

1 Introduction

Ce projet consiste à développer un système de recherche d'information a pour but d'aider l'utilisateur à prendre une décision. Ce système permettra à l'utilisateur d'exprimer ces besoins par une requête et de retourner tous les documents pertinents en comparant les similarités du document dans le corpus utilisé avec la requête.

Dans ce chapitre on va présenter la conception et l'implémentation de ce système.

2 La Recherche d'information dans la langue arabe

Notre but est de trouver des textes pertinents (loi ou collection des lois) par rapport à une requête formulé par l'utilisateur, les textes (corpus) et la requête sont écrits en la langue arabe.

2.1 Le corpus utilisé

Nous avons collecté de 91 articles extraits de la loi algérienne. Les articles sont sélectionnés de trios codes :

- Code de la famille : On a utilisé 46 articles de ce code, extrait du livre I « *le mariage et sa dissolution* ».
- Code pénale : On a utilisé 18 articles de ce code, extrait du Livre I « *les punitions et les procédures de sécurité* », Livre II « *actes et personnes soumis à des sanctions* » et Livre III « *les crimes et les délits et leur pénalités* ».
- Code de procédures civiles et administratives : On a utilisé 27 articles de ce code, extrait du livre II « *procédures pour chaque organe judiciaire* ».

3 Système de recherche d'information SRI

Une SRI comprend plusieurs tâches et concepts (illustré dans le Figure 3.1), dans les sections suivant on va décrit la conception de notre SRI.

Notre SRI comprend trois principales tâches sont :

- La formulation de la requête (besoin en information).
- Indexation.
- L'appariement requête-document.

