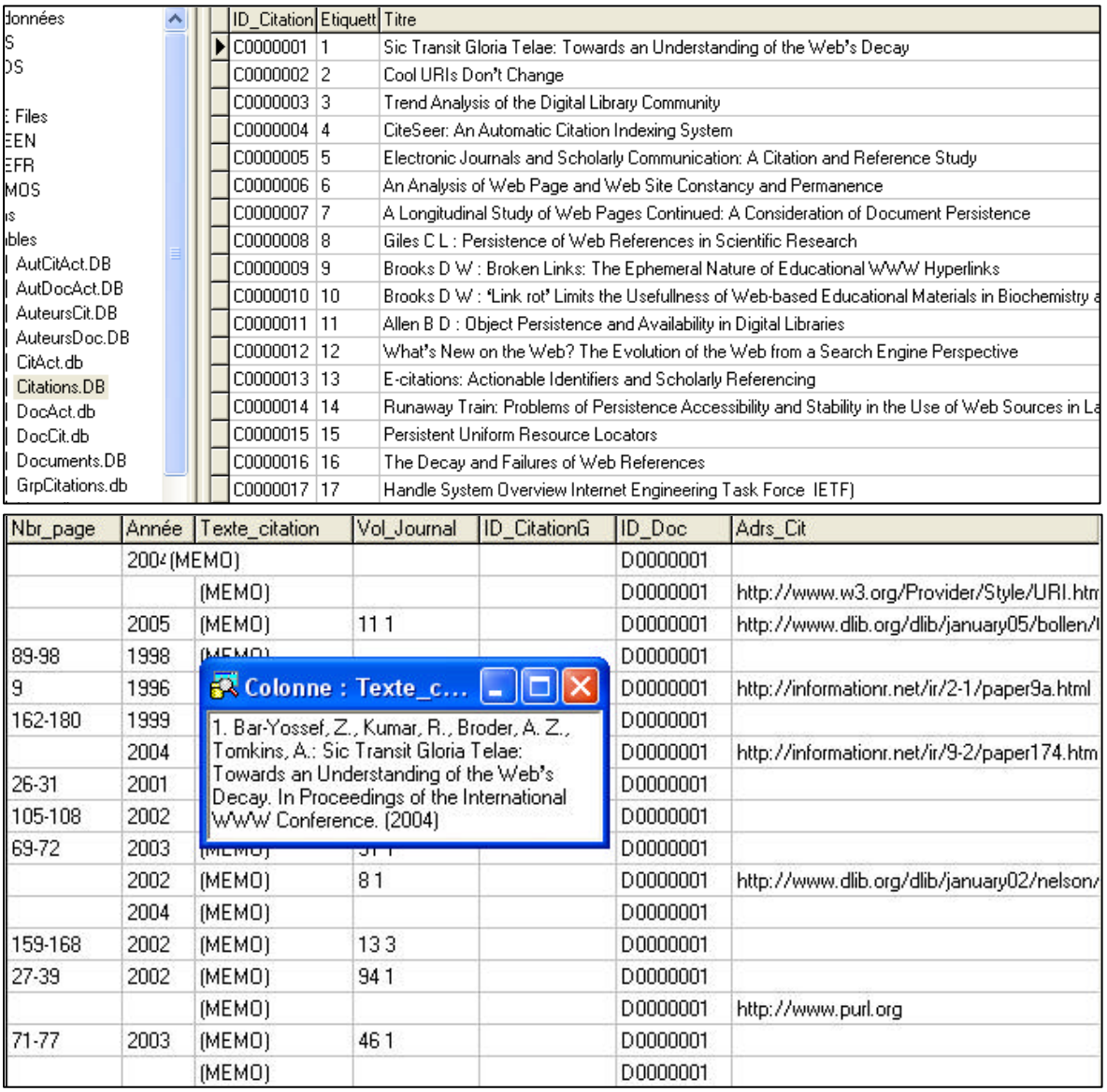


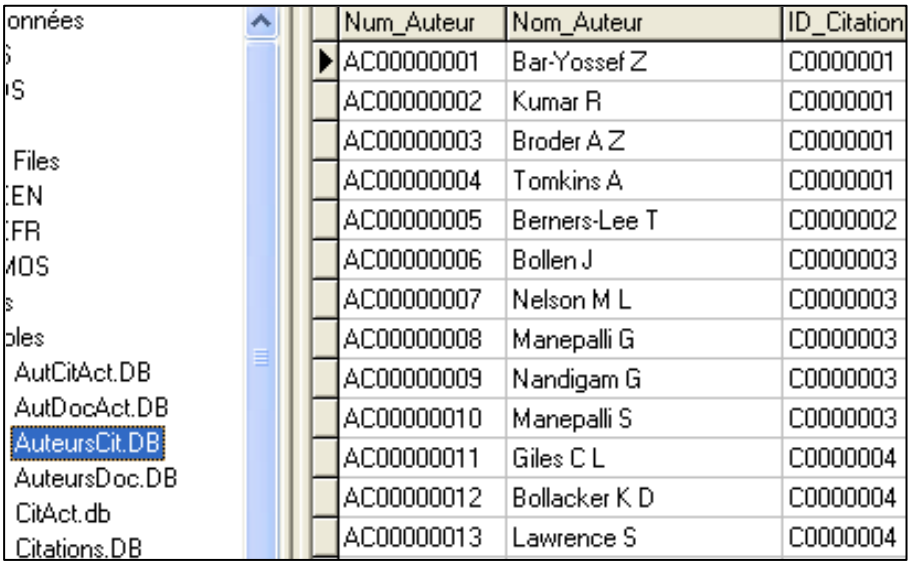
Figure 5.9: La table Auteurs documents



ID_Citation	Etiquett	Titre
C0000001	1	Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay
C0000002	2	Cool URIs Don't Change
C0000003	3	Trend Analysis of the Digital Library Community
C0000004	4	CiteSeer: An Automatic Citation Indexing System
C0000005	5	Electronic Journals and Scholarly Communication: A Citation and Reference Study
C0000006	6	An Analysis of Web Page and Web Site Constancy and Permanence
C0000007	7	A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence
C0000008	8	Giles C L : Persistence of Web References in Scientific Research
C0000009	9	Brooks D W : Broken Links: The Ephemeral Nature of Educational WWW Hyperlinks
C0000010	10	Brooks D W : 'Link rot' Limits the Usefullness of Web-based Educational Materials in Biochemistry a
C0000011	11	Allen B D : Object Persistence and Availability in Digital Libraries
C0000012	12	What's New on the Web? The Evolution of the Web from a Search Engine Perspective
C0000013	13	E-citations: Actionable Identifiers and Scholarly Referencing
C0000014	14	Runaway Train: Problems of Persistence Accessibility and Stability in the Use of Web Sources in La
C0000015	15	Persistent Uniform Resource Locators
C0000016	16	The Decay and Failures of Web References
C0000017	17	Handle System Overview Internet Engineering Task Force IETF)

Nbr_page	Année	Texte_citation	Vol_Journal	ID_CitationG	ID_Doc	Adrs_Cit
	200	(MEMO)			D0000001	
		(MEMO)			D0000001	http://www.w3.org/Provider/Style/URI.htm
