Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme de Licence en Informatique

Thème

Réalisation d'un moteur de recherche

Réalisé par :

- *M lle* SAIDI Ilham
- *M lle* HAMSI Essma

Présenté le 27Juin 2013 devant la commission composée de MM.

- *M er BENTAALAH M.A* *(Examinateur et Encadreur)*
- *M er BENAZZOUZ M* *(Examinateur)*
- *M er HADJILA F* *(Examinateur)*

Année universitaire : 2012-2013

Remerciement

Tout d'abord on remercie DIEU le tout puissant de nous avoir aidé a surmonté tout les moments difficile et de nous avoir donné assez de force pour accomplir notre travail. .

on tiens aussi à exprimer notre profonde reconnaissance et notre gratitude à Mohamed amine Bentaalahtout d'abord pour avoir accepter de nous encadrer, pour nous avoir soutenus, dirigés et orientés toute la période de notre projet.

On remercie aussi nos parents pour avoir crues en nous ainsi que tous les membres de notre famille.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail Et de l'enrichir par leurs propositions.

On remercie aussi les professeurs et les étudiants de notre faculté pour les aides et les encouragements qu'il nous ont apportés.

Dédicace

Je dédie ce travail

A la mémoire de ma mère que j'ai souhaité compter parmi nous et partager avec elle la joie de mes fins d'études, je prie Dieu, le Tout-Puissant, de lui accorder Sa Sainte Miséricorde et de l'accueillir en son vaste Paradis.

A tous les membres de ma famille qui ont sacrifié leurs droits et leurs temps pour que je puisse mener ces études :

À ma sœur et petite princesse Amira ;

À mon frère et ami Ilyas ;

À mon père et ma belle mère Faiza ;

Aux membres de ma grande famille SAIDI et SEFAOUI pour leur précieux soutien et leurs énormes sacrifices.

Je dédie également ce travail à ceux ou celles qui m'ont apporté leur savoir et contribué à ma formation, notamment, a mes enseignants de l'université de Tlemcen.

À mes copains et copines du groupe de l'informatique pour l'agréable bout de chemin qu'on a passé ensemble.

À mes amis(es)

Merci a vous tous

M^elle SAIDI Ilhem

Dédicaces

Je dédie ce travail

A la plus formidable des mamans qui a toujours été présente dans le meilleur et dans le pire et mon très cher père pour son soutien et son courage, m'ont toujours donné la force pour poursuivre mes études.

A mon tres cher frère qui ma toujours apporter son soutien morale et mes merveilleuses sœurs et leurs maris qui ont toujours été présent

A mon binôme ILHEM qui ma supporté tout au long de notre projet

Ainsi qu'a mes amis(es), auprès de qui, j'ai souvent trouvé réconfort et soutien ;

Merci vous tous.

M^Elle Hamsi Esma

Tables des matières

Liste Des figures

Chapitre I
La Recherche D’information

1. Introduction :

Historiquement, la croissance du volume de données textuelles comme les livres et les articles dans les bibliothèques durant des siècles à imposer de définir des mécanismes efficaces pour les localiser. Là où on marque la naissance de la "Recherche d'Information" comme discipline de recherche. Le terme " Recherche d'Information " fut donné par Calvin N. Mooers en 1948 pour la première fois **[1].**

Un SRI peut être défini comme étant un mécanisme de gestion qui joue l'intermédiaire entre un utilisateur et une collection d'informations. Son but est de satisfaire le besoin en information de cet utilisateur.

Nous présentons dans ce chapitre les concepts de base de la recherche d'information et le système de recherche d'information (SRI) ainsi que les principaux modèles existants et les différentes méthodes d'évaluation des performances de SRI.

2. Définition de la recherche d'information :

La recherche d'information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information. [**2]**

La RI traite de la représentation, du stockage, de l'organisation ainsi que de l'accès à l'information. Un SRI est un ensemble de modèles et de processus permettant la sélection d'informations pertinentes dans une ou plusieurs collections en réponse aux besoins d'un utilisateur. Depuis toujours, la recherche documentaire est subordonnée à la RI. Dans la majorité des cas, un utilisateur recherche une information plutôt qu'un document, mais il accepte qu'un système lui renvoie une liste de documents dans lesquels il est supposé trouver l'information dont il a besoin.

3. Processus de RI :

Le processus général de Recherche d'Information (RI) est bien décrit par R.K. Belew **[3]**. Il résume l processus entier en trois (3) processus élémentaires ce qui montre la Figure I.1 : "Poser une Question" (requête), "Construire une Réponse"

(liste des documents pertinents) et "Evaluer la Réponse" (jugement des documents restitués),

