Chapitre 1SOMMAIRE

Chapitre 2

Chapitre 3

Chapitre 4

REMERCIEMENTS

Je tiens à remercier plusieurs personnes qui m'ont beaucoup aidé pour la réalisation de ce mémoire.

Tout d'abord, je remercie Mr. Jean-Claude BIGNON, mon directeur de mémoire pour son encadrement et ses conseils durant la période du stage.

Je remercie Mr. Pascal HUMBERT, mon co-directeur de mémoire et Mr. Gilles HALIN pour leurs encadrements et conseils.

Mes remerciements à Mr. Jean-Marie PIERREL Directeur du laboratoire ATILF; Mr. Azim ROUSSANALY et Mr. Matthieu QUIGNARD du laboratoire LORIA-LED pour leurs conseils et informations.

Je remercie également l'ensemble des étudiants en thèse, plus particulièrement Sabrina pour son aide précieuse, mais aussi l'ensemble des étudiants qui ont participé à l'expérimentation.

De même, je remercie les membres du laboratoire du CRAI, pour m'avoir permis de réaliser ce stage dans les meilleures conditions.

INTRODUCTION

Ce travail s'inscrit dans la continuité des travaux déjà réalisés dans les différents domaines de la recherche de l'image par le contenu [BIG, HAL, & NAK, al 01] appliqués à la conception du projet architectural.
Dans le cas de la conception architecturale l'image joue un rôle majeur. Elle sert à simuler des faits, des idées, et permet la représentation des caractéristiques de l'objet architectural. Des schémas conceptuels sont matérialisés dans des images intermédiaires (dessins esquisses etc.) permettant ainsi l'élaboration et l'exploitation visuelles des hypothèses [AOU 03]. L'image joue un rôle important dans le processus de conception, car elle peut intervenir à m'importe qu'elle phase du projet, par exemple au cours d'un parcours d'un catalogue de bâtiment, à la phase d'une esquisse [NAK 03].
Le laboratoire CRAI(1) a effectué plusieurs travaux d'étude portant sur la recherche d'informations techniques par l'image tel est le cas de BATIMAGE (2). Ces recherches avaient pour objectifs l'assistance à la recherche d'informations relatives aux produits du bâtiment, afin de mettre en place des méthodes et des outils de recherche d'informations qui seront proposées aux architectes.

Dans cette même optique entre notre thème de stage qui s'effectue au sien du CRAI. Il consiste en l'étude des techniques de recherche d'image par le contenu sémantique appliqué au domaine de la construction bois. Cette étude s'effectue sur un corpus de vingt projets de construction bois ; pris sur le site du CNDB(3). Ce corpus sera utilisé comme échantillon, en vue de comparer certaines méthodes d'indexation, qui seront effectuées par deux robots d'indexation qui sont au stade de prototype à savoir :

- EngineSK(4) : indexation effectuée sur l'analyse statistique
- WimexBot (5) « Web Image Extractor Robot »: utilisant une méthode d'indexation sur la base d'un thesaurus spécialisé à la construction bois

Ces derniers seront comparés à l'indexation manuelle spécialisée. Indexation effectuée à l'aide d'un thésaurus relatif à la construction bois ; auquel les descripteurs sont attribués d'un poids relatif à leur pertinence à décrire le contenu d'une image.

(1) Centre de Recherche en Architecture et Ingénierie
(2) http://www.crai.archi.fr:batimage
(3) Comité National pour le Développement du Bois, sa mission est la promotion du bois, le CNDB concentre son action sur le marché du BTP, avec un objectif : faire construire et aménager avec du bois
http: //www.cndb.org
(4) Amélioré par Mr. Pascal HUMBERT
(5) Qui fut réalisé par Mr. Marc WAGNER ingénieur au CNAM(Conservatoire National des Arts et Métiers), dans le cadre de son stage d'étude effectué au CRAI. Actuellement ce dernier est en cour de perfectionnement par Mr. Gilles HALIN

Problématique

La problématique de notre travail s'inscrit dans le cadre de l'assistance à la conception architecturale par l'image.
A l'heure actuelle l'informatique et ses applications occupent une place importante dans les agences d'architecture. Ce qui explique le développement régulier des outils de modélisation qui ont pour objet d'assister l'architecte dans ses différentes tâches.
L'utilisation courante de l'Internet comme source d'information dans le domaine de l'architecture est un moyen d'aide à la conception. Mais ce dernier est confronter à différents problèmes exemple, dans le cadre de la gestion des flux d'information que propose le Web.
Pour palier à ces problèmes, et afin de faciliter l'accès et la gestion de l'information générée par l'Internet. Le CRAI a mis en place des outils d'aide à la recherche d'information et qui se place dans la gamme des outils d'aide à la conception.

Objectifs

- Comparaison des méthodes d'indexation établie par les robots au stade de prototype et celle effectuée par un expert.

Cette comparaison à pour but de voir entre WimexBot (indexation utilisant un thésaurus) et EngineSK (indexation utilisant l'analyse statistique), lequel des deux donnes des résultats pertinents se rapprochant de l'indexation manuelle spécialisée (indexation manuelle effectuée sur la base d'un thésaurus dont les termes sont pondérés).

- Description et expérimentation sur ces robots dont le but est de valider la méthode d'extraction d'images pertinentes, mais aussi de mesurer la performance de chacun.
- Amélioration de ces outils et méthodes d'acquisition existante en fonction des besoins des Architectes.

L'objectif de cette étape est d'apporter des propositions pouvant permettre l'amélioration des outils. Pour ce faire une étude est nécessaire sur les différentes méthodes d'indexation afin de relever leurs limites et de les atténuer.

CHAPITRE 1 : Etat de l'art

1.1- Processus de recherche d'information :

Les systèmes de recherche documentaire ont pour objectif d'assister l'utilisateur lors de ses investigations dans une base d'information, lorsqu'il est à la recherche de documents. Un SRI (6) comprend trois modélisations distinctes [HAL 89] :

- Une modélisation du contenu sémantique des documents (Modèle de document), le processus qui permet d'exprimer le contenu d'un document est l'indexation.
- Interprétation du besoin de l'utilisateur (Un modèle de requêtes), le modèle spécifie la forme que les demandes d'informations peuvent prendre.
- La définition d'une fonction de correspondance entre la requête et la représentation du contenu du document (modèle de correspondance).

