**République Algérienne Démocratique et Populaire**

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**Université 08 Mai 45 Guelma**
**Faculté des sciences économiques, commerciales et des sciences de gestion**


**Département des sciences de gestion**

**Présenté pour l'obtention du diplôme de**
**Master en TIC**

Titre du Mémoire

---

# La recherche d'information dans l'entreprise

---

Présenté par :
**Djeddai bilel**

sous la direction de :
**M-Djelailia Karim**

**Promotion 2011**

# Remerciement

En préambule à ce mémoire, je souhaitais adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire

Je tiens à remercier sincèrement Monsieur Djelailia Karim, qui, en tant que Directeur de mémoire, s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu me consacrer.

Je remercie chaleureusement mes enseignants Fayçal Nouwar et Kelaiaia Abdessalem

Mes remerciements s'adressent également à tous les enseignants de département de science de gestion

J'exprime ma gratitude à tous les consultants et internautes rencontrés lors des recherches effectuées et qui ont accepté de répondre à mes questions avec gentillesse.

Je remercie encore une fois  mon prof Djelailia Karim pour son encouragement et ses grands efforts

Je n'oublie pas mes parents pour leur contribution, leur soutien et leur patience

Je remercie aussi mon amie et mon frère  azaizia soufiane  pour son aide

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.

Merci à tous et à toutes.

# Dédicace

Je dédie ce  modeste travail Aux deux grands héros de ma vie : Ma mère et mon  père sans lesquels je ne serai pas sur terre

A mes frères et ma sœur

# Sommaire

## Résumé :

Notre travail se situe dans le contexte de la recherche d'information(RI), plus particulièrement la recherche d'information dans l'entreprise,
L'objectif principal d'un SRI classique est de retrouver les documents dont le contenu est conforme à une requête donnée. Dans cette optique, les documents sont représentés par un ensemble de mots-clés décrivant leurs contenus. La structure du document n'est pas prise en considération ni au niveau de la requête, ni au niveau de la réponse pour retourner les parties pertinentes : la réponse à une requête reste le document tout entier. Aujourd'hui, l'utilisation de l'information apportée par la structure devient une nécessité dans le domaine d'accès à l'information
L'objectif de notre travail est d'exposer les techniques et méthodes de recherche d'information dans l'entreprise détenant un corpus important de documents en son sein. Nous exposerons aussi les différentes parties d'un système de recherche d'information (SRI)
Nous achèverons notre étude par une enquête sur l'utilisation des systèmes de recherche d'information dans les entreprises algériennes



**Mots clés** : recherche d'information, pertinence, corpus, système de recherche d'information, indexation

# Abstract

Our work is in the context of information retrieval (IR), specifically looking for information throughout the company
The main objective of a classic IRS to find documents whose content conforms to a given query. In this context, documents are represented by a set of keywords describing their contents. The document structure is not taken into consideration either at the request or in response to return the relevant parts: the response to a query is the entire document. Today, the use of information provided by the structure becomes a necessity in the field of access to information
The objective of our work is to present techniques and methods of information retrieval in the company holding a large corpus of documents in it. We also expose the different parts of an information retrieval system (IRS)
We will complete our study by investigating the use of information retrieval systems in the Algerian companies



**Keywords**: information retrieval, relevance, corpus, information retrieval system, indexing

# ملخص:

عملنا هو في سياق البحث عن المعلومات . على وجه التحديد البحث عن المعلومات في المؤسسة
ويتمثل الهدف الرئيسي لنظام البحث عن المعلومات في العثور على الملفات التي تتوافق مع محتوى استعلام معين. وفي هذا السياق تمثل الملفات من خلال مجموعة من الكلمات الرئيسية التي تصف محتوياتها, لا يؤخذ هيكل الملف في الاعتبار إما بناء على طلب أو استجابة لإرجاع الأجزاء ذات الصلة: وردا على استفسار هو المستند بأكمله. اليوم واستخدام المعلومات التي يقدمها هيكل يصبح ضرورة في مجال الحصول على المعلومات
والهدف من عملنا هو تقديم أساليب وطرق البحث عن المعلومات في المؤسسة من خلال الحجم الكبير من الملفات المتوفر لديها. و أيضا كشف مختلف أجزاء نظام استرجاع المعلومات
وسوف نستكمل دراستنا عن طريق التحقيق في استخدام نظم استرجاع المعلومات في الشركات الجزائرية



**الكلمات الدالة**: البحث عن المعلومات، نظام البحث عن المعلومات ، والفهرسة ، استعلام

# Table des figures



# Liste des tableaux

# Abréviation et acronymes

| | |
|---|---|
| ADBS | Association française des documentalistes et bibliothécaires spécialisés |
| CACM | Central American Common Market |
| RD | Recherche documentaire |
| RI | Recherche d'information |
| IA | Intelligence artificielle |
| SRI | Système de recherche d'information |
| BD | Base de donne |
| SQL | Structured Query Language |
| SGBD | Système gestion base de donne |
| QR | Question repense |
| TF | Term Frequency |
| Idf | Inverse Document Frequency |
| RSV | Retrieval Status Value |
| DARPA/ITO | Defense Advanced Research Projects Agency / Information Technology Office |
| NIST | National Institute of Standards and Technology |
| TREC | Text REtrieval Conference |
| HTML | Hypert Text Markup Language |
| XML | Extensible Markup Langage |
| URI | Uniform Resource Identifiers |
| OWL | Ontology Web Language |
| GED | gestion électronique de document |

# Introduction générale :

Avec la reconstruction de l'après-guerre, dans les pays industrialisés, un métier nouveau naît, à quelques pas de ces chercheurs, tout proche des bibliothèques, voisin des laboratoires de recherche : le métier de *documentaliste*. On l'a peut-être oublié, tant le mot s'est banalisé dans notre esprit, le mot même de documentaliste est la contraction de deux autres : document et spécialiste. Le (ou la) documentaliste serait donc le spécialiste du document. L'étymologie, même allégorique, offre parfois de singuliers raccourcis...

Mais la profession allait presque exclusivement se développer autour du traitement et de la recherche de documents et d'informations glanées, traqués, captés *à l'extérieur* de l'entreprise ou de l'organisation au sein de laquelle elle œuvrait. Les centres de documentation, puis les services de veille sont devenus dans les cas les plus achevés, le regard intelligent de l'entreprise sur son environnement.

Vingt ans après... la profession, en mal de reconnaissance, cherche à échapper à ses origines documentaires pour accéder au statut, jugé plus valorisant, de *spécialiste de l'information*. Un pays francophone saute même le pas en adoptant délibérément un nouveau nom pour ce métier : *informatiste* (le Maroc). A l'ADBS (alors *Association française des documentalistes et bibliothécaires spécialisés*) on chuchote que les informaticiens "nous ont pris" le seul terme qui corresponde à notre activité. La réflexion aboutit positivement au changement de développement du nom de l'association, qui garde son sigle, mais s'intitule désormais (depuis 1993) : *Association des professionnels de l'information et de la documentation*. Compromis encore inavoué entre deux tendances : doit-on arborer bien haut la bannière de l'information mise à l'honneur par les penseurs du management, ou doit-on camper sur nos bons vieux documents, même s'ils peuvent paraître poussiéreux, coûteux et improductifs ?

A la même époque, les organisations, enivrées par leurs flux d'information, commençaient à chercher à maîtriser un nouveau problème. Le *"zéro-papier"* promis par les thuriféraires de l'informatique n'avait converti ni les bureaux, ni les usines. Il semblait même que l'inflation informationnelle, due en partie, au morcellement grandissant des domaines scientifiques et techniques et associée à la complexité croissante de l'organisation des entreprises, ait provoqué une explosion du papier et en tout cas une démultiplication des documents dans l'entreprise. Se posait alors la question cruciale de la gestion des documents, de leurs versions successives, de leur maintenance, de leur mise à jour, de leur stockage, de leur sauvegarde... le tout associé bien sûr à des questions de coûts,

de rentabilité, de gains de productivité, de compétitivité, des avantages concurrentiels, pour tout dire.

Dès lors, aux yeux des organisations, les documents allaient retrouver un statut, des fonctions et des missions parées de vertus nouvelles. Les entreprises étaient mures pour rencontrer des professionnels aptes à résoudre ces problèmes.

Or, les spécialistes du document que sont les documentalistes n'étaient pas au rendez-vous dans les premiers temps, confinés qu'ils étaient dans leur activité de documentation externe à l'entreprise et convertis au culte de l'information, là où les managers ne juraient plus que par le concept de gestion des documents. Les documentalistes sont pourtant les plus à même aujourd'hui de résoudre toutes les questions de traitement de l'information et des documents, par les méthodes et techniques intellectuelles qu'ils maîtrisent. Il semble qu'ils intègrent progressivement cette nouvelle dimension, ce nouveau service documentaire à rendre aux entreprises

Le but visé par ce mémoire étant d'exposer les techniques et méthodes de recherche d'information dans l'entreprise détenant un corpus important de documents en son sein
Nous exposerons aussi les différentes parties d'un système de recherche d'information (SRI)
Nous achèverons notre étude par une enquête sur l'utilisation des systèmes de recherche d'information dans les entreprises algériennes

# Organisation de document

Le document est organisé en deux parties :

L'objectif de la première partie est de présenter l'état de l'art de la recherche de l'information traditionnelle. Nous présentons les approches dans la littérature concernant les fondements de base de la recherche d'information traditionnelle dans les documents texte (documents plats). Nous commençons tout d'abord par décrire le rôle et l'utilité de l'information dans l'entreprise. Nous continuons par décrire le processus de recherche d'information, ensuite nous présentons les différents modèles d'appariement document-requête. Nous présentons ensuite les mesures utilisées pour l'évaluation des systèmes de recherche d'information. Enfin nous présentons brièvement comment fournir une information « ciblée » en entreprise?

La seconde partie de ce mémoire, consiste en une enquête au niveau des entreprises Algérienne pour découvrir :
Dune part, le niveau d'introduction des systèmes de recherche d'information dans l'entreprise algérienne, D'autre part, la faisabilité, la stratégie et les outils pour construire un système de recherche d'information

Partie 01 : Etat de l'Art

# Introduction

Parce qu'il est aujourd'hui difficile de trouver l'information pertinente, les entreprises sont peu à peu amenées à reconsidérer le rôle stratégique de la recherche d'information. La recherche d'information en entreprise dresse un panorama des différents chantiers à entreprendre et des approches possibles pour les mener à bien :

- Une analyse, avec une mise en avant de la problématique de pertinence, les trois aspects fondamentaux de la recherche d'information que sont les besoins des collaborateurs, l'organisation de l'information et les technologies utilisées.
- Une présentation des dernières évolutions et tendances technologiques qui feront partie intégrante des solutions de recherche d'information de demain : classifications à facettes, recherche collaborative ou le *social bookmarking*

Dans ce chapitre nous présentons les enjeux de l'information en entreprise, les différents composants d'un système de recherche d'information (SRI) et les modèles de recherche d'information existants

# 1- Le rôle de l'information en entreprise :

La détermination précise de la nature de l'information et de la diversité de ses rôles constitue un préalable pour comprendre en quoi consiste le système de recherche d'information en entreprise. [Die Arnaud 2008 p 13]

## 1-1 La nature de l'information

Le terme information concerne deux réalités. D'un point de vue technique (informaticien), il désigne n'importe quel signe qui puisse être transmis et stocké. La plus petite quantité d'information qui puisse être transmise et stockée est un caractère ou un bit (binary digit), c'est à dire une information qui peut prendre deux valeurs alternatives (0 et 1)

En soi, cette information technique n'est pas utilisable, car n'ayant pas de contenu sémantique (pas de signification). Dans un second sens, l'information est un renseignement qui apporte une connaissance sur un objet ou sur un événement. Dès lors l'information devient significative. Elle va pouvoir faire l'objet de traitements, d'interprétations, permettre de prendre des décisions.

Dans les réseaux d'information, l'information, au sens de signe, peut circuler. Il est essentiel de définir les caractéristiques d'un réseau d'information (capacité de stockage, de transmission, délais d'acheminent…), d'évaluer la qualité de l'information (nombre de bits) à traiter indépendamment du contenu de l'information.

## 1-2. La diversité des informations

L'information n'a de valeur qu'en raison de l'usage qui en est fait. En gestion, l'information est considérée comme la matière première de la décision. Michel Chobron et Robert Reix (1987) ont distingué cinq usages possibles de l'information et des technologies de l'information [Die Arnaud 2008 p 14] :

- *L'information, support des processus de gestion*

Un processus de gestion (gestion d'approvisionnement, de traitement des commandes…) est un ensemble d'activités et de décisions combinées pour produire des résultats souhaités par l'entreprise. Chaque processus lui-même créateur d'information, doit disposer de ressources en informations pour être exécuté.

# Partie 01 : Etat de l'Art

## *- L'information, instrument de communication dans l'organisation*

Des échanges d'informations permettent d'assurer la coordination entre les activités des différents membres de l'organisation. La fonction de communication a acquis un caractère prédominant avec l'émergence des bases de données, de la bureautique (courriers électroniques) et de la télématique (réseaux publics, réseaux d'entreprises). L'information en tant qu'instrument de communication occupe une place importante dans la politique commerciale de l'entreprise (volet communication).

## *- L'information, instrument de liaison avec l'environnement*

Les différentes technologies de l'information sont aussi susceptibles d'utilisations plus directes avec l'environnement de l'entreprise : l'information peut être incorporée au produit (prix, caractéristiques, mode d'emploi…) et devient lisible par un ordinateur (exemple de l'achat d'un logiciel et de la fonction d'aide) , des systèmes d'information inter-entreprises peuvent être mis en œuvre, certains fournisseurs installant des terminaux chez leurs clients destinés à faciliter la prise de commandes (exemple l'industrie automobile).

## *- L'information, support de la connaissance individuelle*

La capacité cognitive de l'organisation est d'abord celle des individus qui la composent. Dans ce domaine de la connaissance individuelle, les technologies informatiques (système de mémorisation, d'aide à la décision, systèmes experts…) fournissent un appui de plus en plus important.

## *- L'information, facteur important de cohésion sociale*

On insistera également sur le fait que l'information est un facteur important de cohésion sociale et de motivation du personnel. Un bon climat social (absence de revendications, de grèves…) repose sur un système d'information efficace. Une information qui remonte et descend la ligne hiérarchique, qui et associée à un processus de décision (délégation, décentralisation…) génère de l'initiative et de la motivation parmi les salariés d'une entreprise. Cette tâche incombera à la fonction des Ressources Humaines de l'entreprise

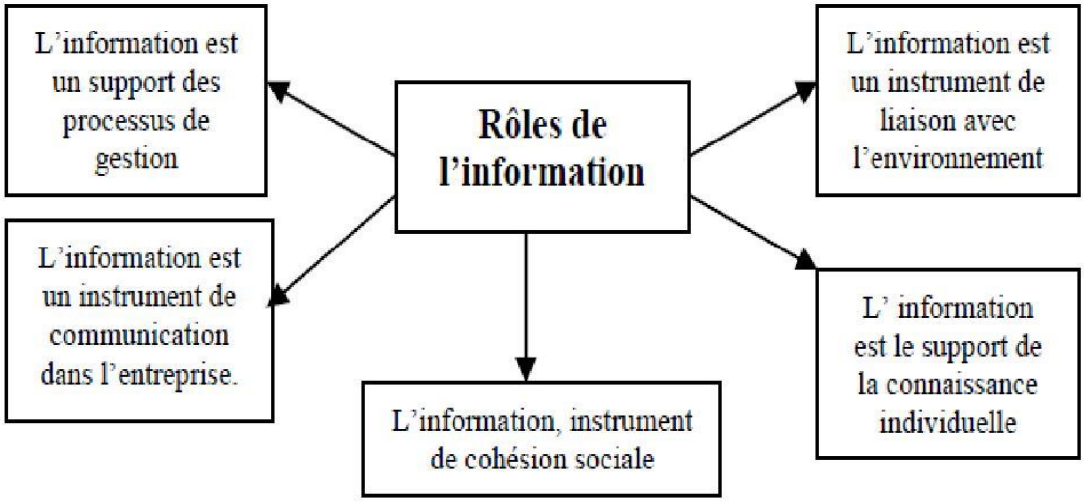



***Figure 1-1 : Les qualités de l'information***

Cours de Mr Diemer Arnaud : Définition et analyse des entreprises partie1 p15

# 2- La recherche d'information et la recherche documentaire

## 2-1 la recherche d'information

la recherche d'information : « Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés ». [Cél Paganelli 2002 p 21]

La Recherche de **l'**information : « Ensemble des méthodes, procédures et techniques ayant pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes ». [Cél Paganelli 2002 p 21]

## 2-2. la recherche documentaire :

*Qu'est-ce-que la recherche documentaire ? :*

La recherche documentaire correspond à l'ensemble d'actions, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents [AFNOR, , p. 99].

En d'autres termes, effectuer une recherche documentaire équivaut à identifier et à accéder à des ressources informationnelles qui ont déjà été traitées et éditées [Bibeau, 1998]

En recherche documentaire (RD), une requête est lancée en vue d'obtenir les documents les plus pertinents (pas comme la recherche d'information où il est question de chercher des informations pertinentes dans un ensemble des documents disponible), ce qui revient à catégoriser les textes en textes pertinents et non pertinents pour une classe donnée.