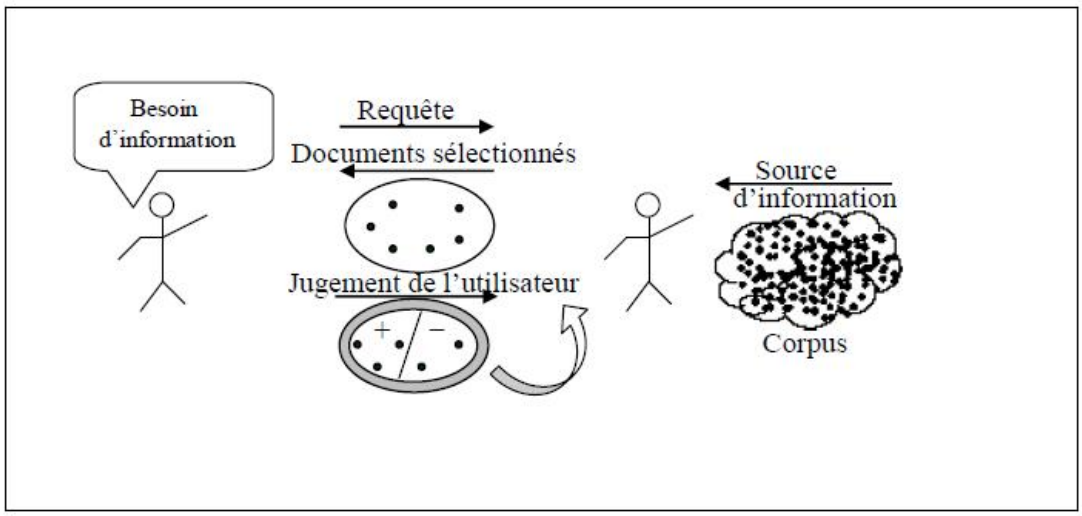

Figure I.1 : Recherche d'information en réponse à une requête

(Schéma inspiré de [3])

Comme nous pouvons représenter schématiquement un SRI, comme illustré par la **figure I.2**

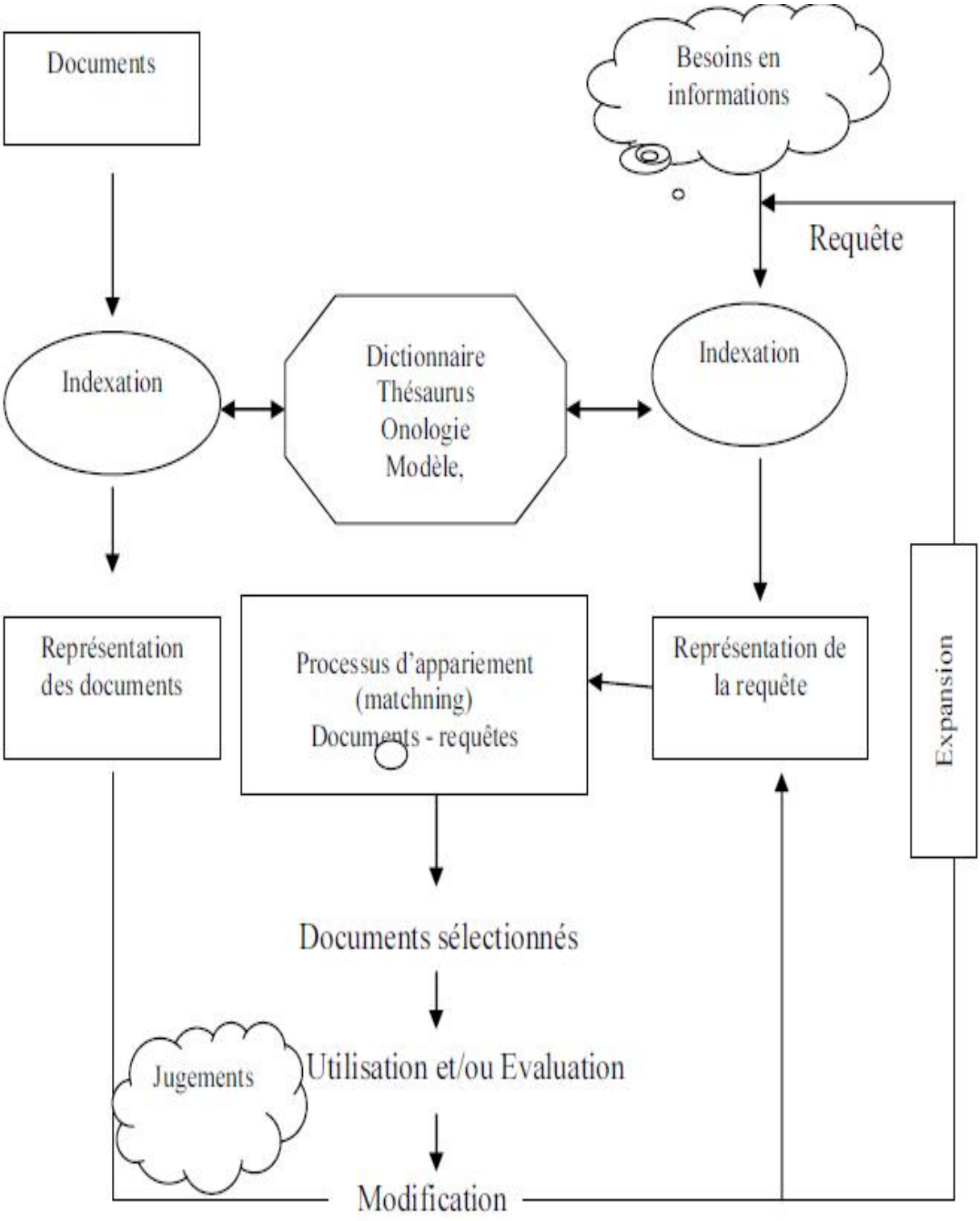

Figure I.2 : Le processus en U de la recherche d'information [8]

Plusieurs éléments clés y sont distingués :

- les documents,
- le besoin en information (requête),
- la représentation des documents et des requêtes (indexation ou analyse),
- l'appariement requête-document,
- expansion

4. l'indexation :

L'indexation consiste à analyser chaque document de la collection afin de créer un ensemble de mots-clés. Ces mots-clés seront plus facilement exploitables par le système lors du processus ultérieur de recherche. L'indexation permet ainsi de créer une représentation des documents dans le système. Son objectif est de trouver les concepts les plus importants du document (ou de la requête), qui formeront le descripteur du document.

L'indexation peut être :

- Manuelle : chaque document est analysé par un spécialiste du domaine ou par un documentaliste, elle permet d'assurer une meilleure pertinence dans les réponses apportées par le SRI, mais le temps nécessaire à sa réalisation est très important.