	2005	(MEMO)	11 1		D0000001	http://www.dlib.org/dlib/january05/bollen/
89-98	1998				D0000001	
9	1996				D0000001	http://informationr.net/ir/2-1/paper9a.html
162-180	1999				D0000001	
	2004				D0000001	http://informationr.net/ir/9-2/paper174.htm
26-31	2001				D0000001	
105-108	2002				D0000001	
69-72	2003				D0000001	
	2002	(MEMO)	8 1		D0000001	http://www.dlib.org/dlib/january02/nelson/
	2004	(MEMO)			D0000001	
159-168	2002	(MEMO)	13 3		D0000001	
27-39	2002	(MEMO)	94 1		D0000001	
		(MEMO)			D0000001	http://www.purl.org
71-77	2003	(MEMO)	46 1		D0000001	
		(MEMO)			D0000001	

Figure 5.11: La table Citations



Num_Auteur	Nom_Auteur	ID_Citation
AC00000001	Bar-Yossef Z	C0000001
AC00000002	Kumar R	C0000001
AC00000003	Broder A Z	C0000001
AC00000004	Tomkins A	C0000001
AC00000005	Berners-Lee T	C0000002
AC00000006	Bollen J	C0000003
AC00000007	Nelson M L	C0000003
AC00000008	Manepalli G	C0000003
AC00000009	Nandigam G	C0000003
AC00000010	Manepalli S	C0000003
AC00000011	Giles C L	C0000004
AC00000012	Bollacker K D	C0000004
AC00000013	Lawrence S	C0000004

Figure 5.12: La table Auteurs Citations

Les mots utilisés comme mots clés dans Ufa*seer* (la recherche à la première fois), après la normalisation de document (enlèvement de traits d'union, les points, point virgule, les mots de liaison, mots d'arrêt …..), touts les mots dans le document sont extraits et après la lemmatisation (que les racines sont enregistrées en utilisant l'algorithme de Porter) la fréquence et le poids de chaque mots sont calculés a l'aide de l'algorithme de TFIDF ((Term Frequency ***X*** Inverse Document Frequency[Salt 73]) décrit précédemment.