*En RD on procède comme suit :* [Lewis 1992b] :

A) *indexer les documents* : c'est la représentation de textes en vue de leur exploitation,

B) Formuler les requêtes par soit :

- un descripteur[1] de thème : ex. « astrophysique »,
- une requête construite à l'aide des mots du langage courant en utilisant des Opérateurs logique, de proximité, de troncature :
  ex. « (Astronomie *et* trous noirs* ou (corps *près* sombre*) »,
- une expression en langage naturel : ex. « tension artérielle très élevée »,
- un document entier, utilisé comme exemple du sujet sur lequel on veut obtenir d'autres informations

[1] Un Mot ou locution destiné a caractérisé les informations contenues dans un document pour faciliter les recherches documentaires

- un graphe de concepts : les concepts, représentés par des termes, peuvent être liés par des relations sémantiques de natures diverse (réseaux sémantiques)

c) comparaison entre la requête et les documents utilisant une fonction de similarité,
d) *feedback* : L'utilisateur reformule sa requête dans le cas ou les documents ne satisfont pas ses besoins

## 2-3.brefs historiques de la recherche d'information :

La RI n'est pas un domaine récent. Il date des années 1940, dès la naissance des ordinateurs. Au début, la RI se concentrait sur les applications dans des bibliothèques, d'où aussi le nom « automatisation de bibliothèques ». Depuis le début de ces études, la notion de pertinence a toujours été son objet.
Dans les années 1950, on commençait de petites expérimentations en utilisant des petites collections de documents (références bibliographiques). Le modèle utilisé est le modèle booléen. Dans les années 1960 et 1970, des expérimentations plus larges ont été menées, et on a développé une méthodologie d'évaluation du système de recherche d'information qui est aussi utilisé maintenant dans d'autres domaines. Des corpus de test (ex. CACM (Central American Common Market)) ont été conçus pour évaluer des systèmes différents. Ces corpus[1] de test ont beaucoup contribué à l'avancement de la RI, car on pouvait les utiliser pour comparer différentes techniques, et de mesurer leurs impacts en pratique. Le système qui a le plus influencé le domaine est sans aucun doute SMART, développé à la fin des années 1960 et au début des années 1970. Les travaux sur ce système a été dirigés par G. Salton, professeur à Cornell. Certains nouvelles techniques ont été implantées et expérimentées pour la première fois dans ce système (par exemple, le modèle vectoriel et la technique de relevance feedback). Du côté de modèle, il y a aussi beaucoup de développements sur le modèle probabiliste.
Les années 1980 ont été influencées par le développement de l'intelligence artificielle. Ainsi, on tentait d'intégrer des techniques de l'IA en RI, par exemple, système expert pour la RI, etc.

[1] Ensemble de documents utilisés pour une étude, spécialement pour une étude linguistique

Les années 1990 (surtout à partir de 1995) sont des années de l'Internet. Cela a pour effet de propulser la RI en avant scène de beaucoup d'applications. C'est probablement grâce à cela que vous entendez parler de la RI. La venue de l'Internet a aussi modifié la RI. La problématique est élargie. Par exemple, on traite maintenant plus souvent des documents multimédia qu'avant. Cependant les techniques de base utilisées dans les moteurs de recherche sur le web restent identiques. [Cél Paganelli p 26]

# 3- Les concepts de base de RI

Un système de recherche d'information a pour rôle de sélectionner, à partir d'un besoin en information utilisateur, les documents qui peuvent l'intéresser, c'est-à-dire ceux qui peuvent être pertinents à son besoin en information. Cette définition fait apparaitre trois notions clés qu'il convient de préciser : documents, besoin en information, pertinence.

## 3-1 Du document à la base documentaire

Un document représente dans un SRI l'élément objectif [T. Saracevic p 49 .] dans le cadre de la RI traditionnelle, c'est du texte libre, qui peut être caractérisé selon trois vues :

_ La vue représentation, c'est la mise en forme d'un document texte (entêtes, paragraphes, alignement. . .)

_La vue logique qui présente la structure logique d'un document, elle porte des informations sur la structure

_La vue de contenu, qui se focalise sur le sens ou la sémantique d'un document le plus souvent par un ensemble de mots.

Dans les SRI traditionnels, la vue de contenu est l'unique intérêt, puisque les utilisateurs forment leurs requêtes en se mettant comme objectif le contenu textuel des documents **[**N. J. Belkin and H. M. Brooks 1997] c'est d'ailleurs la raison même de l'utilisation de tels systèmes.

## 3-1-1 Le prétraitement d'un document :

### 3-1-1-1 Segmentation du document

La tokénisation (ou segmentation) qui consiste à diviser les textes en unités lexicales (token) de plus en plus petites : paragraphes, phrases, mots. C'est une opération de type statistique qui « localise» les chaînes de caractères entourées de « séparateur » (caractère blanc), et les identifie comme étant des mots. - Cette opération est couplée à un traitement linguistique qui permet d'identifier les signes de ponctuation séparant les phrases et les paragraphes. Il permet aussi de procéder à une première correction des fautes d'orthographe25 et des erreurs de typographie.

### 3-1-1-2 Traitement morpholexical et morphosyntaxique :

Cette première opération que l'on nomme aussi « lemmatisation », est indispensable pour pouvoir « *retrouver tous les documents dans lesquels apparaissent différentes formes du même mot* » [IVANCIUC DENIAU, p.66]. Elle consiste à faire correspondre les formes des termes rencontrées dans les textes (féminin, masculin, adjectifs, verbes, adverbe, substantif) à leur « LEMME », c'est à dire la forme fixe et minimale (canonique) du mot.

Le second traitement appelé aussi « étiquetage » ou *tagging* consiste à comparer chaque mot du texte (susceptible d'être ambiguë), avec les termes du dictionnaire intégré (référentiel ou glossaire métier). Ceci, afin de leur attribuer une ou plusieurs étiquettes en fonction du sens qu'ils sont susceptibles d'avoir dans le contexte où ils sont utilisés. Cette opération permet aussi d' « identifier » les mots composés et les expressions toutes faites.

### 3-1-1-3 Le traitement sémantique général de nature lexicale

Il consiste à identifier les réseaux sémantiques qui unissent les concepts en présence dans le corpus indexé. Ces réseaux constituent un graphe (topic map) qui sert de référence pour l'indexation du fonds et peut ensuite servir de « système de guidage » pour l'utilisateur au moment de la requête. En associant les termes présents dans le corpus par « famille », cette opération permet de diminuer les problèmes de silence et de bruit, liés à la synonymie, à l'hyperonymie (meuble/siège), la métonymie (partie de) ou l'association

### 3-1-1-4 Traitements statistiques

A ce stade, les moteurs linguistiques réintroduisent des opérations de type statistiques (cf. p.60), qui permettent de pondérer les termes retenus précédemment.

### 3-1-1-5 Traitement de regroupement / classification

Ces deux opérations consistent à rapprocher les documents similaires en les classant dans des thématiques (catégories) en fonction de leurs degrés de pertinence par rapport à la question posée

## 3-2 Du besoin en information à la requête

Pour exprimer son besoin l'utilisateur compose une suite de mots-clés (requête), souvent séparés par des opérateurs logiques (et, ou et non), ou par des variables linguistiques telles que proche de, contient. . . On distingue trois types de requêtes

-Les requêtes basiques, la requête est un ensemble de mots-clés,

-Les requêtes logiques (booléennes), la base des requêtes est un ensemble d'opérateurs logiques (et, ou et non)

-Les requêtes structurées, ce type de requêtes porte des informations non seulement sur le contenu mais aussi des informations sur la structure des documents telles que en-têtes, titres. .

Dans la RI traditionnelle, le besoin en information de l'utilisateur se compose d'un ensemble de mots-clés, c'est-à-dire que les informations ciblent le contenu textuel d'un document.

## 3-3 Pertinence

La pertinence est une notion fondamentale et cruciale dans le domaine de la RI. Elle est l'objet de tout système de recherche d'information. Définir cette notion complexe n'est pas simple, car elle fait intervenir plusieurs notions. Une définition simple pourrait être :
« La pertinence est la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête ».

On s'est vite aperçu que la pertinence n'est pas une relation isolée entre un document et une requête. Elle fait appel aussi au contexte de jugement. Ainsi, les travaux de recherche [P. Borlund 1998] s'accordent sur la difficulté de la définition de la pertinence et mettent en exergue deux types de pertinence :
*la pertinence système* et *la pertinence*

- la pertinence système est déterministe, objective et elle est définie à travers les modèles de RI. Elle est souvent traduite par un score évaluant l'adéquation du contenu des documents vis-à-vis de celui de la requête

- La pertinence utilisateur [T. Saracevic p 53] quant à elle, est liée à la perception de L'utilisateur sur l'information renvoyée par le système. Elle est subjective, deux Utilisateurs peuvent juger différemment un même document renvoyé pour une même requête. Elle peut évoluer dans le temps d'une recherche. Une information Jugée non pertinente à l'instant t pour une requête peut être jugée pertinente t + 1 car la connaissance de l'utilisateur sur le sujet à évolué.

# 4- La recherche d'information et d'autres domaines :

La RI a des relations fortes avec d'autres domaines, notamment avec les bases de données et avec des systèmes de question-réponse. [Céline Paganelli p 37 ]

## 4-1 la recherche d'information et les bases de données

On peut imaginer un système de RI comme un système de BD textuelles. Cependant, il faut souligner la différence suivante entre les deux types de système: Dans une base de données, on doit d'abord créer des schémas pour organiser les données. Ces schémas définissent des relations, ainsi que les attributs dans chaque relation. C'est en utilisant ces schémas que le système arrive à interpréter une requête de l'utilisateur. Par exemple, on peut définir la relation suivante dans une base de données :

Auteur (Livre, Nom).
Où Auteur est le nom de la relation, Livre et Nom sont ses attributs qui correspondent à l'identification d'un livre et à son (un des) auteur(s).
(Ceci est juste une partie de définition). Pour trouver le livre écrit par "Ali Ben Ali", on peut poser la requête suivante en SQL:

```
Select Livre
From Auteur
Where Nom = "Ali Ben Ali"
```

Cette requête n'est valide que si la relation Auteur a été créée ainsi.
Dans la RI, une partie des spécifications de documents est structurée, notamment les attributs externes. Cette partie peut être organisée assez facilement comme une relation en BD, et ainsi utiliser des SGBD existants pour rechercher des documents selon des critères sur les attributs externes. Mais, comme ce qu'on a dit, cette partie ne représente pas le cœur de la RI. Le cœur se situe dans la recherche selon le contenu. Or, le contenu est en général sans structure, ou avec une structure très souple. Il est très difficile de créer une relation pour représenter la partie contenu de document.
Après l'indexation de document, cependant, la connexion entre la RI et les BD devient plus étroite. Le résultat de l'indexation est d'associer à chaque document un ensemble d'index. Ce résultat peut être vu comme une relation en BD:
Index (Doc, Mot).
Ainsi, il est possible de faire une requête pour sélectionner les documents contenant le mot "recherche" et le mot "information" comme suit:

```
(Select Doc
From Index
Where Mot = "recherche")
Intersect
(Select Doc
From Index
Where Mot = "information")
```

Ce qui signifie que l'intersection de deux sélections sera le résultat. Noter, cependant, que les sélections ne retournent qu'un ensemble de documents sans ordonnancement. En RI, l'ordre de documents dans la liste de réponse est important. Ainsi, les BD ne permettent de réaliser qu'une partie de fonctionnalités de la RI.

## 4-2 La recherche d’information et les systèmes question réponse

Un système QR permet de répondre aux questions relatives à un petit domaine. Par exemple, on peut poser la question "quelle version de Word est disponible sous Windows 2007?" à un système spécialisé sur le marché de logiciel. Pour cela, il faut qu'on crée une modélisation du domaine d'application dans lequel les concepts ou objets sont reliés par des relations sémantiques. Ce modèle permettra de retrouver le concept ou l'objet et ainsi donner une réponse directe à la question. Pour notre exemple, la réponse peut être "Word 2003 et Word 2007", par exemple.
On voit ici qu'il y a une différence sur la nature de réponse entre les deux types de système. Dans la RI, c'est une réponse indirecte à une question: on identifie les documents dans lesquels l'utilisateur peut trouver des réponses directes à sa question. Tandis que dans un système QR, on fournit une réponse directe.
Il y a des tentatives de rapproche la RI vers des systèmes QR, mais cela s'avère très difficile. La raison principale est que la RI s'applique en général à tous les domaines sans restriction. Il est impossible, dans ce cas, de créer un modèle nécessaire pour déduire la réponse directe à une question dans un système QR. Dans certains contextes très spécialisés, la RI incorpore une base de connaissances. Elle utilise aussi des raisonnements pour déduire si un document peut être pertinent ou pas. Donc, le fonctionnement de ce type de RI ressemble un peu plus à celui d'un système QR.
Une tentative plus restreinte consiste à raffiner la notion de document dans la réponse: au lieu de fournir un document complet comme une réponse, on essaie d'identifier un passage dans le document (passage retrieval). C'est une étape qui diminue un peu la distance entre la RI et la QR. Mais la différence fondamentale reste la même. [Céline Paganelli 2002 p 39]

# 5- Processus de recherche d'information :

Pour répondre aux besoins en information de l'utilisateur, un SRI met en œuvre un certain nombre de processus pour réaliser la mise en correspondance des informations contenues dans un fonds documentaire d'une part, et les besoins en informations des utilisateurs d'autre part. Un SRI est défini par ses modèles de représentation des documents, des besoins de l'utilisateur et sa fonction d'appariement. Ce processus est composé de trois fonctions principales :

-*L'indexation des documents et des requêtes* : l'indexation a pour rôle d'extraire à partir d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique.

-*L'appariement requête-document ou l'interrogation* : la comparaison entre le document et la requête revient à calculer un score, supposé représenter la pertinence du document vis-à-vis de la requête.

-*la reformulation de la requête* : qui intervient suite à une itération de recherche, et Consiste à modifier les requêtes en fonction des résultats présentés et le jugement de l'utilisateur.

Le processus général d'un SRI consiste à représenter le besoin en information, et en parallèle de collecter des documents, de déterminer l'appariement entre chaque document et la requête puis de décider si le document est pertinent.
La reformulation de la requête est une procédure de base en RI. Au sens large, la reformulation de la requête peut renvoyer un dialogue interactif entre le système et l'utilisateur ce qui entraîne non seulement une demande adaptée mais aussi une meilleure compréhension par l'utilisateur de ses besoins en information comme illustre la *figure* 3 qui représente l'architecture en U d'un SRI [G. Salton and M. J. McGill 1986] Selon la représentation interne

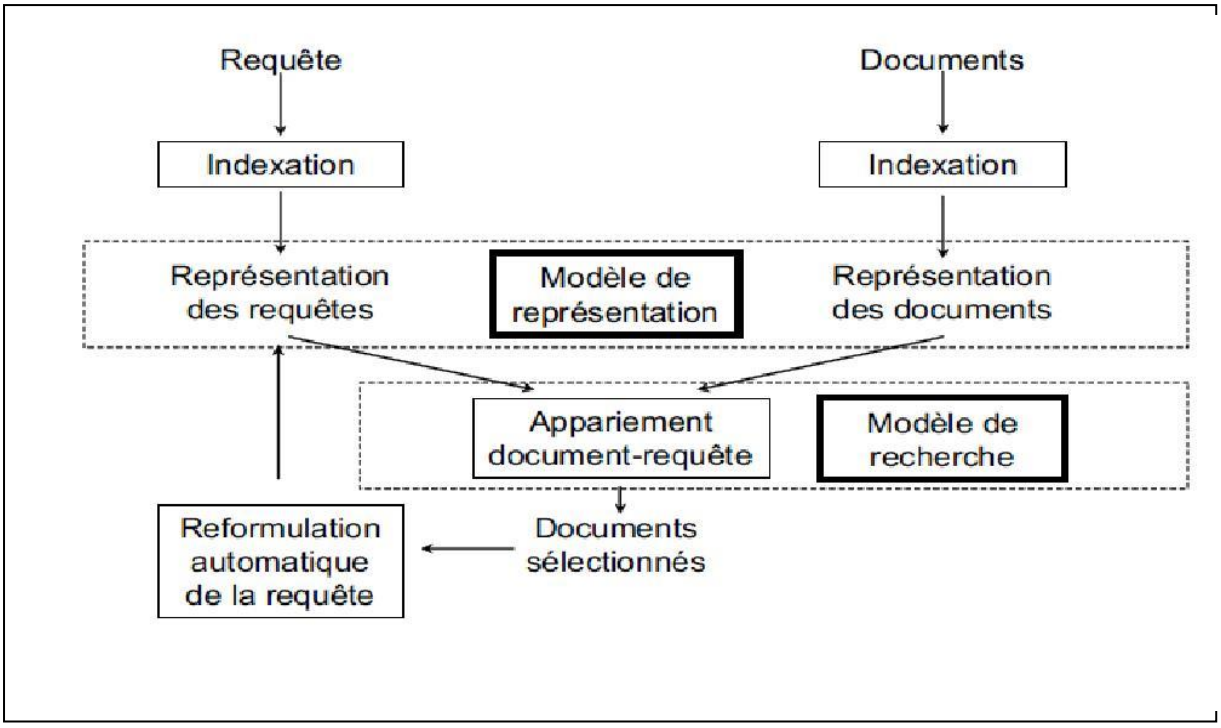



***Figure 3 : Architecture en U d'un Système de Recherche d'Information***

M-Mohamed BEN AOUICHA : Une approche algébrique pour la recherche D'information structurée p28

des requêtes et des documents, le SRI effectue un appariement afin de déterminer les documents potentiellement pertinents (pertinence système). Une fois l'appariement réalisé, le système sélectionne les documents les plus prometteurs[1] et les présentent à l'utilisateur. Sur la base de cette première réponse du système, l'utilisateur peut indiquer au système les documents qu'il juge réellement importants (pertinence utilisateur) et ceux qui ne lui présentent aucun intérêt. A l'aide de ces jugements, le système doit être capable de construire automatiquement une nouvelle requête (bouclage de requête ou reformulation par réinjection de pertinence) afin de présenter une nouvelle liste de référence [M. Boughanem, C. Chrisment, and C. Soulé-Dupuy 1999].Ce processus indique clairement que la RI doit toujours être vue comme un processus itératif.

[1] Les documents les plus pertinents jugés par le SRI.

Mais le point de départ de tout système documentaire est l'organisation du fond documentaire. En effet, et dans des conditions telles que le système devra réagir assez rapidement à son utilisateur, l'organisation des documents collectés joue un rôle primordial à la robustesse du SRI.

## 5-1 L'indexation des documents et des requêtes

Dans un processus de RI le coût de recherche doit être acceptable, il convient alors de procéder à une phase primordiale pour optimiser le temps d'exécution de ce processus. Cette phase consiste à analyser chaque document de la collection, et de créer un ensemble de mots-clés, on parle alors de l'indexation. Le rôle de l'indexation est fournir une présentation synthétique du contenu du document, on distingue trois types d'indexation

*Manuelle* : la représentation du document se fait par un spécialiste (documentaliste)

*Automatique* : le processus d'indexation est complètement informatisé

*Semi-automatique* : le processus d'indexation se fait en premier lieu d'une manière automatique, le documentaliste intervient seulement pour ajouter des mots-clés qu'il trouve intéressants pour représenter un document.