- Automatique : le processus d'indexation est entièrement informatisé, elle regroupe un ensemble de traitements automatisés sur un document. On distingue : l'extraction automatique des mots des documents, l'élimination des mots vides, la lemmatisation (radicalisation ou normalisation), le repérage de groupes de mots, la pondération des mots et enfin la création de l'index.

- Semi-automatique : le choix final revient au spécialiste ou au documentaliste, qui intervient souvent pour choisir d'autres termes significatifs. Les indexeurs utilisent un thésaurus ou une base terminologique, qui est une liste organisée de descripteurs (mots clés) obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques.

Le choix et l'intérêt d'une méthode par rapport aux autres dépend d'un certain nombre de paramètres, dont le plus déterminant est le volume des collections.

5. Pertinence :

La pertinence est la notion centrale dans la recherche d'information (RI) car toutes les évaluations s'articulent autour de cette notion. Mais c'est aussi la notion la plus mal connue, malgré de nombreuses études portant sur cette notion car les utilisateurs d'un système de RI ont des besoins très variés. Ils ont aussi des critères très différents pour juger si un document est pertinent ou pas. Voyons quelques définitions de la pertinence pour avoir une idée de la divergence.

La pertinence est :

- La correspondance entre un document et une requête, une mesure informative du document à la requête.
- Un degré de relation (chevauchement, relativité, …) entre le document et la requête.
- Un degré de la surprise qu'apporte un document, qui a un rapport avec le besoin de l'utilisateur.
- Une mesure d'utilité du document pour l'utilisateur.

6. Modèles de Recherche d'Information :

Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information. Il doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence. Ces modèles peuvent être divisés en deux catégories: les modèles dits « exacts qui ne retournent que des documents répondant exactement à la requête (modèle booléen) ou les modèles dits « partiels (probabiliste, vectoriel...) qui retournent des documents répondant à tout ou partie de la requête.

6.1. Le Modèle Booléen :

Le modèle booléen tire son nom des opérateurs booléens utilisés pour formuler une requête. En effet, une requête est une formule logique, combinant des descripteurs

et les opérateurs **ET, OU**, **NON**. Les documents sont représentés par une liste de descripteurs. Ces descripteurs peuvent appartenir à un langage libre ou contrôlé. Ils peuvent être extraits automatiquement des documents ou manuellement choisis par des documentalistes. La fonction de comparaison retrouve les documents dont les index valident la formule logique de la requête. Donc la base de documents est séparée en deux, les documents qui correspondent à la requête et ceux qui n'y correspondent pas. L'inconvénient majeur de ce modèle est l'absence d'ordonnancement des documents résultats par la fonction de comparaison. **[4]**

6.2. Le Modèle Vectoriel

Ce modèle représente un document ou une requête par un vecteur dans un espace d'indexation construit à partir des entités d'indexation. Les coordonnées des vecteurs sont les poids indiquant l'importance du descripteur par rapport au document. L'ensemble des coordonnées des vecteurs est contenu dans une matrice. La fonction de comparaison évalue la correspondance entre deux vecteurs (document et requête) ce qui permet de classer les résultats. Le schéma suivant (Figure I.3) illustre cette méthode.

T_k est un des vecteurs de base de l'espace de représentation.

Il représente *l'entité d'indexation k*.

D_i est le vecteur désignant le *document i*.

$W_{i,k}$ est le poids de *l'entité k* dans le *document i*.

$Di = (W_{i,1}, W_{i,2}, W_{i,3})$.

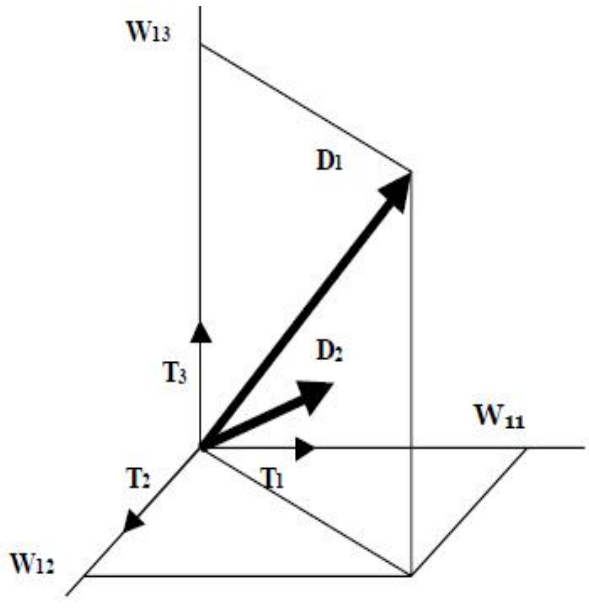

Figure I.3 : Représentation des documents dans un espace vectoriel intitulé espace des termes [8]

Le modèle vectoriel introduit par Salton **[5],** repose sur les bases mathématiques des espaces vectoriels. Dans ce modèle, les documents et les requêtes

sont représentés dans un espace vectoriel engendré par l'ensemble des termes d'indexation.

Chaque document est représenté par un vecteur D_{ij} .

Chaque requête est représentée par un vecteur Q_i.

Avec :

d_{ij} : Poids du terme **t_i** dans le document D,

q_i : Poids du terme **t_i** dans la requête Q .

Les termes de poids nul représentent les termes absents dans un document alors que les poids positifs représentent les termes assignés.

Dans ce modèle, chaque mot à un poids dans chaque document, ce poids représente l'importance du mot dans le document. Le degré de pertinence d'un document relativement à une requête se traduit par la fonction de pondération **tf*idf**. (tf signifie "term frequency" et idf "inverted document frequency").