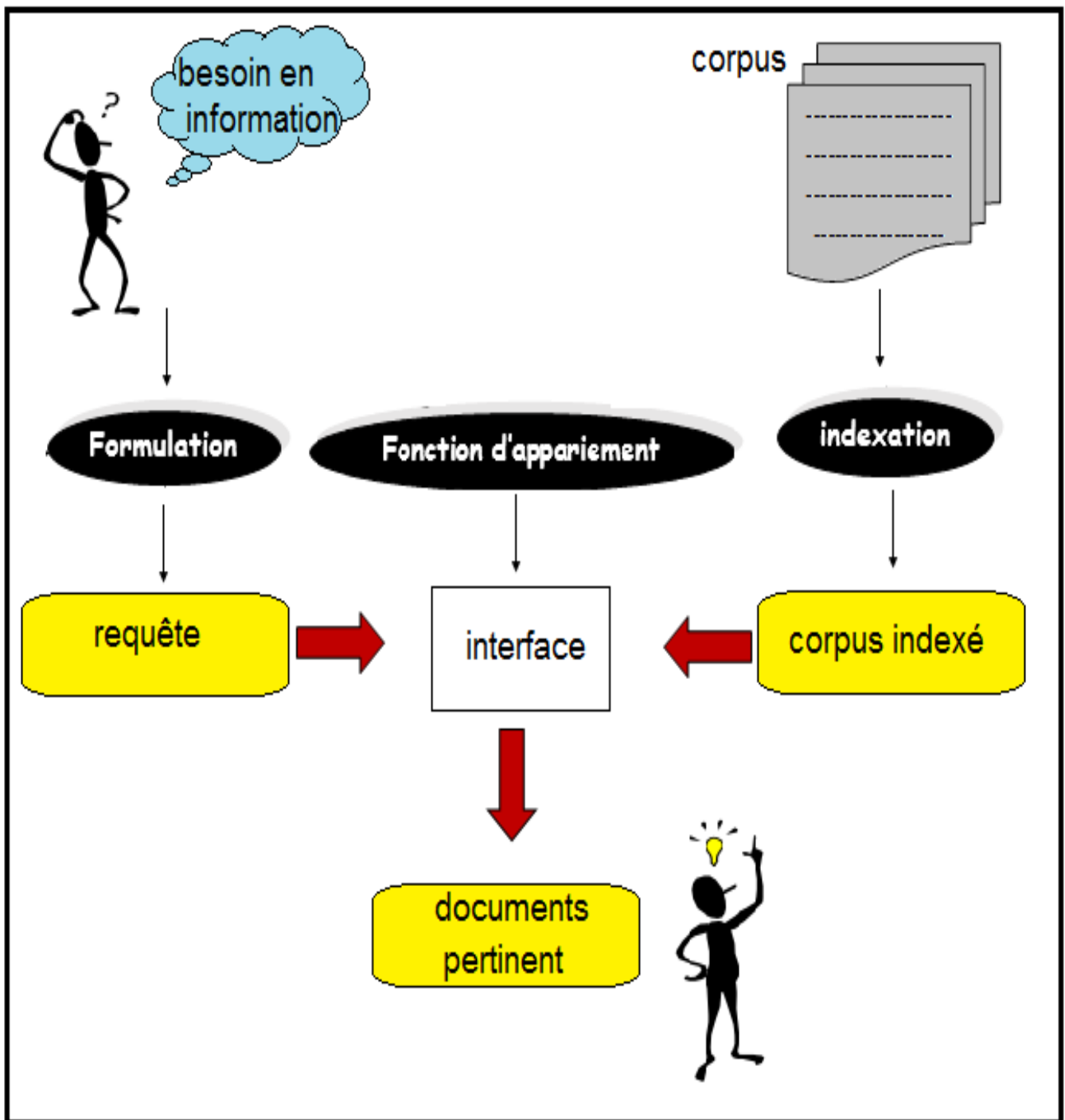


Figure 3.1 : Architecture de notre SRI. Adaptation de [18]

3.1 Formulation de la requête (besoin en information)

L'utilisateur exprime ses besoins par une requête, la formulation de cette dernière permet d'interroger notre SRI.

3.2 Indexation

L'étape d'indexation est le passage d'un document textuel (ou une requête) à une représentation exploitable.

3.2.1 Prétraitements nécessaire

Ce section décrit le processus de prétraitements qui nous avons appliqué sur le corpus.

