Université Abou Bekr Belkaid 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 d'Ingénieur d'État en Informatique

Option : Système d'information avancé

Thème

Représentation d'un document textuel arabe par un nuage de mots

Réalisé par :
Mr. Mokhtar BENDIMERAD

Présenté le 03 Juillet 2011 devant le jury composé de Mrs :

- Abdelkrim BENAMMAR *(Président)*
- Mohammed El Amine ABDERRAHIM *(Encadreur)*
- Mohammed BENAISSA *(Examinateur)*
- Mohammed MERZOUG *(Examinateur)*

Année Universitaire : 2010-2011

TABLE DES MATIERES

Introduction générale :

L'objectif du traitement automatique des langues est la conception de programmes capables de traiter des données exprimées dans une langue naturelle pour lesquels plusieurs phases d'analyse (morphologique, syntaxique, sémantique et pragmatique) sont nécessaires afin d'en extraire des informations.

Avec l'avènement des documents électroniques, des quantités phénoménales d'informations sont générées. Cette montée en volume de textes nécessite la production d'outils informatiques performants dont la tâche est de trouver et d'extraire l'information pertinente sous une forme condensée.

Le nuage de mots-clefs (*tag cloud* en anglais) est une représentation visuelle des mots-clefs (*tags*) les plus utilisés sur un texte. Généralement, les mots s'affichent dans des polices de caractères d'autant plus grandes qu'ils sont utilisés ou populaires.

Ce travail entre dans le cadre d'élaboration d'un outil de visualisation d'un texte sous la forme d'un nuage de mots (on l'applique pour les sourates du coran). Le mémoire est organisé en quatre chapitres.

Dans le premier chapitre nous présentons une description générale sur l'étiquetage collaboratif. Nous commençons d'abord par la définition des tags et l'étiquetage des contenus en ligne et une définition sur les *folksonomies*. Ensuite nous détaillons le concept tag. A la fin nous définissons les deux concepts mots clefs et nuages de mots en citant quelque avantage et inconvénients de l'étiquetage collaboratif.

Dans le second chapitre nous présentons quelques outils existant qui génèrent les nuages de tags.

Dans le troisième chapitre nous présentons le domaine de recherche d'information (RI). Nous commençons par les principaux acteurs du RI, le processus de recherche d'information, les modèles du RI. Nous présentons par la suite la langue arabe avec sa particularité sa morphologie, la structure d'un mot et enfin les catégories des mots.

Le quatrième chapitre constitue le cœur de notre travail. Dans ce chapitre nous commençons par la modélisation du tag ensuite nous présentons le langage de programmation, et nous présentons un moteur de recherche nécessaire à la réalisation du nuage. A la fin nous décrivons les étapes nécessaires pour la réalisation de notre projet.

CHAPITRE I

L'ETIQUETAGE COLLABORATIF ET LES TAGS

1. Introduction :

L'étiquetage collaboratif est une pratique qui consiste, pour les internautes, à indexer des contenus en ligne à l'aide de mots clés appelés *tags*. Lorsque plusieurs internautes réunis par un site communautaire étiquettent les mêmes contenus, une sorte de classification émerge. C'est ce que l'on appelle une *folksonomie*. Elle permet de faciliter les recherches ultérieures de contenus grâce à une représentation visuelle sous forme de nuage de mots (figure 1.1). Ces différents concepts méritent de plus amples explications.

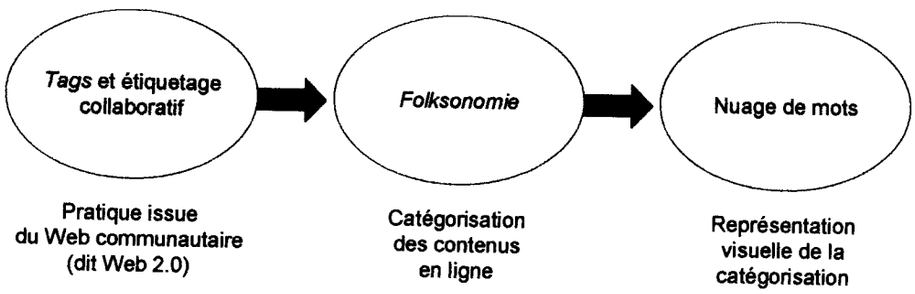


Figure 1.1 : Les principaux concepts

2. Les tags et l'étiquetage de contenus en ligne :

Un *tag* (étiquette) est un mot clé librement choisi par un internaute pour décrire un contenu partagé sur le Web. Ce mot clé peut aussi être une combinaison de plusieurs mots : par exemple, « guerre contre le terrorisme ». Un *tag* peut être associé à tout type d'information : des photos (e.g., www.flickr.com), des articles (e.g., www.digg.com), des blogs (e.g., www.technorati.com), des vidéos (e.g., www.dailymotion.com), et même des pages présentant des produits ou des magasins (e.g., www.eurekster.com). Comme les *tags* sont librement choisis, les contenus peuvent être décrits à partir de tout mot qui définit une relation entre la ressource en ligne et un concept activé dans l'esprit de l'internaute [6].

Par exemple, en voyant la vidéo d'un chien qui tombe dans une piscine, un internaute pourrait choisir les *tags* « chien » et « humour », alors qu'un autre choisira le *tag* « plouf le chien ».

Les *tags* peuvent s'analyser comme des métadonnées (i.e. des données sur des données). Elles peuvent avoir des origines différentes [8] :

- un utilisateur (e.g., un internaute lit un article et lui attribue les *tags* « trop long », et « élections »),
- l'auteur du contenu (e.g., le rédacteur de l'article lui attribue les *tags* « une autre démocratie », « bon choix » et « non partisan »)
- un professionnel (e.g., le responsable d'un centre de documentation attribuera le *tag* « politique » à cet article).

Pour pouvoir attribuer des *tags*, il faut faire partie d'une plateforme qui propose ce service (généralement gratuitement). Cela implique de créer un compte utilisateur (identifiant et mot de passe) sur un site Internet tel que Flickr ou Technorati. Ce genre de sites est qualifié de « communautaire ». En effet, en étiquetant des contenus en ligne (attribution de *tags*) l'internaute partage de l'information et prend part à un projet collaboratif. Cette philosophie participe de ce que l'on appelle le Web 2.0, bien qu'il n'en existe pas vraiment de définition claire. Par exemple, les *blogs* sont une autre manifestation de cette philosophie collaborative : au travers de leurs réactions à un message, les lecteurs d'un *blog* contribuent à la création et l'échange d'information. Au final, on peut définir l'étiquetage comme le processus par lequel plusieurs utilisateurs ajoutent des métadonnées sous la forme de mots clés à un contenu partagé en ligne [3].

Lorsqu'un contenu donné (e.g., une image) est étiqueté, on dispose de l'ensemble des *tags* proposés par les internautes ayant accédés à ce contenu. Pour chaque *tag* attaché à ce contenu, on connaît en outre : la date de création du *tag*, le nombre de fois où il a été proposé pour ce même contenu, et le nom des utilisateurs (ou pseudonyme) qui l'ont choisi pour décrire ce contenu. En cliquant sur un *tag*, on peut aussi découvrir les autres contenus qu'il sert à décrire (le cas échéant). L'accès à toutes ces informations est libre et gratuite. La fonction principale des *tags* est donc d'aider les internautes à mieux organiser et retrouver des contenus en ligne. Par exemple, si je recherche une photo comique mettant en scène un chat, je peux lancer une recherche sur le site Flickr en utilisant les mots clés « *cat* » et « *funny* ». J'obtiens alors une liste de photos qui ont été principalement étiquetées par les internautes qui m'ont précédé en utilisant les *tags* « *cat* » et « *funny* » [13].

3. Folksonomie :

Dans sa définition initiale des folksonomies, Thomas Vander Wal1(1) soulignait leur complète dépendance vis-à-vis des tags, notant aussi leur préséance comme en échos lointain au fait qu'il à fallut attendre prêt de deux ans après la création de Muxway, l'ancêtre de del.icio.us, pour qu'émergeât une appellation correspondant aux résultats socialement partagés de la pratique du tagging. Pourtant, en dépit de cette primauté unanimement admise, force est de constater la relative absence de toute caractérisation précise des tags, les chercheurs ayant souvent préféré étudier les enjeux afférant aux folksonomies. Sans doute faut-il lire dans ce diagnostic l'effet d'une saturation due au vocabulaire existant, le recours incessant à la notion pour le moins confuse et protéiforme de « mot clef » ayant achevé d'obscurcir les discussions autour du tagging en y associant pêle-mêle les balises <meta> des pages HTML, les requêtes en langage naturel formulées par l'entremise des moteurs de recherche ou encore les langages documentaires et leurs multiples déclinaisons lexicales : mots clef, nous l'avons dit, mais aussi vedettes matière ou descripteurs. Esquisser une caractérisation des tags suppose de prendre en compte le contexte technique les ayant vu naître, avant tout lié au Web et à ses technologies en constante évolution, de même que les multiples déclinaisons auxquelles cette dynamique a donné naissance. C'est à la seule condition d'accorder suffisamment d'attention à ce milieu technique qu'il sera possible de dégager, par contraste, la part proprement « symbolique » des tags, pour enfin penser l'entrecroisement de ces deux dimensions [9].

Une *folksonomie* est un système de catégorisation qui émerge de l'étiquetage collaboratif sur Internet. Il s'agit donc simplement d'un ensemble de mots que les membres d'une même communauté informelle (e.g., les utilisateurs de Flickr) utilisent pour étiqueter des contenus en ligne. Une *folksonomie* est donc liée à un site communautaire bien particulier : par exemple, la *folksonomie* de Flickr est différente de celle de Dailymotion. Ce néologisme est la francisation du terme *folksnonomy*, qui provient lui-même de la contraction des termes *folks* (gens) et *taxonomy* (bien qu'il ne s'agisse pas d'une taxonomie) qui évoque l'idée d'une catégorisation par les gens. On trouve parfois des termes synonymes (mais de moins en moins utilisés) : *mob indexing*, *mobdexing*, *folk categorization* (qui est d'ailleurs plus juste), *social tagging*, etc.

L'étiquetage, en lui-même, ne constitue pas une *folksonomie*. Il est tout à fait possible d'attribuer des *tags* sans créer une *folksonomie*. Celle-ci repose plutôt sur la

capacité d'agrégation des *tags* : l'internaute peut les organiser autour de modèles empiriques observés à travers l'usage que font les autres membres de la même communauté informelle. Sans cette agrégation, les *tags* restent de simples étiquettes sans aucun sens collectif ; elles reflètent alors uniquement le sens individuel que chaque internaute lui donne.

Malgré son nom, une *folksonomie* est différente d'une taxonomie car, d'une part, elle n'est pas contrainte par des relations hiérarchiques, et d'autre part, elle n'est pas conçue par des experts. Il ne s'agit pas non plus d'une ontologie (au sens informatique, [9] ; [14]. Une ontologie est un ensemble structuré de concepts, alors qu'une *folksonomie* ne possède qu'une structure émergente, floue, et non contraignante (e.g., un internaute peut utiliser un *tag* dans un sens totalement différent des autres utilisateurs) [13].