$$Wds \;?\; \frac{(0.5 \;?\; 0.5\;\frac{fds}{fd\max})\;(\log\frac{ND}{ns})}{\sqrt{?_{\,j?d}\,(\;(0.5 \;?\; 0.5\frac{fdj}{fd\max})^{2}\;(\log\frac{ND}{nj})^{2}\;)}}$$

fds Fréquence d'une racine de mot **s**
fdsmax La fréquence la plus élevée d'un mot
ND Nombre de tous les documents
ns Nombre des documents ayant la même racine s
nj indice des documents partageant la même racine
Wds Poids d'une racine de mot s

La figure ci-dessous présente une partie de la table **Mots** et **Mots_Documents**, après extraction et lemmatisation des mots du résumé et de l'introduction du document présenté ci-dessus.

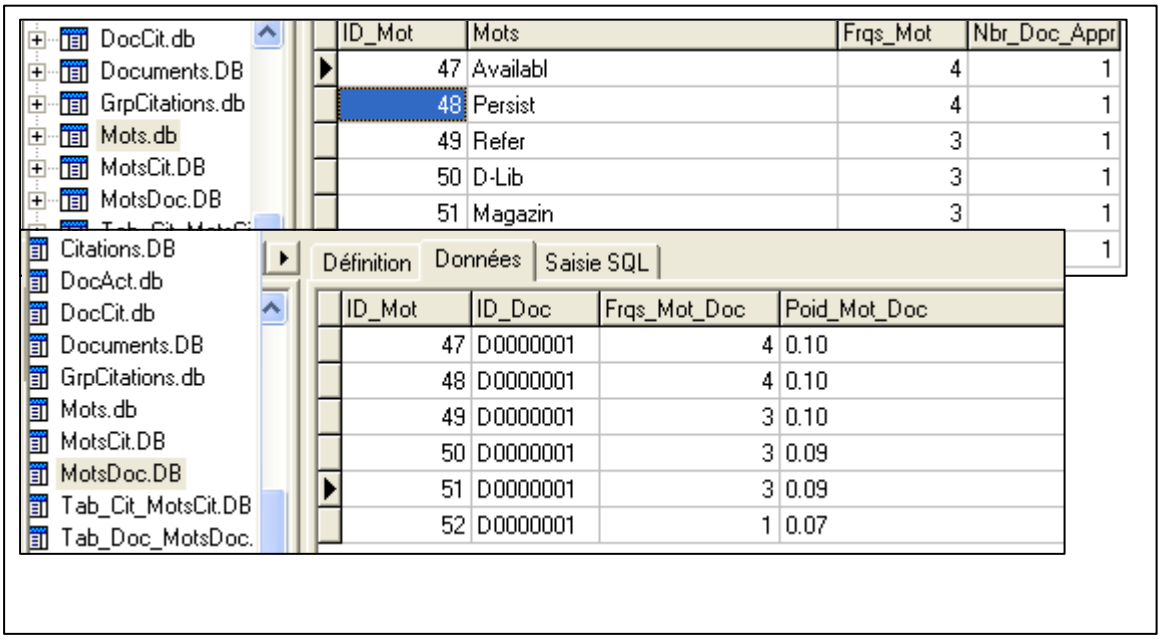


Figure 5.13: Les tables Mots et Mots_Documents

5.5.3 L'interface Web UFA*Seer*

La bibliothèque Ufa*seer* a pour objectif de mettre à la disposition de l'utilisateur une interface Web simple à utiliser. Celle-ci donne la possibilité d'interroger la base de données et lui fournit plusieurs possibilités de recherche, recherche par mots clés sur les documents, les citations et les auteurs.

L'utilisateur lance la première fois la recherche par mots clés, une fois qu'il obtient des résultats, il peut commencer le parcours (documents- citations), selon la disponibilité des documents dans la base de données.