L'indexation manuelle permet d'assurer une bonne représentation des documents, cependant elle présente un inconvénient majeur vu que la tâche du documentaliste[1] se veut souvent influencée par son point de vue, or celui-ci est très subjectif : aucune règle universelle ne peut être appliquée à cette fin [G. Salton. 1968] .L'indexation semi-automatique est un processus partiellement automatisé, le documentaliste intervient pour apporter les rectifications qu'il voit nécessaires [C. Jacquemin, B. Daille, J. Royanté, and X. Polanco p 76 ]
L'indexation automatique est la plus communément utilisée. Ce type d'indexation regroupe un ensemble de traitements automatisés sur un document l'extraction automatique des termes du document, l'élimination des mots vides, la lemmatisation ou la radicalisation des mots, la pondération et enfin la création de l'index.

[1] Archiviste chargé de recueillir, classer, conserver et communiquer la documentation nécessaire à une recherche

## 5-1-1 méthode automatique d'indexation

### 5-1-1 -1 Extraction des termes

L'analyse lexicale constitue la première étape du processus d'indexation. Sa fonctionnalité principale est de connaître une unité lexicale ou un radical **[**C. Fox 1992]. La mission de l'analyse lexicale est alors de transformer une suite de caractères en une suite de mots reconnaissables.

### 5-1-1 -2 Elimination des mots vides

L'une des étapes dans le processus d'indexation permettant d'améliorer la fiabilité d'un SRI au sens de qualité logicielle (temps d'exécution) et de performance, est l'élimination des mots vides (pronoms personnels, prépositions. . .). Ce sont des mots ne traitent pas le sujet d'un document. On distingue deux techniques pour filtrer les mots vides :

• L'utilisation d'une liste prédéfinie de mots vides (aussi appelée anti-dictionnaire ou
stop-List)

• Le comptage du nombre d'apparition d'un mot dans un document de la collection. On supprime les mots ayant une fréquence qui dépasse un certain seuil et qui Deviennent alors vraisemblablement des mots vides. L'élimination de ces mots permet de réduire l'index, on gagne alors en espace mémoire, mais aussi le non traitement des mots vides fait gagner un SRI en temps d'exécution. Par ailleurs et dans le cas où la recherche est très ciblée, cette opération peut baisser la performance du SRI. Pour certains types de requêtes spéciales telles qu'une requête contenant le titre d'une chanson, qui bien évidemment peut contenir des mots vides, la recherche perd sa précision ; mais ces cas se présentent très rarement.

### 5-1-1-3 Radicalisation

Pour des raisons grammaticales, les documents utilisent différentes formes d'un mot, comme flexible, flexiblement, flexibilité. . . En outre, il y a des familles de dérivation des relations associées à la même signification. Dans bien de cas, il semble que ça serait utile pour une recherche sur l'un de ces mots de retourner les documents qui contiennent un autre mot dans la série.

L'objectif de la lemmatisation (radicalisation) est de rendre l'ensemble des formes des mots de la même famille représentées par un seul mot pour toute la famille. C'est la forme commune entre eux, qui est la forme de base (radical, par exemple flexibl pour flexible, flexiblement, flexibilité. . .).

Bien que le passage à la forme canonique présente un avantage, puisqu'elle permet d'indexer en un seul mot (lemme ou radical) sa famille morphologique, cette opération supprime la sémantique originale. Mais ceci ne présente aucun inconvénient puisque l'indexation passe totalement inaperçue par l'utilisateur : il s'agit d'un moyen de codage de l'information sans perte. Les plus courants algorithmes pour la radicalisation des mots de la langue anglaise est l'algorithme de Porter.

La méthode de radicalisation de Porter permet de construire progressivement le radical en inférant les modifications grammaticales potentielles qui se manifestent le plus souvent par des postfixes ou préfixes. Cette méthode n'est par ailleurs utilisable que pour la langue anglaise. L'indexation de documents écrits en langue française est souvent réalisée par la troncature à 7 caractères. En langue arabe on trouve l'algorithme alstem de Kareem Darwish et stem light de S. Khodja. Pour bien d'autres langages, aucune méthode de lemmatisation n'a été popularisée [G. W. Adamson and J. Boreham p 84]
.

### 5-1-1-4 Pondération des termes

La pondération d'un terme peut être utilisée de diverses manières. Elle peut être simplement la fréquence du terme. Mais celle-ci est abusivement simpliste. D'un autre côté, la pondération est décisive de la performance du système, pour mener à bien les phases de recherches ultérieures. La communauté de RI a longtemps investi dans l'identification de la formule la plus discriminative[1] d'un terme radical. Toutes les formules proposées à cet effet reposent sur un facteur très populaire en RI : le facteur tf × idf.

[1] Celle qui attribue à un terme le poids idéal qu'il pourrait quantifier son importance réelle dans un document

***Approches basées sur la fréquence locale (tf)*** : L'objectif de ces approches est de trouver les mots qui représentent le mieux le contenu d'un document. Il est généralement admis qu'un mot qui apparaît souvent dans un texte représente un concept important. Ainsi, il convient de choisir les mots représentatifs selon leurs fréquences d'occurrence. La façon la plus simple consiste à définir un seuil de fréquence : si la fréquence d'occurrence d'un mot dépasse ce seuil, alors il est considéré comme important pour le document.
Cependant, les mots les plus fréquents sont des mots usuels et ne présentent paradoxalement aucun intérêt pour un utilisateur potentiel. Quand on fait une statistique d'occurrences, si on classe dans l'ordre décroissant des mots par leurs fréquences, et on leur donne un rang (1,2 .. .) alors le rang serait inversement proportionnel à la fréquence : c'est la loi de Zipf [G. K. Zipf 1949]
Selon cette loi, la fréquence est inversement proportionnelle à son rang la distribution des mots suit la courbe illustrée par la *figure* 3 Il devient évident qu'on ne peut pas garder tous les mots dans l'index en définissant un seuil maximal, ni un seuil minimal pour les mots qui ne peuvent pas représenter les documents. L'utilisation de ces deux seuils mesure l'informativité ou encore le sens d'un mot, la correspondance entre l'informativité et la fréquence est illustrée par la *figure* 4. Ainsi, en choisissant

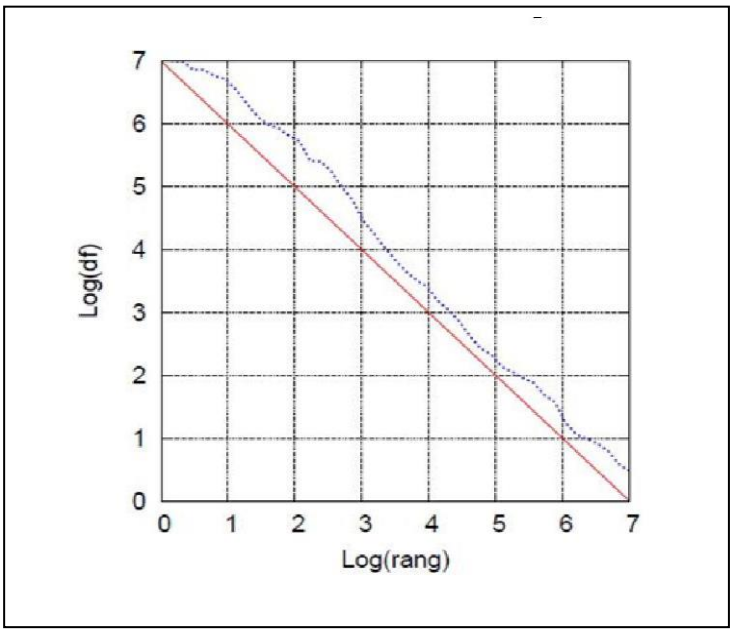



***Figure 3 : Fréquence d'un terme en fonction de son rang***
M-Mohamed BEN AOUICHA : Une approche algébrique pour la recherche D'information structurée p32

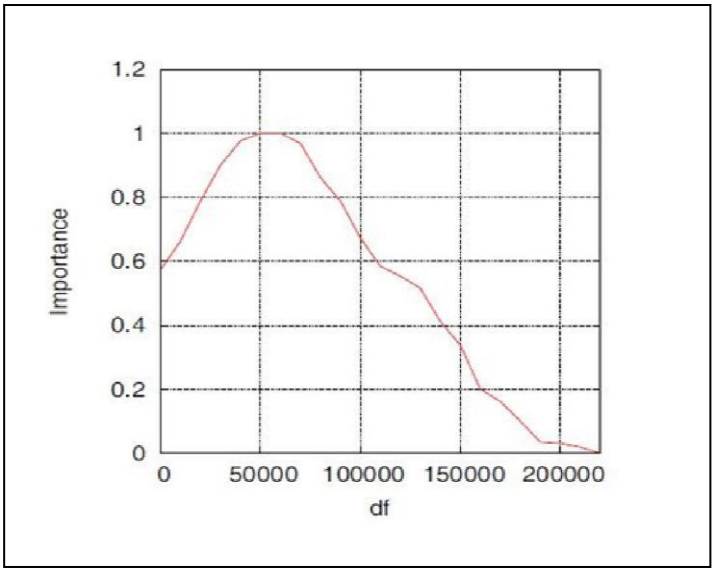



***Figure 4: Importance d'un terme en fonction de sa fréquence***

M-Mohamed BEN AOUICHA : Une approche algébrique pour la recherche D'information structurée p32

Les mots se situant entre ces deux seuils, on peut mieux représenter un document.

**Approches basées sur la fréquence globale (idf)** : La pondération d'un terme par discrimination reflète qu'un terme distingue bien un document des autres documents. C'est-à-dire, un terme qui a une valeur de discrimination élevée doit apparaître seulement dans un petit nombre de documents : un terme qui apparaît dans tous les documents n'est pas discriminant. La capacité de discrimination d'un terme est importante dans le choix des termes d'indexation à retenir. L'idée est de garder seulement les termes discriminants, et éliminer ceux qui ne le sont pas les approches basées sur tf ainsi que ceux qui sont basées sur idf sont efficaces, par le choix des termes discriminants. La pondération est néanmoins communément basée sur la combinaison des deux mesures. Les raisons pour lesquelles ses deux paramètres ont l'intérêt de coexister est qu'ils sont très révélateurs de l'importance d'un terme, et aussi pour se livrer de l'obligation d'identifier des seuils et de migrer vers l'identification de l'importance d'un terme par une mesure plus statistique.

**Approches basées sur tf ×idf** : La mesure tf ×idf en RI désigne un ensemble de schémas de pondération de termes. tf désigne une mesure en rapport avec l'importance d'un terme pour un document. En général, cette valeur est déterminée par la fréquence du terme dans le document. Par idf, on mesure si le terme est discriminant (un terme est discriminant s'il apparaît peu dans d'autres documents [G. Salton, A. W., and C. S. Yang p 77.]. Cette mesure est utilisée en RI pour calculer la pertinence d'un document par rapport à une requête, ceci traduit le fait qu'un document est pertinent à une requête s'ils partagent assez de termes importants.
La mesure tf × idf est communément utilisée en RI, vu que cette mesure donne une bonne approximation de l'importance du terme dans le document. Cependant, elle ne donne aucune importance à la longueur du document. Pourtant cette caractéristique est utile dans le processus de recherche. En effet pour les documents longs, la fréquence des termes augmente, de ce fait la similarité entre ces documents et la requête augmente et les documents de petites tailles se trouvent souvent défavorisés. L'introduction de ce facteur (longueur du document) qu'on appelle facteur de normalisation s'avère indispensable.

### 5-1-1-5 Organisation de l'index

Au terme de toutes les différentes étapes expliquées ci-avant (analyse lexicale, élimination des mots vides, radicalisation et choix de la technique de pondération d'un terme),il devient nécessaire d'organiser et de stocker les informations sélectionnées. Le stockage de données peut se faire dans des fichiers structurés ou non structurés. Le fichier de base dans lequel sont stockées les données est appelé fichier maître. L'opération peut

| Document | Contenu |
|---|---|
| $d_1$ | La(1) recherche(4) d(14) '(15) information(16) gère(28) des(33) textes.(37) |
| $d_2$ | Un(1) système(4) de(12) recherche(15) d(25) '(26) information(27) doit(39) restituer(44) l(54) '(55) information(56) pertinente(68) à(79) l(81) '(82) utilisateur.(83) |
| $d_3$ | Une(1) information(5) est(17) pertinente(21) si(32) elle(35) satisfait(40) l(50) '(51) utilisateur.(52) |

***Tableau 1: Exemple de collection (fichier maître)***

M-Mohamed BEN AOUICHA : Une approche algébrique pour la recherche D'information structurée p34

| terme | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|
| recherche | 4 | 15 | 3 |
| information | 16 | 27, 56 | 5 |
| gère | 28 | | |
| textes | 37 | | |
| système | | 4 | |
| restituer | | 44 | |
| pertinente | | 68 | 21 |
| utilisateur | | 83 | 52 |
| satisfait | | | 40 |

***Tableau2: Exemple de fichier inverse***

M-Mohamed BEN AOUICHA : Une approche algébrique pour la recherche
D'information structurée p34

Durer quelques secondes sur un fichier maître de quelques centaines d'enregistrements, cependant, elle peut se révéler très lente si la base atteint des milliers de documents. Les fichiers inverses sont créés autour du fichier maître. [A. Moffat and L. Stuiver P 91 ]. Ces fichiers comme leur nom l'indiquent, sont le résultat de l'inversion du fichier maître. Plus exactement, au lieu de donner pour chaque document les mots et les fréquences qui le constituent, on donne pour chaque mot les documents qui le contiennent et sa fréquence dans chacun. Le *tableau* 2 illustre un exemple de fichier inverse construit à partir de la collection illustrée par le *tableau* 1

### 5-1-2 Méthode manuelle d'indexation

L'indexation manuelle est ajustée par des corrections humaines et principalement adoptée par les outils de recherche de type thématique (les annuaires). il s'agit d'un contrôle manuel des informations récoltées soit par soumission des utilisateurs. L'information est ensuite classée et cataloguées par grandes catégories. Les catalogues thématiques ainsi créés sont construits à la main par des personnes filtrant les sites en fonction de leur qualité, pertinence et fiabilité. Cette méthode manuelle apporte une valeur ajoutée certaine mais la mise à jour est moins rapide et la couverture beaucoup moins large.

Le rôle de l'indexeur est d'attribuer au document un certain nombre de descripteurs ou scripteurs suivant

### 5-1-2-1 Mots –clés uni termes

Ces descripteurs sont formés d'un seul mot par conjonction de plusieurs unités, on obtient des expressions composées.

### 5-1-2-2 Descripteurs composés

Ils sont constitués d'expression de deux ou trois termes. On peut utiliser des expressions de différents types :

- Nom -adjectif (ex : Droit social),
- Nom –préposition -nom (ex : Histoire de la musique),
- Des termes possédant un trait d'union (ex : Libre -échange),
- Des termes avec rejet (ex Boole, algèbre) bien adapté pour les catalogues manuels,
- Des termes avec un qualificatif (ex : Mercure (métal), Mercure (planète)).

### 5-1-2-3 Descripteurs structurés

Un descripteur structuré contient plusieurs informations sous une même entrée dite « vedette ». On fait succéder les descripteurs dans l'ordre suivant :

- Tête de vedette, significative du sujet,
- Sous-vedette de point de vue,
- Sous-vedette de localisation géographique,
- Sous-vedette de localisation chronologique,
- Sous-vedette de forme (dictionnaire, bibliographie, congrès.)

*Exemple :*

Titre : la recherche d'information dans l'entreprise

Genre : informatique

Partie : 01

Page : 210 pages

Année : 2006

Résumé : Parce qu'il est aujourd'hui difficile de trouver l'information pertinente, les entreprises sont peu à peu amenées à reconsidérer le rôle stratégique de la recherche d'information……

## 5-1-2-4 Codes numériques

Les descripteurs peuvent être codés pour représenter simplement des notions difficiles à réduire en quelques mots.

## 5-1-2-5 Indices de classification

### A / *Classification hiérarchiques ou non*

On considère que la connaissance est constituée d'éléments emboîtés à différents niveaux.
La maintenance de la classification est délicate car elle nécessite de trouver à quel niveau de connaissance correspond le document à insérer. Elle entraîne l'utilisation d'indice très grand.
**Exemple** : en classification Dewey on a :
5 Sciences

51 Mathématiques

512 Algèbre

513 Arithmétique

***B / Classification à facettes***

On essaye de regrouper les connaissances sans que chacune soit dépendante d'une autre. Par exemple on utilise le principe de « Raganathan » que l'on peut décomposer chaque sujet en éléments inclus dans des domaines sémantiques pré -définis :

- Personnalité
- Matière
- Energie
- Espace

*Exemple :*

-Alcool= substance chimique
-Liquide = état de cette substance
-Volatilité = propriété de cette substance
-Combustion = réaction de cette substance
-Analyse = opération de l'homme sur cette substance

### *5-1-3 Mise à jour des index :*

Afin de garantir les performances de réponse d'un moteur de recherche, les index doivent être régulièrement mis à jour. L'évolution des outils de recherche permet actuellement une mise à jour en temps réel ou en léger différé. Le procédé de mise à jour diffère entre un moteur de recherche sur internent et un moteur d'une base de données. [Cél Paganelli p 67 ]

Dans le premier cas, la mise à jour consiste simplement à parcourir la base de données du moteur, d'indexer les nouveaux documents et de supprimer les entrées d'index correspondant à des documents qui n'existent plus.

Dans le deuxième cas, les documents nouveaux ou modifiés peuvent être mis à jour à la volée Au besoin, les indexes ne doivent pas seulement être mis à jour ou entièrement reconstruits mais également optimisés afin de garantir des taux de réponses fiables et rapides.

## 5-2 Appariement document-requête

L'objectif du SRI est le calcul de la pertinence d'un document par rapport à une Requête, c'est une fonction d'appariement qui détermine le degré de ressemblance d'un document par rapport à une requête, et permet éventuellement de classer les documents par ordre de pertinence. Cette fonction est notée rsv (q, d) (Retrieval Status Value), où q est une requête et d est un document de la collection

La fonction d'appariement est indépendante de l'indexation et de la pondération des termes, par contre elle caractérise le SRI plus que le modèle d'indexation : la plupart des approches de recherche inspirent leurs noms à partir de la façon dont ils entament l'appariement.