4. Le tag : étiquette matérielle et libellé, entre accès et référence :

4.1. Les trois dimensions canoniques :

S'il ne faut pas chercher de définition canonique précise du tag, paradoxalement, du fait du rapprochement opéré de bonne heure entre folksonomies et ontologies, longtemps perçues, de prime abord dans un rapport exclusif de pure et simple opposition, nous disposons d'ontologies susceptibles de nous éclairer quant à ses propriétés constitutives. A cet égard, un quasi consensus se dégage de la littérature dévolue à cette question. Une ontologie du tag comme celle de Richard Newman par exemple (c'est également vrai de la *TagOntology* de Thomas Gruber[4]) se propose avant tout de décrire un processus individuel de tagging en opérant une distinction entre:

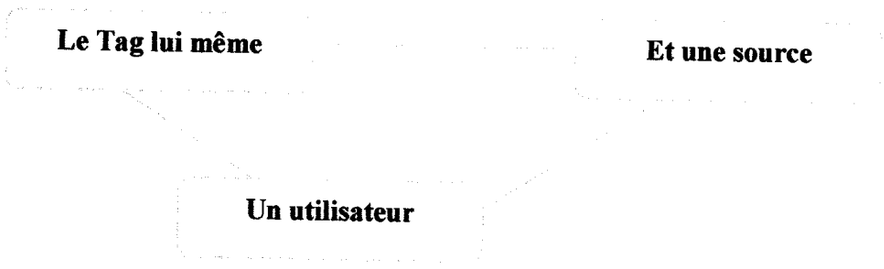


Figure 1.2 Les trois axes de la définition/réification courante du tag

Une telle tripartition, à laquelle on serait bien en peine de chercher des alternatives radicales dans la littérature dévolue à ces questions, pourrait cependant sembler limitée en ce qu'elle oblitère la nature duale du tag : à la fois instance *matérielle* (à l'instar de l'étiquette concrète à laquelle son nom est attaché) mais également *symbolique*. Associer ces deux aspects c'est oublier que le lien symbolique usuel entre mots et choses ne nécessite aucunement de se voir implémenté d'une quelconque manière. Nul besoin d'avoir recours à des moyens d'ordre techniques pour qu'un mot atteigne son objet, aucun artefact n'y pourvoira ; autrement dit, la référence ressortit à la seule sémantique.

A l'inverse, chaque site qui emploie des folksonomies définit, selon ses besoins propres, les règles encadrant le tagging (qui a le droit de tagger ? quoi ?, comment ?, etc.), complémentant, *de facto*, la relation de *référence* par une relation associant matériellement (la dépendance de cette relation vis-à-vis d'un réseau informationnel physique en atteste) le tag à une ressource, fondée sur la notion *d'accès*. Ses tenants et aboutissants sont à chercher du côté du design informationnel des interfaces et de la réalité technique des réseaux, en particulier l'architecture du Web, et non plus simplement de l'analyse du langage. Or, l'absence de la dichotomie référence/accès au coeur de l'ontologie de Newman, ne va pas sans entraîner de sérieuses conséquences. Au premier rang desquelles, celle-ci : une telle ontologie vaut autant par ce qu'elle précise et explicite (d'où son incorporation à d'autres ontologies plus vastes à l'instar de SIOC) que pour ce qu'elle offusque et qu'il nous semble impératif de restituer.

Une vision du tag tournant invariablement autour de trois axes, sans que la relativité des interfaces ne modifie cette donnée essentielle, apparaîtra pour le moins discutable du point de vue de la *description*. En revanche, une fois implémentée, elle fournira un support adéquat pour réaliser (*prescrire*) une interopérabilité entre services utilisant le tagging, implémentant, par ce fait même, une définition unifiée du tag indépendamment de toute autre considération.

4.2. Du tag au machine-tag : entre accès et libellé :

La relation d'accès diffère de la relation de référence en ceci qu'elle est indissolublement d'ordre causal, et par conséquent matériel, et relie, par une relation d'ostension, l'utilisateur à la chose taggée. L'ostension, comme n'ont cessé de le montrer les philosophes, en particulier depuis Wittgenstein, se caractérise par un rapport

d'indétermination intrinsèque à l'objet, ou, pour préciser les choses d'une manière plus conforme aux usages et à la réalité technologique du Web, à la ressource ainsi désignée.

Le libellé n'est rien d'autre, quant à lui, que la suite de caractères inscrite *à même le tag*, lui-même conçu à la manière d'une étiquette. En ménageant un accès à la ressource (informationnelle ou non), cette étiquette permet à l'utilisateur d'associer à celle-ci le texte qu'il désire. Il devient dès lors loisible d'indexer, d'évaluer, de partager ou encore de retrouver des objets qui échappaient jusqu'alors à ces possibilités d'annotations. Précisément ce que permettait depuis longtemps, dans l'univers analogique, le traditionnel post-it : produire une surface matérielle accueillant du texte là où celle-ci faisait défaut. Illustration de ce constat, l'application *Lignes de temps*, développée au sein de l'IRI, autorise un accès technique à des séquences filmiques qu'il devient possible d'annoter (à l'instar, sur un versant cette fois-ci collaboratif, de la toute récente (*VideoTagGame* de Yahoo!).

Ultime avatar de cette logique de mise à disposition de nouveaux supports '*ad hoc*' : l'application *Gallery* de Windows Vista. En permettant de tagger ses photos localement, elle ne propose plus d'accoler une étiquette à une ressource numérique mais de l'y injecter directement. A la différence du post-it qui est souvent appelé à jouer le rôle de pense-bête mais dont la perte menace d'entraîner avec elle celle de l'objet étiqueté. *Gallery*, en conservant la trace de tous les tags présents dans le système, garantit un accès pérenne à l'ensemble des ressources taggées. La logique n'est plus celle du raccourci, susceptible de pointer dans le vide pour peu que le chemin d'accès vers la ressource visée se soit modifié, mais de *l'incorporation* au système des images, dès lors accessibles au même titre que les tags eux-mêmes ; tags et ressources formant un nouvel ensemble disposant de ses caractéristiques propres, immunisé contre la perte.

Rien, et l'analogie avec les étiquettes matérielles le confirme, ne contraint l'utilisateur à inscrire sur le tag une chaîne de caractère formant un mot – sans parler de sa forme lemmatisée, et ce, en dépit des contraintes syntaxiques minimales qu'imposent les différent systèmes existants. Il ressort de ce constat l'impossibilité d'assimiler les tags à quelque formes linguistiques précises que ce soit ; d'où une réalité incomparablement plus complexe que ne le laissait percevoir le précédent schéma à la structure ternaire où l'entité « *tag* », pour réifiée qu'elle fût en ses relations avec un *utilisateur* et une *ressource*, n'était paradoxalement pas interrogée, bien que sertie d'attributs destinés à la rendre manipulable.

Ce faisant, la figure unitaire du tag cède la place à une entité biface, à la fois réalité matérielle attachée à une ressource de par leur insertion à toutes les deux dans un système technique donné, et relation sémiotique et langagières portée par des items de natures variables.

Avec, au surplus, et du fait de cette absence de contrainte qui constitue l'essence même du tagging, un mélange éventuel des deux : ce sont les fameux *triple tags* ou *machine-tags*, popularisés par Flickr. Structuration légère des contenus produits par les utilisateurs, ils articulent trois dimensions : un espace de nom, un prédicat et une valeur associée.

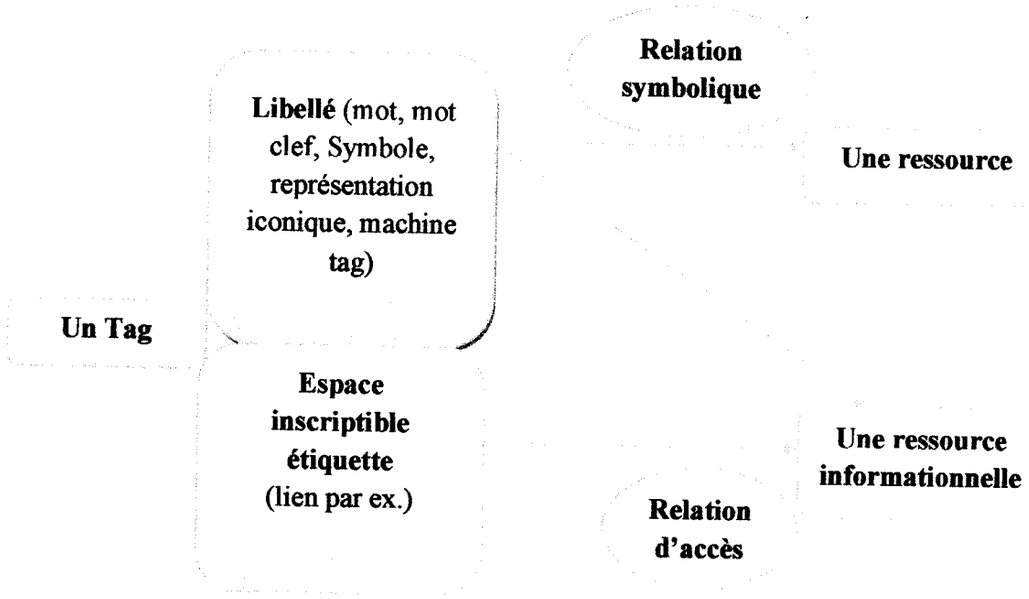


Figure 1.3 La bipartition du tag

En scindant l'information de la sorte, et en la répartissant conséquemment au sein d'une base de données spécifiquement dédiée à chacune de ces trois facettes, il devient possible de poser des requêtes sur un espace de nom donné, un prédicat, une valeur, ou l'une quelconque des combinaisons associant ces trois aspects.

Ces tags représentent un cas limite qui illustre néanmoins la dualité accès/référence inhérente au modèle proposée dans la figure 1.3. Historiquement, deux cas de figure se présentent. Dans le premier, le machine-tag, baptisé *triple tag* à l'origine, est utilisé en guise de libellé : sa syntaxe, calquée sur les langages de type XML et influencée par le développement des micro-formats et de RDF, permet

d'exprimer des relations complexes à l'aide d'un seul et unique tag. Celui-ci, une fois inséré dans une URL, donne alors accès à une ressource quelconque.

Dans le second cas de figure, le machine-tag n'est associé à aucune URL en ce qu'il se suffit à lui-même, en vertu de sa syntaxe interprétable par les machines, pour ménager un accès à la ressource numérique par l'intermédiaire des requêtes lancées via les API des sites où furent développées des applications susceptibles de traiter ce type d'informations (c'est le cas de Flickr qui, sous l'impulsion des utilisateurs qui usaient déjà des machine-tags en guise de simples libellés, ajouta à son API de nouvelles fonctionnalités permettant de les parser, *reconduisant ainsi du même coup la dichotomie accès/référence au niveau du libellé lui-même*).