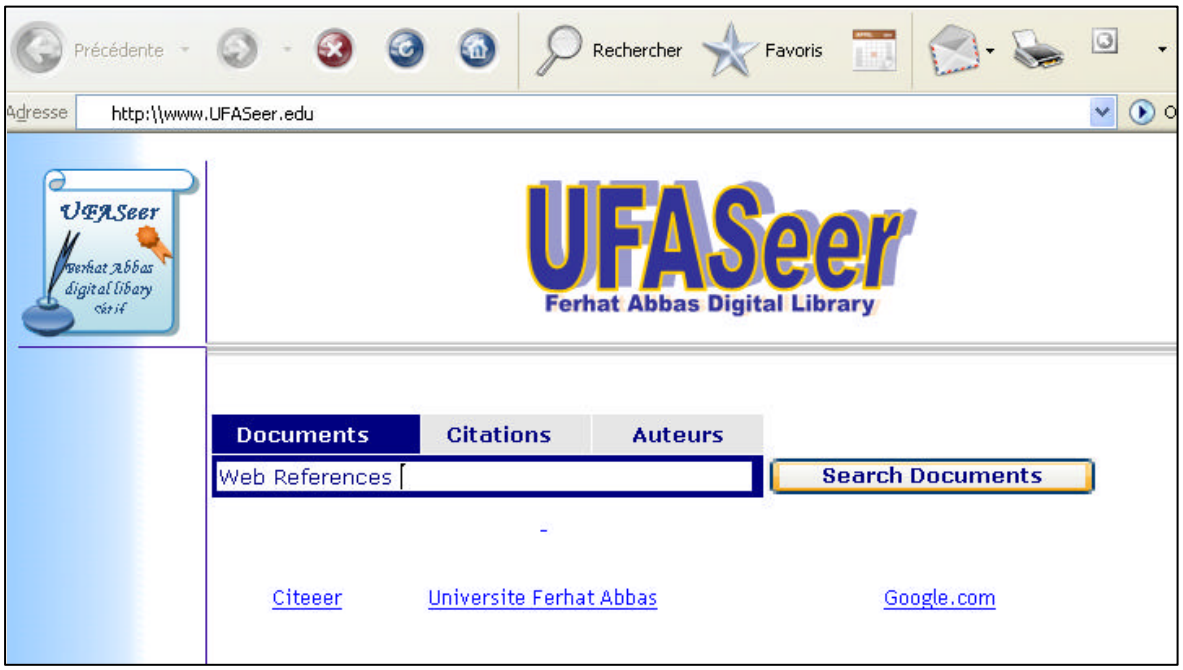


Figure 5.14: L'Interface Web le la bibliothèque UFA*Seer* .

Le site UFA*seer* permet j l'utilisateur de télécharger les documents trouvés, voir leurs détails (résumé, introduction, Auteurs…….), consulter les références bibliographiques et les citations cités par ces documents.

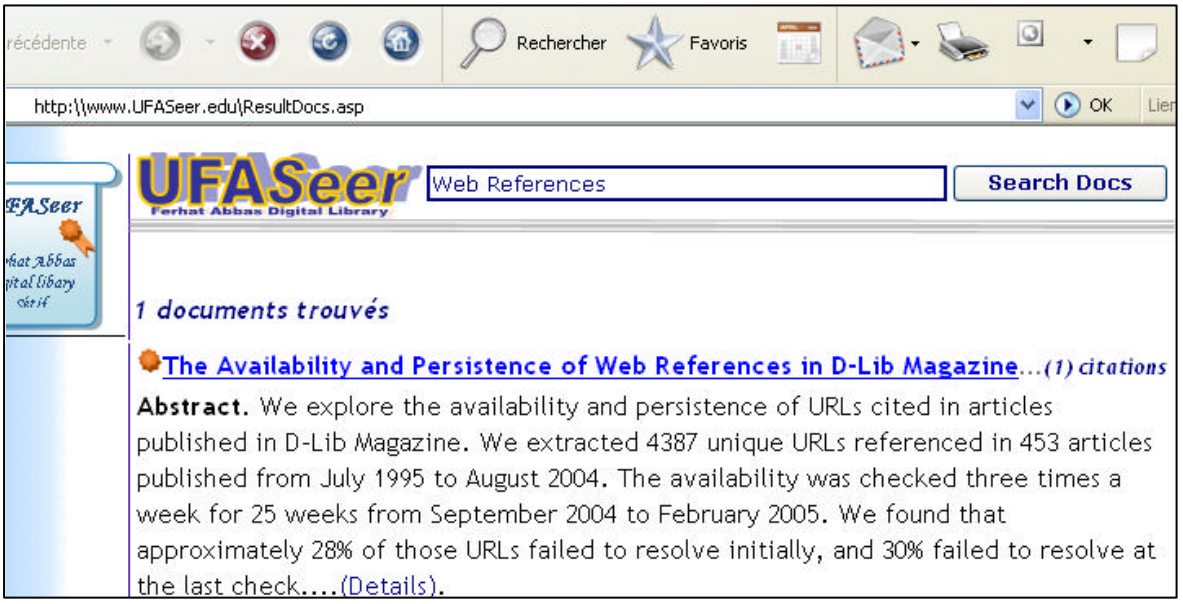


Figure 5.15: Résultat après la recherche par mots clés sur les documents UFA*Seer* .

La figure ci-dessous montre les détails concernant le document *« The Availability and Persistence of Web References in D-Lib Magazine »*.
Le lien hypertexte des références et les citations disponibles dans la base de données et activé. Par contre les références et les citations non disponibles représentent des liens désactivés.