1.2- Analyse du contenu sémantique des documents : l'indexation

L'indexation d'un document est l'étape fondamentale qui donne au document un statut conceptuel dans la base gérée par le système de recherche d'information [HAL 89]. Elle a pour objectif d'identifier l'information contenue dans tout texte et de la représenter au moyen d'un ensemble d'entité appelé mots clés(7), pour faciliter la comparaison entre la représentation d'un document et sa requête [DUP & ERM 00].
Le processus d'indexation peut-être définit comme étant le transfert de l'information contenue dans le texte vers un autre espace de représentation traitable par un système informatique. Les critères d'une bonne indexation dans le cadre des résultats SRI dépendent de la qualité de leur processus d'indexation. Elle est jugée sur deux critères à savoir la cohérence et l'adéquation entre les représentations des requêtes et des documents.

1.2.1- Cohérence :

Une bonne indexation doit être cohérente, c'est-à-dire que deux textes traitant du même sujet, sans utiliser le même vocabulaire, sont indexés avec les mêmes descripteurs. Il est à noter que la probabilité que deux personnes différentes puissent choisir le même terme (mots clés) pour décrire un objet est faible. D'où la difficulté d'obtenir une cohérence dans l'indexation humaine; par contre, l'indexation automatique est forcément cohérente car elle utilise toujours le même processus d'indexation (si deux documents traitant d'un même sujet sont indexés avec les mêmes descripteurs, cas réservé à notre étude).

(6) Système de recherche d'information.
(7) Mots-clé : terme en langage naturel, choisi généralement dans le titre ou le texte d'un document pour en caractériser le contenu et en permettre la recherche. Il est à distinguer d'un descripteur, qui est un terme normalisé dans un thésaurus [DEG & MEN 01].

1.2.2- L'adéquation entre les représentations :

L'indexation doit vérifier le critère d'adéquation entre les représentations de la requête et du corpus. Si nous devons retrouver un document, il faut que ses descripteurs appartiennent au même vocabulaire que ceux qui sont utilisés pour décrire la requête [DUP & ERM 00]. Il semble plus facile de vérifier les critères de cohérence et d'adéquation en indexant automatiquement, car le vocabulaire utilisé par l'auteur du document à de grandes chances d'être différent de celui de l'utilisateur du SRI.

1.3- Les différents types d'indexations :

L'opposition entre indexation manuelle et automatique, a montré que ces deux modes de travail ont chacun leurs points forts et leurs points faibles.

- L'indexation automatique(8) est plus régulière, plus exhaustive, permet les mises à jours de thesaurus, à cela elle permet d'éliminer les mots vides de sens d'où l'utilisation de la fréquence d'occurrence de mots qui joue un rôle important dans le critère des sélections de mots. Mais elle est confrontée aux erreurs dues aux ambiguïtés de polysémie dans les textes.
- L'indexation humaine est moins régulière, mais plus synthétique, et moins sujette aux erreurs d'ambiguïté [DUP &ERM 00]. Mais elle présente l'inconvénient d'être fastidieuse et pose le problème de l'homogénéité des mots clé lors d'une indexation effectuée par plusieurs usagers.
- L'indexation semi-automatique est un compromis entre les deux méthodes précédentes. Elle utilise en premier une indexation automatique qui donne les premiers éléments d'indexation. Puis l'indexation manuelle complète et corrige les informations obtenues par l'indexation automatique [NAK 03].

1.3.1- Technique d'indexation manuelle :

L'indexation manuelle se caractérise de manière générale par trois étapes [HAL 89] :

- La prise de connaissance du contenu du document,
- Le choix des concepts,
- La traduction des concepts en descripteurs

L'indexeur peut choisir ces mots clés dans une liste de vocabulaire contrôlé «Thésaurus», ce qui permettra d'assurer l'uniformité de la représentation du document.

(8) Indexation effectuée exclusivement par des systèmes informatiques. Le système d'indexation peut être opéré par sélection de mots ou termes extraits du texte [DEG & MEN 01].

- Techniques d'indexation manuelle :

  - L'indexation dite à plat :

Elle consiste à l'analyse d'un document en considérant que les descripteurs entretiennent le même statut entre eux. Dans cette technique, un document est représenté par une liste non ordonnée de descripteurs [DUP & ERM 00]. Elle est généralement obtenue en traitant le contenu du document par des techniques d'identification des entités à base de traitement linguistique. Dans le cas d'une indexation en langage contrôlé, un système à base de connaissance, propre à un domaine, peut être utilisé pour trouver les descripteurs à partir des termes du document.
L'avantage considérable de ce système est la cohérence de l'indexation, c'est-à-dire que des textes de même sujets utilisant le même vocabulaire, auront les même mots clé ; mais les systèmes à base de connaissances ne sont capables d'identifier les descripteurs ou les concepts que si ceux-ci sont préalablement enregistrés dans une base de connaissances.

  - L'indexation pondérée :

Dans cette représentation, un poids est affecté au descripteur pour représenter l'importance du descripteur vis-à-vis du contenu du document [DEG & MEN 01]. Ce poids peut être calculé de différentes manières. Ce genre de technique affecte un poids aux termes proportionnels à leur fréquence d'apparition dans les textes. Plus le poids est important, plus le terme est jugé apte à appartenir à l'index du document. Un descripteur avec une forte pondération est un terme fréquent dans un document et absent dans d'autres documents, par conséquent les termes rares qui risquent d'être peu utilisés pour la recherche sont privilégiés. Cette pondération amplifie considérablement l'importance des termes étrangers, des noms propres et des mots mal orthographiés.