La fonction de pondération est donnée par :

tf*idf(t,d) = tf(t, d) * log (N/df(t))

Où :

tf : est la fréquence du terme 't' dans le document 'd'.

df : c'est le nombre de document contenant le terme 't'.

et **N** : c'est le nombre totale de documents dans la collection.

Un terme qui a une valeur de tf*idf élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents. C'est le cas où un terme correspond à une caractéristique importante et unique d'un document.

Ainsi dans ce modèle on trouve les fonctions de similarité qui permettent de mesurer la ressemblance des documents et de requête. Ces fonctions sont détaillées dans le deuxième chapitre.

6.3. Le Modèle Probabiliste :

Ce modèle comprend :

- **Le modèle probabiliste général :** Le modèle de recherche probabiliste utilise un modèle mathématique basé sur la théorie de la probabilité. Le processus de recherche se traduit par calcul de proche en proche, du degré ou probabilité de pertinence d'un document relativement à une requête. **[6]**
- **Le modèle de réseau de document ou d'inférence (Document Network) :** Le réseau de document comprend les nœuds de document (un pour chaque document dans la collection), les nœuds de représentation de texte et les nœuds de représentation de concepts. Les nœuds de représentation de texte sont à l'intersection des deux niveaux de représentation. Ils synthétisent l'information sur la manière dont le document est représenté, notamment lorsqu'il s'agit de documents non textuels, tels le son et la vidéo. Un document peut avoir différentes représentations. Les nœuds représentant les concepts décrivent les différents concepts identifiés dans le texte des documents et des requêtes. **[7]**

7. Objectif de la Recherche d'Information :

La recherche d'information a pour objectif de :

- Identifier en vue d'exploiter de l'information contenue dans des documents et des bases de données (son, texte, image) par rapport à une requête formulée par un utilisateur.
- Le SRI devra nous retourner le moins possible de documents non pertinents.
- Les contenus des documents peuvent être non structurés ou semi structurés.

8. système de recherche d'information :

Les systèmes de recherche d'information (SRI), servent d'interface entre une collection contenant des quantités considérables de documents et des utilisateurs cherchant des informations susceptibles de se trouver dans cette collection, en utilisant

des requêtes. Ils intègrent un ensemble de techniques permettant de sélectionner ces informations. Elles peuvent être résumées en quatre fonctions, qui sont :

- Stockage des informations.
- Organisation de ces informations (processus d'indexation).
- Recherche d'informations : en réponse à des requêtes utilisateurs.
- Restitution des informations pertinentes pour ces requêtes.

L'intérêt de plus en plus croissant porté au domaine de la recherche d'information du à l'explosion du volume d'informations dans les entreprises et d'autres organismes (Internet, bibliothèques, presse etc.) a conduit les chercheurs à proposer des modèles de représentation, des mécanismes d'appariement entre les requêtes et les documents de la base et des modes d'interfaces différents pour améliorer les performances de leurs systèmes. **[7]**

9. Evaluation des systèmes de recherche d'information :

On peut évaluer les SRI selon plusieurs métriques, d'une façon générale, tout SRI a deux objectifs principaux : retrouver tous les documents pertinents, et rejeter tous les documents non pertinents. Ces objectifs sont évalués par les mesures de rappel et précision comme illustré dans la **figure I.4** :

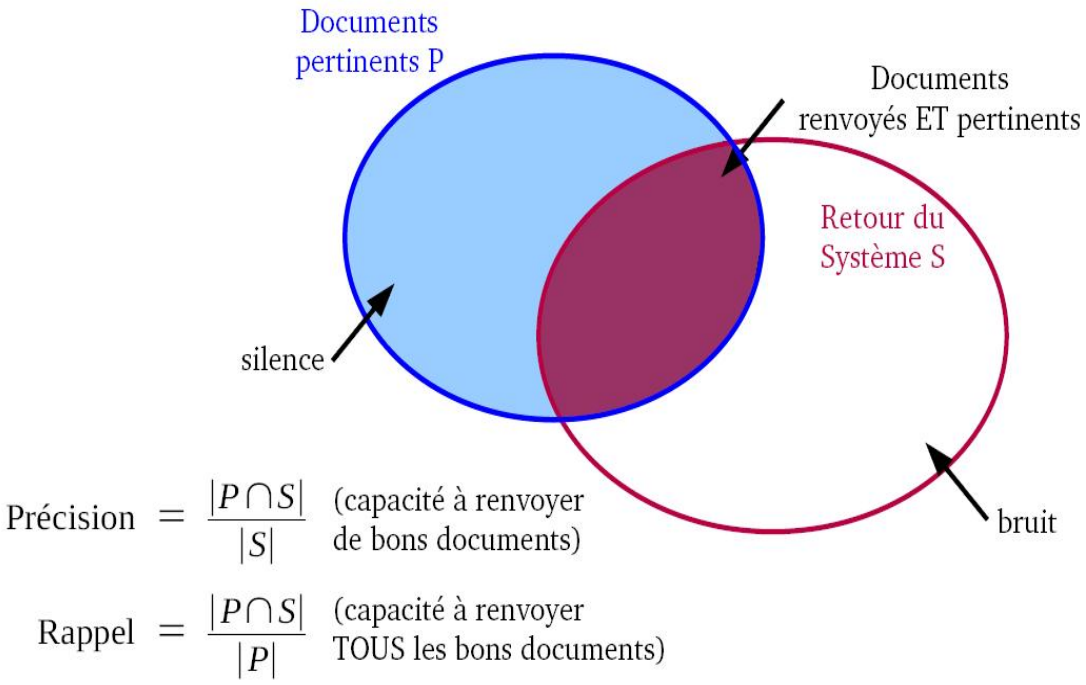

Figure I.4 : précision et rappel

- **Précision :** Elle mesure la proportion de documents pertinents relativement à l'ensemble des documents restitués par le système.
- **Rappel :** Il mesure la proportion de documents pertinents restitués par le système relativement à l'ensemble des documents pertinents contenus dans la base documentaire. Il est exprimé par :
- **La mesure F :** La moyenne harmonique F combine le rappel et la précision en un nombre compris entre 0 et 1.