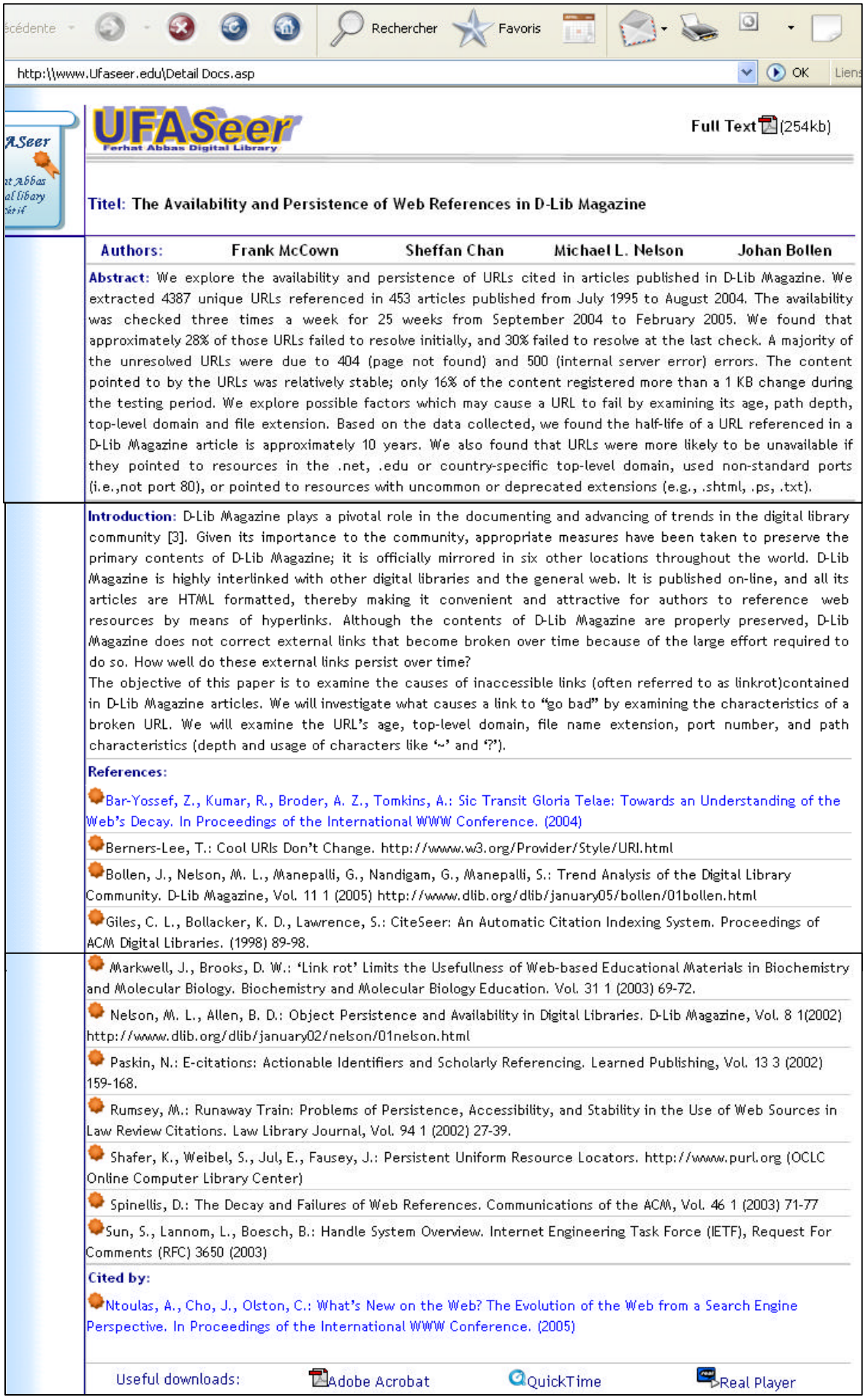


Figure 5.16: Détail document affiché par Ufa*seer*.

Les figures ci-dessous montrent un exemple d'une recherche par citation.