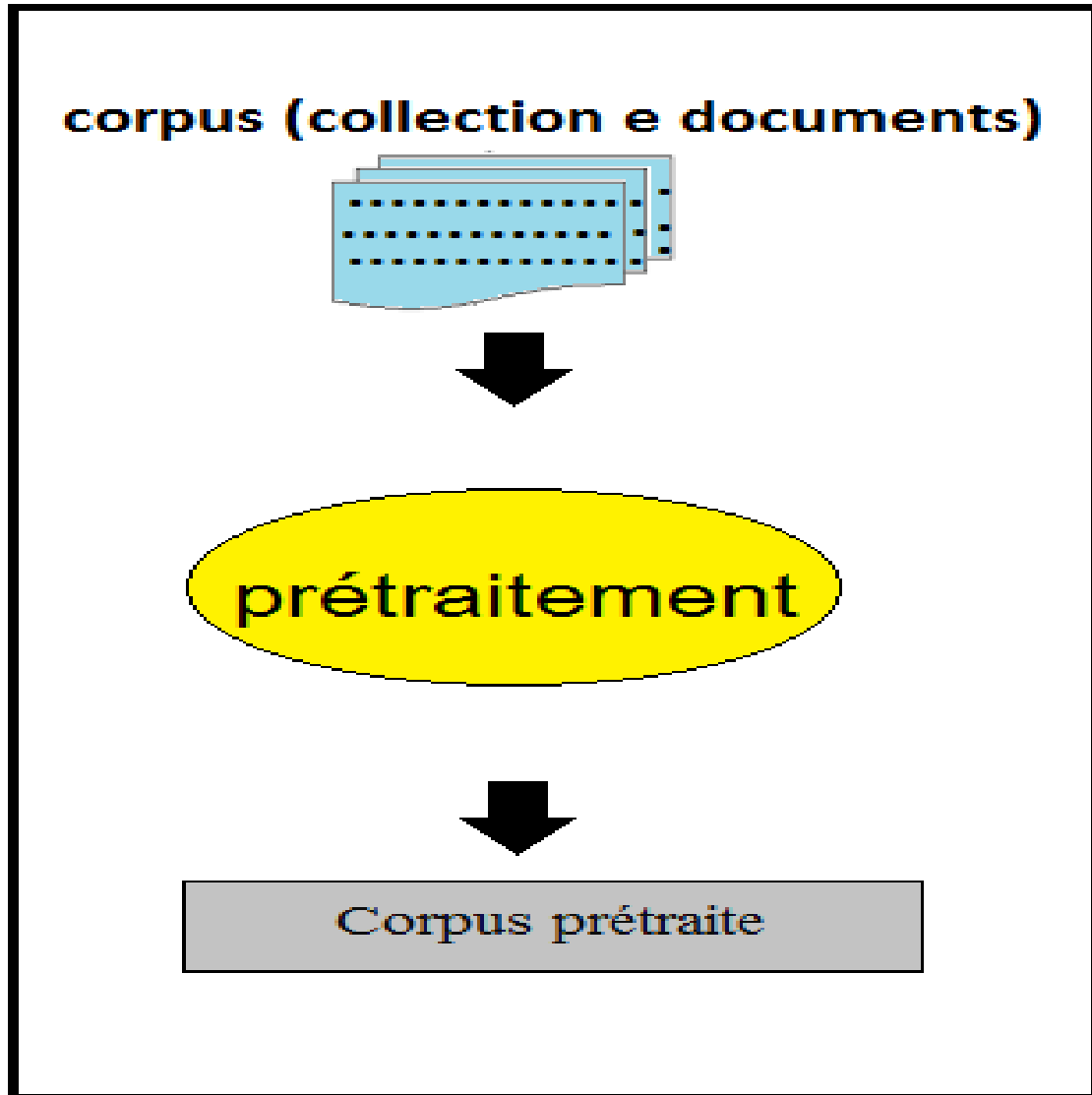


Figure 3.2 : processus de prétraitement

Pour établir le processus de prétraitement nous avons utilisé khoja stemmer :

3.2.1.1 Segmentation

La segmentation (tokenisation) est une étape importante dans le processus de prétraitement de la langue arabe, cette étape est difficile car le système d'écriture de la langue arabe est mixte, elle combine une écriture segmentée et une écriture non segmentée, ou dans certains cas les pronoms, sujets et compléments sont attachés aux verbes et une seule chaîne de caractères représente une phrase. [25] par exemple, « قلته : je l'ai dit ».

Dans notre SRI nous utilisons la segmentation lexicale basée sur la séparation des mots par espace blancs et les signes de ponctuation.

3.2.1.2 Normalisation

Dans cette étape les voyelles sont supprimées

3.2.1.3 Suppression des mots vides

Éliminer tous les mots non significatifs : sélectionner une liste des mots vides. Si un mot en fait partie, il ne sera supprimé. La liste regroupe généralement les particules (chapitre 2)

Exemple :

الى في من حتى به على تلك حول دون مع هذا ثم هذه أنه قد كان لهم لم فإن فيه ذلك لو أن ومع فقد هو في تحت او و ما لا الي إلى ظل بات صار ليس

3.2.1.3 Stemming

Cette étape nous permettant d'enlever les affixes des mots pour ne conserver que la partie racine.