## 5-3 Reformulation de la requête

Il est souvent difficile, pour l'utilisateur, de formuler son besoin exact en information. Par conséquent, les résultats que lui fournit la RI ne lui conviennent parfois pas. Retrouver des informations pertinentes en utilisant seule la requête initiale de l'utilisateur est quasi-impossible. Afin de faire correspondre au mieux la pertinence utilisateur et la pertinence système, une étape de reformulation de la requête est souvent utilisée. La requête initiale est traitée comme une tentative pour retrouver de l'information. Les documents initialement présentés sont examinés et une formulation améliorée de la requête est construite, dans l'espoir de retrouver plus de documents pertinents. La reformulation de la requête met en œuvre un algorithme de modification de la requête en termes, en poids ou les deux simultanément, moyennant des critères de choix de termes d'expansion et des règles de calcul des nouveaux poids.

En RI, savoir reformuler une idée par des termes différents est une des clefs pour L'amélioration des performances des SRI existants. L'un des moyens pour résoudre ce problème est d'utiliser des ressources sémantiques spécialisées et adaptées à la base documentaire sur laquelle les recherches sont faites [M. Gilloux, E. Lassalle, and J. M. Ombrouck p 47]. La stratégie de reformulation de la requête est la plus populaire. On la nomme communément réinjection de la pertinence ou relevance feedback [G. Salton and C. Buckley p 65 ]
La reformulation par réinjection de pertinence est une forme de recherche évolutive et interactive ; elle procède à la modification de la requête initiale en termes et en poids, sur la base des jugements de pertinence de l'utilisateur sur les documents restitués par le SRI. Son principe fondamental est d'utiliser

la requête initiale, puis exploiter itérativement les jugements de pertinence de l'utilisateur afin d'ajuster la requête par expansion, repondération ou combinaison des deux procédures, en direction des documents pertinents.

# 6- les modèles de RI

Un modèle de RI modélise la fonction d'appariement qui joue un rôle central dans la RI. C'est le modèle qui détermine le comportement clé d'un SRI. On distingue trois principaux modèles pour la RI, qui sont cependant étroitement liés.

## 6-1 Le modèle booléen ou ensembliste et ses dérivés

Dans ce modèle, un document est représenté par un ensemble de termes. Une requête est une expression logique composée de termes assemblés par les opérateurs logiques et, ou et non.

La formulation de la requête se base sur les trois opérateurs booléens :
• La conjonction et (^), exige que les termes soient présents simultanément dans la description d'un document,
• La disjonction ou (_), exige qu'au moins un des termes soit présent dans la description des documents à retourner,
• La négation non (¬), utilisée pour écarter les documents qui contiennent un terme.

### 6-1-1 Le modèle booléen basique

Le modèle booléen [G. Salton and C. Buckley p 87] est le modèle le plus ancien dans la RI, les documents réagissent suivant la présence ou l'absence des termes utilisés. L'approche booléenne utilise le mode d'appariement exact qui consiste à ne restituer que les documents répondant exactement à la requête. Le modèle booléen considère dans l'index qu'un terme est présent ou non dans le document, par conséquent le poids d'un terme noté wi,j 2 {0, 1}. Considérons une requête contenant trois termes (ta, tb et tc), et l'expression logique de la requête définie par
q = ta ⋁ (tb ^ ¬tc), la similarité entre un document et la requête est définie par :

$$Rsv(q, d) = \begin{cases} 1 \text{ si d contient ta} \\ 1 \text{ si d contient tb et ne contient pas tc} \\ 0 \text{ si non} \end{cases}$$

Le modèle booléen considère qu'un document est soit pertinent soit non pertinent. L'inconvénient de ce modèle est qu'il rend la tâche de l'utilisateur pour formuler son besoin en information plus complexe, à cause de sa manière d'appariement De plus, il est incapable de fournir une liste ordonnée de documents car la perception de la pertinence selon le modèle booléen est très différente de celle d'un utilisateur novice. Par exemple, si l'utilisateur formule une requête t1 ^ t2 ^ t3 ^ t4, et si aucun document ne répond exactement à la requête, il convient tout de même de présenter un panorama des documents qui répondent partiellement à celle-ci en admettant qu'un document qui contient les termes t1, t2 et t3 est plus pertinent qu'un autre qui ne contient que les termes t1 et t2, ou un autre qui ne contient aucun des termes de la liste. Tout l'art est de trouver le moyen permettant une utilisation plus flexible des opérateurs booléens. Le modèle booléen étendu est l'un des premiers modèles qui ont été proposés à cette fin.

## 6-1-2 Le modèle booléen étendu

Le modèle booléen étendu est introduit en 1983 par Salton, Fox et Wu
**[**G. Salton, E. Fox, and H.Wu 1983], l'idée est de permettre l'utilisation des opérateurs logiques tout en proposant une pertinence graduée. Ce modèle peut être vu comme une combinaison des modèles booléen et vectoriel (section 1.3.2). Au lieu d'estimer le poids d'un terme par son absence ou présence, la pondération des termes dans un document est basée sur tf ×idf normalisé c'est-à-dire que le poids d'un terme dans un document se trouve entre 0 et 1. Ce modèle consiste à calculer la distance entre les coordonnées d'un document et les coordonnées d'une requête. Le modèle booléen étendu considère que les opérations booléennes ont une influence sur la façon dont il faut entreprendre la requête. La représentation d'un document contrairement au modèle booléen basique, tient compte des poids des termes. Chaque document est représenté par un vecteur de termes pondérés. Selon l'opérateur booléen utilisé, le modèle booléen étendu calcule le score d'un document selon deux cas principaux :

• Requête de type disjonction : ka ou kb, où ka et kb sont deux termes. Si on considère un document d dont les poids respectifs de ka et kb sont wa et wb alors le score de d par rapport à la requête ka _ kb est estimé selon sa distance par rapport au point d'origine (0, 0) :

$$score(k_a \ ou \ k_b, d) = \sqrt{\frac{score^2(k_a,d)+score^2(k_b,d)}{2}} = \sqrt{\frac{w_a^2+w_b^2}{2}}$$

**La figure 5** montre que le document d2 est plus pertinent que le document d1 malgré qu'ils contiennent les termes ka et kb (les poids de chacun sont différents). Les quarts de cercles concentriques représentent les documents ayant le même scor

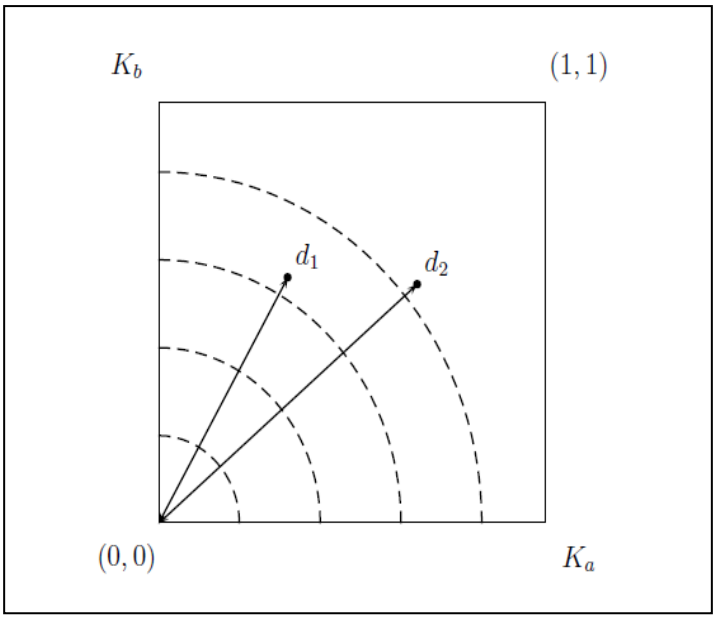



***Figure 5 : Mesure de similarité entre un document et une requête de type ou***

Céline Paganelli : la recherche d'information et l'interface humaine p 75

• Requête de type Conjonction : ka et kb, où ka et kb sont deux termes. Si on considère un document d dont les poids respectifs de ka et kb sont wa et wb alors le score de d par rapport à la requête ka ^ kb est estimé selon sa distance par rapport au point (1, 1). **La figure 6** montre que le document d1 est plus pertinent que le document d2 malgré qu'ils contiennent les termes ka et kb (les poids de chacun sont différents). Les quarts de cercles concentriques représentent les documents ayant le même score. Le point (1, 1) représente la situation où les deux termes ka et kb sont présents dans le document. La mesure de similarité de cette requête par rapport à un document est

$$score(k_a \text{ et } k_b, d) = 1 - \sqrt{\frac{(1-score(k_a,d))^2+(1-score(k_b,d))^2}{2}}$$
$$= 1 - \sqrt{\frac{(1-w_a)^2+(1-w_b)^2}{2}}$$

• Requête de type négation : non ka, le score d'un document est estimé par : score(non ka, d) = 1 − score(ka, d)
Les requêtes complexes sont traitées récursivement par exemple, le score d'un document d par rapport à une requête ka^ (kb ^ kC^¬kd) est calculé comme suit :

$$
\begin{aligned}
score(d, k_a \wedge (k_b \vee k_c \wedge \neg k_d)) &= 1 - \sqrt{\frac{(1-score(k_a,d))^2+(1-score(k_b \vee k_c \wedge \neg k_d,d))^2}{2}} \\
score(k_b \vee k_c \wedge \neg k_d, d) &= \sqrt{\frac{score^2(k_b,d)+score^2(k_c \wedge \neg k_d,d)}{2}} \\
score(k_c \wedge \neg k_d, d) &= 1 - \sqrt{\frac{(1-score(k_c,d))^2+(1-score(\neg k_d,d))^2}{2}} \\
score(\neg k_d, d) &= 1 - score(k_d, d) \\
score(k_x, d) &= k_x \forall x \in \{a, b, c, d\}
\end{aligned}
$$

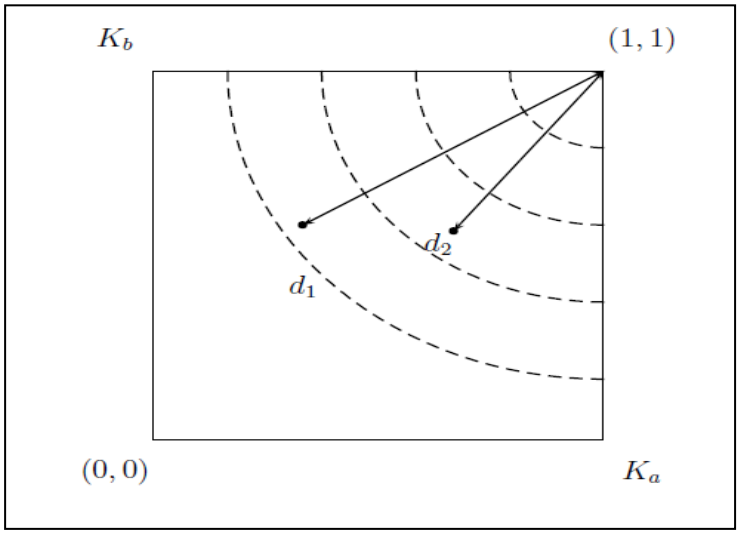



***Figure 7 : Mesure de similarité entre un document et une requête de type <u>et</u>***

Céline Paganelli : la recherche d'information et l'interface humaine p 76

❖ **Avantages et inconvénients de Modèle booléen**

## Avantage :

– Le modèle est transparent et simple à comprendre pour l'utilisateur :

• Pas de paramètres "cachés"
• Raison de sélection d'un document claire : il répond à une formule logique
– Adapté pour les spécialistes et les vocabulaires contraints

## Inconvénients :

– Il est difficile d'exprimer des requêtes longues sous forme booléenne
– Le critère binaire peu efficace
• Il est admis que la pondération des termes améliore les résultats

– Il est impossible d'ordonner les résultats

• Tous les documents retournés sont sur le même plan
• L'utilisateur préfère un classement lorsque la liste est grande

## 6-2 Le modèle algébrique et ses dérivés

Les modèles algébriques proposent une représentation vectorielle pour le document et la requête. La mise en correspondance entre le document et la requête consiste à calculer la similarité entre les vecteurs représentant les documents et les requêtes.
Le modèle vectoriel est l'ancêtre de tous les modèles algébriques. Les premiers travaux de Salton [G. Salton p 87 ] avaient pour finalité de concevoir la fonction d'appariement selon les propriétés et les opérations associées au concept d'espace vectoriel. Bien que simpliste, ce modèle reste un des plus utilisés et des plus efficaces.

Les modèles dérivés du modèle booléen préconisent une habilité importante à fournir des listes ordonnées de résultats. D'un point de vue statistique, les requêtes qui ne contiennent que des mots clés (sans les opérateurs ensemblistes et, ou et non), qui sont d'ailleurs les plus fréquemment composées, prétendent une synthèse plus approfondie. Plusieurs initiatives ont été proposées pour l'estimation de la fonction d'appariement, son comportement et son habilité à se rapprocher de l'appariement perçu par l'utilisateur, mais elles sont toutes dérivées soit de la logique vectorielle soit de la logique probabiliste. Nous détaillons dans la section suivante ces modèles et leurs modèles dérivés.

## 6-2-1 Le modèle vectoriel basique

Le modèle vectoriel (nommé aussi VSM pour Vector Space Model), a été popularisé par Salton en 1971. Ce modèle propose de représenter les documents et les requêtes par des vecteurs d'indexation dans un espace engendré par les termes d'indexation. Le modèle vectoriel représente les requêtes et les documents sous forme de vecteurs dans un même espace vectoriel. La mesure de similarité entre le document représenté par un vecteur ~d = (d1, d2 . . . dn), où di est le poids d'un terme i dans le document, et la requête définie par ~q = (q1, q2 . . . qn) où qi est le poids (souvent 0 ou 1 selon que le terme appartient ou pas à la requête) du terme i dans la requête, est estimée selon les propriétés et les mesures populaires issues de la théorie des espaces vectoriels. Ils existent plusieurs mesures pour calculer la similarité entre le document et la requête, la plus simple est le produit scalaire :

$$rsv(\vec{d},\vec{q}) = \sum_{k=1}^{n} d_k \times q_k$$

Si les composantes des deux vecteurs sont binaires (1 si le terme existe dans le document, 0 si non) alors la mesure de similarité entre le document et la requête est égale au nombre de mots partagés entre eux. Le produit scalaire est très sensible à la norme des vecteurs documents et requête (leurs longueurs). D'autres mesures ont été proposées, elles sont tout de même basées sur le produit scalaire. La mesure du cosinus, qui mesure le cosinus de l'angle formé par le document et la requête, est utilisé dans le modèle vectoriel : plus l'angle est petit, plus la requête est proche du document et par conséquent plus le cosinus de l'angle est élevé. La mesure cosinus est donnée par :

$$rsv(\vec{d},\vec{q}) = cos(\vec{d},\vec{q}) = \frac{\vec{d} \times \vec{q}}{|\vec{d}| \times |\vec{q}|} = \frac{\sum_{k=1}^{n} d_k \times q_k}{\sqrt{\sum_{k=1}^{n} d_k^2 . \sum_{k=1}^{n} q_k^2}}$$

**La figure 8** illustre le cosinus de l'angle formé par une requête et deux documents d1 et d2. Le document d2 est différent de la requête, le cosinus de l'angle résultant est égal à cos ($\theta$) < 1. Le document d1 est contrairement à d2, très proche de la requête, son rsv est égal à 1 = cos(0). D'autres mesures sont utilisées pour percevoir le score

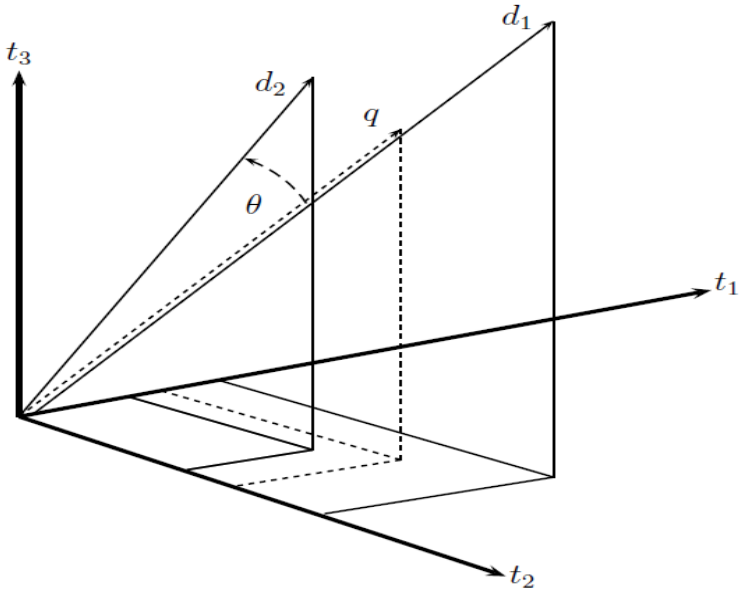



**Figure 8 : Modèle vectoriel**

Céline Paganelli : la recherche d'information et l'interface humaine machine p 80

D'un vecteur document par rapport à un vecteur requête :

- Jaccard : $rsv(d,q) = \frac{\sum_{k=1}^{n} d_k.q_k}{\sum_{k=1}^{n} d_k^2 + \sum_{k=1}^{n} q_k^2 - \sum_{k=1}^{n} d_k.q_k}$
- Dice : $rsv(d,q) = \frac{2 \times \sum_{k=1}^{n} d_k.q_k}{\sum_{k=1}^{n} d_k^2 + \sum_{k=1}^{n} q_k^2}$

- Overlap : $rsv(d,q) = \frac{\sum_{k=1}^{n} d_k.q_k}{\min\left(\sum_{k=1}^{n} d_k^2, \sum_{k=1}^{n} q_k^2\right)}$

Toutes ces mesures ont l'avantage de profiter des propriétés de l'espace vectoriel pour la perception de l'appariement utilisateur. Le principal intérêt porté à leur application est leur habilité à retourner des listes ordonnées de documents. Le principal inconvénient du modèle vectoriel est le fait qu'il suppose que les termes d'indexation forment une base. Or ils existent énormément de relations sémantiques qui font qu'un terme pourra s'exprimer en fonction des autres. Par ailleurs il est très difficile voire impossible de traduire des relations par des combinaisons linéaires de termes, or ceci s'avère indispensable à la construction de vraie base de termes d'indexation. La représentation vectorielle procure une autre fonctionnalité très importante : le degré de discrimination des termes d'indexation. Le modèle LSI a été proposé pour percevoir ce degré et d'exprimer un terme en fonction des autres.