Aucune règle expresse ne commandant l'inscription d'un mot ou d'un signe humainement compréhensible, rien n'interdit non plus le recours à du code informatique en guise de libellé. Mieux, une vague syntaxe que des utilisateurs reconstruisaient aisément sous forme d'énoncés du langage naturel donna naissance à des micro-formats à partir du moment où les machines furent reprogrammées pour l'« interpréter » de manière adéquate. Bien qu'extrêmement sommaire, de pseudosyntaxe calquée sur des langages informatiques existants, elle acquît *ipso facto* le statut parallèle de syntaxe informatique de plein droit. Sous cet angle, les machines tags résultent bien d'un développement logique qui tire partie de la caractéristique définitionnelle fondamentale des tags.

5. Les mots clef :

Il est difficile, voire impossible, de s'accorder aujourd'hui sur une définition consensuelle du mot clef, celui-ci connaissant un extraordinaire succès depuis son emploi pour signifier des requête effectuées (en plein texte) via des moteurs de recherche, jusqu'à sa caractérisation par les normes documentaires. C'est à cet usage bien établi, et à lui seul, que nous nous référerons pour l'occasion. « Mot ou groupe de mots choisi soit dans le titre ou le texte d'un document, soit dans une demande de recherche documentaire, pour en caractériser le contenu », selon la définition de la norme AFNOR NF Z 47-102, le mot clef, bien qu'issu du langage naturel, se conçoit la plupart du temps comme extrait directement d'un document analysé. Or, dans le cas précis qui nous occupe, les types de documents soumis au tagging varient dans des proportions suffisamment importantes pour qu'il ne soit tout simplement pas possible de

définir les libellés uniquement en ces termes : qu'ont de commun, en effet, une photo, un événement, un plan séquence ou un enregistrement audio ? S'agissant de ce qui retient ici notre attention, leur nature de document non-textuel n'offre guère de prise à l'extraction directe de mot clef. Inversement, la force du tagging, face à ce type d'objets, est précisément de nous permettre d'ajouter du texte (entendu au sens large, cf. *supra*). A une logique d'extraction à laquelle se prêtent tout particulièrement les documents de nature textuelle, succède une autre logique, *expressive* et non seulement descriptive, que sont incapables de capter les applications proposant des « tags » générés automatiquement. Tagger c'est aussi ajouter un contenu absent d'un document (ou d'une ressource) ; en d'autres termes, lui adjoindre un contenu *extrinsèque*.

6. Les nuages de mots :

Les représentations visuelles des *folksonomies* permettent aux internautes de les utiliser pour leur recherche et leur activité d'étiquetage. Il en existe plusieurs : diagrammes, réseaux sémantiques (où chaque nœud représente un *tag*), etc. Mais la représentation la plus commune sur les sites Web est le nuage de mots. Son succès provient probablement de sa facilité d'utilisation et de sa capacité à fournir de manière simple un assez grand nombre d'informations. Un nuage de mots est une présentation visuelle en deux dimensions des *tags* utilisés pour décrire les contenus d'un site Web particulier, ou des contenus extérieurs mais indexés sur ce site (voir la figure 1.4 pour un exemple).

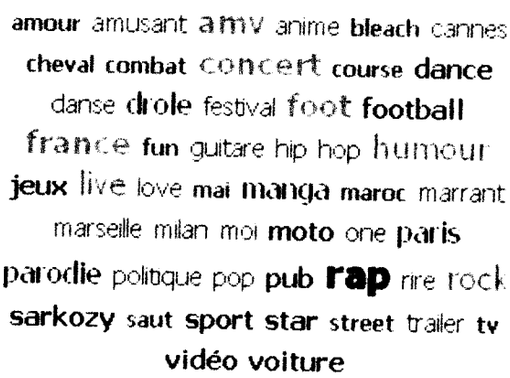


Figure 1.4 Exemple d'un nuage de mots

Pour faciliter la recherche de contenus, un nuage de mots peut être construit selon divers paramètres :

- L'ordre des *tags* : il peut être alphabétique ou en fonction de la popularité des mots.

- La couleur de *tags* : elle peut indiquer l'origine des *tags* (l'utilisateur lui-même, les autres, les utilisateurs les plus populaires, etc.).

- Le contraste des *tags* : les plus foncés sont les plus récents.

- La taille des *tags* : les plus populaires sont affichés en taille plus grande. Deux standards sont utilisés : soit la taille représente le nombre de fois que le *tag* en question a été attribué à un contenu donné, soit elle représente le nombre de contenus qui ont été étiquetés avec chaque *tag* (la taille devient alors un indicateur de la popularité du *tag*, et donne aussi une indication sur les centres d'intérêt de la communauté).

Les nuages de mots sont interactifs : chaque *tag* affiché est un lien vers une page de résultats qui contient une liste des contenus qui ont été indexés avec le mot en question. Il existe des nuages de mots améliorés, comme par exemple celui du moteur de recherche www.quintura.com qui mixe réseau sémantique et nuage, le tout d'une manière dynamique.

7. Avantages et inconvénients :

L'étiquetage collaboratif est un phénomène relativement récent qui commence à susciter de nombreuses recherches en informatique et en sciences de l'information. On commence donc à en identifier les principaux avantages et inconvénients.

L'étiquetage collaboratif de contenus Web, avec son outil de base (les *tags*) et sa principale conséquence (l'émergence d'une catégorisation représentée par un nuage de mots) présente plusieurs avantages. D'abord, la catégorisation qui émerge de cette pratique est plus intuitive pour les utilisateurs, car elle est le fruit du codage des internautes eux-mêmes. Les recherches sont ainsi plus aisées que lorsqu'il s'agit de naviguer dans une arborescence pré établie à la recherche d'un contenu particulier (comme c'est souvent le cas dans les sites marchands en ligne). Puisque l'étiquetage est continu, la catégorisation est dynamique : elle est mise à jour automatiquement et en permanence. En outre, les recherches peuvent conduire l'utilisateur à découvrir des items auxquels il ne pensait pas *a priori*, mais qui sont néanmoins liés aux thèmes qui l'intéressent.

Ensuite, la catégorisation découlant de l'étiquetage collaboratif est ouverte : elle peut être extrêmement réactive aux changements de perception de contenus des internautes, ce qui améliore la probabilité d'obtenir des résultats pertinents lors des recherches. Par ailleurs, la catégorisation de ressources en ligne par l'étiquetage collaboratif est peu chère à créer et à maintenir, car elle résulte du travail gratuit des internautes. Ceux-ci ne fournissent d'ailleurs qu'un effort cognitif limité, car ils n'ont pas besoin de comprendre et d'apprendre une taxonomie imposée, avec une nomenclature pré établie et des règles de classification potentiellement complexes, avant de pouvoir indexer et rechercher des contenus. Au contraire, c'est eux qui sont à l'origine d'une classification émergente en constante évolution. Les internautes sont ainsi actifs et impliqués dans le processus, que l'on peut qualifier de démocratique. En outre, les utilisateurs ont la possibilité d'utiliser un *tag* particulier pour signaler des contenus inappropriés, ce qui rend le système autocensuré.

Enfin, l'étiquetage fournit des informations explicites sur la manière dont chaque internaute impliqué indexe les contenus auxquels il accède, ce qui peut être utile pour étudier sa perception et sa façon de raisonner. D'un point de vue collectif, on obtient une perception de l'ensemble de la communauté sur un contenu, ce qui peut être fort instructif. Néanmoins, l'étiquetage de contenus sur Internet présente également des inconvénients.

La première limite de ce système est l'ambiguïté des *tags*, leur caractère personnel, et leur possible inexactitude en l'absence d'un contrôle. L'ambiguïté tient à l'existence de synonymes (mots différents, même sens) ou d'homonymes (même mot, sens différents). En outre, la précision du système est amoindrie parce que les internautes utilisent en même temps des formes plurielle et singulière de mêmes mots, des verbes conjugués, des acronymes, des mots techniques, voire des termes incompréhensibles en dehors d'un sous-groupe de référence. Quant à la possible inexactitude des *tags*, (Guy et Tonkin) ont observé qu'environ un tiers des *tags* utilisés sur Flickr et Delicious (http://del.icio.us) est erroné [6]. Le système est également sensible à la pollution : des internautes malveillants peuvent *spammer* le système.

En outre, celui-ci peut être encombré par des *tags* sans intérêts car trop idiosyncrasiques tels que « ne pas oublier » ou « moi » qui ne sont d'aucune utilité pour les autres utilisateurs. Notons enfin que les nuages de mots peuvent être difficiles à lire pour des personnes dont la vue est faible (certains mots étant écrits en petite taille).

Contrairement à un texte affiché en HTML, l'internaute n'a pas la possibilité de modifier la taille d'affichage des mots du nuage.

Malgré ces quelques limites, il nous semble que l'étiquetage collaboratif pourrait être avantageusement utilisé par les sites marchands. A titre d'exemple, on peut noter qu'il a été récemment utilisé avec succès par le *Metropolitan Museum of Art* (New York) pour permettre un accès plus intuitif aux objets exposés [12]. Dans le cas d'un site marchand, cet emploi suscite beaucoup d'interrogations, tant au niveau de la mise en œuvre d'un tel système, que de ses effets possibles.

8. Conclusion :

Cette étude a été rendue difficile par le manque de littérature adéquate. Le sujet du tag n'est que peu traité. Et la majorité des ouvrages qui l'abordent préfèrent agrémenter leurs pages de superbes photos de graffitis ou de brûlures, ce chapitre décrit tout d'abord ce que c'est un tag et l'étiquetage de contenus en ligne ensuite les mots clés. Enfin il introduit le concept de nuage de mots et cite quelque avantage et inconvénients de l'étiquetage collaboratif.

CHAPITRE 2

LES OUTILS EXISTANTS

2.1 Introduction :

Pour avoir un nuage de mots il faut construire d'abord un outil qui permet de le réaliser, dans ce chapitre on va voir quelques outils déjà existant qui génèrent des nuages de tags.

2.2 Essais de quelques générateurs :

Voici quelques générateurs de Tag Cloud que nous avons recensé :

- MozbotMozclouds : http://web.mozbot.info/mozclouds/
- Outils de référencement : http://www.referencement-page1.fr/tag-cloud/tag,cloud.php
- TagCrowd generator : http://www.tagcrowd.com/
- Clusty cloud generator : http://cloud.clusty.com/
- Tag cloud generator 1 : http://www.tag-cloud.de/
- Tag cloud generator 2 : http://www.tagcloud-generator.com/
- Korpus : http://www.lepotlatch.org/korpus/index.php
- Wordle.net : http://www.wordle.net/
- TextTagCloud : http://www.artviper.net/texttagcloud/e

Après avoir testé les générateurs ci-dessous, on a choisi de parler sur 3 générateurs.

2.2.1 Générateur n°1 : « MozbotMozclouds » :

MozbotMozclouds est un générateur de nuages de mots clés par l'analyse d'une page web.

Voici ce que propose ce dernier :

- ✓ Choix des zones à analyser : balise Title, balises meta, options "Alt" des images, options "Title" des liens et texte de la page.
- ✓ Possibilité de fixer une fréquence minimale des mots : seuls les mots présents au moins "X" fois seront affichés.
- ✓ Possibilité d'indiquer une taille minimale en termes de nombre de caractères : seuls les mots faisant au moins "Y" caractères seront affichés.
- ✓ Recherche d'expressions (suiddtes de deux ou trois mots clés).