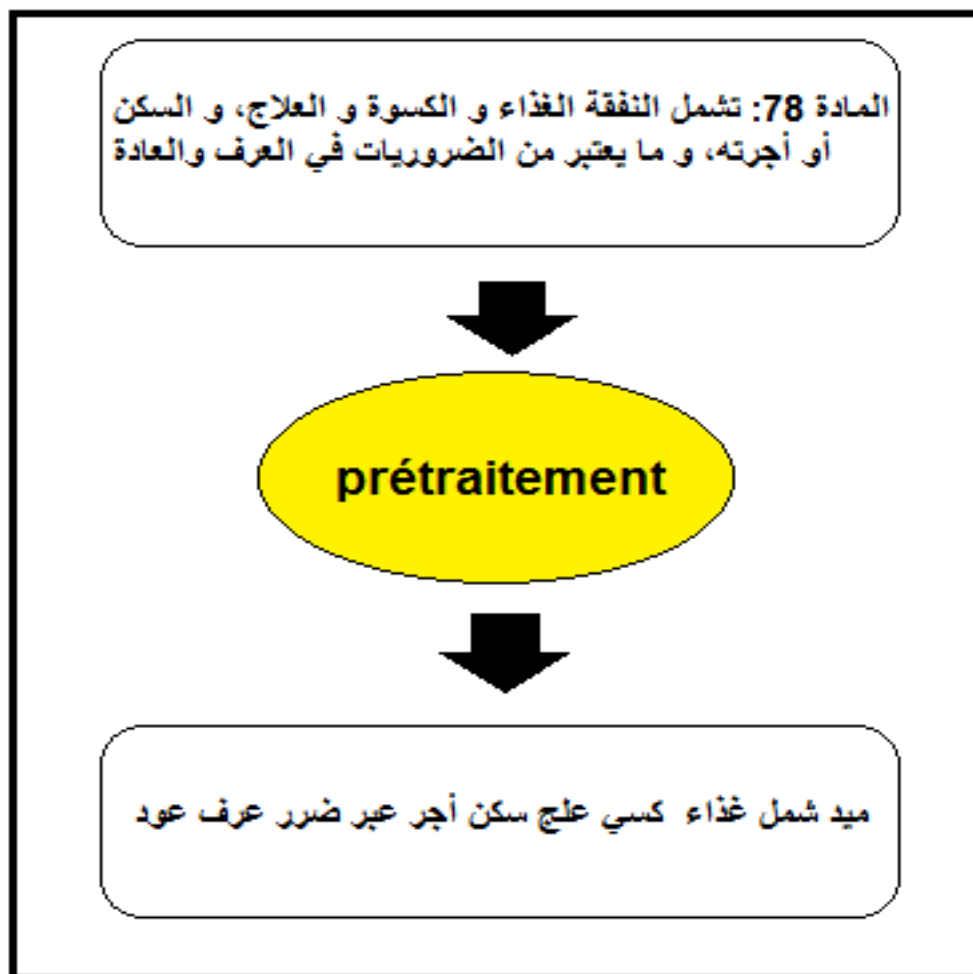


Figure 3.3 : Exemple du processus de prétraitement en utilisant khoja stemmer.

Notons que le même processus de prétraitement du corpus doit être appliqué sur la requête.

3.2.2 Pondération des termes

Afin de mesurer l'importance de chaque terme dans les documents où il apparaît nous avons utilisé la fonction de pondération *TF-IDF*.

3.2.3 L'index

L'index cette structure permet de sélectionner, pour n'importe quel terme, tous les documents où il apparaît.

3.3 L'appariement requête-document

Pour sélectionner l'information pertinente pour l'utilisateur une mesure de correspondance entre la requête et les documents est calculée.

Les documents pertinents retournés par un SRI peuvent être définis comme les plus proches de la requête selon la mesure *la similarité cosine*.

4 Implémentation

Nous avons développé une Système de Recherche d'Information (Figure 3.4) en utilisant le langage de programmation java dans la plate-forme NetBeans.

4.1 NetBeans

La plate-forme NetBeans est un cadre générique d'application pour les applications de bureau Java. La plate-forme NetBeans fournit la plomberie infrastructurels qui, sans elle, chaque développeur doit écrire eux-mêmes, comme des solutions pour la persistance état de l'application ; connexion actions à des éléments de menu, les éléments de la barre d'outils et les raccourcis clavier ; la gestion des fenêtres, et bien plus encore. La plate-forme NetBeans fournit tous ces hors de la boîte de sorte que vous n'avez pas besoin de coder manuellement ces ou d'autres caractéristiques de base vous-même. Au lieu de cela, vous pouvez vous concentrer sur ce que vos clients se soucient : le logique métier spécifique au domaine. [12]