### 6-2-2 LSI (Latent Semantic Indexing)

La représentation des documents et des requêtes dans le modèle vectoriel souffre de son incapacité de gérer les synonymes (voiture et automobile) les hyponymes. . . Or l'espace vectoriel de représentation ne parvient pas à saisir la relation entre les termes, ce qui influe sur le comportement du SRI. Prenons un document di contenant deux termes (automobile et voiture), et un autre document dj qui contient uniquement un terme (voiture). Il est clair que le document di est plus pertinent à une requête q qui ne contient qu'un seul terme (voiture), pourtant ils possèdent le même contexte des documents potentiellement pertinents. L'intérêt du modèle LSI est de remédier à ce problème. Pour ce faire l'idée consiste à transformer les vecteurs de termes de l'index dans un autre espace vectoriel de dimension associée aux concepts. Le modèle LSI utilise la SVD (Singular Value Décomposition) pour créer un nouvel espace vectoriel :

$$A = U \times \Sigma \times V^T$$

où A est la matrice documents-termes originale m × n, U est une matrice m × r, est une matrice diagonale r × r (seulement les éléments en diagonale sont non nuls) et V T est une matrice r × n. La valeur r est telle que r = inf.(n,m) où m est le nombre de termes et n est le nombre de documents. On trie les valeurs dans $\Sigma$ dans l'ordre décroissant. Il existe une seule décomposition de cette façon :

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_r \end{pmatrix}$$

$$\text{avec } \sigma_1 \geq \sigma_2 \dots \sigma_r.$$

La représentation par des mots-clés contient beaucoup de bruit. Typiquement, ce bruit se traduit par les valeurs les moins élevées de $\Sigma$ . Ainsi, la technique de LSI tend à supprimer ce bruit en écartant ces valeurs, ce qui ramène la dimension de $\Sigma$ à k, cette matrice réduite est notée $\Sigma$ k (**voir figure 9** pour illustration). Quand une requête est

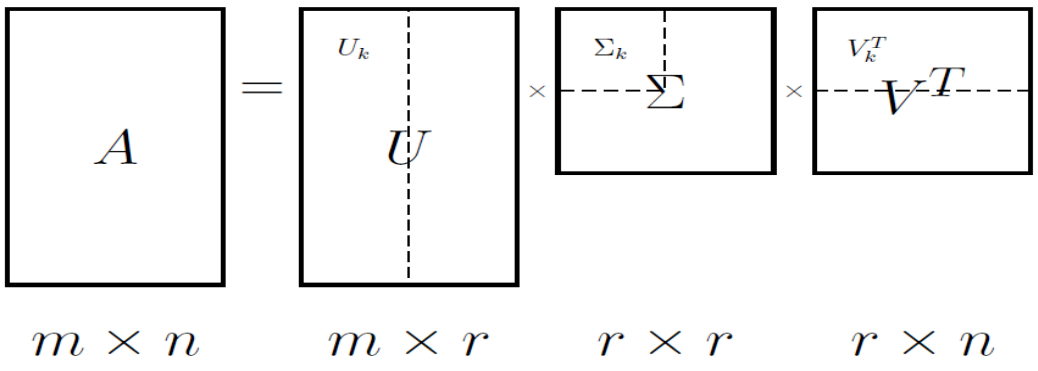



**Figure 9 : Modèle LSI**

Céline Paganelli : la recherche d'information et l'interface humaine machine p 85

soumise, elle est aussi traduite dans ce nouvel espace (changement de base) :

$$q = q^T \times U_k \times \Sigma_k^{-1}$$

La réduction de la dimension de l'espace vectoriel se traduit par une compression

sans perte considérable. La quantité de l'information préservée par la décomposition de A en Uk × ∑ k × V Tk peut être évaluée comme suit :

$$contraste = \frac{\sum_{i=1}^{k} \sigma_i}{\sum_{i=1}^{r} \sigma_i}$$

## Avantages et inconvénients de model vectoriel

**Avantage :**

– Le langage de requête est plus simple (liste de mot-clé)
– Les performances sont meilleures grâce à la pondération des termes
– Le renvoi de documents à pertinence partielle est possible
– La fonction d'appariement permet de trier les documents

**Inconvénients :**

– Le modèle considère que tous les termes sont indépendants
(Inconvénient théorique)
– Le langage de requête est moins expressif
– L'utilisateur voit moins pourquoi un document lui est renvoyé

## 6-3 Le modèle probabiliste

Le modèle probabiliste est basé sur la théorie de la décision. Le but est de calculer la Probabilité de la pertinence d'un document d par rapport à une requête q.

Dans ce modèle un document et une requête sont représentés par un vecteur comme dans le modèle vectoriel, mais les poids des termes sont binaires, soient D1,D2 . . .Dn (Di est égale à 1 si le terme ti est présent dans le document et 0 si non). L'idée de base dans ce modèle est de tenter de déterminer les probabilités p(L = 1|d) (la probabilité que le document d soit pertinent pour la requête q) et p(L = 0|d) (la probabilité que le document d ne soit pas pertinent pour la requête q). Si le document d contient les termes D1,D2 . . .Dn, alors p(L = 1|d) qui mesure la probabilité que d soit pertinent pourra se ramener à la probabilité qu'un document soit pertinent sachant qu'il contient les termes (D1,D2 . . .Dn), cette probabilité sera notée P(L = 1|D1,D2 . . .Dn) [S. E. Robertson and S. Walker 1994 p 232 ]
La fonction de similarité entre un document d et une requête q est donnée par

$$rsv(d,q) = \log \frac{P(L=1|D_1, D_2 \ldots D_n)}{P(L=0|D_1, D_2 \ldots D_n)}$$

Cependant, il y a très peu de documents, voire aucun, qui contiennent exactement le même contenu que d quel qu'il en soit, d'où P(L = 1|D1,D2 . . .Dn) devient presque nul, et par la même P(L = 0|D1,D2 . . .Dn) = 1 − P(L = 1|D1,D2 . . .Dn) devient presque égal à 1. Or, le calcul de P(L = 0|D1,D2 . . .Dn) donne paradoxalement une valeur presque nulle également. On utilise la règle de transformation de Bayes :

$$
\begin{aligned}
rsv(d,q) &= logitP(L=1|D_1,D_2\ldots D_n) \\
&= log\frac{P(L=1|D_1,D_2\ldots D_n)}{P(L=0|D_1,D_2\ldots D_n)} \\
&= log\frac{\frac{P(L=1)P(D_1,D_2\ldots D_n|L=1)}{P(D_1,D_2\ldots D_n)}}{\frac{P(L=0)P(D_1,D_2\ldots D_n|L=0)}{P(D_1,D_2\ldots D_n)}} \\
&= log\frac{P(D_1,D_2\ldots D_n|L=1)}{P(D_1,D_2\ldots D_n|L=0)} + logitP(L=1)
\end{aligned}
$$

L'hypothèse d'indépendance entre les termes est supposée pour simplifier le calcul (P (ti|tj) = P (ti)) ainsi :

$$P(D_1,D_2\ldots D_n|L=1) = \prod_{i=1}^{n} P(D_i|L=1)$$

et :

$$P(D_1,D_2\ldots D_n|L=0) = \prod_{i=1}^{n} P(D_i|L=0)$$

ainsi :

$$
\begin{aligned}
rsv(d,q) &= log\frac{\prod_{i=1}^{n} P(D_i|L=1)}{\prod_{i=1}^{n} P(D_i|L=0)} + logitP(L=1) \\
&= log\prod_{i=1}^{n}\frac{P(D_i|L=1)}{P(D_i|L=0)} + logitP(L=1) \\
&= \sum_{i=1}^{n} log\frac{P(D_i|L=1)}{P(D_i|L=0)} + logitP(L=1)
\end{aligned}
$$

Le calcul de la pertinence d'un document par rapport à une requête nécessite une base d'apprentissage dans laquelle la pertinence de quelques documents est connue ainsi :

$$P(D_i|L=1) = \frac{|Rel \cap doc_i|}{|Rel|}$$

$$P(D_i|L=0) = \frac{|NRel \cap doc_i|}{|NRel|}$$

$$P(L=1) = \frac{|Rel|}{|Rel|+|NRel|}$$

$$P(L=0) = \frac{|NRel|}{|Rel|+|NRel|}$$

où $doc_i$ est l'ensemble des documents contenant le terme $t_i$, Rel et NRel représentent respectivement les ensembles de documents pertinents et non pertinents de la base d'apprentissage. Outre la supposition de l'indépendance entre les évènements, le modèle probabiliste ne tient pas compte de la fréquence d'un terme dans un document, or ce critère est indispensable pour quantifier le score d'un document. Cependant, la prise en compte de cette mesure peut avoir des conséquences très négatives sur le calcul des scores. En effet, si un terme apparaît 15 fois dans un document et 14 fois dans un autre, le modèle probabiliste considère en revanche que les fréquences sont différentes malgré qu'elles sont sensiblement les mêmes. Le modèle 2-Poisson a été proposé pour pallier à cette ambiguïté.

## 6-4 Modèle de langue

- *Idée de base* :

Par « modèle de langue », on désigne une fonction de probabilité *P* qui assigne une probabilité *P*(*s*) à un mot ou à une séquence de mots *s* en une langue. Une fois cette fonction définie, il est possible d'estimer la probabilité d'une séquence de mots quelconque dans la langue, ou d'un point de vue générative, d'estimer la probabilité de générer cette séquence de mots à partir du modèle de la langue. [M. Boughanem, W. Kraaij, and J-Y Nie p 182]

Considérons la séquence *s* composée des mots suivants : $m_1$, $m_2$,..., $m_n$. La probabilité *P*(*s*) peut être calculée comme suit :

$$P(s) = \prod_{i=1}^{l} P(m_i \mid m_1 \ldots m_{i-1}) \quad (1)$$

Si on utilise la règle de chaîne en théorie de probabilité pour calculer cette probabilité, il y a souvent trop de paramètres (c'est-à-dire $P(m_i|m_1 \ldots m_{i-1})$) à estimer, et ceci est souvent impossible de réaliser. Ainsi, dans les modèles de langue utilisés en pratique, des simplifications sont souvent faites. En général, on suppose qu'un mot $m_i$ ne dépend que de ses *n*-1 prédécesseurs immédiats, c'est-à-dire :

$$P(m_i \mid m_1 \ldots m_{i-1}) = P(m_i \mid m_{i-n+1} \ldots m_{i-1}) \quad (2)$$

On utilise, dans ce cas, un modèle de langue n-gramme. En particulier, les modèles souvent utilisés sont les modèles uni-gramme, bi-gramme et tri-gramme comme suit :

Uni-gramme : $P(s) = \prod_{i=1}^{l} P(m_i)$

Bi-gramme : $P(s) = \prod_{i=1}^{l} P(m_i \mid m_{i-1}) = \prod_{i=1}^{l} \frac{P(m_{i-1} m_i)}{P(m_{i-1})}$

Tri-gramme : $P(s) = \prod_{i=1}^{l} P(m_i \mid m_{i-2} m_{i-1}) = \prod_{i=1}^{l} \frac{P(m_{i-2} m_{i-1} m_i)}{P(m_{i-2} m_{i-1})}$

Ce que l'on doit estimer sont les probabilités :
$P(m_i)$ (un-gramme), $P(m_{i-1}m)$ (bi gramme) et $P(m_{i-2}m_{i-1}\, m_i)$ (tri-gramme)
Pour la langue. Cependant, il est difficile d'estimer ces probabilités pour une langue dans l'absolue. L'estimation ne peut se faire que par rapport à un corpus de textes *C*. Si le corpus est suffisamment grand, on peut faire l'hypothèse qu'il reflète la langue en général. Ainsi, le modèle de langue peut être

approximativement le modèle de langue pour ce corpus – $P(\bullet|C)$. Selon les fréquences d'occurrence d'un n-gramme α, sa probabilité $P(\alpha|C)$ peut être directement estimée comme suit :

$$P(\alpha) = \frac{|\alpha|}{\sum_{\alpha_j \in C} |\alpha_j|} = \frac{|\alpha|}{|C|} \qquad (3)$$

où | α | est la fréquence d'occurrence du n-gramme α dans ce corpus, $\alpha_i$ est un n-gramme de la même longueur que α, et $|C|$ est la taille du corpus (c'est-à-dire le nombre total d'occurrences de mots). Ces estimations sont appelées les estimations de vraisemblance maximale (*Miximum Likelihood*, ou ML). On désignera aussi ces estimations par $P_{ML}$.

Nous donnons ici un exemple simple pour illustrer l'estimation de la probabilité uni-gramme ainsi que son utilisation pour calculer la probabilité d'une phrase.

Supposons un petit corpus contenant 10 mots, avec les fréquences comme montrées dans Table 1

| **Mot** | le | un | prof | ML | dit | aime | de | langue | modèle | RI |
|---|---|---|---|---|---|---|---|---|---|---|
| **Fréq.** | 3 | 2 | 2 | 1 | 2 | 1 | 4 | 2 | 1 | 2 |
| ***P*(•|*C*)** | 0,15 | 0,1 | 0,1 | 0,05 | 0,1 | 0,05 | 0,2 | 0,1 | 0,05 | 0,1 |

Table1. Exemple d'estimation de vraisemblance maximale
Céline Paganelli : la recherche d'information et l'interface humaine p 91

En utilisant l'estimation de vraisemblance maximale, nous obtenons les probabilités comme illustrées dans la table (note : la fréquence totale de mots dans ce corpus est $|C| = 20$).

En utilisant ces probabilités estimées, nous pouvons calculer la probabilité de construire la séquence
*s* = « **le prof aime le ML** » dans cette langue comme suit :

$P(s) \approx P(s|C) = P(\textbf{le}|C)*P(\textbf{prof}|C)*P(\textbf{aime}|C)* P(\textbf{le}|C)* P(\textbf{ML}|C) =$ 0,15*0,1*0,05*0,15*0,05.

# 7-Evaluation de système recherche d'information

L'évaluation des SRI est un problème majeur sur lequel la communauté de la RI a Investi beaucoup d'efforts [C. W. Cleverdon, J. Mills, and M. Keen 1997]
. L'intérêt est de pouvoir comparer des SRI entre eux à l'aide des critères objectifs (les réponses idéales que l'utilisateur souhaite recevoir). Plusieurs quantités mesurables ont été proposées pour l'évaluation d'un SRI : le temps de réponse, la pertinence, la qualité et la présentation des résultats

## 7-1 Hypothèses d'évaluation

Plusieurs hypothèses sont considérées ou faites dans les différentes mesures d'évaluation dont les principales sont [C. W. Cleverdon, J. Mills, and M. Keen 1997] :
• Présentation : les documents sont ordonnés par scores décroissants,
• Ordre de parcours : l'utilisateur parcourt la liste des documents en partant du premier jusqu'au dernier présenté (ne procède jamais de façon aléatoire),

• Jugement absolu : un document reste pertinent même s'il contient exactement la même information qu'un autre document déjà présenté à l'utilisateur,
• Le non additivité : deux documents non pertinents ne pourront jamais former une unité d'information pertinente,
• Pertinence binaire : un jugement de pertinence doit pouvoir se ramener au mieux à un nombre réel généralement borné. Cette hypothèse est réduite à un jugement de pertinence binaire.

## 7-2 Mesures d'évaluation

Les mesures les plus courantes pour mesurer la performance d'un SRI sont le temps et l'espace. Les plus rapides en temps de réponse et les moins gourmands en espace utilisé sont les meilleurs SRI. Dans le cadre des SRI, on s'intéresse plutôt aux résultats retournés par ce dernier, sans négliger les deux premiers critères de performance. L'évaluation d'un SRI se mesure indépendamment de la méthode d'indexation ou du modèle qu'il l'implante. Pour cela, ces techniques s'appuient essentiellement sur l'estimation de la qualité des informations retrouvées par le SRI.
Ils existent plusieurs facteurs d'évaluation, les deux principaux facteurs permettant d'évaluer un SRI sont le rappel et la précision. Le rappel mesure la capacité du système à retrouver tous les documents pertinents et la précision mesure son habilité à ne retrouver que des documents pertinents. Ces mesures

peuvent être quantifiées en pourcentage ou par des valeurs entre 0 et 1.
On désigne par R l'ensemble des documents pertinents et par L l'ensemble des documents retrouvés.

### 7-2-1 Le rappel

Le rappel mesure la capacité du SRI à sélectionner tous les documents pertinents. Il donne une indication sur le nombre de documents pertinents trouvés par rapport au nombre total de documents pertinents pour la requête. La valeur de rappel est entre 0 et 1 :

$$\text{rappel} = \frac{|R \cap L|}{|R|}$$

Le rappel mesure aussi la probabilité qu'un document d soit sélectionné sachant qu'il est pertinent :

$$\text{rappel} = P(d \in L | d \in R)$$

### 7-2-2 La précision

La précision mesure la capacité du système à rejeter tous les documents non pertinents. Elle donne une indication sur la proportion des documents pertinents renvoyés par le SRI. La précision est d'une part un indicateur pour la qualité du résultat à la demande de la recherche, et d'autre part, elle sert à ne pas distribuer des documents non pertinents afin de ne pas saturer la capacité d'un système. La valeur de la précision est entre 0 et 1. La précision se calcule alors par :

$$\text{précision} = \frac{|R \cap L|}{|L|}$$

La précision mesure aussi la probabilité qu'un document d soit pertinent sachant qu'il est sélectionné :
:

$$\text{précision} = P(d \in R | d \in L)$$

Le SRI n'a pas à décider de la pertinence d'un document : il range juste les documents de la collection selon leurs potentiels de répondre au besoin en information de l'utilisateur. La mesure de précision et de rappel ne peuvent alors avoir un sens que si l'on considère les x premiers documents de la liste ordonnée. La précision et le rappel ne sont alors que des mesures relatives de performance

**La précision exacte ou la R-précision** La précision à x documents est souvent reliée à ce que l'on appelle la précision exacte ou la R-précision. La précision exacte représente celle obtenue à l'endroit où elle vaut le rappel. Si la requête admet x documents pertinents, la précision exacte est celle calculée pour les x premiers documents de la liste ordonnée des documents restitués (top x)

**La précision moyenne** On peut faire bouillir les valeurs informant sur la précision d'un SRI à quelques chiffres ou même à un seul chiffre. La précision moyenne consiste à interpoler la précision mesurée à différentes positions, nous calculons alors la moyenne arithmétique des valeurs interpolées de la précision. La précision moyenne est la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée. Elle tient compte à la fois de la précision et du rappel. En effet, si un document est pertinent et apparaît loin dans la liste, la précision à ce document et par conséquent la précision moyenne, tend à devenir nulle.

$$\text{précision moyenne} = \sum_{d \in R} \frac{\text{précision}(d)}{|R|}$$

où précision(d) est la précision calculée sur la base de tous les documents classés avant le document d :

$$\text{précision}(d) \models \frac{|\{d' \in R, rsv(d', q) \geq rsv(d, q)\}|}{|\{d', rsv(d', q) \geq rsv(d, q)\}|}$$

Idéalement, on voudrait qu'un système donne de bons taux de précision et de rappel en même temps. Un système qui aurait 100% pour la précision et pour le rappel signifie qu'il trouve tous les documents pertinents, et rien que les documents pertinents. Cela veut dire que les réponses du système à chaque requête sont constituées de tous et seulement les documents idéaux que l'utilisateur a identifiés. En pratique, cette situation n'arrive pas. Plus souvent, on peut obtenir un taux de précision et de rappel aux alentours de 30%.