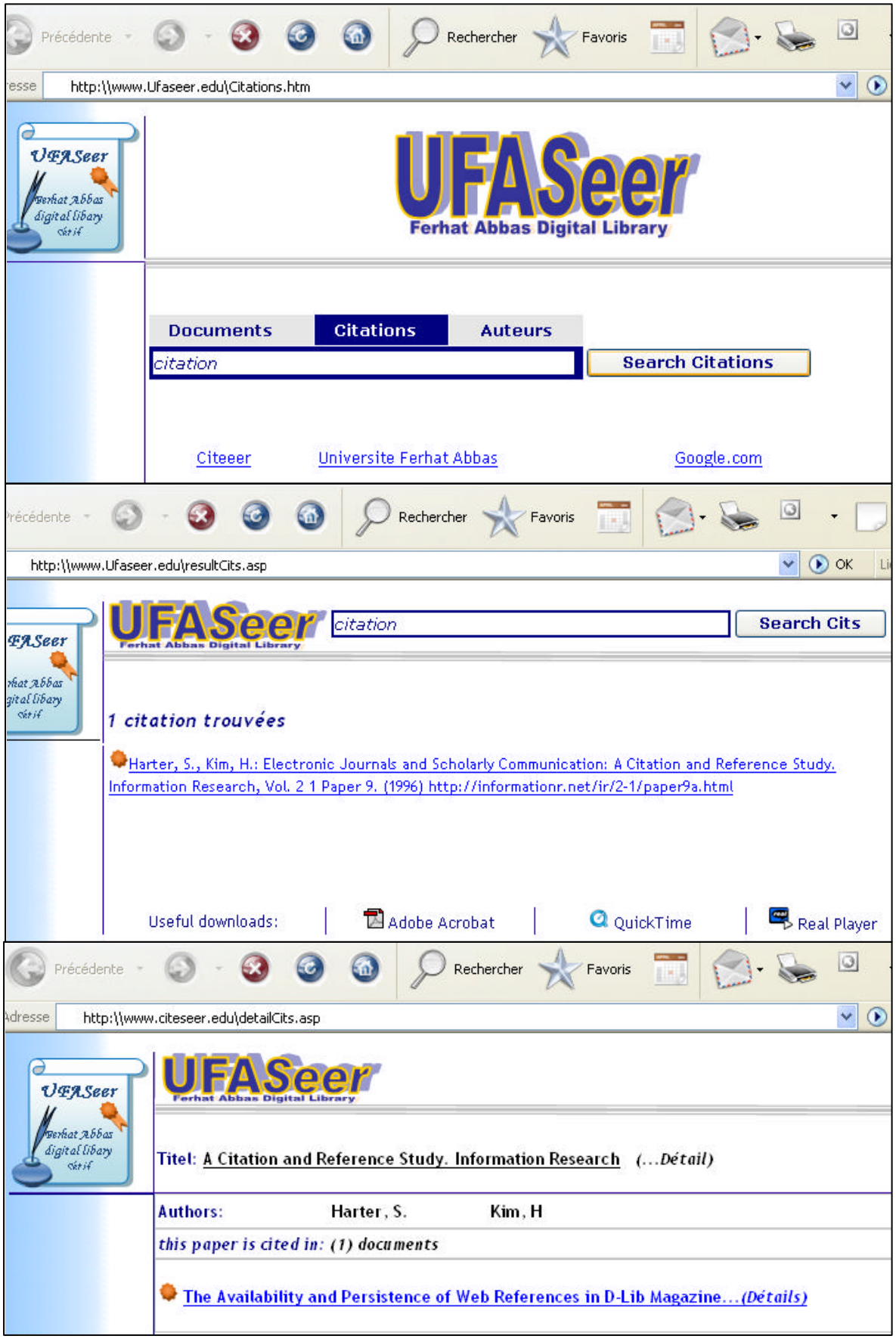


Figure 5.17: Résultat de recherche par citations.

5.6 Conclusion

Dans ce Chapitre nous avons décrit la réalisation d'un prototype de bibliothèque qui utilise l'indexation des citations. Cette bibliothèque se base sur le modèle de *Template mining* pour construire la base de données. Ce modèle possède quelques atouts pour créer automatiquement les bases de données des citations. Cependant *Template mining* a rencontré plusieurs obstacles lors l'extraction des informations, en particulier les irrégularités qui existent entre les citations et les modèles contradictoires des journaux. Les modèles de citation sont suivis strictement par les journaux imprimés (standards/house), et en fait, si notre base de données contient les modèles des grands journaux et des plus publiés sur le web, on pourrait alors traiter la plupart des articles disponibles sur le Web. L'amélioration des résultats et aussi lié aux auteurs qui sont requis de suivre les modèles des journaux ou ils ont publié leurs articles.

CONCLUSION

Nous avons présenté dans ce travail un système d'indexation automatique de citation. Le processus d'indexation est complètement autonome. Il localise, analyse, et classe les articles trouvés sur le Web.