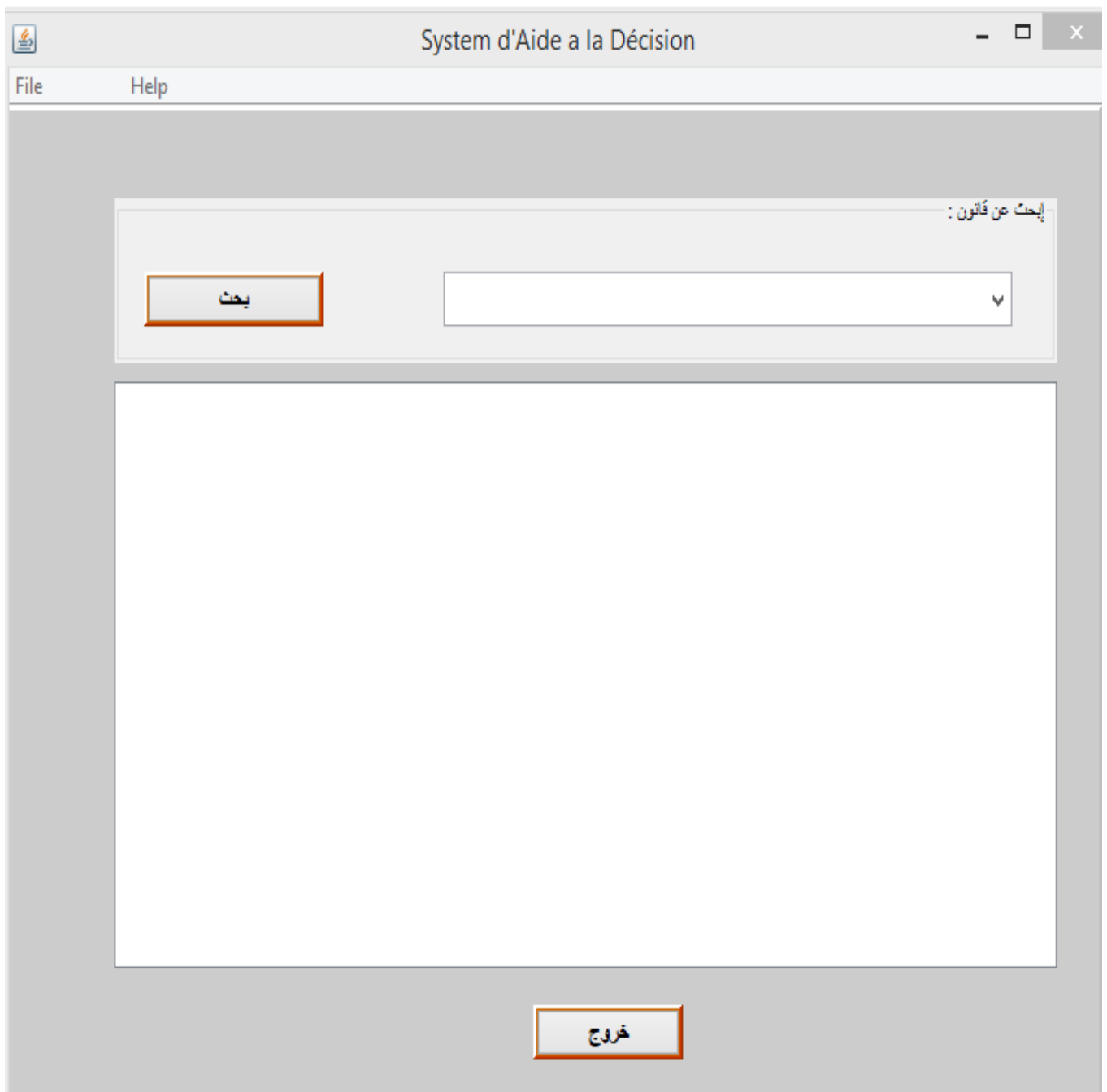


Figure 3.4 : l'interface de notre SRI

5 Conclusion

Dans ce chapitre, nous avons présenté la conception et l'implémentation de notre Système de Recherche d'Information, en commençant tout d'abord par présenter les différentes tâches de notre SRI. Ensuite, nous avons mentionné l'outil qui nous avons utilisé pour développer ce système.