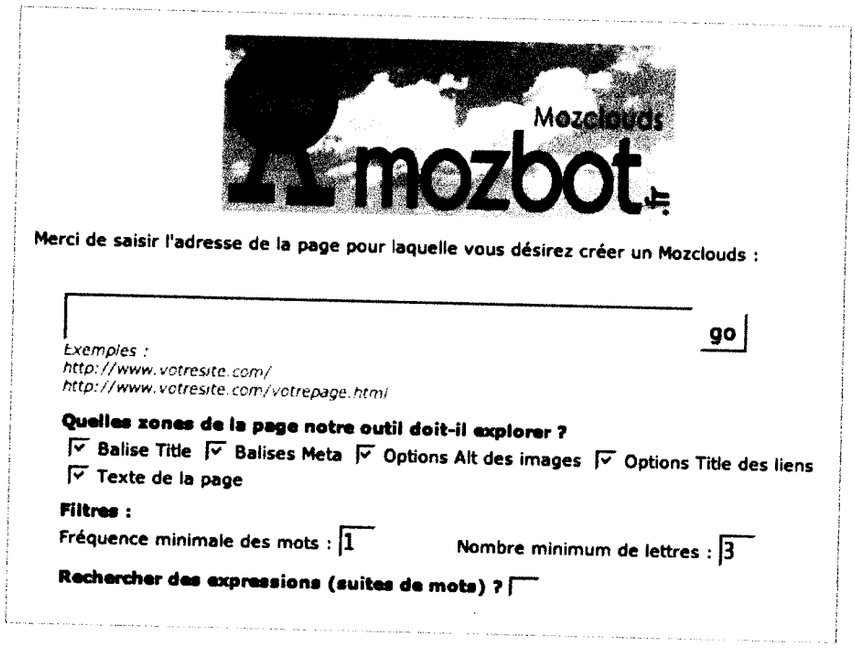


Figure 2.1 : Capture d'écran, le générateur mozbot

Fonctionnement :

Après avoir passé l'étape présentée sur la capture d'écran n°1, on doit sélectionner les mots clés les plus souvent présents dans la page, ensuite faire nos paramètres (couleurs d'affichage des tags si c'est multi couleurs ou couleur unique, police de caractère, tri des mots clés, type de recherche …) et enfin on peut visualiser le tag Cloud ainsi généré.

Notons qu'il nous génère aussi le code HTML du tag Cloud, ce qui permet de le modifier à notre convenance et pouvoir l'insérer une page web par exemple.

2.2.2 Générateur n°2 : « wordle.net » :

Le programme travaille à partir d'un texte, d'une page web ou d'un nom d'utilisateur delicious. La fréquence des mots est utilisée pour leur donner leur taille. Le fonctionnement est très fluide et rapide, et il est possible d'obtenir de très beaux résultats très rapidement.

Une touche ***"randomize"*** permet de changer aléatoirement les paramètres pour proposer un tout autre nuage.

Les options de personnalisation sont nombreuses.

- Filtre sur les mots (en fonction de la langue)

- Limite du nombre de mots
- Choix de la police
- Choix de l'agencement (ordre des mots, orientations, forme du nuage)
- Modification de la palette de couleur.

Ce nuage peut constituer une très bonne base de travail à un graphiste, ou peut servir à créer très rapidement un joli nuage de mot. L'inconvénient est que ce générateur de tags ne nous permet pas de créer un nuage de tags dynamique et esthétique en même temps.

Nuage de tags d'un compte delicious : icetamango

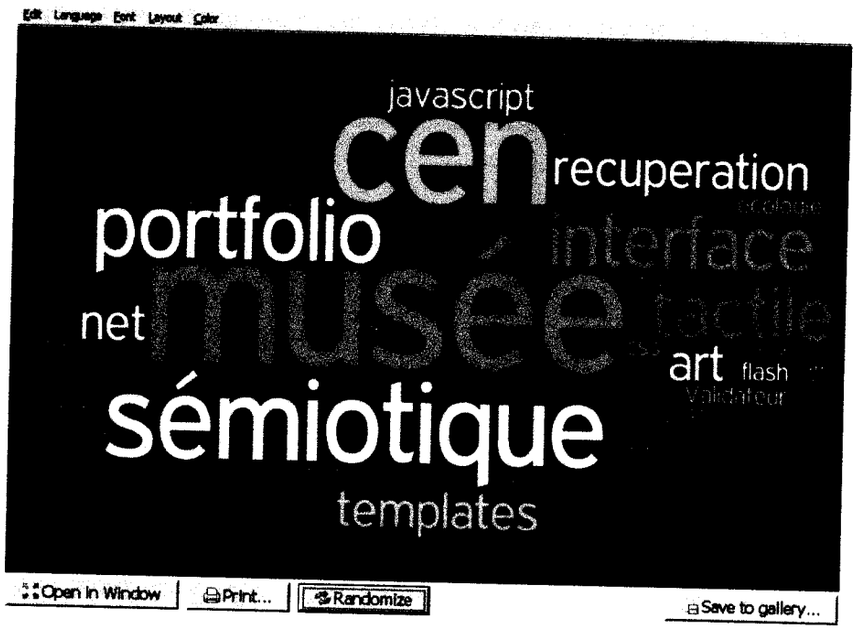


Figure 2.2 : Capture d'écran, le générateur wordle

2.2.3 Générateur n°3 : « Clusty Cloud generator » :

L'outil **Clusty cloud** permet de générer à la volée un nuage de mots reliés à un mot ou une expression de notre choix.

Le résultat est, comme par exemple ci-dessous pour l'expression "Obama", un nuage de mots qui, lorsque l'on clique dessus, conduisent aux résultats (en temps réel) de la recherche sur ce mot à l'aide du moteur Clusty. Concrètement, si par exemple nous voulons afficher dans notre blog ou site web un nuage de liens, on peut utiliser cet outil.

Notons que tout comme le générateur n°1, celui-ci nous génère aussi le code HTML du tag Cloud, ce qui permet de le modifier à notre convenance et pouvoir l'insérer dans une page web par exemple.

Exemple :

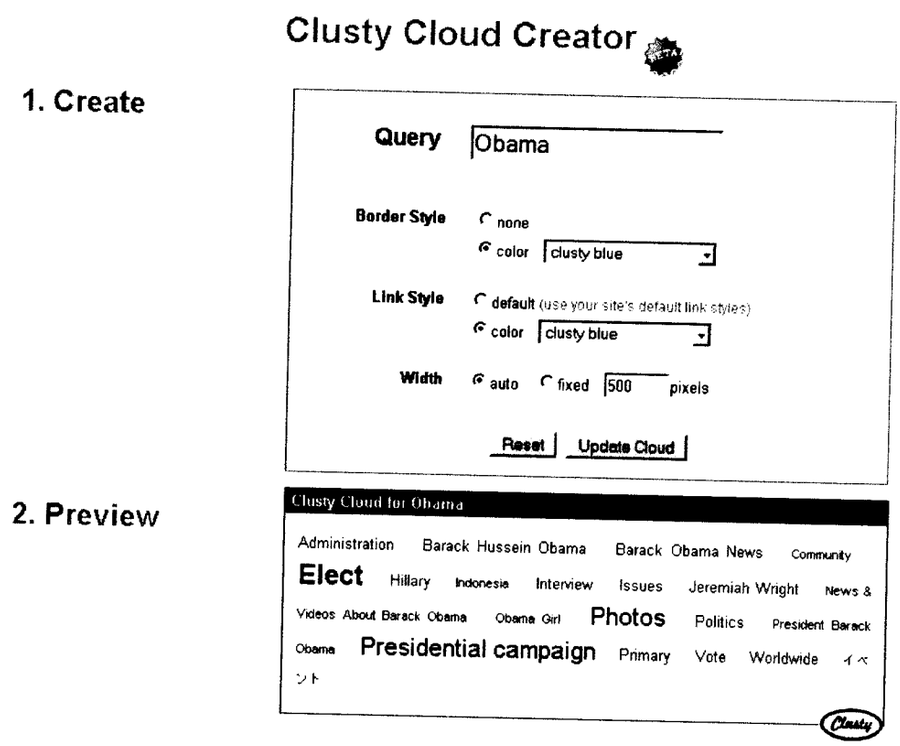


Figure 2.3 : Capture d'écran, le générateur Clusty.

2.2.4 Le générateur le plus intéressant :

Le générateur de tags qui nous a le plus attiré lors de nos nombreux tests est donc le Tag Cloud generator 2 : http://www.tagcloud-generator.com/ . Il est très intéressant dans la mesure où il nous donne la possibilité de générer un nuage de tags à la fois esthétique et dynamique.

Ce générateur nous donne la possibilité de choisir la couleur de nos liens mots clés, le type et la taille de la police de caractère, le positionnement du paragraphe, etc.

Outre cela, il donne la possibilité d'obtenir nos tags automatiquement à partir de l'URL du site en question, mais il nous donne aussi la possibilité d'ajouter nous même nos tags manuellement. Et dans ce cas, nous devons joindre à chaque fois les URL contenant nos tags.

Et tout comme précédemment, il y a une génération du code HTML, ce qui permet de modifier à notre convenance et pouvoir insérer le nuage de tags dans une page web par exemple.

Voici donc un exemple avec les quatre étapes de génération du tag Cloud avec cet outil :

Etape n°1 : Ajout des mots clés

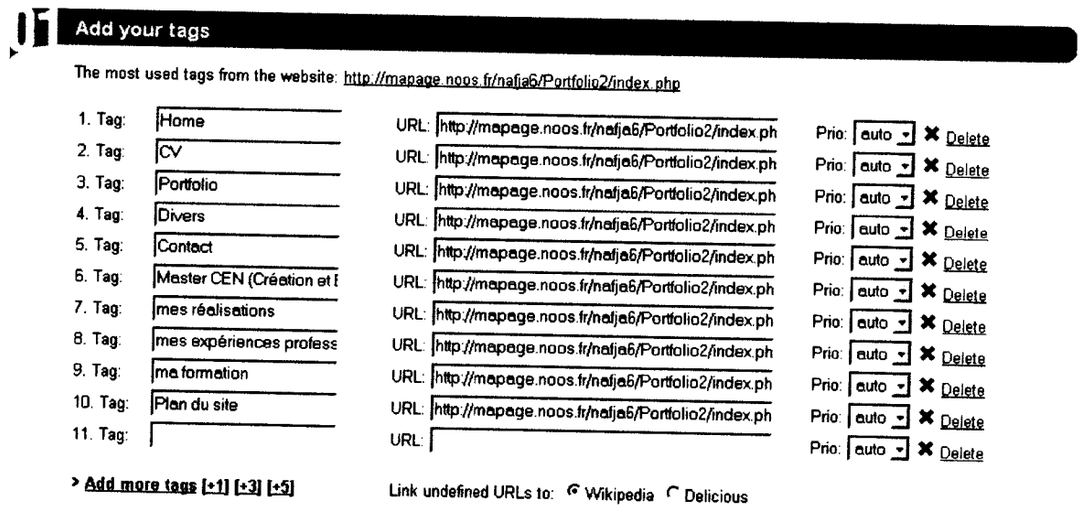


Figure 2.4 : Capture d'écran, ajout des mots clés dans le générateur tagcloud.