10. Domaines d'application :

- La RI est un domaine vaste qui se situe dans les frontières de plusieurs disciplines tel que:
- 1. Recherche ad hoc

- 2. Classification /catégorisation (*clustering*),Question-réponses (*Query answering*),
- 3. Filtrage d'information (*filtering/recommendation*)
- 4. Méta-moteurs (*data-fusion, Meta-search*)
- 5. Résumé automatique (*Summarization*)
- 6. Croisement de langues (*cross language*)
- 7. Fouille de textes *(Text mining*)

11. conclusion :

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information. Nous y avons développé les principales étapes d'un processus de recherche d'information, à savoir, la représentation ou indexation de l'information, la comparaison de l'information et du besoin en information. Les principaux modèles existants, ainsi que les différentes méthodes d'évaluation des performances des systèmes de recherche d'information. Le chapitre suivant est consacré à la présentation de notre SRI dont nous nous intéressons beaucoup plus aux mesures de similarité.

Chapitre II
La Réalisation de l’application

1. Introduction :

Après avoir donné un aperçu général sur la recherche d'information, nous allons consacrer ce chapitre à la présentationd'une notre SRI développépour répondre aux requêtes des utilisateur nous allons exploiter le choix de distances de similarité dans notre application qui sont importantes pour mesurer la ressemblance afin de retournerdes documents potentiellement pertinents par ordre décroissant de similarités.

2.Environnement de travail :

Nous avons développé notre application sous l'environnement de développement **NetBeans ,** placé en open source ,en plusdu **JAVA , NetBeans**permet de supporter différents autres langages comme Python , C , C++…,il comprend toutes les caractéristiques d'un IDE moderne, **NetBeans** constitue par ailleurs une plate-forme d'exécution des applications autonomes.

3. Le Système :

Comme n'importe quel SRI, notre système passe par les étapes suivantes :

3.1-Prétraitement :

Cette étape consiste à effectuer un prétraitement afin de rendre les documents exploitable par la machine, ces prétraitements sont :

- Segmentation des documents : un document est une chaine de caractére, il est utile de localiser les mots, pour cela on utilisé les séparateurs suivants :

{ "( ) [ ] ^ , ; \ n ! ? ' 1 2 3 4 . 5 6 7 8 9 0 <> + - * : / ", }.

Prenant un exemple (qui va être utilisé tout au long de chapitre).

Entrée : (une organisation retenue, pour la présentation de nos : travaux 58 ^^dans une ORGANISATION)

Résultat :

une	Organisation	retenue	pour	La	Présentation	de	Travaux	dans	une	ORNANISATION

- Formater les mots en miniscule : cette étape est nécessaire car elle permet dereconnaitre le mot en minuscule et en majuscule comme un seul moi.

Résultat :

une	organisation	Retenue	Pour	La	présentation	De	Travaux	dans	une	organisation

- Supprimer la redandence : dans un corpus, il exisite plusieurs document qui contient le même mot , il est inutile de les stocker dans plusieurs cases, pour cela on propose de les suprimer et les mettre dans une seul case.

- **Résultat :**

une	organisation	retenue	pour	la	Présentation	de	Travaux	dans

- Eliminer les mots vides : il est toujours utile de supprimer les mots qui neportent aucun sens, ces mots se trouvent dans toutes les langues et comportent les articles, les conjonctions…etc, voici un exemple de certains mots vides de la langue française : La, Les, Dans,Une….etc. Dans notre application nous allons utiliser //250 un fichier comptenent 250 mot vide de françai mots téléchargés à partir :
  http://perso.univ-lyon2.fr/~maniezf/stop_list_fr.txt

Le résultat de prétraitement sera une matrice (documents* mots) où les lignes sont les documents et les colonnes sont les mots :

	organisation	retenueprésentation		travaux
$doc1$	.	.	.	.
$doc2$	.	.	.	.
$doc3$	.	.	.	.
$doc4$	.	.	.	.

3.2-Pondération :

Dans cette étape nous allons calculer le poids d'un mot dans chaque document. pour cela nous utilisons la fonction de pondération suivante :

tf*idf(t,d) = tf(t, d) * log (N/df(t))

où :

tf : est la fréquence du terme 't' dans le document 'd'.

df : c'est le nombre de document contenant le terme 't'.

et N : c'est le nombre totale de documents dans la collection.

Prsq la plus utilisé est donne mieux de resultat

Résultat de ceeteetape de notre exemple est representé dans la matrice suivante :

	organisation	retenueprésentation		travaux
1	0.24	0	036	1
2	0	0	0	0.2
3	0.1	0.3	1	0
4	0	1	0.2	0

3.3-le prétraitement de la requête :

L'utilisateur exprime son besoin à travers une requête, il s'agit d'appliquer les mêmes prétraitements de documents sur la requête (Segmentation, formater les mots en minuscule, élimination des mots vides), le résultat de cette étape sera un vecteur.