  - Thésaurus ; vocabulaire de langage d'indexation :

Dans un langage contrôlé, un terme d'indexation peut être soit un **terme préférentiel**, soit un terme **non préférentiel**.
Le terme préférentiel est un terme régulièrement utilisé lors de l'indexation pour représenter un concept donné, **descripteur** (technique utilisée par WimexBot)**.** Le terme non préférentiel est le synonyme ou quasi-synonyme d'un terme préférentiel appelé aussi **non-descripteur**. Il n'est pas utilisé à l'indexation, mais il sert de renvoi ou d'entrée permettant à l'utilisateur de s'orienter, vers le terme préférentiel approprié [ZER 01]. Exemple le mot *poutre* est considéré comme étant un descripteur et le terme *poutraison* est pris comme un non-descripteur, car il est le synonyme de ce dernier.

1.3.2- Technique d'indexation automatique

- Technique du texte intégral :

Cette technique à pour objectif de retrouver tous les mots présents dans les textes du corpus à l'exception des mots dits vides (articles, prépositions, etc.), les différentes étapes pour l'analyse des textes sont les suivantes :
La segmentation: découper le texte en phrases et identifier les mots. Les mots reconnus comme une suite de caractères entourés de deux séparateurs (espaces, guillemet, virgule, etc.) et les mots fonctionnels de la catégorie lexicale telle que les articles, pronoms, etc. ; sont stocker dans une liste considérée comme étant des mots vide inutile au cour de l'indexation [DUP & ERM 00]. Cette technique très simple choisit comme descripteurs tous les mots reconnus comme non vides (méthode utilisée par un filtre avant l'utilisation de EngineSK).

- Les méthodes linguistiques :

Elles reposent sur différents niveaux d'analyse linguistique du texte intégral pour extraire les unités de langage porteuses de sens [AMA 98]. Au final, un système d'acquisition des connaissances identifie les concepts du domaine caché sous ces unités de langage. Actuellement des laboratoires de recherches effectuent des études sur l'indexation documentaire dans le cas de la langue française, tel est le cas des laboratoires : ATILF(9) et LORIA-LED(10). Les techniques linguistiques s'appuient sur des automates de traitement du texte, qui utilisent des technologies diverses d'algorithmes, mais aussi, et surtout d'intelligence artificielle fondée sur des réseaux neuronaux. Elle comprend :

- L'analyse morphologique

C'est la première étape de l'analyse d'un texte elles sont appelées aussi lemmatisation, il s'agit d'associer à la forme fléchie d'un mot sa forme canonique (singulier pluriel, infinitif masculin..). Actuellement des logiciels tel que TREE-TAGGER(11), utilisé au LORIA-LED permettent ce genre de transformation, et en ce moment Mr. Azim Roussalyni du laboratoire LORIA-LED tente de mettre en place un lemmatiseur. Ce qui semble être prometteur dans l'avancement de la recherche (dans le traitement automatique de la linguistique).

- L'analyse syntaxique :

Elle permet de lever les ambiguïtés au niveau grammatical, en décrivant les catégories grammaticales d'un mot selon sa structure dans la phrase (nom, verbe, article, ..).

- L'analyse sémantique :

Elle s'applique dans le cas où l'on voudrait traiter le sens d'un texte, on s'appui sur des systèmes sémantique de type Réseaux Sémantique, pour extraire des relations implicites entre les termes rendant compte du sens du texte.

- Les méthodes statistiques :

A la fin des années cinquante les études statistiques connaissent un grand succès, amorcées par H.P Luhn. Elles reposaient au début, sur de simple fréquence d'occurrence des mots dans les documents à indexer. Seuls les termes moyennement fréquents étaient retenus. Au Fil du temps d'autres propriétés statistiques ont considérées, la nécessité d'ajouter des traitements morphologiques et sémantiques au moyen de thésaurus [DUP & ERM 00]. Il faut noter que les méthodes statistiques sans traitement linguistique sont très peu performantes (dans le cas de notre étude (Cf : Ch 4).

(9) Analyse et Traitement Informatique de la Langue Française http://www.atilf.fr
(10) Laboratoire Lorrain de Recherche en Informatique et ses Applications -langue et Dialogue http://www.loria.fr/equipes/led/
(*) ATILF et LORIA-LED sont deux sous branche du CNRS (Centre National de Recherche Scientifique)
(11) Outil pour annoter le texte avec la partie du discours et l'information de lemme qui à été développé dans le projet de comité technique à l'institut pour l'informatique linguistique de l'université de Stuttgart http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Decision/TreeTagger.html

➢ La méthode de pondération par fréquence inverse :

Elle repose sur l'hypothèse qu'il existe une relation inversement proportionnelle entre l'importance d'un terme pour l'indexation d'un document et le nombre total de documents contenant ce terme dans la base documentaire.

➢ La méthode de pondération par valeur discriminatoire (SALTON) :

Elle consiste à définir la valeur d'un terme dans la base documentaire selon des cosinus de SALTON(12). Cette méthode calcule la similarité globale (entre deux documents) et la valeur discriminatoire d'un terme ou descripteur.

Cette technique est utilisée en complément des techniques linguistiques d'identification des unités du langage (mot à syntagme), pour filtrer les entités d'indexation. La difficulté réside ici dans le choix des descripteurs (mots clé). Toutes ces méthodes sont basées sur l'étude de la répartition des mots sur l'ensemble du corpus pour définir les critères de sélection des descripteurs.

Un des critères possibles peut être l'aptitude d'un mot à discriminer le corpus. En partant du principe qu'un mot rare, au contraire d'un mot fréquent, discrimine le corpus, c'est-à-dire qu'il le sépare en deux groupes : groupe de documents le contenant ou non.

[Sal 75] SALTON filtre les mots du corpus pour trouver les descripteurs qui discriminent le mieux les documents. Il calcule la similarité entre les documents du corpus à l'aide de cette fonction de comparaison mathématique (appelé fonction de similarité) pour définir un espace de représentation moins dense possible c'est-à-dire qui minimise au mieux la similarité entre les documents.

1.4- Analyse de l'image :

1.4.1- les différents systèmes d'indexation des images

Bien longtemps avant que l'image ne soit numérisée, l'accès à l'image était géré par les bibliothécaires ou les archivistes, en associant à l'image des termes descriptifs et une référence. Lorsque la gestion de la base d'images numérique apparues, la recherche d'image fondée sur la sémantique fut une des méthodes développées [NAK 03].