Etape n°2 : Personnalisation du nuage de tags

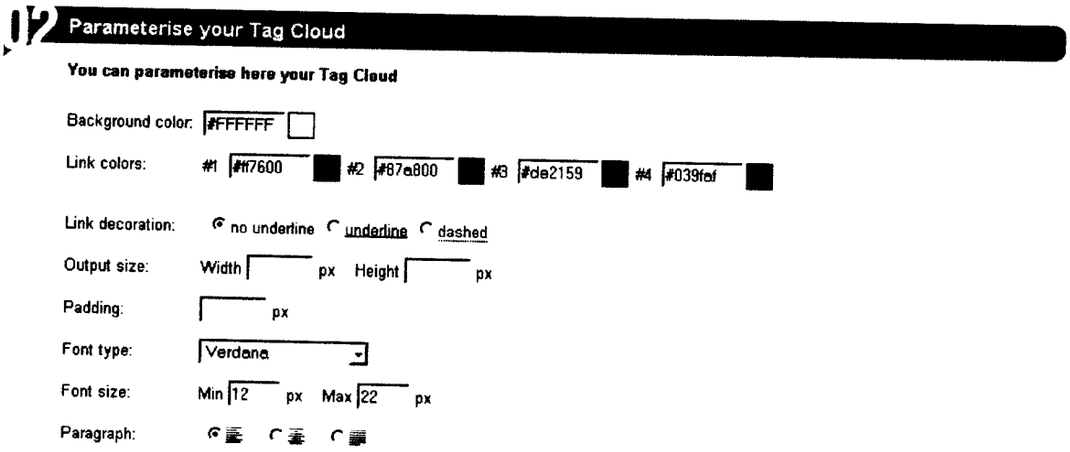


Figure 2.5 : Capture d'écran, personnalisation du nuage de tags dans le générateur tagcloud.

Etape n°3 : Prévisualisation de notre nuage de tags ainsi généré

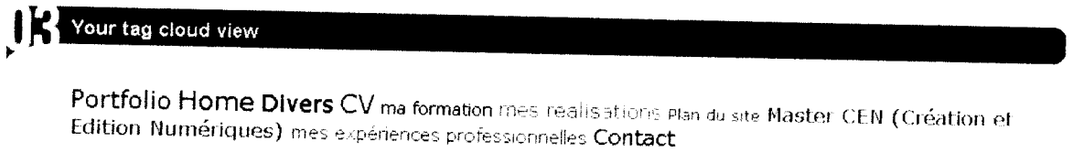


Figure 2.6 : Capture d'écran, prévisualisation du nuage de tags dans le générateur tagcloud.

Etape n°4 : Le code HTML pour notre site Web

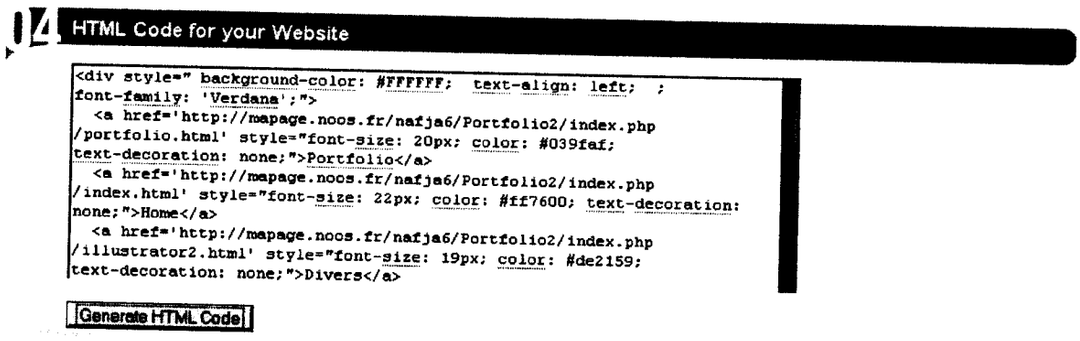


Figure 2.7 : Capture d'écran, le code HTML généré par le générateur tagcloud.

2.3 Conclusion :

Ce chapitre a été très instructif, dans la mesure où on a testé des outils très intéressants qui génèrent des nuages de tags et aussi de voir leur façon d'afficher les nuages de tags. On ne s'arrêtera pas à ce stade, on à bien envie de poursuivre et de voir les possibilités qui s'offrent à nous pour construire un outil semblable qui permet de traiter la langue arabe qui est notre langue maternelle.

CHAPITRE 3

RECHERCHE D'INFORMATION ET LA LANGUE ARABE

3.1 Introduction

Le but de la recherche d'information (RI) est de développer des systèmes capables de retrouver parmi un ensemble de documents ceux qui répondent au mieux à la requête d'un utilisateur. Pour cela, il est important de constituer une représentation du contenu du document et de la requête afin de procéder à un appariement plus pertinent entre eux. L'approche souvent adoptée en RI textuelle est plutôt de chercher des représentants qui correspondent généralement, dans le cadre de l'indexation automatique, à un ensemble d'unités lexicales extraits des documents et requêtes, nommés termes d'indexation.

L'indexation consiste donc à associer à chaque document (ou à chaque requête) un descripteur (également nommé index) formé de l'ensemble des termes d'indexation extraits de son contenu. Pour établir une correspondance entre documents et requêtes, représentés par des descripteurs, les SRI se basent sur des modèles de RI. Ils permettent :

- d'offrir une interprétation aux descripteurs en donnant une représentation interne des textes et des questions basée sur les termes d'indexation ;
- de définir les stratégies à adopter pour comparer les représentations des documents et des requêtes. Leur comparaison donne lieu à un score qui traduit leur degré de ressemblance ;
- de proposer éventuellement des méthodes de classement des résultats retournés à l'utilisateur.

Une fois les représentations des documents et des requêtes mises en correspondance, le système retourne à l'utilisateur la liste des documents répondant à sa requête. Ainsi, des méthodes et des mesures d'évaluation sont nécessaires pour estimer la validité des résultats retournés par le système. Une partie de ce chapitre y est consacrée [11].

3.2 Les principaux acteurs de RI :

L'objectif principal d'un système de recherche d'information (SRI) est de sélectionner dans une collection de documents ceux qui sont susceptible de répondre au besoin en information de l'utilisateur exprimé à travers une requête.

Dans cette définition on distingue trois notions clés : document, requête et pertinence qui sont les principaux acteurs de RI.

-Document : Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc. On appelle document toute unité ou granule documentaire qui peut constituer une réponse à une requête d'utilisateur.

- Requête : Une requête exprime le besoin d'information d'un utilisateur écrite sous plusieurs formes

- Pertinence : La notion de pertinence est très complexe. De façon générale, dans le document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin.

C'est sur cette notion de pertinence que le système doit juger si un document doit être donné à l'utilisateur comme réponse [7].

3.3 Processus de recherche d'information :

Le processus de RI a pour but d'établir une correspondance pertinente entre l'information recherchée par l'utilisateur, représentée généralement par le biais d'une requête, et l'ensemble des documents disponibles. Il s'articule autour de deux étapes essentielles : les phases d'indexation et de recherche. Le processus complet est représenté en figure (3.1).

L'étape d'indexation se base sur l'analyse des documents et des requêtes afin de créer une représentation de leur contenu textuel qui soit utilisable par le SRI. Chaque document (et requête) est alors associé à un descripteur représenté par l'ensemble des termes d'indexation extraits.

La phase de recherche a pour objectif d'apparier les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Elle se base sur un formalisme précis défini par un modèle de RI. Les documents présentés en résultat à l'utilisateur, et considérés comme les plus pertinents, sont ceux dont les termes d'indexation sont les plus proches de ceux de la requête [13].

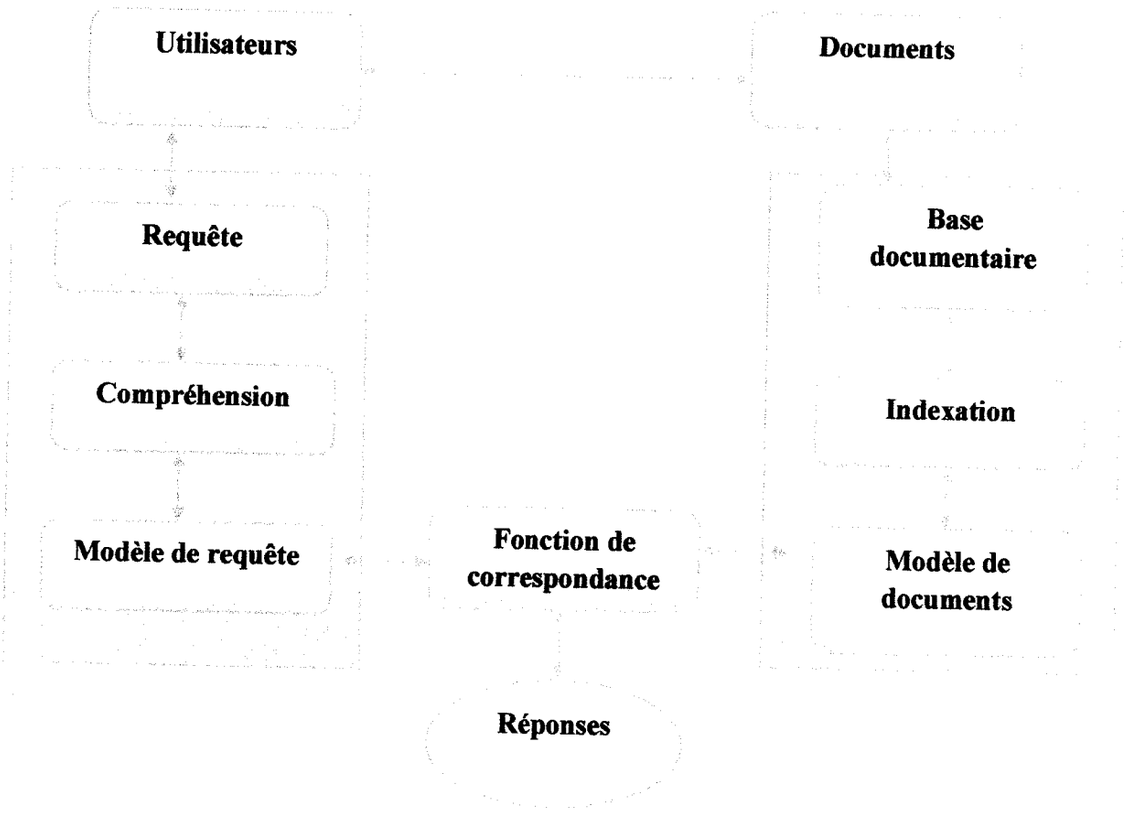


Figure 3.1 – Système de recherche d'information

3.4 Modèles de RI :

Comme nous l'avons vu, le but d'un SRI demeure dans sa capacité à établir une correspondance entre un document et une requête. De nombreux modèles ont été proposés en RI, ils sont généralement regroupés autour des trois familles suivantes [3]:

• les modèles ensemblistes qui considèrent le processus de recherche comme une succession d'opérations à effectuer sur des ensembles d'unités lexicales contenues dans les documents,

• les modèles algébriques au sein desquels la pertinence d'un document par rapport à une requête est envisagée à partir de mesures de distance dans un espace vectoriel,

• les modèles probabilistes qui représentent la RI comme un processus incertain et imprécis où la notion de pertinence peut être vue comme une probabilité de pertinence [13].