Les deux métriques ne sont pas indépendantes. Il y a une forte relation entre elles: quand l'une augmente, l'autre diminue. Il ne signifie rien de parler de la qualité d'un système en utilisant seulement une des métriques. En effet, il est facile d'avoir 100% de rappel: il suffirait de donner toute la base comme la réponse à chaque requête. Cependant, la précision dans ce cas-ci serait très basse. De même, on peut augmenter la précision en donnant très peu de documents en réponse, mais le rappel souffrira. Il faut donc utiliser les deux métriques ensemble.

Les mesures de précision-rappel ne sont pas statiques non plus (c'est-à-dire qu'un système n'a pas qu'une mesure de précision et de rappel). Le comportement d'un système peut varier en faveur de précision ou en faveur de rappel (en détriment de l'autre métrique). Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme suivante:

### Comment évaluer Précision-Rappel?

La liste de réponses d'un système pour une requête peut varier en longueur. Une longue liste correspond à un taux de rappel élevé, mais un taux de précision assez basse, tandis qu'une liste courte représente le contraire. La longueur de la liste n'est souvent pas un paramètre inhérent d'un système. On peut très bien le modifier selon le besoin. Mais cette modification ne modifie pas le comportement global du système et de sa qualité. Ainsi, on peut varier cette longueur pour estimer les différents points de précision-rappel pour constituer une courbe de précision-rappel pour le système. Le processus d'évaluation est donc comme suit[C. W. Cleverdon, J. Mills, and M. Keen 1997] :
Pour i = 1, 2, … #document_dans_la_base faire:

- évaluer la précision et le rappel pour les i premier documents dans la liste de réponses du système

Par exemple, soit une requête qui a en tout 5 documents pertinents dans la base. La liste de réponse du système à cette requête est comme suit:

| **Liste de réponses** | **Pertinence** |
|---|---|
| Doc1 | (*) |
| Doc2 | |
| Doc3 | (*) |
| Doc4 | (*) |
| Doc5 | |
| … | |

Où (*) signifie que c'est un document pertinent (selon l'évaluation de l'utilisateur). On considère d'abord le premier document Doc1 comme la réponse du système. À ce point, on a retrouvé un document pertinent parmi les 5 existants. Donc on a un taux de rappel de 0.2. La précision est 1/1. Le point de la courbe est (0.2, 1.0).
On considère ensuite les deux premiers documents comme la réponse
(Doc1 et Doc2). À ce point, on a le même rappel (toujours 1/5), mais la précision devient 1/2. Ainsi le point est (0.2, 0.5).
On considère Doc1, Doc2 et Doc3, on a un rappel de 2/5, et une précision de 2/3: (0.4, 0.67).

Ce processus est continué jusqu'à l'épuisement de toute la liste de réponse du système (qui peut être très longue, jusqu'à inclure tous les documents de la base). Les premiers points de la courbe est comme dans la figure suivante:

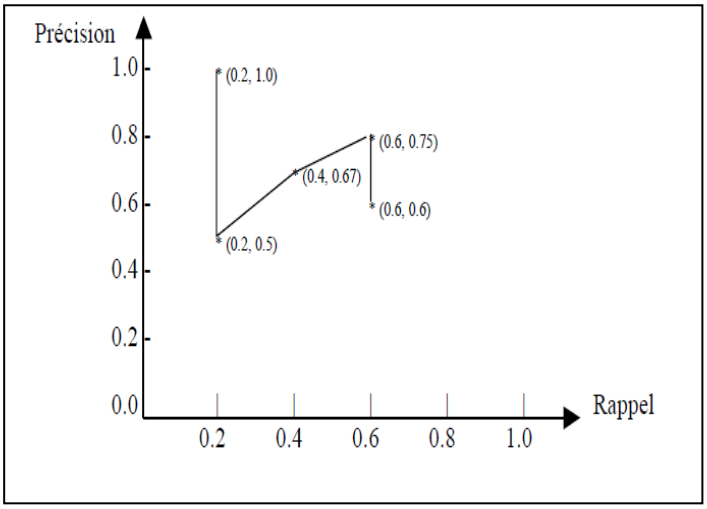



**Figure 10 : exemple d'évaluation Précision-Rappel**
Céline Paganelli : la recherche d'information et l'interface humaine p 95

Cette courbe ne correspond pas tout à fait à la forme générale. Mais c'est juste pour une seule requête. Si on calcule la moyenne sur un ensemble de requêtes, la courbe sera plus lisse, et ressemble davantage à la forme générale. Il arrive fréquemment qu'on applique la *interpolation* sur la courbe de chaque requête. La polarisation vise à créer une courbe qui descend (comme la forme générale). Le traitement est le suivant: Soit i et j deux points de rappel, et i<j. Si au point i, la précision est inférieure à la précision au point j, alors, on augmente la précision du point i à celle du point j. Concrètement, cela signifie qu'on remplit un creux

de la courbe par une ligne horizontale comme suit:

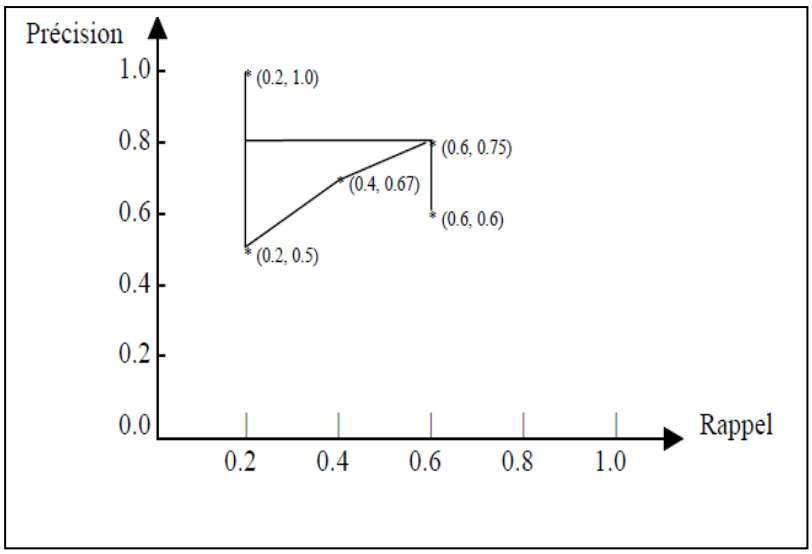



**Figure 11 : exemple d'évaluation Précision-Rappel**
Céline Paganelli : la recherche d'information et l'interface humaine p 95

On obtient donc une courbe en forme d'escalier. L'idée qui motive la polarisation est que les creux de la courbe ne représentent pas vraiment la performance du système. S'il existe un point à un rappel et une précision plus élevés, on peut toujours donner plus de documents dans la réponse pour augmenter la performance. Donc, le creux est surmontable.

Évidemment, on peut discuter sur cette motivation, et être en désaccord. Ce n'est pas important. L'important est qu'on compare les systèmes sur la même base. Si tous les systèmes sont mesurés avec une courbe polarisée, alors la polarisation ne donne pas plus d'avantage à un système qu'à un autre. Donc, la courbe polarisée est aussi une base équitable pour comparer des systèmes.

## Comparaison de systèmes et Précision moyenne

Si on veut comparer deux systèmes de RI, il faut les tester avec le même corpus de test (ou plusieurs corpus de test). Un système dont la courbe dépasse (c'est-à-dire qu'elle se situe en haut à droite de) celle d'un autre est considéré comme un meilleur système.

Il arrive parfois que les deux courbes se croisent. Dans ce cas, il est difficile de dire quel système est meilleur. Pour résoudre ce problème, on utilise aussi la *précision moyenne* comme une mesure de performance. La précision moyenne est une moyenne de précision sur un ensemble de points de rappel. On utilise soit la précision moyenne sur 10 points de rappel (0.1, …, 1.0), ou celle sur 11 points de rappel (0.0, 0.1, …, 1.0). Cette dernière est possible seulement avec la polarisation.

La précision moyenne décrit bien la performance d'un système. C'est la mesure souvent utilisée en RI.

Pour comparer deux systèmes ou deux méthodes, on utilise souvent l'amélioration relative qui est calculée comme suit :
Amélioration de méthode 2 sur méthode 1 = (performance de méthode 2 – performance de méthode 1) / performance de méthode 1.

## 6-3 TREC

TREC est un projet international qui a été lancé en 1992 par le NIST (National Institute of Standards and Technology) aux Etats-Unis. Il est aujourd'hui co-sponsorisé par le NIST et DARPA/ITO (Defense Advanced Research Projects Agency – Information Technology Office). Son objectif est de proposer un standard pour comparer les différents modèles de RI, indépendamment de la méthode de l'indexation ou bien du modèle qu'ils implémentent. Afin de mesurer l'efficacité des SRI de manière standard, il suffit de fournir une collection de test, des requêtes et les jugements à ces requêtes.

### *6-3-1 Collection de test*

Dans ce cadre, il y a eu de nombreuses pistes sur une gamme d'essais différents, le plus connu est TREC ad hoc. Les documents sont représentés par du texte intégral, ils sont faiblement structurés (titre, paragraphe. . .), ces documents sont sous forme électronique. La plupart de ces documents possèdent une longueur entre 300 mots (documents courts) et 3000 mots (documents longs).

### 6-3-2 Collection d'entraînement

Afin de caractériser l'habilité des systèmes à s'adapter aux goûts des utilisateurs, une gamme d'essais est également fournie pour l'entraînement. Généralement, elle est de très petite taille comparée à celle de test (de l'ordre de 10000 documents contre plus que 100000 pour le test).

### 6-3-3 Les requêtes (Topics)

Les requêtes représentent le besoin en information de l'utilisateur, mais dans TREC elle représente aussi un moyen pour évaluer un SRI. La forme des requêtes a notamment évolué. Au début la forme était très structurée, elle comportait des champs permettant la structuration des requêtes (un titre, le thème, une description et l'objet de la recherche). Pour alléger la forme des requêtes, seuls le titre et une description très brève (d'une phrase ou deux) sont conservés.

### 6-3-4 Les jugements de pertinence

Parce que les collections de test sont si volumineuses, les jugements ont été recueillies uniquement pour les documents qui ont été classés parmi les premiers. La pertinence d'un document pour une requête est codée par une valeur binaire (pertinent ou non).L'affectation des valeurs de pertinence pour chaque document se fait donc par une personne, cette étape est très pénible vu le nombre important de documents dans la collection.

# 8- Comment fournir une information « ciblée » en entreprise ?

Ou comment transférer aux logiciels, les fonctions de reformulation des requêtes, de sélection des documents pertinents, et de classement des résultats, exercées autrefois par le documentaliste ?

**Deux axes sont à prendre en compte :**

## 8-1 Les utilisateurs et l'activité de l'entreprise

« La difficulté, pour une unité d'information, d'appréhender l'adéquation entre l'information qu'elle collecte et organise, et la satisfaction de ses clients, est qu'une bonne part d'aléatoire entre dans l'appréciation de la situation de ceux-ci. Nous avons vu que l'information n'est pas but un en-soi, mais qu'elle dépend de la personne qui en a besoin et de l'exploitation qu'elle en fera dans le cadre de son travail. Il faut donc en permanence tenir compte de la situation et de l'activité des personnes susceptibles d'avoir besoin d'information déjà collectée ou à collecter. » [GUYOT 2005]

# Partie 01 : Etat de l'Art

L'information est devenue un élément décisionnel incontournable pour le personnel des entreprises. Mais, comment concevoir des systèmes de recherches d'information (et par conséquent d'indexation) adaptés à des cultures professionnelles différentes, cohabitant dans une même structure ? En effet, un commercial et un technicien appartenant à la même entreprise n'ont pas besoin de la même information. Par ailleurs, la fréquence et la forme sous laquelle est fournie cette information est conditionnée par le rythme et le cadre de leur activité professionnelle.

Il convient, avant tout, de réaliser un « référentiel culturel » des métiers de l'entreprise ou du service concerné. Pour mener à bien cette démarche, Eric Sutter propose d'envisager la question selon trois angles principaux **[**SUTTER p 88 ]

• La nature des activités des professionnels de l'entreprise en termes de finalité.

C'est à dire comprendre les objectifs et les missions confiés à chaque corps de métier, afin d'identifier les informations qui leurs seront utiles.

• La nature des besoins et des usages de l'information demandée.

Savoir comment et dans quel cadre l'information sera utilisée permet de circonscrire le domaine de recherche :

- ❖ informations internes : conditions tarifaires, état des stocks, etc.…
- ❖ informations externes : étude de marché, veille concurrentielle

• Les « us et coutumes » de chaque corps de métiers

Chaque profession a ses particularités, qu'il s'agisse des conditions de travail (en équipe ou seul), du lieu (fixe ou itinérant), du rythme (de nuit, de jour ou décalé). Tous ces éléments ont des répercussions sur le temps que les employés peuvent consacrer à la recherche d'information ou même simplement à sa prise de connaissance. Il est donc nécessaire pour le documentaliste de bien étudier les habitudes de chaque profession afin d'adapter au mieux la forme sous laquelle il transmet l'information aux utilisateurs.
La réalisation de ce « référentiel » peut prendre plusieurs formes. Le sujet de ce mémoire n'étant pas axé sur ce point, je me contenterai de les évoquer rapidement, sans développer plus avant les techniques nécessaires à leur mise en œuvre.

De même qu'on procède à une enquête de « besoins » ou « d'usages » sur l'ensemble d'un service pour définir les stratégies documentaires utiles, le documentaliste peut recourir à ce type de démarche pour analyser les méthodes de travail et les processus de recherche d'un groupe de salariés exerçant le même métier (ou les mêmes fonctions). Cette enquête prend en général la forme d'un questionnaire ou d'une série d'entretiens semi - directifs menés avec les salariés Toutefois, cette méthode a l'inconvénient d'être longue et coûteuse à mettre en place, ce qui la rend statique et difficile à mettre à jour.

Dans d'autres cas, le salarié peut configurer son profil en remplissant lui-même un formulaire. Cette méthode est plus souple mais elle ne donne qu'une vision interne de la démarche de l'utilisateur. Ce dernier manquant parfois de recul par rapport à leur propre fonction.

En dernier lieu, les spécialistes de l'information peuvent avoir recours à des profils générés dynamiquement. Un sous- système de modélisation observe l'utilisateur de derrière l'interface et découvre son profil à partir de ses actions. Cette méthode permet d'observer le cheminement de sa recherche sur le web (sur les sites Internet et sur le portail de l'entreprise), et elle corrige les caractéristiques de son profil au fur et à mesure de leurs évolutions. Ces dernières sont enregistrées dans une liste de variables, qui sont ensuite réutilisées par les logiciels, pour personnaliser les thématiques de classement des documents.

## 8-2 Le contenu (les documents à organiser)

« Une inflation galopante des nouveaux textes [qui] rendront de plus en plus évident le besoin d'outils pour guider la recherche et décanter les données » [**MANIEZ 2002]**

L'un des problèmes récurrents dans les grandes entreprises tient à la coexistence de plusieurs systèmes de gestions de contenus. Cette situation est le plus souvent la résultante de campagnes d'informatisation successives, parfois menées dans l'urgence afin de rattraper un retard technologique handicapant pour la survie de l'entreprise.

Un nouveau système implanté depuis peu, se présente comme un concurrent de taille en ce qui concerne l'accès aux documents : le portail Intralignes et son moteur de recherche Verity. Ce système de gestion de contenu, propose aussi de nombreux documents HTML créés spécialement pour les sites. Il s'agit pour la

plupart d'articles issus des journaux électroniques consultables sur les portails métiers.

On constate aisément à quel point ces situations engendrent un foisonnement de fichiers électroniques de nature extrêmement variée : fichiers en format Word, Excel, PDF, RTF, format propriétaire Lotus, HTML, etc.…Or, lors d'un processus d'indexation automatique, le format d'un document peut jouer un rôle capital. En effet, les documents « numérisés » (cf.p.47) peuvent bénéficier d'une description de leur contenu à l'aide de balises intégrées dans la structure du document, c'est le cas des fichiers HTML, qui comportent des « métadonnées » invisibles pour le lecteur, mais lisibles par les moteurs de recherche. Mais, encore faut-il que les utilisateurs soient sensibilisés et formés à l'utilisation de ce langage…..
Néanmoins, l'avenir de l'indexation sur les portails web passe par ces techniques de représentation du contenu des documents. Celles qui offrent aujourd'hui le plus d'avenir, sont basées sur le langage XML (eXtensible Markup Langage ou Langage Extensible de Balisage).

« Comme HTML (Hypertext Markup Language) c'est un langage de balisage (markup), c'est-à-dire un langage qui présente de l'information encadrée par des balises. Mais contrairement à HTML, qui présente un jeu limité de balises orientées présentation (titre, paragraphe, image, lien hypertexte, etc.), XML est un métalangage, qui va permettre d'inventer à volonté de nouvelles balises pour isoler toutes les informations élémentaires (titre d'ouvrage, prix d'article, numéro de sécurité sociale, référence de pièce…) ou agrégats d'informations élémentaires, que peut contenir une page Web ».