Il existe plusieurs systèmes qui ont pour but d'indexer des images comme par exemple WimexBot qui entre dans le cadre de notre étude. Chaque système adopte une approche différente, ils tendent tous vers un même objectif, faire en sorte que l'indexation soit la meilleure possible afin que la gestion soit des plus facile et efficace.

L'indexation d'image s'effectue actuellement sous deux formes l'indexation graphique et l'indexation sémantique (Fig: 1).

(12) ftp: //ftp.es.cornell.edu/pub/smart/

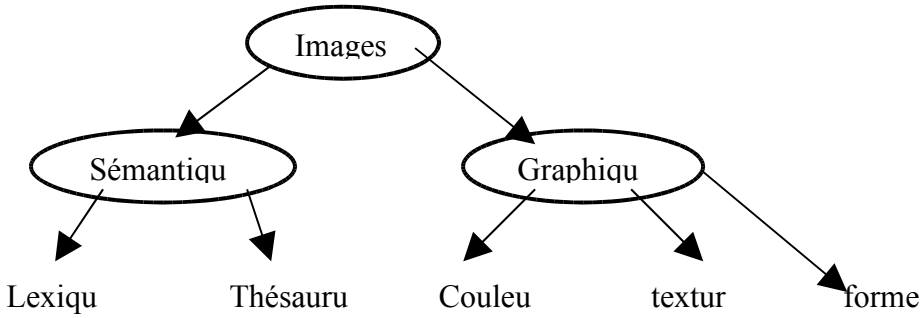


Figure 1 : Les différentes manière d'indexation d'images

- Indexation d'images par le graphisme :

Cette méthode d'indexation est adaptée aux grandes bases ou l'indexation manuelle ou semi-automatique serait fastidieuse. Elle consiste à extraire une ou plusieurs caractéristiques physiques des images de la base :

- Couleur: la couleur est une des caractéristiques physiques qui est couramment utilisée dans la recherche d'images.
- Texture: (Travail effectué par ZIDANE [ZIN 02]) la texture résulte de la présence de plusieurs couleurs ou de plusieurs intensités de pixels qui composent une surface dans l'image.
- Forme: la forme permet de détecter un objet sur une image.

Ces caractéristiques extraites sont présentées sous formes de données numériques. Ces dernières, que l'on appèle la signature, sont généralement représentées sous la forme de vecteurs.

- Indexation d'images fondée sur la sémantique :

Plusieurs recherches ont permis de caractériser divers aspects de l'image, liées au sens de l'image, qui peuvent être indexée et décrit par des mots. A l'heure actuelle plusieurs laboratoire de recherche se penchent sur le sujet de l'indexation d'images sur la sémantique, tel est le cas du CRAI, MRIM(13), qui tentent de mettrent en place des outils permettant l'indexation d'images de manière automatique, fondée sur l'utilisation des mots se rattachant à l'image. Dans le cadre du processus d'indexation, les termes d'indexation peuvent être fournis de deux manières :

(13) Modélisation et Recherche d'information Multimédia

- Soit manuellement à l'aide d'un thésaurus, afin de faciliter et d'harmoniser l'indexation.

- Les termes d'indexation peuvent être aussi extraits automatiquement lorsque l'image est accompagnée d'un texte qui le décrit (appelé **texte collatéral,** légende, titre, sous-titre, titrage, commentaire projet.) tel est le cas du robot WimexBot. Cependant, les termes d'indexation extraits des collatéraux ne sont jamais aussi pertinents sémantiquement que ceux fournis manuellement; c'est pourquoi, un processus d'indexation semi-automatique d'images est souvent préconisé.

➢ Méthodes d'indexation basée sur un thésaurus :

Le thésaurus est une liste organisée de termes contrôlés et normalisés (descripteurs et non-descripteurs), servant à l'indexation des documents et des questions dans un système documentaire[DEG & MEN 01].

1.5- Bilan

Les documents images fixes, et plus particulièrement les photographies, ont comme caractéristique principale, le manque d'un langage permettant d'en exprimer la sémantique, à l'inverse des textes et de la parole. Il en résulte une très grande difficulté pour proposer des systèmes de recherches d'images à base de concepts.
Indexer les images n'est pas une tâche simple, une image à l'inverse d'un texte ne se décrit pas naturellement à l'aide de mot. L'indexation des documents textuels a largement été pratiquée, alors que l'indexation des images reste problématique. Une première approche consiste à donner des attributs textuels aux images (mots clés) c'est une façon de renvoyer le problème à celui de l'indexation textuelle, seulement une image se prête à des interprétations très subjectives et approximatives [ZER 01].
Le but de l'indexation d'un document image est d'extraire et de représenter le contenu nécessaire et suffisant pour qu'il soit retrouvé par l'utilisateur. Cette indexation se base donc sur une représentation (supportée par un modèle) et sur un processus d'extraction. Dans le cas de documents images, on peut différencier deux types d'approches :

- Celles qui se basent sur le contenu extrait de l'information « brute » des documents (par exemple la matrice de pixels d'une image)
- Celles qui considèrent une interprétation sémantique du document comme son contenu (par exemple des mots clés)

Notre étude vise à développer des modèles de représentation symbolique du contenu sémantique des images aptes, d'une part, à supporter les processus de recherche d'information, et d'autre part, à automatiser l'indexation de ces images.
Un document possède un signal propre : le signal d'un texte est une chaîne de caractères, le signal d'une image est une matrice de pixels [NAK 03].
Dans ces deux cas, les modèles et les processus d'indexation diffèrent. Dans le deuxième cas il est courant de recourir à un processus d'indexation uniquement manuel. Notre démarche consiste donc à étudier les modèles nécessaires, dans certains contextes, basés sur une représentation de connaissance suffisante, et de déterminer comment intégrer des processus d'extraction automatique.

Nous avons, pour l'instant, considéré deux façons d'indexer: une indexation entièrement manuelle et une indexation complètement automatique.