3.5 La Langue Arabe :

L'arabe (al 3arabiya en transcription traditionnelle) est la langue parlée à l'origine par les Arabes. C'est une langue sémitique (comme l'akkadien et l'hébreu). Au sein de cet ensemble, elle appartient au sous groupe du sémitique méridional. Du fait de l'expansion territoriale au Moyen Âge et par la diffusion du Coran, cette langue s'est répandue dans toute l'Afrique du nord et en Asie mineure.

Dire langue arabe, c'est donc parler d'un ensemble complexe dans lequel se déploient des variétés écrites et orales répondant à un spectre très diversifié d'usages sociaux, des plus savants aux plus populaires. Mais au delà de cette diversité, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène. Elles sont farouchement attachées à l'intégrité de leur langue, d'où l'importance de l'ASM (Arabe Standard Moderne) qui constitue le terrain commun pour cette large population.

Par ses propriétés morphologiques et syntaxiques, le traitement automatique doit faire face à :

• la nature agglutinante de la langue : l'ensemble des morphèmes collés à l'unité lexicale véhiculent plusieurs informations morphosyntaxiques.

• la richesse flexionnelle de l'arabe

• l'absence de voyellation de la majorité des textes arabes écrits : ce phénomène entraîne un nombre important d'ambiguités morphologiques. En arabe, chaque lettre doit prendre un signe de voyellation et de surcroît les voyelles finales sont porteuses de certains traits morpho-syntaxiques comme la déclinaison, le mode, le cas.

En outre des propriétés linguistiques, l'arabe recense un nombre de ressources linguistiques comprenant des lexiques monolingues et multilingues ainsi que des corpus de langue générale et des corpus de spécialité consacrés à une situation de communication ou à un domaine de la connaissance. L'arabe compte aussi un certain nombre d'outils linguistiques à savoir les analyseurs morphologiques ainsi que les racineurs basés essentiellement sur une procédure de désuffixation qui consiste à supprimer les suffixes qui différencient les flexions des unités lexicales (les formes conjuguées d'un verbe par exemple) [13].

3.5.1 Particularité de la langue arabe :

L'alphabet de la langue arabe compte 28 consonnes (Tableau 3.1). L'arabe s'écrit et se lit de droite à gauche les lettres changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Le Tableau 3.2 montre les variations de la lettre ع (Ayn). Toutes les lettres se lient entre elles sauf (ا, و, ر, ز, د, ذ) qui ne se joignent pas à gauche.

Lettre arabe	Correspondant français	Prononciation
ا	a	Aief
ب	b	Ba'
ت	t	Ta'
ث	th	Tha'
ج	j	Jim
ح	h	Hha'
خ	kh	Kha'
د	d	Dal
ذ	d	Thal
ر	r	Ra
ز	z	Za
س	s	Sin
ش	sh	Shin
ص	s	Sad

Lettre arabe	Correspondant français	Prononciation
ض	d	Dad
ط	t	Tah
ظ	z	Zah
ع	‘	Ayn
غ	gh	Ghayn
ف	f	Fa
ق	q	Qaf
ك	k	Kaf
ل	l	Lam
م	m	Mim
ن	n	Nun
ه	h	Ha
و	w	Waw
ى	y	Ya

Tableau 3.1 : Les 28 lettres arabes [Leclerc 2000]

à la fin d'une lettre non joignable	à la fin	au milieu	au début
ع	ـع	ـعـ	عـ

Tableau 3.2 : Exemple de variation de la lettre ع Ayn

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres (ً, ٌ, ٍ, َ, ُ, ِ, ْ). Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation. Le Tableau 3.3 donne un exemple pour les mots كتب et مدرسة. Cependant, les voyelles ne sont utilisées que pour des textes sacrés et didactiques. Les textes courants rencontrés dans les journaux et les livres n'en comportent habituellement pas. De plus certaines lettres comme ا Alef peuvent symboliser le آ, أ ou إ; de même que pour les lettres ى et ة qui symbolisent respectivement ي et ة [15].

Mot sans voyelles	1ère Interprétation		2ème Interprétation		3ème Interprétation	
كتب	كَتَبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتُبٌ	Des livres
مدرسة	مَدْرَسَةٌ	Ecole	مُدَرِّسَةٌ	Enseignante	مُدَرَّسَةٌ	Enseignée

Tableau 3.3 : ambiguïté causée par l'absence de voyelles pour les mots كتب et مدرسة

3.5.2 Morphologie arabe :

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales [10]. Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est caractéristique à la morphologie arabe. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine. Le Tableau 3.4 donne quelques exemples de schèmes appliqués aux mots كتب écrire et حمل porter. On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux.

Schèmes	KTB	كتب	Notion d'écrire		HML	حمل	Notion de porter
R_1â- R_2i-R_3	KâTiB	كَاتِب	écrivain		HâMil	حَامِل	porteur
R_1ā- R_2ă-R_3a	KaTaBa	كَتَبَ	a écrit		HaMaLa	حَمَلَ	a porté
maR_1R_2aR_3	maKTaB	مَكْتَب	bureau		maHMal	مَحْمَل	brancard
R_1u R_2i R_3a	KuTiBa	كُتِبَ	A été écrit		HuMiLa	حُمِلَ	a été porté
...							

Tableau 3.4 : Exemple de schèmes pour les mots كتب écrire et حمل porter

Les lettres en majuscule (Ri) désignent les consonnes de base qui composent la racine.

Les voyelles (â, a, i,..) désignent les voyelles et les consonnes en minuscule (m,..) sont des consonnes de dérivation utilisées dans les schèmes. La majorité des verbes arabes ont une racine composée de 3 consonnes. L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux. Une autre caractéristique est le caractère flexionnel des mots : les terminaisons permettent de distinguer le mode des verbes et la fonction des noms [10].

3.5.3 Structure d'un mot :

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Post fixe	Suffixe	Corps schématique	Préfixe	Antéfixe

- Antéfixes sont des prépositions ou des conjonctions.

- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,...)

- Postfixes sont des pronoms personnels.

Exemple : أتتذكرونا

Ce mot exprime la phrase en français : "*Est ce que vous vous souvenez de nous?*" La segmentation de ce mot donne les constituants suivants :

أ | ت | تذكّرُ | ونـــ | نا

Antéfixe : أ conjonction d'interrogation

Préfixe : ت préfixe verbal du temps de l'inaccompli.

Corps schématique: تذكر dérivé de la racine: ذكر selon le schème

taR1aR2aR3a

Suffixe : ون suffixe verbal exprimant le pluriel

Post fixe : نا pronom suffixe complément du nom

3.5.4 Catégories des mots :

L'arabe considère 3 catégories de mots

• Le verbe : entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.

• Le nom : l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.

• Les particules : entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte.

3.5.5 Le verbe :

La plupart des mots en arabe, dérivent d'un verbe de trois lettres. Chaque verbe est donc la racine d'une famille de mots. Comme en français, le mot en arabe se déduit de la racine en rajoutant des suffixes ou des préfixes.

La conjugaison des verbes dépend de plusieurs facteurs :

– Le temps (accompli, inaccompli).

– Le nombre du sujet (singulier, duel, pluriel).

– Le genre du sujet (masculin, féminin).

– La personne (première, deuxième et troisième)

– Le mode (actif, passif).

Par exemple : ب + ت + ك *K+T+B* donne le verbe كتب *KaTaBa*. (écrire).

Dans tous les mots qui dérivent de cette racine, on trouvera ces trois lettres K, T, B (Voir Tableau 3.4).

La conjugaison des verbes se fait en ajoutant des préfixes et des suffixes, un peu comme en français.

La langue arabe dispose de trois temps.

• L'accompli : correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a كتبن KaTaBna, *elles ont écrit* et pour le pluriel masculin on a كتبوا KaTaBuu, *ils ont écrit*).

• L'inaccompli présent: présente l'action en cours d'accomplissement, ses éléments sont préfixés (يكتب yaKTuBu *il écrit*; تكتب taKTuBu, *elle écrit*).

• L'inaccompli futur : correspond à une action qui se déroulera au futur et est marqué par l'antéposition de س *sa* ou سوف *sawfa* au verbe (سيكتب sayaKTuBu *il écrira*, سوف يكتب sawfa yaKTuBu *il va écrire*).

3.5.6 Les noms :

Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne le sont pas comme les noms propres et les noms communs. Dans le premier cas, le fait que le nom soit dérivé d'un verbe, il exprime donc une certaine sémantique qui pourrait avoir une influence dans la sélection des phrases saillantes d'un texte pour le résumé.

La déclinaison des noms se fait selon les règles suivantes:

• Le féminin singulier: On ajoute le ة, exemple صغير *petit* devient صغيرة *petite*

• Le féminin pluriel : De la même manière, on rajoute pour le pluriel les deux lettres ات, exemple صغير *petit* devient صغيرات *petites*

• Le masculin pluriel: Pour le pluriel masculin on rajoute les deux lettres ين ou ون dépendamment de la position du mot dans la phrase (sujet ou complément d'objet), exemple : الراجع *revenant* devient الراجعين ou الراجعون *revenants*

• Le Pluriel irrégulier: Il suit une diversité de règles complexes et dépend du nom. Exemple : طفل *un enfant* devient أطفال *des enfants*

Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative, mais aussi parce que son analyse dépend fortement de la structure comme pour les verbes irréguliers.

Certain dérivés nominaux associent une fonction au nom :

- Agent (celui qui fait l'action),
- Objet (celui qui a subit l'action),
- Instrument (désignant l'instrument de l'action),
- Lieu.

Pour les pronoms personnels, le sujet est inclus dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. On distinguera entre singulier, duel (deux) et pluriel (plus de deux) ainsi qu'entre le masculin et féminin.

3.5.7 Les particules :

Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence, ...). Elles jouent un rôle important dans l'interprétation de la phrase [16]. Elles servent à situer des faits ou des objets par rapport au temps ou au lieu, elles jouent également un rôle clé dans la cohérence et l'enchaînement d'un texte.

Comme exemple de particules qui désignent un temps منذ , قبل , بعد *pendant, avant, après*, un lieu حيث *où*, ou de référence الذين *ceux*,....

Ces particules seront très utiles pour notre traitement à deux niveaux :

- Elles font partie de l'anti dictionnaire qui regroupe les termes à ne pas prendre en considération lors de calcul de fréquence de distribution des mots,
- Elles identifient des propositions composant une phrase.

Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification.

3.6 Conclusion :

Le but de ce chapitre était de présenter un tour d'horizon sur le domaine de Recherche d'information (RI). Il a décrit tout d'abord Les principaux acteurs de RI, le processus général de RI, les modèles de RI. Enfin Il introduit la langue arabe avec sa particularité, sa morphologie, la structure et les catégories des mots.