Entrée : « 123 Organisation !?++ etétudiant… »

Résultat :

(organisation etudiant)////**table**

3.4-Le calcul de degré de similarité :

Cette étape consiste à calculer le degré de pertinence entre le vecteur de la requête et le vecteur du document et permet de mesurer// la ressemblance. Il existe plusieurs fonctions de similarité.Les mesures utilisées dans notre système sont : leproduit scalaire, cosinus, mesure de Dice, mesure de Jaccard, et overlap.

- **Produit scalaire :**

$$\boldsymbol{RSV}(Q_i, D_{i\ j}) = \sum_{i=1}^{t} q_i * d_{i\ j}$$

Où :**q:**la requette

d:la matrice

t :le nombre de mot

la **figure(II.1**) illustre un exemple de calcul de produit scalaire entre une requête et plusieurs documents.

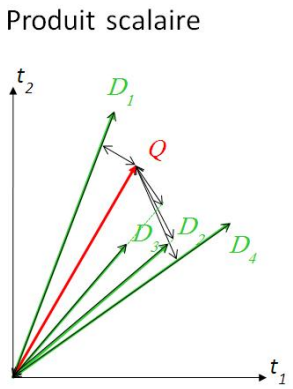

Figure II.1 : Produit scalaire

D’autres mesures ont été proposées, elles sont, tout de même, basées sur le produit scalaire. La mesure la plus utilisée, c’est celle du cosinus qui mesure le cosinus de l’angle formé par le document et la requête, 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 :

- **Mesure de cosinus :**

$$\boldsymbol{RSV}(Q_i, D_{i\ j}) = \frac{\sum_{i=1}^{t} q_i * d_{i\ j}}{\sqrt{\sum_{i=1}^{t} q_i^{\mathbf{2}} * \sum_{i=1}^{t} d_{i\ j}^{\mathbf{2}}}}$$

La **figure II.2** illustre le cosinus de l’angle formé par une requête et deux documents d1et d2, le document d1 est contrairement à d2,parce qu’il est très proche de la requête Q.

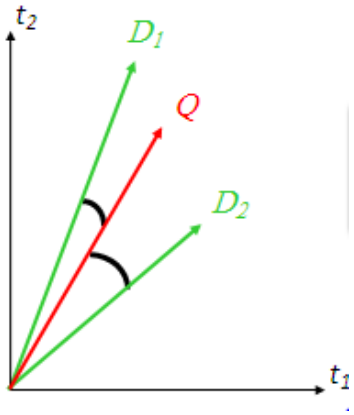

Figure II.2 :l’angle formé par une requête

Parmi les mesures utilisés, les plus connues citons le coefficient de JACCARD etcelui de DICE :

- **Mesure de JACCARD :**
  La formule de JACCARD est définie par :

$$\boldsymbol{RSV}(Q_i, D_{i\ j}) = \frac{\sum_{i=1}^{t} q_i * d_{i\ j}}{\sum_{i=1}^{t} (q_i^2) + \sum_{i=1}^{t} (d_{i\ j}^2) - \sum_{i=1}^{t} q_i d_{i\ j}}$$

Et celle de DICE est dérivée du coefficient du JACCARD en donnant plus d'importance aux éléments partagés.

- **Mesure de DICE :**la formule de DICE est definie par :

Et celle de DICE est dérivée du coefficient du JACCARD en donnant plus d'importance aux éléments partagés.

$$\boldsymbol{RSV}(Q_i, D_{i\ j}) = \frac{\mathbf{2} * \sum_{i=1}^{t} q_i * d_{i\ j}}{\sum_{i=1}^{t} q_i^2 + \sum_{i=1}^{t} d_{i\ j}^2}$$

Lesmesures de JACCARD et de DICE ont été initialement construites pour des analyses écologiques, et comme il existe d'autres mesures exemple la méthode OVERLAP qui est définie par :

- **Mesure de OVERLAP :**

$$\boldsymbol{RSV}(Q_i, D_{i\ j}) = \frac{\sum_{i=1}^{t} q_i * d_{i\ j}}{\mathbf{Min}(\sum_{i=1}^{t} q_i{}^2, \sum_{i=1}^{t} d_i{}^2)}$$

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 formant une base, en plus Le langage de requête est moins expressif.

Le Résultat de cette étape c'est un tableau qui contient les valeurs de pertinence :

Résultat :

Doc 01	**Doc 02**	**Doc 03**	**Doc04**
0.25	**0.43**	**0.75**	**0.66**

3.5 Tri des documents:

Cette étape consiste à trier le tableau de degré de pertinence afin de localisé les documents pertinent, du plus pertinent vers le moins.//par ordre coissant :

Résultat :

Doc 03	**Doc 04**	**Doc 02**	**Doc01**
0.75	**0.66**	**0.43**	**0.25**

4. Architecture fonctionnelle du système :

- l'ouverture de l'application : la **figure *II*.1** represente l'interface de notre notre SRI