Il existe des logiciels ou des sites spécialisés qui permettent une indexation automatique (presque gratuitement). Pourtant, elle montre vite ses limites. Réalisée sans contrôle, elle ne garantit ni la pertinence des résultats, ni même la validation de l'enregistrement du site.
Pourquoi indexer manuellement ? L'indexation manuelle permet une intervention personnalisée sur chacun des outils que l'on souhaite utiliser. Ceux-ci développent leurs propres modes de fonctionnement. Ce qui est apprécié par les uns, peut être contesté par les autres. L'indexation manuelle permet de choisir les mots clés adaptés à chaque moteur.

A l'heure actuelle aucun outil n'indexe de façon totalement autonome des textes. C'est la raison pour laquelle on parle davantage d'indexation assistée par l'ordinateur ou d'indexation semi-automatique (Compromis entre les deux modes précédents). Souvent le système d'indexation applique d'abord une indexation automatique qui donne les premiers éléments d'indexation. Puis l'indexation manuelle complète et corrige les informations obtenues par l'indexation automatique) [NAK 03].

CHAPITRE 2 : Élaboration du travail préalable

Notre stage s'oriente vers la recherche d'images par le contenu sémantique, adapté au domaine de la construction bois. Il fut mené en plusieurs étapes qui seront énumérées ci- après et qui ont permis de mieux diriger nos travaux.

2.1- Choix d'un corpus

Il fut réalisé par le choix d'un magasine relatif au domaine de la construction bois, à cet effet, nous avons retenu notre attention sur la revue « *séquence bois »* spécialisée dans ce domaine. Dans un premier temps, nous avions pris connaissance d'un grand nombre de ces revues en vue de voir la composition de chacun des projets présentés. Ces derniers sont composés par un ensemble d'élément important à savoir (Fig : 2):

- Un ensemble d'images (dessins techniques et photographies) relatives à la construction bois
- Commentaire associé à chaque image, qui décrit plus ou moins le contenu de l'image
- Un titre (faisant référence au thème du projet)
- Un sous-titre
- Un résumé du projet (décrivant brièvement le projet)
- Un commentaire général du projet (présente l'ensemble du projet en détail)
- Des renseignements relatifs à la réalisation du projet (nom de l'architecte, lieu de réalisation, non de l'entreprise etc.)

➢ La pertinence des images :

Étant donné que notre travail a pour but d'indexer des images, il nous est paru important que les images retenues soient numérisées pour cette étape qu'est « l'indexation ». À cet effet le choix des projets s'est effectué directement sur le site du CNDB, car ce dernier a l'ensemble de ces projets numérisés au format PDF (14), ce qui est un atout positif en matière de temps et de qualité des images. Ne sont prises en considération que les images photographiques, les images relevant du dessin technique sont rejetées, elles sont jugées non intéressante pour notre étude (cas de EngineSk qui les prend pas en compte).

(14) Portable Document Format, c'est un format de dossier de multi-plateforme développé par Adobe systems. Pour visualiser un fichier de PDF, vous pouvez utiliser le logiciel « Acrobat » d'Adobe, une application libre distribué par Adobe systems

➢ La pertinence sémantique :

Une des propriétés de l'image est qu'elle peut être associer à un texte. Ce phénomène n'est pas négligeable puisque le texte associé à l'image devient son contexte immédiat. Par conséquent, le texte influence son interprétation d'une manière considérable [NAK 03]. Dans le cadre de notre étude, la pertinence sémantique d'une image peut être distinguée, dû fait qu'une image est pertinente si elle à un texte qui lui est proche, si ce dernier décrit de manière explicite le contenue de l'image que l'on veut indexer, c'est l'une des exigences de WimexBot.

Les éléments cités ci-dessus furent les critères de sélection d'un projet. A ce titre, vingt projets ont été retenus en prenant en compte la diversité des thèmes.