CHAPITRE 4

IMPLEMENTATION

4.1 Introduction :

Pour générer un nuage de tags (ou nuage de mots clefs) à partir d'un flux de données textuelles. On a présenté un outil qui utilise un moteur de recherche afin d'afficher les mots ou expressions de 2 ou 3 termes les plus fréquents dans les textes. Tous les termes ou expressions ne sont pas à conserver dans le nuage de tag. Une des étapes consiste en un filtrage selon des règles définies dans les fichiers de règles : suppression des mots vides (فيه, منه, هو, ...), suppression des expressions commençant ou se terminant par un mot vide ("ولكن", "وهذا", ...), suppression des nombres, ...

4.2 Modélisation du tag :

Pour la modélisation de notre système, nous avons utilisé UML. UML est un langage conçue pour représenter, spécifier, construire et documenter les systèmes logiciels. Ses deux principaux objectifs sont la modélisation de systèmes utilisant les techniques orientées objet, depuis la conception jusqu'à la maintenance, et la création d'un langage abstrait compréhensible par l'homme et interprétable par les machines. UML s'adresse à toutes les personnes chargées de la production, du déploiement et du suivi de logiciels (analystes, développeurs, chefs de projets, architectes...), mais peut également servir à la communication avec les clients et les utilisateurs du logiciel. Il s'adapte à tous les domaines d'application et à tous les supports. Il permet de construire plusieurs modèles d'un système, chacun mettant en valeur des aspects différents : fonctionnels, statiques, dynamiques et organisationnels. UML est devenu un langage incontournable dans les projets de développement.

Diagramme de classes :

Le diagramme de classes est considéré comme le plus important de la modélisation orientée objet, il est le seul obligatoire lors d'une telle modélisation.

Il s'agit d'une vue statique car on ne tient pas compte du facteur temporel dans le comportement du système. Le diagramme de classes modélise les concepts du domaine d'application ainsi que les concepts internes créés de toutes pièces dans le cadre de l'implémentation d'une application. Chaque langage de Programmation Orienté Objets donne un moyen spécifique d'implémenter le paradigme objet (pointeurs ou pas, héritage multiple ou pas, etc.), mais le diagramme de classes permet de

modéliser les classes du système et leurs relations indépendamment d'un langage de programmation particulier.

Les principaux éléments de cette vue statique sont les classes et leurs relations : association, généralisation et plusieurs types de dépendances, telles que la réalisation et l'utilisation.

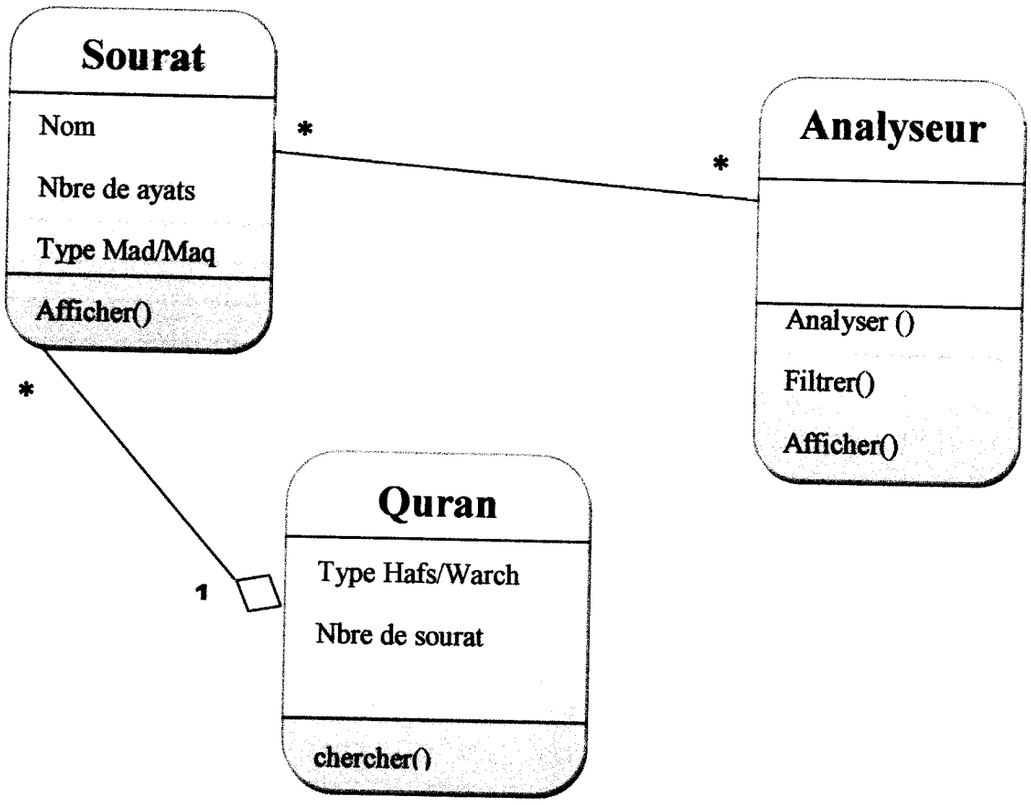


Figure 4.1 : Diagramme de classe de notre système

4.3 Choix du langage de programmation:

Pour coder notre application nous avons opté pour le langage java. Apparu en 1991, le langage JAVA a commencé à être intéressant à partir de 1995 avec sa prise en charge par le navigateur phare de l'époque, Netscape. Ce langage ne cesse de se développer. Il s'agit d'un langage orienté objet dont la syntaxe est très proche de celle du C++.

C'est également un langage portable, c'est à dire qu'il s'adapte à une foule de plateformes différentes. C'est là l'une des qualités de JAVA.

Grâce à une machine virtuelle, JAVA peut s'exécuter sur une quantité incroyable de plateformes. Le rôle de la machine JAVA est simple : lorsque nous installons la machine virtuelle JAVA, elle permet d'exécuter sur l'architecture de l'ordinateur courant des instructions. Une machine virtuelle JAVA pour MAC et pour PC est totalement différente mais fera la même chose au final (du point de vue utilisateur).Le programme JAVA restera donc identique, d'où la très bonne portabilité de ce langage.

4.4 Les API utilisées :

Pour la partie indexation des documents, nous avons utilisé « Lucene » qui est une API libre.

4.4.1 Lucene :

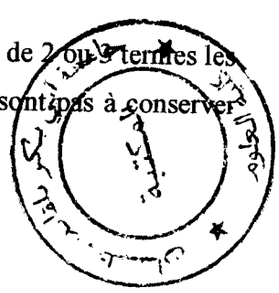
Lucene est une librairie open source en Java permettant d'ajouter des fonctionnalités de recherche plein-texte aux applications. Le projet Lucene est dirigé par « The Apache Software Foundation ». D'autres projets très connus et de grande qualité de la fondation sont : Apache HTTP server, Tomcat, Cocoon, Ant, ...

Il s'agit bien d'une librairie avec laquelle il n'est pas fourni d'outils permettant l'indexation de données en quelques clics de souris et quelques paramétrages. Il faut donc en passer par du code Java afin de mettre en place une solution sur mesure de recherche plein-texte.

4.4.2 Principe :

Lucene indexe et retrouve des « documents ». Par document, on ne parle pas de fichiers Excel, Word, PDF ou HTML, mais d'une structure de données constituée de champs. Un champ est une donnée possédant un nom (titre, auteur, date de publication, contenu, ..) et à laquelle est associé du texte. C'est ce texte qui est indexé, recherchable et affichable. Les documents indexés sont regroupés au sein d'une collection de documents appelée « index ». Un index peut contenir plusieurs centaines, milliers ou millions de documents et il est possible de créer autant d'index différents que le nécessite l'application. Physiquement, un index est un répertoire (a spécifier par le développeur) hébergeant un nombre variable de fichiers.

Le but de notre projet est d'afficher les mots ou expressions de 2 ou 3 termes les plus fréquents dans les textes. Tous les termes ou expressions ne sont pas à conserver



dans le nuage de tag. Une des étapes consiste en un filtrage selon des règles définies dans fichiers de règles : suppression des mots vides (إلى, حيث, كأن, ...), suppression des expressions commençant ou se terminant par un mot vide ("وأبو", "و كان", ...), suppression des nombres, ...

Les étapes sont donc comme suivant:

- Etape 1 : indexation du texte dans Lucene
- Etape 2 : extraction des termes ou expressions de l'index
- Etape 3 : filtrage des termes ou expressions
- Etape 4 : affichage du nuage.

Les 3 premières étapes sont réalisées en java avec Lucene et donne comme résultat un fichier de termes ou expressions retenues dans le nuage de tags avec leur fréquence d'apparition.

La 4ème étape est réalisée en java aussi, elle exploite le fichier résultat des étapes précédentes afin de mettre en forme le nuage de tags.

4.4.3 Utilisation de Lucene (étapes 1 et 2) :

La génération du nuage de tags nécessite 2 étapes que Lucene est à même de réaliser : analyser et découper le texte servant à générer le nuage de tags et indiquer la fréquence d'apparition de chaque mot ou expression.

Lucene utilise des analyzers afin d'indexer le texte. Ces analyzers ont pour rôle principal de découper le texte en « token ». Selon les analyzers utilisés les tokens sont séparés par des espaces, des ponctuations, des caractères spéciaux (#, @, ...), des chiffres, ...

Par exemple, pour la phrase suivante : « ٱللَّهُ لَآ إِلَٰهَ إِلَّا هُوَ ٱلۡحَيُّ ٱلۡقَيُّومُ», les tokens seraient sans doute : «ٱلۡقَيُّومُ», «ٱلۡحَيُّ », «هُوَ», «إِلَّا», «إِلَٰهَ», «لَآ », «ٱللَّهُ».

Dans notre nuage de tags, nous ne voulons pas uniquement des mots simples, mais également des expressions de 2 ou 3 mots. Il nous faut donc extraire toutes les combinaisons possibles de tokens de 1, 2 ou 3 mots, c'est-à-dire : «ٱلۡقَيُّومُ», «ٱلۡحَيُّ »,

«هُوَ ٱلْحَىُّ», «إِلَٰهَ إِلَّا هُوَ», «إِلَّا هُوَ», «إِلَٰهَ إِلَّا», «لَآ إِلَٰهَ», «ٱللَّهُ لَآ», «ٱللَّهُ», «لَا», «إِلَٰهَ», «إِلَّا», «هُوَ» et «هُوَ ٱلْحَىُّ ٱلْقَيُّومُ»

4.5. Le filtrage des termes ou expressions (étape 3) :

Dans l'exemple indiqué ci-dessus, les termes ou expressions obtenues sont : «ٱلْقَيُّومُ», «ٱلْحَىُّ », «هُوَ», «إِلَّا», «إِلَٰهَ», «لَا», «ٱللَّهُ», «ٱللَّهُ لَآ», «لَآ إِلَٰهَ», «إِلَٰهَ إِلَّا», «إِلَّا هُوَ», « إِلَٰهَ » «إِلَّا هُوَ», «هُوَ ٱلْحَىُّ» et «هُوَ ٱلْحَىُّ ٱلْقَيُّومُ», «هُوَ ٱلْحَىّ» et «ٱلْحَىُّ ٱلْقَيُّومُ». On remarque immédiatement des éléments que l'on ne souhaite pas voir apparaître dans le nuage : «لَا », «إِلَّا», «هُوَ», «إِلَّا هُوَ», «هُوَ ٱلْحَىُّ», «لَآ إِلَٰهَ», «ٱللَّهُ لَآ» et «إِلَٰهَ إِلَّا». Il s'agit des mots-vides ou expression commençant ou se terminant par un mot-vide. Il reste donc les éléments suivants : «ٱلْقَيُّومُ», «ٱلْحَىُّ », «إِلَٰهَ», «ٱللَّهُ», «ٱلْحَىُّ ٱلْقَيُّومُ».