Ce système de balisage est à la base des langages RDF (Ressource Description Framework) qui exprime des faits à l'aide de triplet d'URI (Uniform Resource Identifiers), et OWL (Ontology Web Language) qui fournit une syntaxe pour exprimer des relations logiques de type : union, intersection, inverse de, etc.… La combinaison de ces deux langages constituent les bases du web sémantique, et contribuent à aider les ordinateurs à mieux comprendre le sens des informations qu'ils traitent.

En résumé, le langage XML permet de structurer le contenu sémantique des documents dès leur conception. Toutefois, l'utilisation systématique de cette technique n'est pas encore à l'ordre du jour dans les grandes entreprises.

## 8-3 Quels sont les clés de succès pour réussir son projet de recherche d'information interne ?

La recherche d'information en entreprise est une problématique complexe à la croisée de nombreuses disciplines et préoccupations. Celle-ci repose sur trois éléments fondamentaux qui constituent les véritables piliers de la recherche d'information en entreprise : les individus et leurs besoins d'accéder à une information pertinente, l'information qui est au cœur de ces besoins, et les technologies qui vont orchestrer le processus de recherche d'information. [Gil almisse ]

Il faut donc tout d'abord bien comprendre les besoins des individus en segmentant les différentes populations qui peuvent être notamment les responsables de contenus qui veulent valoriser leurs informations, les gestionnaires de l'information qui tentent d'avoir une vision globale des flux d'information dans l'entreprise et les experts sui souhaitent valoriser leur expertise.

Ensuite, dans la mesure où il ne peut pas y avoir de bonne recherche d'information sans une bonne gestion voire gouvernance de l'information, il convient de mener un audit de l'information disponible et de leurs métadonnées. D'autres bonnes pratiques seront également recommandées comme l'indexation de l'information uniquement actualisée et valide ou encore la gestion homogène et transversale des droits d'accès. Jouant un rôle de médiateur entre les individus et les informaticiens, il faudra enfin veiller à choisir des solutions de recherche d'information permettant de faciliter à la fois la mise en contexte de l'information et des interactions sociales qui ne devront pas se faire au détriment de l'interface utilisateur.

# 9 - Les portails et les moteurs de recherche

## 9-1 Qu'est ce qu'un portail ?

### *Définition*

« Un portail Internet [ou intranet] est un site qui offre la possibilité de disposer, à partir d'un point d'accès unique, de données issues de sources multiples. Il permet d'accéder rapidement à une information qualifiée, organisée et structurée, personnalisée en fonction des centres d'intérêt des usagers ». [CHARVET p 70 ]

Avec le succès d'Internet et la multiplication exponentielle des sites Web, il s'est rapidement avéré nécessaire d'organiser l'accès aux informations et aux services proposés sur la toile. Ce rôle a été dévolu aux portails qui ont du s'adapter à la demande du public en fournissant des services de plus en plus spécialisés. Par conséquent, nous avons vu apparaître plusieurs familles de portails aux fonctionnalités bien définies.

## 9-2 Typologie des portails

### 9-2-1 Les portails généralistes

Ils sont destinés au « grand public », c'est à dire à une multitude de « profils » utilisateurs, qui représentent autant de centres d'intérêt. Le rôle du portail est de faciliter et d'organiser l'accès à ces contenus informationnels, de manière à ce que tous les publics puissent localiser les informations qu'ils recherchent dans les meilleurs délais. Par conséquent, ces portails sont le plus souvent munis d'un annuaire de sites (organisé selon une classification thématique « généraliste ») et un moteur de recherche en texte intégral du type Google. En parallèle, ils proposent un certain nombre de service (météo, flash actualités, programme culturel de la semaine, etc.…) destinés à fidéliser les utilisateurs potentiels, qui pourront ensuite, faire de ce portail leur page d'accès par défaut au réseau. Free et Yahoo font partis de cette famille de portail, fournisseurs d'accès à Internet, ils proposent de surcroît des accès sécurisés à des boites de messagerie personnelle, des « lieux » d'échanges communautaires (Yahoo group) et des espaces de publications web personnel (pagesperso, blog). La tendance actuelle est la personnalisation de ces portails, l'utilisateur à la possibilité de « customiser » la page d'accès en sélectionnant les rubriques et services qui correspondent à ses besoins.

### 9-2-2 Les portails thématiques ou spécialisés

Fondamentalement, leur principe de fonctionnement est le même que celui des portails généralistes, pour ce qui est de l'accès aux informations. En revanche, ils ne proposent pas ou rarement des services du type messagerie ou météo, mais plutôt des prestations en corrélation avec la spécialisation et les attentes de leur public : newsletter, forums de discussions en rapport avec la spécificité du site, offres promotionnelles, etc.…En effet, suivant que le site est destiné aux professionnels , au grand public (portail sur le cinéma, le droit, le sport, etc.…) ou à un public d'amateurs avertis (portail du surf, portail des métiers d'art), les services proposés et leur mode de diffusion seront différents. [ACCART p 155]

### 9-2-3 Les portails des portails

La multiplication des portails spécialisés a naturellement appelé la création de « super » portail destiné à organiser et mutualiser les adresses des autres portails présent sur le web, au sein d'un annuaire ou par l'intermédiaire d'un moteur de recherche. Cette tendance se retrouve sur les intranets des grands groupes internationaux.

### 9-2-4 Le portail d'entreprise

Il peut être une simple « vitrine » proposant la carte d'identité de la société ou un véritable lieu d'échanges, où l'internaute peut procéder à des achats en ligne, consulter le catalogue de produits ou les archives ouvertes de l'organisation, dialoguer avec certains services (après-vente), ou accéder aux offres de recrutement. Certains de ces portails offrent même un accès sécurisé au réseau intranet du groupe, pour le personnel habilité. C'est le cas d'Air France, ce qui est particulièrement utile aux PNC et PNT21 qui ne disposent pas de bureau au sein de l'entreprise de pouvoir se connecter de chez eux ou lors des escales, au réseau via  De même qu'il existe une typologie des portails Internet, il existe aussi une typologie des réseaux intranet. Elle tient principalement à la nature et aux fonctions que chaque organisations leurs assignent,

## 9-3 Les portails Intranet des entreprises : des informations variées destinées à un public hétérogène

Contrairement aux portails Internet qui donne accès à un stock de documents ouvert, constamment renouvelé, destinés à un public non défini, les portails Intranet sont destinés à gérer des lots de documents validés, correspondant aux activités de l'entreprise et consultables par un public professionnel.

Toutefois, la cohabitation de plusieurs corps de métiers au sein d'une même entreprise, implique des besoins informationnels et fonctionnels différents, ce qui a entraîné la création de trois grandes « familles » de réseau intranet.

Les réseaux Intranet répondent à 3 grandes exigences des entreprises :
-La transversalité de l'information
-La mutualisation des applications informatiques
-L'homogénéisation des accès à l'information et aux outils applicatifs

Selon l'activité, la taille et les infrastructures de l'entreprise déjà en place, certaines de ces exigences prendront le pas sur d'autres, et détermineront la nature du réseau mis en place :

## 9-3-1 Un intranet documentaire :

Ils sont destinés à organiser la production, la publication et l'administration des collections documentaires. La qualité de ce type de réseau réside dans l'efficacité et la qualité du référencement des documents et des modes de recherches proposés aux utilisateurs (taxinomies, recherche en texte intégral, interface personnalisée, glossaire métier, etc.). Ils sont tout particulièrement indiqués pour les entreprises où cohabitent plusieurs directions indépendantes et dispersées géographiquement, ce qui induit une redondance des bases documentaires et une mauvaise connaissance des documents produits par chaque entité. La réussite de ce type d'intranet repose sur la mise en place de processus à la fois souples et rigoureux, permettant de « valider » la production et la publication des documents (workflow). Dans la phase de production, il est capital de définir les critères qui permettront d'identifier chaque document dès leur création et de les gérer tout au long de leur vie au sein de l'entreprise : auteur, mot-clés, durée de vie, n° d'archivage, N° de la version, nom du destinataire etc.….Ces données peuvent soient être saisies dans une base de données (logiciel de Geide) et/ou être intégrées directement dans le document (XML : eXtensible Markup Language).

## 9-3-2 Un intranet applicatif

« La migration d'applications centrales ou client/serveur traditionnelles vers un système intranet se justifie donc souvent car cela permet d'une part de profiter des avantages liés au déploiement et d'autres part de bénéficier d'un niveau de standardisation ouvrant sur l'interconnexion de services applicatifs » [ALERI, 1998] L'intérêt de ce type d'intranet est de proposer un accès centralisé à des outils ou à des sources d'informations produites par un service mais pouvant être utile à tout ou partie des salariés de l'entreprise : planning du personnel, formulaire de congés, banques de données internes, offres de mobilité interne,

veille stratégique et économique, règlement interne, outils de workflow, groupware, etc.…

## 9-3-3 Un intranet d'intégration

Il est destiné à « fédérer les applications hétérogènes existantes, par l'intégration des progiciels et des développements spécifiques dans un environnement graphiques homogène » [ALERI, 1998]

Cette démarche permet d'offrir aux usagers un accès homogène aux applications proposées par l'entreprise, quel que soit le point d'accès utilisé par le salarié. Son profil, géré à partir de l'annuaire d'entreprise et identifiable grâce à système d'accès sécurisé, lui permet de bénéficier d'un « bureau », toujours identique, quel que soit le lieu où l'ordinateur qu'il utilise. Il offre en outre l'avantage de faciliter la gestion et la mise à jour des logiciels informatique (nouvelles versions de Word, mise à jour des anti-virus, etc.…). Ces applications étant hébergées dans un lieu unique, le serveur, leurs « maintenance » n'imposent une intervention individuelle sur chaque poste de l'entreprise.

A un moment ou un autre, tous ces besoins organisationnels ont été successivement ou simultanément ressentis au sein des grandes entreprises.
Au point de provoquer un développement anarchique des bases de données et autres outils applicatifs, sur les réseaux Internes des sociétés. Face à cette situation, les entreprises ont adopté le principe du portail, pour fédérer et structurer l'accès aux sources informationnelles publiées sur ces réseaux. Mais sans de bons outils de recherche, adaptés à la structure fonctionnelle de l'entreprise et aux besoins des utilisateurs, un portail devient une simple page d'accueil enrichie de quelques services, rien de plus. C'est pourquoi, ce type de CMS est systématiquement fourni avec un ou plusieurs systèmes de recherches complémentaires, le plus répandu étant le moteur de recherche en texte intégral.

# Conclusion

Dans ce chapitre nous avons présenté le processus de RI traditionnelle, et les modèles utilisés pour construire un SRI, ainsi que les facteurs d'évaluation pour comparer les différents systèmes, qui n'exploitent que le contenu sémantique des documents. L'apparition des ordinateurs a joué un rôle central dans le développement des SRI. Différentes tâches qui étaient manuelles ont été automatisées. L'utilisation massive de larges ressources d'information rend les techniques classiques incontournables pour la RI. Une des conséquences de cette évolution a été la structuration de l'information. Actuellement, Internet a favorisé l'essor des documents, par la structuration des documents par des liens entre eux et par la structure interne. Les SRI actuels s'intéressent à cette évolution. Nous réalisons dans la partie suivante une enquête sur la recherche d'information dans les entreprise algérienne dans l'objectif de savoir Dune part, le niveau d'introduction des systèmes de recherche d'information dans l'entreprise algérienne. D'autre part, la faisabilité, la stratégie et les outils pour construire un système de recherche d'information

Partie 02 : enquête et réalité de la RI en Algérie

# Introduction

L'enquête a été réalisée au moyen d'un questionnaire et d'entretiens, auprès d'un échantillon d'entreprises. Une grille d'analyse a permis de synthétiser les réponses et de dégager quelques résultats. Les observations et analyses sont effectuées pour découvrir :

- Dune part, le niveau d'introduction des systèmes de recherche d'information dans l'entreprise algérienne
- D'autre part, la faisabilité, la stratégie et les outils pour construire un système de recherche d'information.

## La méthode utilisée : une approche qualitative

L'approche choisie est une approche de type qualitatif plutôt que quantitatif : les résultats obtenus ne peuvent donc être retenus comme représentatifs du monde de l'entreprise aujourd'hui. Ils vont servir à illustrer un certain nombre d'hypothèses de travail par rapport à la recherche d'information
La méthode de travail employée est la méthode de l'enquête sur entretien avec un questionnaire. L'objet de notre enquête, les dispositifs de capitalisation des connaissances, concerne un nombre limité d'entreprises (dix-huit entreprises) et cela peut paraître limitatif. Cependant l'échantillon est assez large puisque sont représentées des petites et moyennes entreprises (PME-PMI), ainsi que des très grandes entreprises, dont certaines sont des filiales de groupes internationaux

## Les questions posées

En préalable à notre enquête, un questionnaire a été élaborée afin de pouvoir diriger l'étude et analyser les observations. Ce questionnaire porte essentiellement sur les points suivants :

- La nature de l'entreprise
- L'effectif de l'entreprise
- le degré de numérisation de document en entreprise
- les moyennes de partages de l'information
- la caractéristique de réseau utilisé
- la stratégie de sécurité de l'information
- la quantité ou le nombre de documents disponibles dans l'entreprise
- le degré de consultation des documents
- degré de difficulté pour la consultation de documents

En résumé, les questions posées cherchent à :

- analyser la situation de la documentation dans l'entreprise,
- les moyens de partage de l'information,
- analyser pour savoir s'il est opportun de mettre en place un système de recherche d'information

## Les entreprises étudiées :

Cette enquête touche les petites entreprises ainsi que les grandes entreprises (une entreprise comptant plus de 1000 salariés), des moyennes entreprises (cinq entreprises), mais surtout des petites entreprises (onze entre 20 et 100 salariés). Malgré un nombre relativement restreint d'entreprises prises en compte dans cette enquête, l'éventail est cependant relativement large et va permettre de dégager un certain nombre de tendances et de constats.

## 1- l'analyse des donnes et l'affichage des résultats :

Ont été distribues des questionnaires a 18 entreprises et on n a alors analysé les résultats des questionnaires

## Quelques caractéristiques de l'échantillon:

- Sur les 18 entreprises on constate que 13 entreprises sont des entreprises publiques représentant un taux de 72.20%.

- Nous remarquons que 13 entreprises possèdent des documents non numérisés et 5 entreprises seulement ont une documentation numérisée en partie à cause de l'archivage électronique.

- Nous avons constaté que 14 entreprises possèdent un réseau local de différentes caractéristiques et uniquement 04 entreprises sont dépourvues de réseaux.

- Les différentes caractéristiques des réseaux disponibles :

| **Les caractéristiques des réseaux** | **Effectifs** | **Pourcentage** |
|---|---|---|
| Nombre d'entreprises qui n'ont pas de réseaux | 4 | 22,2 |
| Intranet | 9 | 33,3 |
| intranet plus réseau WIFI, serveur HP ML 350 | 1 | 5,6 |
| intranet, salon de support de transmission, fibre optique, type LAN WIFI WAN application | 1 | 5,6 |
| réseau sous forme de grappes pour partage de fichier et impression | 1 | 5,6 |
| serveur pour partage le fichier et l'impression | 1 | 5,6 |
| un réseau local, un serveur de fichier FTP, une connexion VPN | 1 | 5,6 |
| Total | 18 | 100,0 |

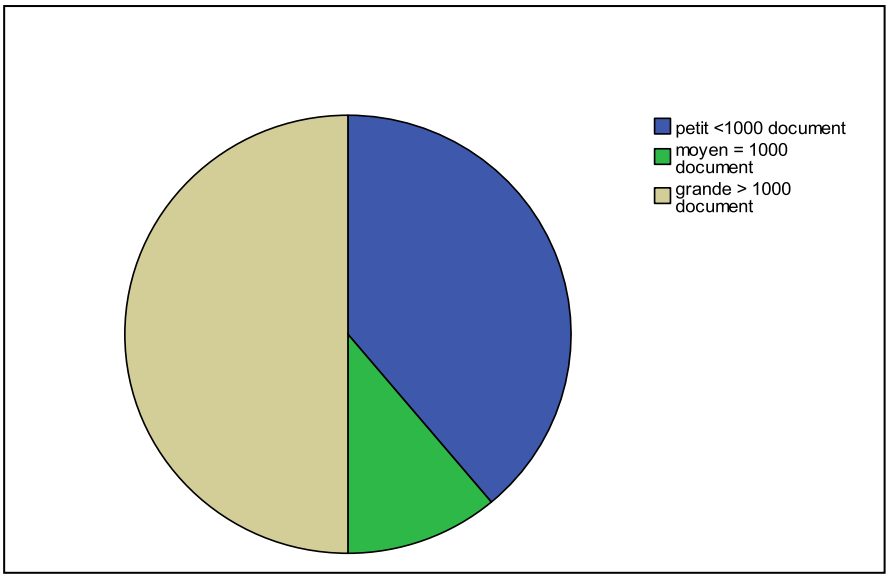
- pour la stratégie de sécurité d'information on a trouvé que 14 entreprises appliquent différentes stratégies de sécurité de l'information les 04 restantes n'en appliquent aucune à cause de la non disponibilité de réseau.

- différentes stratégies de sécurité de l'information :

| **Les stratégies de sécurité de l'information** | **Nombre** | **Pourcentage** |
|---|---|---|
| Entreprises n'ayant aucune stratégie | 4 | 22,2 |
| antivirus pour la sécurité de fichiers, cd sauvegarde périodique pour la sécurité d'information, | 1 | 5,6 |
| fire well | 3 | 5,6 |
| les mots de passe+identifiant | 5 | 5,6 |
| TCP/IP | 5 | 5,6 |
| Total | 18 | 100,0 |

- le nombre de documents traités :

| **Volume documentaire** | **nombre** | **Pourcentage** |
|---|---|---|
| petit <1000 document | 7 | 38,9 |
| moyen = 1000 document | 2 | 11,1 |
| grande > 1000 document | 9 | 50,0 |
| Total | 18 | 100,0 |




**Le nombre de documents utilise**

- 15 entreprises ont estimé que leur flux est important et seulement 03 ont jugé que le flux informationnel est dérisoire

- Disponibilité d'information en entreprise :

| disponibilité | nombre | Pourcentage |
|---|---|---|
| souvent | 11 | 61,1 |
| quelques fois | 7 | 38,9 |
| Total | 18 | 100,0 |

-Seulement 03 entreprises estiment qu'ils ont un taux de consultation des documents élevé. Les 15 autres déclarent que la consultation est aléatoire.