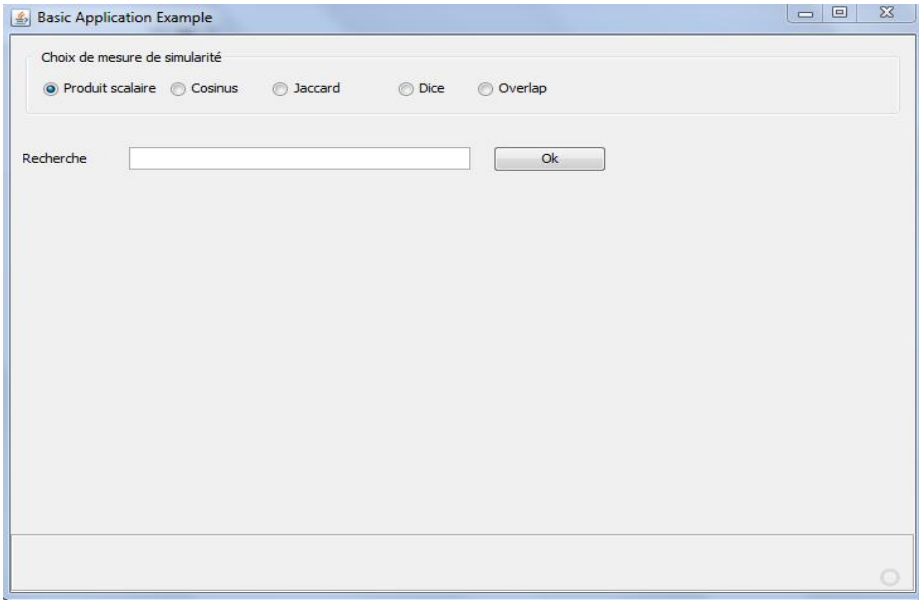

Figure *II*.3 : l'interface d'entré de l'application

- Le choix de mesure de similarité : la **figure *II*.2** montre comment l'utilisateur exploite le système dont il fait le choix de la méthode de mesure de similarité avant l'écriture de la requête.

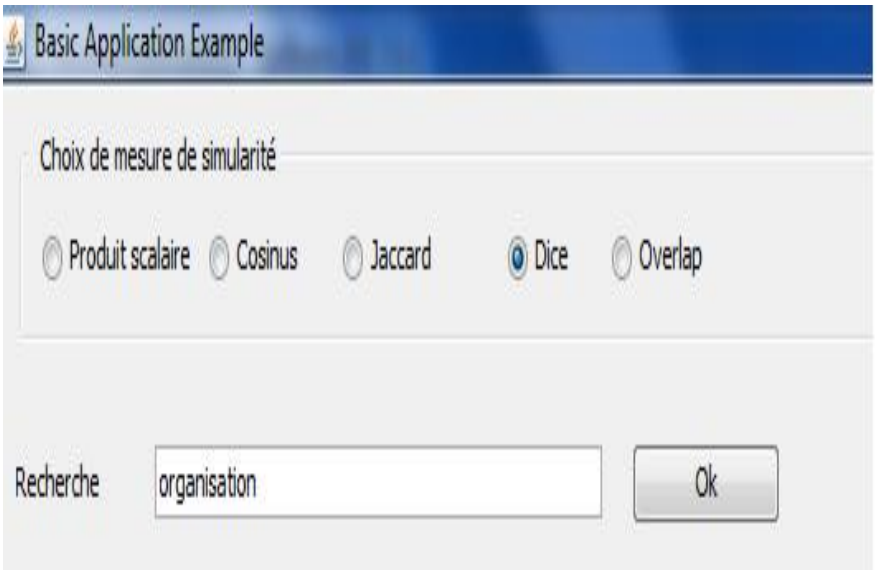

Figure *II*.4 : choix de mesure de similarité

- Le fonctionnement :la **figure *II*.3** montre comment le système affiche le résultat final après avoir appliqué les étapes précédentes de notre SRI, le choix de document selon le degré de pertinence puis l'affichage du documment choisi textuel.