Les avantages d'un tel système:

- L'opportunité, l'automatisation, et la lecture rapide du contexte de citation.
- La possibilité d'améliorer l'effort scientifique.
- Traitement automatique (i.e. n'exige pas d'effort humain dans le processus d'indexation).
- Capacité d'analyser des citations des articles et d'identifier les citations des papiers identiques qui peuvent différer en syntaxe. Ceci permet la génération des statistiques concernant les citations et de grouper les travaux qui citent un article donné.
- Capacité d'extraire et de montrer le contexte des citations par rapport a un article donné.
- Capacité de trouver les articles pertinents basés sur des citations communes.

En raison de l'organisation du Web, un système d'indexation automatique de citation ne peut pas actuellement fournir un index aussi complet que les systèmes traditionnels parce que beaucoup de publications ne sont pas disponible actuellement en ligne. Mais cet inconvénient disparaîtra avec le temps parce que de plus en plus de documentation est mise en ligne. Dans cette éventualité, l'indexation automatique dépasserait alors de très loin l'indexation manuelle traditionnelle.

Perspectives

Il y a beaucoup directions dans lesquelles il serait utile de développer le présent travail:

- Utilisation de la notification par Email au cas où un nouveau document pertinent serait téléchargé. Si un nouveau document est assez semblable à un document utilisateur, un email est envoyé a l'utilisateur.
- Les mesures de distance sémantiques peuvent aider dans la recommandation de nouveaux documents intéressants.
- Amélioration de la présentation des statistiques de la base de données. Par exemple, le nombre de fois un papier, un auteur, ou un journal est cité peut donner une certaine indication de son influence dans la communauté scientifique.
- Classement des documents sur la base du nombre de citations. On peut alors proposer un classement basé sur les auteurs, les journaux, etc.
- Mise à jour régulière des statistiques. Car ces statistiques changent chaque instant, ceci peut être un indicateur pour la direction des recherches.
- Amélioration du système en employant les bases de données Bib-TeX sur le Web. L'information de BibTeX est beaucoup plus précise que celle analysée à partir d'un dossier Postscript.

Ces contributions seraient les bienvenues et contribueraient à un système beaucoup plus performant.

Références bibliographiques

[Alia 06] Aliaksandr B., Enrico B., Paolo G., *"A multi-agent system that facilitates scientific publications search"*. May 2006 AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. Publisher: ACM.

[Boll 99] Bollacker K., Lawrence S., Giles C., *" Indexing and Retrieval of Scientific Literature"*. Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2–6, pp. 139–146, 1999.

[Bonh 97] BONHOMME S., ROISIN C. , "*Transformations de documents électroniques*", In Documents numériques, vol.1, n°4.

[Boui 01] BOUILLON P., VIEGAS E., « *Lexiques sémantiques* », Traitement automatique des langues, volume 42 n°3, Hermès, Paris, 2001.

[Brad 69] Bradford S. C., "*Bradford's Law and the bibliography of science*," Nature, 224(5223), 953-956, 1969.

[Buck 97] BUCKLAND M. K.**,**"*What is a document* ", In Journal of the American Society for Information Science, 48 (9): 804-809, 1997.

[Came 97] CAMERON R. D., *"A universal citation database as a catalyst for reform in scholarly communication"*, First Monday, February 1997., http://www.firstmonday.dk/issues/issue2_4/ca.

[Caro 95] CARO S., "*Rôle des organisateurs para-linguistiques dans la consultation des documents électroniques"*, Thèse de doctorat en Sciences de l'Information et de la Communication, Université Stendhal Grenoble 3, 1995.

[Char 97] CHARTRON G., CASSEYRE P., MARANDIN C., "*L'accès à la presse scientifique médicale : évolutions en cours*". In Journées d'études de la Société Française de Bibliométrie Appliquée, 12-16 mai 1997.

[Deer 90] DEERWESTER S., DUMAIS S., LANDAUER T., FURNAS G., HARSHMAN R., *"Indexing by latent semantic analysis"*, Journal of the American Society for Information science, 416(6), p. 391-407, 1990.

[Gard 72] GARDIN J. C., *"Document analysis and linguistic theroy"*, Journal of documentation, 29 n°2 juin 1972

[Garf 88] Garfield E., *"Announcing the SCI® Compact Disc Edition"*: CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery a reality. *Current Contents*® (22):3-13, 30 May 1988. (Reprinted in: *Essays of an information scientist: science*

literacy, policy, evaluation, and other essays. Philadelphia: ISI Press®, Vol. 11. p 160-70, 1990.

[Gils 98] Giles C., Lawrence S., " *Digital Libraries 98*: *Third ACM Conf. On Digital Libraries"*, ACM Press New York, 1998.