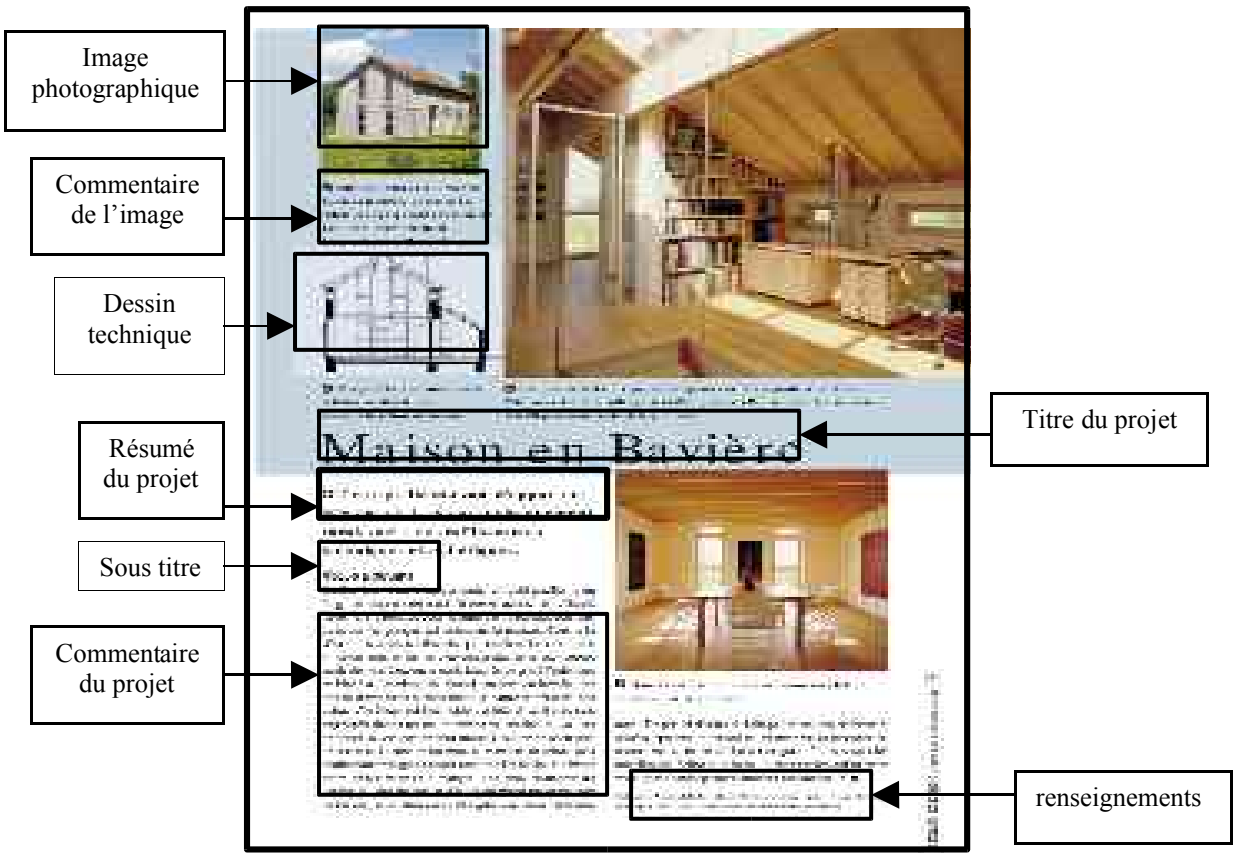


Figure 2 : Composition d'une page de projet issue du site CNDB

2.2- Choix des mots clés

Le choix des projets étant effectué, l'étape suivante consiste à sélectionner des mots clé dans les composantes textuelles (titre, sous titre, commentaire d'images et projet, résumé).

La première étape consistait à choisir des mots se rattachant à l'image, ces mots doivent à eux seuls décrire le contenu de l'image sans aucune ambiguïté. Les mots retenus seront considérés comme étant des mots clé qui seront utilisés lors de l'indexation d'images. Pour le choix de ces derniers, nous nous sommes heurtés à plusieurs obstacles à savoir :

Quels sont les critères qui nous permettent de juger qu'un mot a plus de valeur qu'un autre, du point de vue de son sens et de la symbolique qu'il émet. C'est-à-dire décrit-il un concept, ce concept est-il bien représenté dans l'image, ce sont autant d'éléments qui ont rendu difficiles le choix de certains mots.

Le choix de nos mots clé s'est effectué tout d'abord, par un ensemble de termes retenus par nos connaissances personnelles sur des concepts dans le domaine de l'architecture. Ce choix arbitraire ne permettant pas la crédibilité de notre démarche ; nous avons jugé important que chacun de ces mots doivent figurer non seulement dans le DICOBAT(15), mais aussi que ces derniers se trouvent dans le thésaurus relatif à la construction bois. Les mots jugés important pour nous et décrivant au mieux les images n'étant pas inscrit dans le thésaurus furent ajoutées.

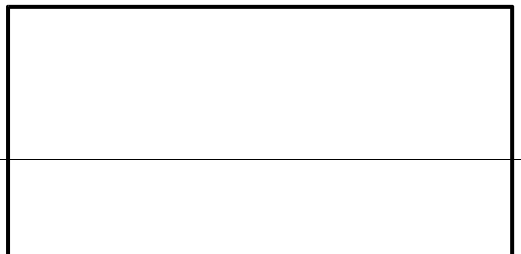
Et à cela d'autres critères furent instaurés afin de rendre plus crédible nos choix : un tableau récapitulatif des mots retenus fut établi. Lequel avait pour objectif de créer des relations entre les mots et la fréquence d'apparition dans le document auquel ils appartiennent, mais aussi leur fréquence dans l'ensemble du corpus (Cf annexe n° 1). Nous constatons que les mots dont la fréquence est élevée ne décrivent pas forcément le mieux une image et qu'ils peuvent être un obstacle dans la recherche d'un document précis (par exemple les termes : bois, métal, ouvrage, structure, qui se retrouve dans l'ensemble du corpus avec une fréquence d'apparition assez important). Les mots dont leur fréquence est assez faible décrivent le mieux une image et permet de distinguer un document d'un autre (par exemple les termes éclairage zénithal, salle à manger, que l'on ne retrouve que dans un seul document et représenté une seule fois).

2.3- Autres travaux réalisés

2.3.1- Cas de EngineSk

Dans le cadre de EngineSk, la grande difficulté réside dans le fait que le robot n'est qu'au stade embryonnaire comparer à WimexBot. Il présente de nombreuses limites. Celui-ci ne pouvant indexer uniquement que du texte. D'où la création de fichiers textes, dans lequel figure le commentaire d'une image, et à chaque fichier le code de l'image auquel il appartient lui est attribué (Fig: 3). Cela étant fait, établir des listes: de mots vides (Fig: 4), de mots normalisés, de mots composés, toutes ces listes devant permettre d'orienter le robot afin de répondre à nos attentes.

(15) Dictionnaire spécialisé dans les termes techniques de la construction en général