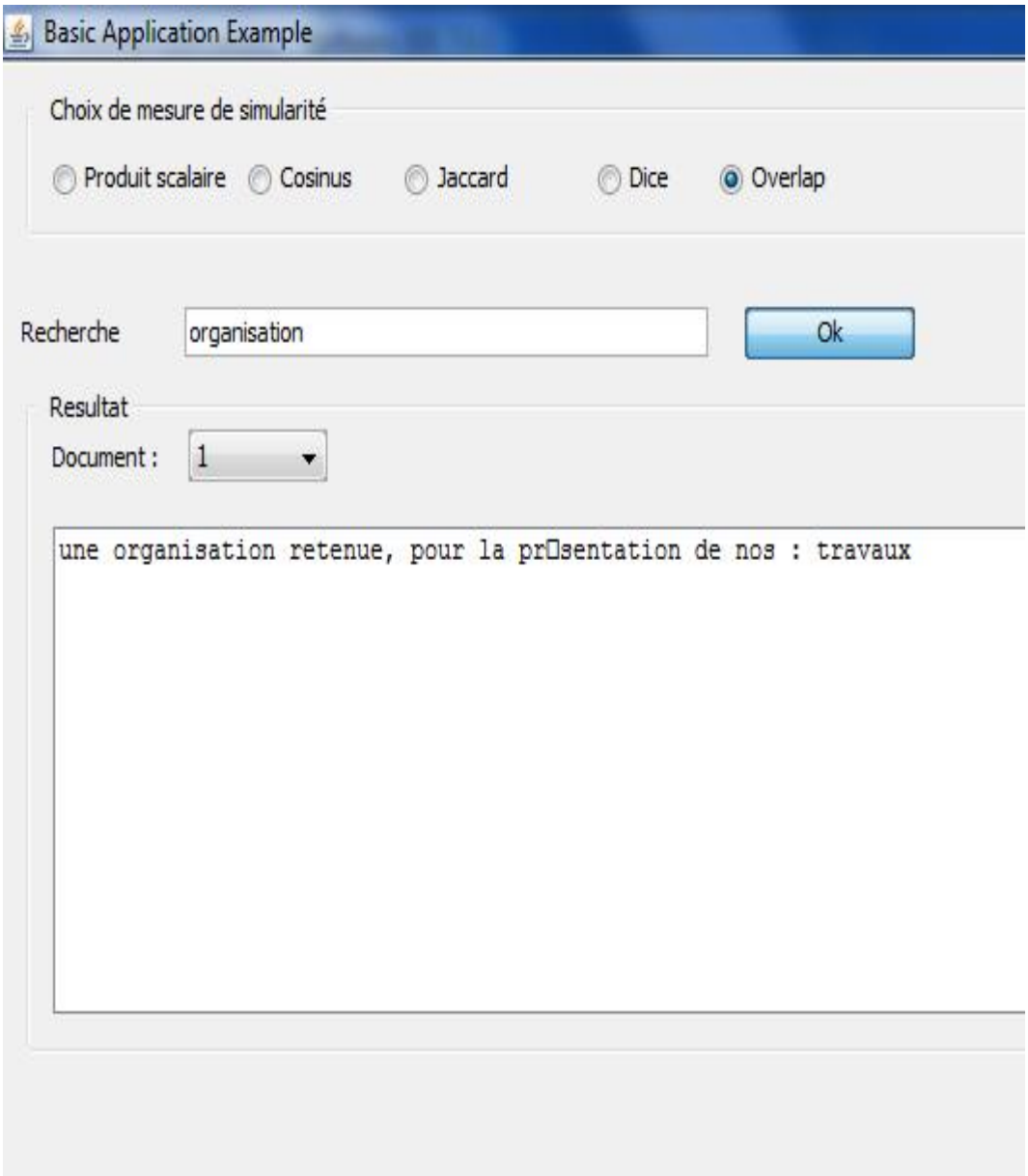

Figure *II*.5 : l'affichage de document le plus patinent

5. Conclusion :

Dans ce chapitre, nous avons présenté les étapes détaillées qui ont été suivies pour la constructionde notre interfacedans le but de répondre au besoin de l'utilisateur.Nous avons fait le prétraitement de notre base documentaire et la requête puis on a calculé la pondération et la pertinence où nous avons défini en détail les différentes mesures de similarité dont l'utilisateur peut choisir une parmi ces méthodes.

Conclusion Générale

Dans notre travail nous avons réalisé un système de recherche d'information basé sur les mesures de similarité dans la SRI. Le memoire s'articule sur deux chapitre, le premier chapitre a permis de présenter un bref historique d'Internet et du Web. Nous avons présenté ce que peut être un SRI et les différents modèles et nous avons rappelé ce qu'est l'indexation et leur type utilisé. En résumé, les moteurs de recherche permettent d'accéder à un grand nombre de documents par une recherche basée principalement sur les mots du texte. Dans le second chapitre nous avons proposé quelques fonctions de similarité qui mesurent le degré proximité entre un document et une requête, en vue d'ordonner les documents les plus similaires à la requête au moins similaire.

Conçues comme réponse aux problèmes posés par l'intégration de connaissances au sein des systèmes de recherche d'information, les différentes méthodes de calcul de similarité apparaissent désormais comme une clef importante et facile pour la manipulation automatique de l'information.

Dans le souci d'élargir et d'améliorer notre travail nous proposons d'enrichir notre application avec l'aide des experts du domaine et envisager une éventuelle intégration de cette dernière dans un moteur de recherche dédié aux domaines d'informatiques.

Bibliographies

[1] N.Moouers c.n.Mooers.Application of random codes to the gathering of statiscal information, 1948.

[2] **K.**sauvagnat and M.Boudganem .a la recherche de nœuds informatifs dans des corpus de documents xml.CORIA, pages 119/134,2005

[3] Richard K.Belew. Review published in information retrieval, vol .5, April _july2002.

[4] SALTON, G. A Simple Blue Print for Automatic Boolean Query Processing. Information Process Management, Vol. 24, N°3. Pp.269-280.1988.

[5] G.Salton, The smart Retriveal system Experiments in automatic document
processing, Perntice Hall Inc, Englewood Cliffs, New Jersey 1971.

[6] G**.**Salton, E.A, Fox,H, Wu ,extended Boolean information retrieval system CACM26(11), pp,1022-1036,1983.

[7] Mustapha baziz, a fuzzy logic approach ti information retrieval using an otology based representation of documents, 2005.

[8] Projet fin d'etude,conception et realisation d'un moteur de recherche 2010

Sites Web:

http://alain-defrance.developpez.com/articles/Java/J2SE/micro-rmi/#LII-A

http://www.wikipedia.fr/

http://gfx.developpez.com/tutoriel

Ouvrage :

Bougeault, Jérôme, Java : la maîtrise : Java 5 et 6, Eyrolles : Tsoft éd, 2008.

Résumé:

Notre projet consiste à développer une application qui a pour but de mesurer le degré de la similarité et la ressemblance entre une requête et un document au sein des systèmes de recherche d'information . Ceci facilitera au utilisateur de trouver unrésultat selon son besoin afin de définir une méthode de comparaison entre une représentation de document et de requête pour déterminer leur degré de correspondance.

Nous avons adopté la conception orientée objet avec la langage JAVA pour la conception de notre application .

Abstarct :

Our Projerct consist of devlopping an application which its goal of mesuring the degree of similarity and ressemblance between a query and document within systems of information's research. This facilate to the user to fond a result according to his need in order to dfine comparison method between a representation of document and query determinate their degree of correspondence , we have adopted the conception object using JAVA .

ملخص :

بحثنا يهدف الى حساب درجة التشابه بين الطلب و قاعدة بيانات الوثيقة ضمن نظام البحث المعلوماتي

حيث يسهلل على المستعمل ايجاد نتيجة حسب طلبه .

لتحديد منهجية المقارنة بين قاعدة بيانات الوثيقة و طلب المستخدم لايجاد درجة التشابه

اعتمدنا تصميم التطبيقة على لغة JAVA لتسهيل استعمالها للمستخدم .