[Hui 03] Hui H., Lee Giles C., Manavoglu E., Zha H., Zhenyue Z., Edward., "*Automatic document metadata extraction using support vector machines"*. A. Fox May 2003 **JCDL '03:** Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. Publisher**:** IEEE Computer Society

[Isaa 05] Isaac G. Councill, C. Lee Giles, Hui Han., *"Automatic acknowledgement indexing: expanding the semantics of contribution in the CiteSeer digital library"*, Eren ManavogluOctober 2005 **K-CAP '05:** Proceedings of the 3rd international conference on Knowledge capture. Publisher: ACM

[Isaa 06] Isaac G. C.l, Huajing L., Ziming Z., Sandip D., Levent B., *"Learning metadata from the evidence in an on-line citation matching scheme"* Wang Chien Lee, Anand Sivasubramaniam, C. Lee Giles June 2006 **JCDL '06:** Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries

[Kurt 99] Kurt B., Lawrence S., Giles C., *"A system for automatic personalized tracking of scientific literature on the Web"*, Proceedings of the fourth ACM conference on Digital libraries Publisher: ACM, 1999.

[Law 98] Lawrence S., Bollacker K., L. Giles C., *"CiteSeer: An autonomous Web agent for automatic retrieval and identification of interesting publications*" in Proceedings of the Second International Conference on Autonomous Agents (K. P. Sycara and M. Wooldridge, eds.), New York, pp. 116–123, ACM Press, 1998.

[Lawr 98] Lawrence S. and Giles C., *"Searching the World Wide Web"*. Science, 280(5360):98–100, 1998.

[Lawr 99] Lawrence S. and Giles C., *" Accessibility of information on the web"*. Nature, 400(6740):107–109, 1999.

[Laws96] Lawson M., Kemp N., Lynch M.F., and Chowdhury G.G., *"Automatic extraction of citations from the text of English language patents: An example of template mining"*. Journal of Information Science, 22, 6,pp. 423-436, 1996.

[Leve 65] LEVENSHTEIN V., *"Binary codes capable of correcting spurious insertions and deletions of ones"*, Russian Problemy Peredachi Informatsii,12–25, January 1965,

[Maro 60] MARON M., KUHNS J., "*On relevance, probabilistic indexing and information retrieval*", Journal of the ACM, n°3, p. 216-244, July 1960.

[Melv 83] Melvin W., " *Citation Indexes "*. Institute for scientific Information ,Vol:6, p.548-572, 1983.

[Pier 00] PIERREL J. M., " *Ingénierie des langues* ", Hermès, Paris, 2000.

[Port 80] Porter M. F., *"An algorithm for suffix stripping*", 14:130–137. 3, 1980.

[Port 80] SALTON G., *"Term weighting approaches in automatic text retrieval*". Tech Report 87-881, Dept of Computer Science, Cornell University, 1997.

[Pric 86] Price D.J, *"Little science, big science...and beyond"*, New York: Columbia University Press, 301 p, 1986.

[Reve 98] REVELLI Carlo (1998). Intelligence stratégique sur Internet : Comment développer efficacement des activités de veille et de recherche sur les réseaux : Moteurs de recherche, réseaux d'experts, agents intelligents. Paris : Dunod, 1998.

[Salt 97] SALTON G., *"Term weighting approaches in automatic text retrieval*". Tech Report 87-881, Dept of Computer Science, Cornell University, 1997.

[Salt 71] SALTON G., *"Automatic indexing using bibliographic citations*", Journal of Documentation 27, 98–110, 1971.

[Salt 72] SALTON G., "*Experiments in automatic thesaurus construction for information retrieval*", Congrès IFIP, Ljubljana., 1972.

[Salt 73] SALTON G., YANG C., *"On the specification of term values in automatic indexing"*, Journal of Documentation 29, 351–372, April 1973.

[Salt 75] SALTON G., WONG A., YANG C. S., *"A vector space model for automatic indexing"*, Communication of the ACM, 18, p.613-620, 1975.

[Salt 89] SALTON G., *"Automatic text processing, the transformation, analysis, and retrieval of information by computer"*, Addison-Wesley, Reading, 1989.

[Tlet 63] Tletze K., *"On the Problems of Seasonal Fluctuations in the Dates of Birth and Conception*," Therapie der Gegenwart, 102, 955-962, 1963.

[Vink 94] Vinkler P., *"The origin and features of information referenced in pharmaceutical patents*". Scientometrics, 30, 1,pp. 283-302,1994.

[Wout 99] Wouters, P. (1999) the citation culture. Thesis, faculty of science, University of Amestrdam.

[YIAN 93] YIANILOS P., *"Data structures and algorithms for nearest neighbor search in general ametric spaces*". In Proceedings of the 4th ACM-SIAM Symposium on Discrete Algorithms, pp. 311–321, 1993.

[Yian 97] YIANILOS P., " *The LikeIt intelligent string comparison facility*", NEC Institute Tech Report, 97-093, 1997.