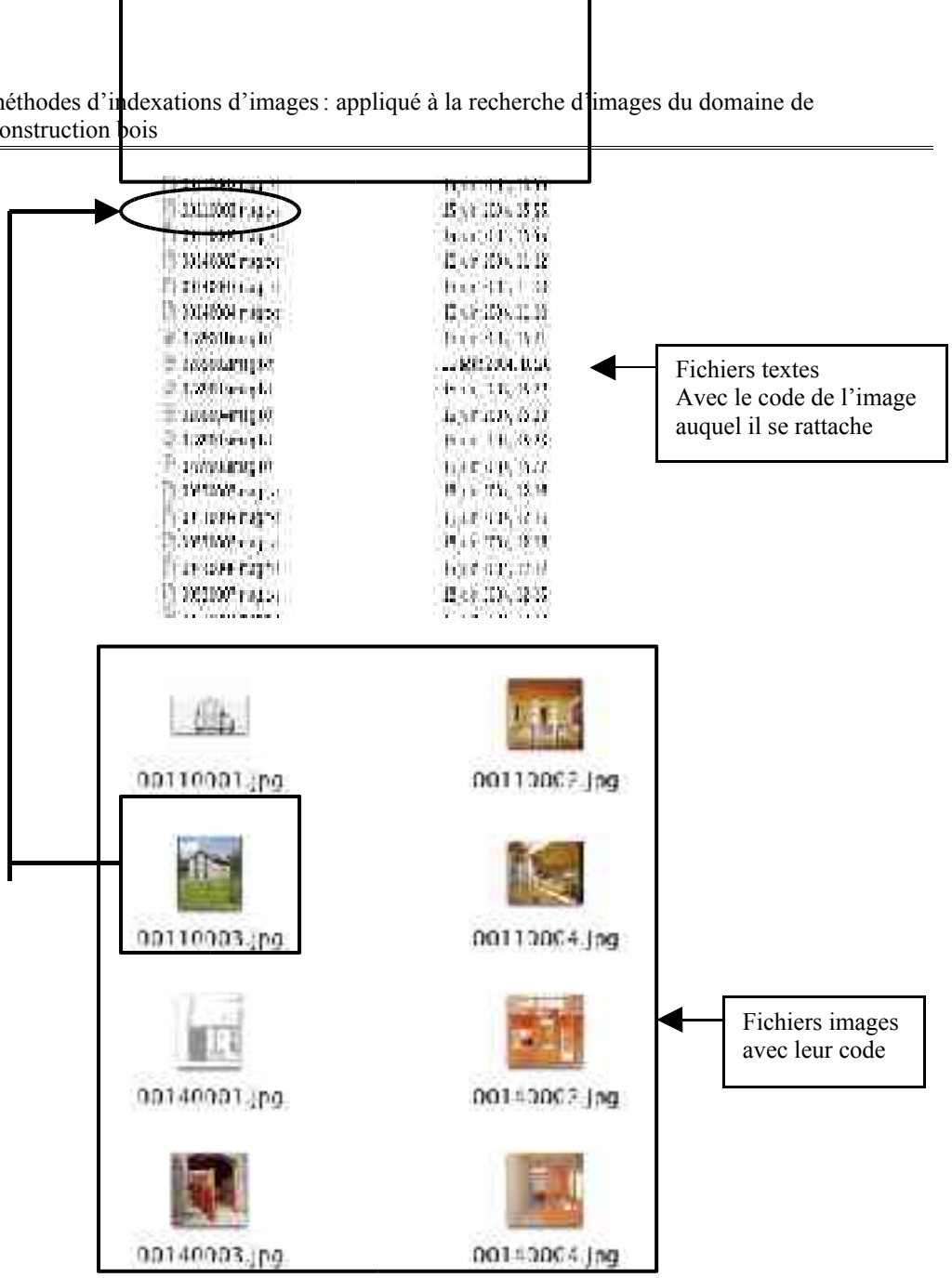


Figure 3 : Etablissement des fichiers textes

➢ Suppression des mots vides :

Les termes dit *mots vides* sont des pronoms, articles, etc (Fig :4). Ils ne sont pas pris en compte, car jugés non pertinent dû fait qu'il n'aide pas dans la description d'images. Il faut

savoir que nous envisage une indexation à texte intégral avec EngineSK. Si nous les prenons en compte, cela pourra enduire notre indexation à l'erreur au cours d'une requête. Nous aurons non seulement les pronoms, déterminants, etc; mais il essayera de retrouver ces derniers dans le contenu d'un mot (exemple dans le cas de « ***par*** » il effectue une similitude sur le mot « ***par****quet* » ou « ***la*** » similitude sur « *p****la****fond* »).

La liste de mots vides fut établie pour être utilisée par un filtre, afin de les supprimer et le résultat obtenu (fig : 15) sera utilisé après par EngineSk pour l'indexation.

autre	celui	depuis	et	leur	outre	quant	seuls	vous
autres	celui-ci	des	façon	leurs	par	que	son	y
aux	cependant	dès	grâce	lors	parcontre	quel	sous	voici
auxquelles	ces	dessous	hors	lorsque	parfaitement	quelle	sur	voilà
avant	c'est-à-dire	dessus	ici	lui	parfois	quelles	surtout	telle
avec	cet	donc	il	m'	part	quels	tant	tel
beaucoup	cette	dont	jusqu'	mais	pendant	qui	tard	uns
bien	ceux-ci	du	l'	même	plus	s'	tôt	unes
c'	dedans	effet	la	mêmes	plutard	sa	tous	certain
ça	chacun	en	là	moins	plutôt	sans	tout	certains
car	chacune	enfin	laquelle	n'	pour	se	toute	certaine
ce	chaque	ensuite	le	nous	pourtant	selon	toutes	certaines
ceci	comme	autrement	lequel	on	puis	ses	très	
celle	d'	entre	lequel	ou	puisque	seul	un	
celle-ci	dans	environ	les	où	qu'	seule	une	
tant	de	environs	cependant	outre	quand	seules	vers	

Figure 4 : Liste de mots vides

- Normalisation des mots :

La normalisation des mots avait pour objectif de rendre les mots plus réguliers par leur forme canonique (mettre au singulier, forme canonique des mots).

Lorsqu'il s'agit de transformer un mot au pluriel en singulier le problème est très simple, mais lorsque que nous sommes face à des termes plus complexes, où nous devrons retrouver la forme canonique nous nous heurtons à des obstacles. La forme canonique d'un mot sous-entend que le mot dérivé doit avoir le même concept que le mot canonique dont il est la source. A cet effet, nous nous sommes heurtés par moments avec un mot dérivé dont leur forme canonique ne décrive pas le même concept, (exemple *aggloméré* et *agglo*) dans de telle situation nous prenons ce terme en tant que descripteur à part entière, en considérant qu'il n'est pas issu d'un autre terme.

Au-delà de du traitement de la forme canonique des termes, résoudre le problème des mot-composés. Ces derniers n'étant pas reconnus de manière singulière par le robot EngineSK. Une liste de ces derniers fut établie et utilise au préalable par un filtre pour leur reconnaissent et qui par cet intermédiaire permettraient à EngineSk de les reconnaître à son tour.

2.3.2- Cas de WimexBot :

➢ Création des pages HTML :

Dans le cas de WimexBot, l'indexation les images n'étant possible que sous le format HTML(16) et non sous le format PDF. La transformation des fichiers PDF en HTML fut possible de manière automatique par l'intermédiaire d'un logiciel sur Internet, mais il a comme inconvénient majeur : que les images sont désolidarisées de leur texte immédiat (comparer Fig : 2 & Fig: 5), ce qui se traduit par une dislocation des images d'une part et des contenus textuels d'autre part (Fig: 5).