Nous pousserons le nettoyage jusqu'à supprimer les mots simples ou expressions apparaissant dans une expression constituée de plus de mots. En effet, imaginons des données à analyser contenant une ou plusieurs fois les expressions suivantes : « ٱللَّهَ غَفُورٌ رَّحِيمٌ», «ٱللَّهُ سَمِيعٌ عَلِيمٌ» et «ٱللَّهُ غَفُورٌ حَلِيم», mais jamais ou peu de fois les mots «غَفُور», «سَمِيعٌ », «عَلِيم» et «حَلِيم» hors de ces expressions. Sans ce nettoyage, le nuage de tags contiendrait certainement «غَفُور», «سَمِيعٌ », «عَلِيم» et «حَلِيم» (très fréquents) et sans doute mais sans certitude «ٱللَّهَ غَفُورٌ رَّحِيمٌ», «ٱللَّهُ سَمِيعٌ عَلِيمٌ» et «ٱللَّهُ غَفُورٌ حَلِيم» (moins fréquents). Le nuage de tags serait alors complètement faux.

Ainsi nettoyé, le nuage devient «ٱللَّهَ غَفُورٌ رَّحِيمٌ», «ٱللَّهُ سَمِيعٌ عَلِيمٌ» et « ٱللَّهُ غَفُورٌ حَلِيم» et est beaucoup plus pertinent que ce que nous avions obtenu avant le filtrage.

Le fichier de règles est principalement constituer de la liste des mots vides mais commence par 5 paramètres pouvant être désactivés (en mettant les lignes en commentaires avec un #). Il s'agit de :

- smallwords : pour retirer les mots de 3 caractères ou moins
- numbers : pour retirer les nombres
- dashes : pour retirer les mots contenant un tiret (« - »)
- period : pour retirer les mots contenant un point
- include : pour retirer les mots ou expressions incluent dans une autre expression. On a mis cette option pour le principe, mais en la désactivant les résultats sont généralement très décevants.

4.6 Code source des étapes 1, 2 et 3 :

Le code source java est disponible sous forme d'un projet. Il contient les fichiers suivants :

- NGramAnalyzerWrapper.java et NGramFilter.java (analyzer Lucene)
- TagCloud.java (classe principale avec l'algorithme de filtrage)
- input.txt (les données exemples à analyser)
- rules.txt (le fichier de règles)

Pour lancer l'analyse d'un fichier de données, l'usage est le suivant :

Usage: org.apache.demo.TagCloud -input [inputfile] -output [outputfile] -rules [rulesfile] -count [count] -boost [boost] -minfreq [minfreq] -maxterm [maxterm]

input - input file with data to be clouded

output - output file with cloud data

rules - rules file

count - max number of items within the cloud (**default** = 1000)

boost - boost value **for** multi-terms tags (**default** = 1)

minfreq - minimum frequence **for** a ngram (**default** = 2)

maxterms - maximum number of terms in expressions in the cloud (**default** – 3)

Example: java org.apache.demo.TagCloud -input /tmp/input.txt –output /tmp/output.txt -rules /tmp/rules.txt -count 100 -boost 1 -minfreq 3 -maxterm 4

4.7. Etape 4 : affichage du nuage en java :

La mise en forme du nuage de tags utilise le fichier généré par le programme précédent. Une classe java réalise donc cette mise en forme, elle permet la lecture du fichier et la génération des styles d'affichage.

Code source des étapes 4 :

Le code source java est disponible sous forme d'un projet. Il contient les fichiers suivants :

- Tag.java

Aperçu de l'exécution :

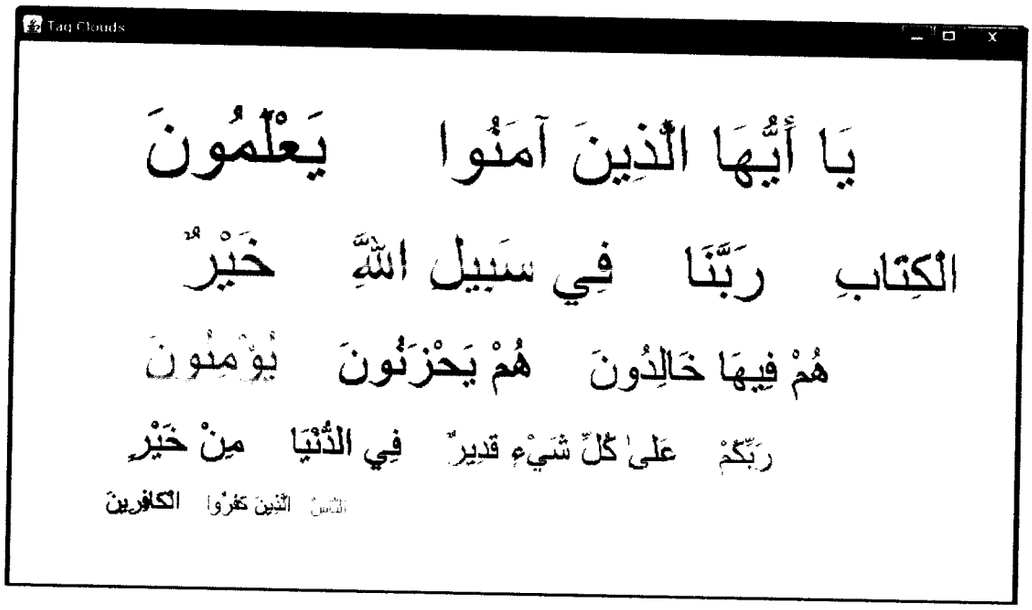


Figure 4.2 : Copie écran, résultat d'exécution de notre outil de visualisation.

4.8 Conclusion :

Dans ce chapitre, on a présenté la réalisation de notre outil qui permet de générer des nuages de tags (ou nuages de mots clefs) à partir d'un flux de données textuelles. On a commencé par la modélisation du tag ensuite on a présenté le langage de programmation, et le moteur de recherche (Lucene) utilisé pour la réalisation de notre outil. A la fin on a décrit les étapes de la réalisation de notre projet.

Conclusion Générale

Les nuages de mots sont un moyen de classification graphique différent des tableaux, secteurs, listes, ..., c'est un outil nouveau en pleine évolution avec son parrain le web. Ce dispositif graphique permet de s'orienter, permet aussi de hiérarchiser et en tous cas d'attirer l'attention sur des indices en produisant des différences.

Dans ce mémoire nous avons donné un aperçu sur l'étiquetage collaboratif, des définitions sur ces composants tels que les tags, les folksonomies, les mots clé et les nuages de mots. Nous avons cité quelque avantage et inconvénients de l'étiquetage collaboratif.

On a fait un tour d'horizon sur quelques outils existant qui permettent de générer les nuages de mots.

Nous avons aussi présenté le domaine de recherche d'information (RI) et la langue arabe avec sa particularité, sa morphologie, la structure d'un mot et les catégories des mots.

Dans notre projet nous avons développé un programme permettant de créer des nuages de mots a partir d'un document textuel arabe, avant d'arriver au résultat final il a fallu passer par quatre étapes. Les trois premières étapes sont réalisées en java avec Lucene qui est une librairie open source en Java. Ces trois étapes donnent comme résultat un fichier de termes ou expressions retenues dans le nuage de tags avec leur fréquence d'apparition. La quatrième étape réalisée en java aussi, exploite le fichier résultant des étapes précédentes afin de générer le nuage de tags.

Ce projet peut être intégrer dans d'autre applications tel que les applications qui traitent la langue arabe et surtout le coran. D'ailleurs, nous proposons comme perspective d'étudier la façon la plus adéquate de l'intégration de notre outil dans les applications relevant du domaine de la recherche d'information.

BIBLIOGRAPHIE :

[1] B. PIWOWARSKY. Techniques d'apprentissage pour le traitement d'informations structurées : *Application à la recherche d'information*. Thèse de Doctorat, Université de Paris VI, France, 2003.

[2] Cf. VANDER WAL, T. (2007), Folksonomy Coinage and Definition, *vanderwal.net* : « *Folksonomy is the result of personal free tagging of information and objects (...) for one's own retrieval. The tagging is done in a social environment (*usually [Ndr : je souligne] *shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information* ». On notera que l'aspect collaboratif n'y est pas explicitement posé comme une condition nécessaire à la constitution d'une folksonomie.

[3] GOLDER S. A. et HUBERMAN B. A. (2006), Usage patterns of collaborative tagging systems, *Journal of Information Science*, 32, 2, 198-208.

[4] Gruber T. R., A translation approach to portable ontology specification, *Knowledge Acquisition*, 5, 2, 199–220, 1993.

[5] GRUBER, T., TagOntology - a way to agree on the semantics of tagging data, http://tomgruber.org/GRUBER, T. (2007), Ontology of Folksonomy: A Mash-up of Apples and Oranges, *Int'l Journal on Semantic Web & Information Systems*, 3(2). 2005.

[6] GUY M. et TONKIN E., Folksonomies: Tidying up tags?, D-Lib Magazine, 12, 1 (http://www.dlib.org/dlib/january06/guy/01guy.html) (2006).

[7] K.MAMMERI. Recherche d'information par croisement de média texte et image. Mémoire de Magister, Université M'hamed BOUGARA de BOUMERDES, 2009.

[8] MATHES A., Folksonomies – cooperative classification and communication through shared metadata, computer mediated communication – LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html), 2004.

[9] MONNIN A. (2009), From Game Neverending to Flickr. Tagging systems as ludic systems and their consequences. In: *Proceedings of the WebSci'09: Society On-Line*, 18-20, Athens,Greece. (In Press), March 2009

[10] S. BALOUL, M. ALISSALI, M. BAUDRY, P. BOULA DE MAREÜIL: Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe, 24es Journées d'Étude sur la Parole, 24-27 juin2002 Nancy

[11] S. BOULAKNADEL. Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation. Thèse de Doctorat, Université de Nantes,2008.

[12] TRANT J. Exploring the potential for social tagging and folksonomy in art museums: Proof of concept, *New Review of Hypermedia and Multimedia,* 12, 1, 83-105., 2006.

[13] Violeta ROXIN. et Yohan BERNARD, Etiquetage collaboratif et usages de mots : quels apports pour les sites marchands ? ,COMMUNICATION, 2007.

[14] W. N. Borst, *Construction of engineering ontologies*, Centre for Telematica and Information Technology, University of Tweenty, Enschede, The Netherlands 1997

[15] Xu J., Fraser A., Weischedel Ralph, Empirical Studies in Strategies for Arabic Retrieval, *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), August 11-15, 2002*

[16] Y. Kadri, A.Benyamina, Système d'analyse syntaxicosemantique du langage arabe, *mémoire d'ingénieur, université d'Oran Es-sénia, 1992*