- 07 entreprises ont déclaré ne pas avoir de difficulté pour l'accès aux documents stockés par contre les 11 restantes estiment qu'ils en ont.

- causes de difficultés d'accès aux documents

| Causes de difficultés d'accès | Nombre | Pourcentage |
|---|---|---|
| Importance du nombre de documents | 7 | 5,6 |
| Absence d'application automatisée de gestion des archives | 1 | 5,6 |
| Difficulté de la recherche manuelle | 1 | 5,6 |
| Archivage hasardeux des documents à consulter | 1 | 5,6 |
| les documents ne sont pas numérisés et traité de manière automatique | 1 | 5,6 |
| Total | 18 | 100,0 |

-15 entreprises souhaitent la mise en place d'un système de recherche d'information.

# Partie 02 : enquête et réalité de la RI en Algérie

## 2-1 Analyse des causes de souhait de mise en place d'un SRI

### Les caractéristiques de ces entreprises :

La nature de l'activité :

Ce sont les entreprises de production de bien et services et les entreprises commerciales qui sont les plus demandeuses de la mise en place d'un SRI

Les moyens de partage de l'information :

Il y a une utilisation abusive de la communication sur support papier malgré la disponibilité de moyens de communication numériques. Le souhait des entreprises est d'utiliser plus efficacement les moyens numériques de communication en vue de diminuer la lourdeur de la communication classique (papier) tout en souhaitant la mise en place de moyen adéquats de sécurisation de l'information
La communication orale qui, à notre sens restera toujours présente car elle complète la communication écrite, doit être prise en compte dans l'historisation de l'information dans l'entreprise. Les moyens à partir de rapports de réunion jusqu'à la visio-conférence sont souhaitable pour ne pas perdre les informations

Le volume des documents disponibles :

| | | petit <1000 document | moyen = 1000 document | Grand > 1000 document |
|---|---|---|---|---|
| Mise en place d'un SRI | non | 3 | 0 | 0 |
| | oui | 4 | 2 | 9 |

C'est le volume des documents traités par l'entreprise qui influence le plus le besoin de mise en place d'un SRI.
Le choix de mise en place d'un SRI est motivé principalement par :

- Les coûts de traitement et production des documents
- Les coûts de stockage des documents (local, équipements de sécurité, …)
- le risque de perte des documents durant leur transit d'un service à un autre

La difficulté d'accès aux documents stocké

| | | Volume documentaire | | |
|---|---|---|---|---|
| | | petit <1000 document | moyen = 1000 document | grande > 1000 document |
| Difficulté de stockage | non | 4 | 1 | 2 |
| | oui | 3 | 1 | 7 |

On remarque que le volume des documents stockés est la plus grande source de difficulté d'accès.

Disponibilité de l'information au sein de l'entreprise :

| | | Disponibilité de l'information | |
|---|---|---|---|
| | | Souvent | quelques fois |
| Difficulté d'accès aux documents | Non | 4 | 3 |
| | Oui | 2 | 9 |

C'est la difficulté d'accès aux documents qui est source principale de l'indisponibilité de l'information.

## En conclusion :

à travers l'analyse ci-dessus, nous remarquons que c'est principalement le volume des documents stockés et la difficulté d'accès à ces documents qui sont la source principale du besoin de mise en place d'un SRI.

# 2-2 Etude d'impact

Pour étoffer l'importance de l'utilisation de la recherche d'information dans l'entreprise, nous avons jugé utile de comparer les effets de son utilisation en comparant deux entreprises, l'une utilisant un SRI et l'autre ne l'utilisant pas.

## 1. l'entreprise ne disposant pas de SRI :

La CNR (Caisse Nationale de retraite) qui compte plus de 80 employés et dont la mission principale est la gestion des dossiers des retraités. Ce choix est motivé par le grand volume de document dont dispose cette entreprise et la masse importante d'information dont elle dispose.

L'information en général, l'information documentaire en particulier, est la clé de voûte de toute organisation désireuse de réussir et de se promouvoir. Ne dit on pas que : « qui maîtrise l'information, maîtrise le pouvoir ! » .Alvin TOFFLER, dans son ouvrage, intitulé POWER SHIFT, cité par MARTINET Bruno et al.

Disait que « l'information joue un rôle de plus en plus important dans notre civilisation et qu'elle devient parfois plus importante pour le fonctionnement des entreprises que l'accès au capital »

-Elle est un élément central sur lequel se fonde en effet toute prise de décision ou innovation
-Elle constitue de nos jours un véritable enjeu de développement.

Pour que cette information documentaire dont dispose la Caisse soit pertinente, fiable, efficace et utile, il faut que la recherche d'information soit la plus souple possible. Le processus de recherche doit être capable de prendre en charge tous les types de documents existants dans l'institution afin d'en faire un élément de premier plan dans l'amélioration de la gestion et de la productivité de l'entreprise.

## L'organisation du fonds documentaire est caractérisée par :

<u>L'inexistence d'une structure de gestion documentaire :</u>

Un grand volume des documents inorganisé et traité d'une manière manuelle en plus il n'ya pas un logiciel pour gérer ce volume de documents rendant difficile le repérage et la prise de décisions adéquates dans le traitement des dossiers sensibles

Les pertes et la difficile circulation et repérage de l'information administrative

Au niveau des différentes entités organiques (au sein d'un même service ou d'une même direction, les documents disparaissent ou sont tout simplement égarés), ce qui est source de lenteur dans les prises de décisions, éternel recommencement dans les études de dossiers et la perte fréquente de procès par manque d'informations fiables et élémentaires

De même, il y'a une rétention de l'information au niveau de certains responsables : les dossiers qui doivent faire l'objet de traitement du point de vue hiérarchique sont conservés dans les tiroirs et par manque d'instruments de repérage et de manuel de procédures, ils restent introuvables ou égarés

Le coût élevé de la reproduction des documents

Chaque bureau est équipé d'un ordinateur, dispose d'une imprimante et chaque service dispose d'une photocopieuse. Ce suréquipement est la conséquence directe de l'utilisation abusive de grandes quantités de copies des documents manipulés.

Exemple : Le service gestion du personnel de la Direction des Ressources Humaines prend 400 décisions par an qui doivent être signées par le directeur des ressources humaines (congés annuels, autorisations d'absence...) et 1500 décisions par an qui doivent être signées par le directeur général (embauches de nouveaux agents, engagements d'agents temporaires ; reclassement d'agents, intérims...).

L'établissement de ces décisions représente environ 50 rames de 500 papiers soit 25000 feuilles de papiers par année. De même, la secrétaire particulière du secrétaire général de l'institution, utilise quant à elle 20 à 25 rames par an de papier : en effet, service de coordination des différentes directions de l'institution, le secrétariat du secrétaire général est chargé de faire des copies des différents documents et décisions pour les directions et agents concernés par les dites décisions .

La mauvaise gestion des documents essentiels de l'institution :

La gestion de certains documents laisse à désirer : certains dossiers de personnel et de prestations sont incomplets rendant souvent difficile la détermination des droits des agents et des assurés

La mauvaise gestion de l'information administrative Peut avoir des conséquences néfastes sur le schéma directeur informatique qui est en œuvre : en effet, une bonne informatisation de l'existant passe par la mise à la disposition

d'informations fiables, pérennes qui authentifient la qualité des données saisie ( il n'y a pas eu de réel rapprochement entre les dossiers papiers et informatiques ) Les dysfonctionnements engendrés par la saisie de données peu fiables auront des coûts très élevés sur le processus d'informatisation. De même, la non-implication des archivistes de l'institution au processus d'informatisation en cours est à déplorer la documentation existante est obsolète et dispersée au sein de certaines entités.

## 2. l'entreprise disposant d'un SRI :

ERCE entreprise publique qui produit le ciment adopte les NTIC dans la gestion. On trouve la GED qui est déjà mise en œuvre au sein de l'entreprise depuis 3 ans.

## L'organisation du fonds documentaire est caractérisée par :

La structure de gestion documentaire

Tous les documents (fichiers, factures, commandes, compte rendue, décisions …) sont bien indexés et stocké d'une manière numérisé (archivage électronique) dans une base documentaire sur un support électronique rendant possible et facile, l'accès à ces documents. La consultation se fait à l'aide de l'utilisation un moteur de recherche hébergé dans le réseau interne (intranet). Le temps ou la durée de repense du système de recherche d'information est quasi nul (1mn).

La circulation de l'information au sein de l'entreprise :

Le moyen de partage de l'information au sein de l'entreprise c'est la messagerie Il faut savoir que les NTIC mettent à disposition la facilite de création d'un portail d'entreprise qui aura les avantages suivants : faciliter la circulation, le traitement et les échanges de dossiers et d'informations entre services aussi entre l'entreprise et les autres groupes par l'application du workflow (ce qui permet de faciliter un processus de travail ex processus d'achat ou de production , par sa systématisation et son informatisation en totalité ou en partie). Ceci autorise à générer des statistiques sur les étapes des processus de travail, à des fins de rationalisation et de contrôle.) du groupeware qui sont des composantes de la gestion électronique des documents, rendant ainsi fluides les décisions prises.

De même, cela permettra la consultation non seulement du catalogue de la documentation et des produits documentaires, mais aussi de certaines décisions (en tenant compte des règles de conservation qui seront édictées) à partir du portail qui sera une fenêtre ouverte dans l'entreprise et l'environnement extérieur (clients , fournisseurs , les entreprises financières …..)

## Conclusion

Cette comparaison nous interpelle qu'il serait plus judicieux sur le plan de la qualité de service et sur le plan économique de mettre en place un système d'archivage électronique accompagné d'un module de recherche d'information en vue de pallier à tous les surcoûts générés par les effets énumérés ci-dessus.

# 2-3 Les NTIC et les entreprises algériennes

L'introduction des NTIC Dépend de plus en plus de la capacité d'anticipation et d'adaptation de l'entreprise. L'intelligence économique est de fait, une bouée de sauvetage indispensable pour se maintenir permettant à l'entreprise de détecter et d'interpréter des signes d'alerte précoces concernant les changements, les mutations ou tout simplement les ruptures pouvant se produire dans l'évolution de son environnement social, économique, culturel et technologique. Pour s'insérer dans le marché, les entreprises de production et de services, privées ou publiques, sont appelées à entamer immédiatement, simultanément et en continu l'amélioration de la valeur de leurs produits, de leurs systèmes de production, de leurs services et de leur gestion en vue d'offrir le meilleur rapport qualité/prix dans les meilleurs délais. Les TIC (technologies de l'information et de la communication) sont soutenues par un discours très optimiste quant à leurs potentialités techniques et aux améliorations de gestion qu'elles peuvent procurer aux entreprises. Dans le cadre du développement des PME et l'ouverture du marché algérien à l'international, cette promesse technologique suscite un grand intérêt. Il semblerait en effet possible de combler certaines carences des PME algériennes : Difficile connaissance des marchés intérieurs et extérieurs, collecte de l'information onéreuse, mauvaise connaissance en techniques de gestion, impossibilité de se déplacer de façon fréquente sur les marchés intérieurs et extérieurs par manque de moyens humains et financiers. Par ailleurs, les nouvelles technologies de l'information et de la communication de par, le monde ont permis aux entreprises d'entrevoir de nouvelles techniques de gestion et un gain de productivité. Cependant, même si ce gain de productivité n'est toujours pas facile à mesurer, le paradoxe de Solow disant : « Les outils informatiques sont présents dans toutes les entreprises, mais pas dans les statistiques », comme toute nouvelle technologie, il lui faut un temps d'adaptation et de diffusion pour ensuite pouvoir mesurer son impact sur la productivité marginale de l'entreprise. Les PME occupent une place de plus en plus importante au sein du tissu industriel algérien, notamment par le biais de l'élargissement du secteur industriel privé et l'arrivée en masse, ces dernières années, d'entreprises étrangères par le biais des IDE. Cette réalité impose à ces entreprises de se préoccuper davantage de l'utilisation des TIC dans son management quotidien, afin de bénéficier de leurs bienfaits, et surtout s'aligner

sur la concurrence nationale et internationale pour une meilleure productivité et réactivité aux changements du marché. Une avancée considérable menée par les autorités concernées ces dernières années pour la diffusion et la démocratisation des nouvelles technologies sur l'ensemble des activités économiques. A titre d'information, nous pouvons citer quelques exemples sur les projets lancés ces dernières années, 62% investis dans les NTIC (nouvelles technologies de l'information et de la communication), ce qui représente 6,451 milliards de dinars, 2 520 milliards de dinars dans les technologies, spatiales etc. Cependant, le secteur des TIC en Algérie reste insuffisant, et nous dirons même dérisoire par rapport aux besoins du marché, il ne représente que 1% du PIB. Pour atteindre cet objectif, il s'agit, d'une part, d'entreprendre un ensemble d'actions multidimensionnelles relatives au parachèvement du processus de réforme et de restructuration industrielle, l'organisation du marché du libre-échange à la recherche des partenaires économiques étrangers, l'adaptation du cadre intervention des exportateurs aux règles et pratiques du commerce international, à la réhabilitation de l'outil de production et, d'autre part, en recours à des stratégies manufacturières qui assureraient plus de flexibilité, plus de rapidité d'exécution, une plus grande sensibilité aux besoins du marché mondial et plus d'indépendance à l'égard des économies d'échelle. Ce deuxième aspect ne peut se concrétiser que par une réorganisation des activités de production, en tenant compte des contraintes de flexibilité, de rapidité, de qualité et de sécurité imposées par les conditions du marché. Quel serait l'impact de l'insertion des TIC sur le processus de modernisation et de redynamisation de nos entreprises ? Améliorer les performances : l'introduction des technologies de l'information et de la communication au sein des opérateurs économique, permettrait les échanges rapides d'information. Et, l'utilisation optimale de l'outil informatique dans la gestion de l'entreprise permet de réduire les coûts et les délais de production ou de logistique. Les TIC apporteront donc, un gain de temps et de productivité. Elles permettront d'augmenter la réactivité dans toutes les activités de l'entreprise : commerciale, achats, approvisionnements, services administratifs, fabrication, expédition, études ; s'aligner et rivaliser face à la concurrence ; elles permettent aussi de rivaliser grâce à la connaissance et de ne pas se laisser distancer par la concurrence. En effet, les TIC, par l'échange rapide de gros volumes d'informations, permettent de rester toujours bien informées sur ce que fait la concurrence et même de la devancer, grâce, par exemple, à l'améliorations des échanges d'information avec les partenaires extérieurs : c'est le concept de l'entreprise communicante ou étendue. Se faire connaître : nous commencerons bien ce point par cette belle expression : « N'attendez pas qu'on vous cherche, faites de telle sorte que l'on vous trouve. » Les prodiges de ce merveilleux outil qu'est Internet, en particulier et les TIC, en général, permettent à l'entreprise de se faire connaître et d'opérer même au-delà des frontières du pays. Son image de marque peut en être grandement améliorée. On peut communiquer en temps réel des informations avec le monde entier et,

élément non négociable, on peut toucher des clients potentiels que l'on ne pourrait pas atteindre autrement : on augmente la part de marché de l'entreprise, on fidélise les clients et on leur offre de nouveaux services. Cette possibilité et cette capacité virtuelle à être présent sur les marchés mondiaux simultanément et d'une manière interactive ouvrent à l'entreprise algérienne d'autres horizons, une possibilité d'atteindre une clientèle plus importante de par le monde. Cette vitrine est aussi une possibilité incontournable et inespérée de s'aligner aux côtés des entreprises étrangères de différentes nationalités. La question qui se pose maintenant est : « Sommes-nous en mesure de nous adapter à ces nouvelles réalités citées plus haut ? » Car aujourd'hui, la concurrence ne se fait pas seulement par le rapport qualité/prix, mais par la capacité à capter, analyser et transformer l'information en facteur déterminant dans la productivité. Nous ouvrons un autre débat pertinent et primordial, celui de l'intelligence économique ! Qu'en est-il en Algérie ?

# Conclusion

A travers ce travail on conclure que le NTIC Pour une entreprise, correspondent à l'ensemble des technologies qui s'appuient sur l'informatique et les réseaux électroniques pour lui permettre de communiquer, stocker, gérer , échanger de l'information en son sein ou avec son environnement (clients , fournisseurs, partenaires...). Toutefois, ce ne sont pas ces technologies en elles-mêmes qui rendent possible la réalisation de ces différentes tâches, amis bien un "système d'information et de communication" au sein duquel les TIC occupent un place importante mais ne sont rien sans le "personnel", les "données" et els "procédures

Comme l'enfant qui commence par apprendre à reconnaître les lettres de l'alphabet avant de les associer pour former des syllabes puis des mots, le système De recherche d'information grâce à ses outils un élément d'une importance capital pour l'avenir de notre société numérique : « le bon sens » !

Ranger, trier, classer, organiser, rien de plus logique pour faciliter l'accès à un document. Si le passage au tout numérique, et l'amélioration grandissante des performances des moteurs de recherche ont pu laisser croire à la disparition programmée des langages documentaires, on s'aperçoit aujourd'hui, que ces langages ont encore un avenir. Liftés, remodelés, les technologies de l'intelligence artificielle leurs ont donné un coup de jeune qui leurs ont permis de retrouver peu à peu leur place dans l'univers numérique, entraînant dans leurs sillages un repositionnement du métier de documentaliste. Considérés comme des médiateurs entre les utilisateurs et l'information recherchée, les documentalistes ont depuis toujours été des concepteurs d'outils documentaires destinés à faciliter l'accès aux documents. Or, comme nous l'avons vu dans cette étude, face à l'autonomie croissante des « chercheurs d'information », ces outils, autrefois destinés à l'usage exclusif des documentalistes, sont devenus indispensables au grand public. Par conséquent, leur transposition dans un contexte numérique n'a pu se faire sans provoquer une hybridation du métier de documentaliste. Experts en gestion des fonds documentaires, mais aussi concepteurs d'outils permettant la consultation et la communication de l'information contenue dans ces documents, les spécialistes de l'information ont du intégrer de nouvelles connaissances qui ne faisaient pas partie de leur cœur de métier : langages informatiques, graphisme, ergonomie, linguistique, marketing, sociologie, etc.…

# Partie  02 : enquête et réalité de la RI en Algérie