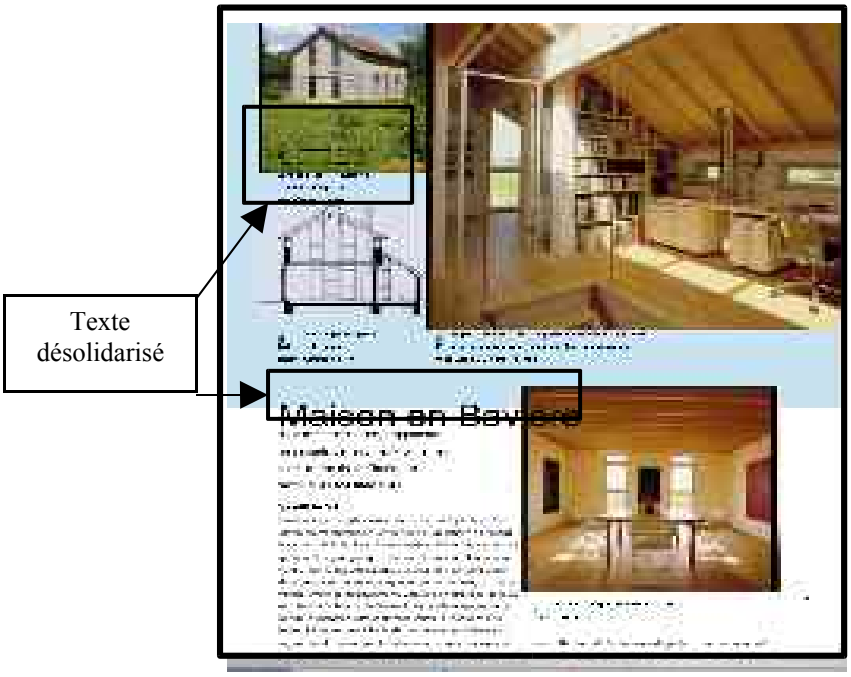


Figure 5 : Transformation du fichier PDF en HTML

(16) Une page HTML comporte comme dans l'ensemble des langages de programmation, une indication d'ouverture et une autre de fermeture

Figure 6 : Transformation d'une page PDF en HTML manuellement

Or la fiabilité d'une bonne indexation ne peut s'effectuer dans de bonnes conditions que si le texte se rattachant à l'image lui est proche (une des propriétés exigées par wimexBot) ; ce qui ne fut pas le cas lors de notre transformation. N'ayant plus recours à d'autre transformation automatique, cet inconvénient nous conduit irréversiblement à effectuer cette conversion de manière manuelle les vingt projets en pages HTML (Fig: 6).

La présentation d'une page en HTML est régit sous certaines contraintes bien définies, c'est-à-dire le dimensionnement des images, la proximité du texte se rattachant à l'image. Ces exigences doivent être respecter, afin que les pages réalisées soient en conformité avec la balise(17) « Cf: chapitre 5, au 5.1.2 ». De manière générale nous avons essayer de reconstituer les pages de projet de telle sorte qu'il puisse être le plus proche possible de l'original.

(17) Les balises sont des instructions HTML.Elles sont reconnaissables à leur forme particulière. En effet, chacune est encadrée par les signes inférieurs à (<) et supérieur à (>).De plus, les balises permettent l'ouverture et la fermeture d'une instruction HTML.Une barre oblique, le slash (/) en l'occurrence, différencie les deux types de balises.

La balise <…> marque le début de l'instruction.

La balise </. >marque la fin de l'instruction

- Problème de synonymie :

L'établissement de termes synonyme s'est effectué en deux étapes ; regrouper les termes dont la synonymie est évidente (exemple *poutraison* qui envoie à *poutre*, ou *solivage* qui renvoie à *solive*). La deuxième approche concernait les termes plus complexes, dont la synonymie n'est pas rattacher au fait qu'ils ont la même racine. Mais que ces termes doivent être synonymes avec les termes du thésaurus uniquement et doivent avoir le même concept (exemple *triangulé* qui renvie à *treillis*, ou *éclairage zénithal* qui renvoie à *éclairage en toiture*), cas réservé uniquement à WimexBot qui utilise les termes du thésaurus pour l'indexation des images.

mots	**Thesaurus bois**	**N°**
accolée	assemblage	13
acrotère	finition de toiture	131
cathédrale	bâtiment religieux	29
chape	dalle de protection	127
chevillées	cheville	466
cintrage	cintré	136
collage	collé	55
contrefort	saillie de façade	72
contremarche	escalier	214
copeaux	aggloméré	389
ctbx	contreplaqué	389
dallage	dalle de protection	127
dédoublé	moisant	485
duplex	habitation individuelle	25
éclairé zénithalement	éclairement en toiture	151

Figure 7 : Synonymie des termes et classification par rapport au thésaurus

Pour créer les liens entre les termes retenus et ceux du thésaurus, l'utilisation du DICOBAT, nous a permit de trouver le sens des termes choisis et de les relier aux termes du thésaurus correspondant. Ceci étant fait, nous avons dû attribuer un code à chacun des mots choisis (Fig: 7) ,code qui correspond à leur synonyme (ou non-descripteur)dans le thésaurus de WimexBot (Cf annexe n° 2). Ce dernier ne reconnaissant pas les synonymes lors de l'indexation, ce lien permettra d'effectuer une correspondance «ou renvoi » au cour de l'indexation.

2. 3.3- Insertion d'une rubrique

En vue d'apporter plus de renseignements pouvant permettre la description d'images, nous avons jugé important d'insérer une nouvelle rubrique au thésaurus à savoir : « Espace » qui se subdivise en deux ensembles, les espaces extérieurs, et les espaces intérieurs (Fig: 8). Pourquoi insérer cette rubrique qui ne figure pas dans le thésaurus ? Elle jour un rôle très important du fait que l'on trouve des termes se rattachant à des concepts qui décrivent bien le contenu d'une image.

Espaces	
Espaces intérieurs	**Espaces extérieurs**
salle à manger	aire sportive
salle d'exercice	blocs techniques
salle de lecture	local technique
salon	entrée
séjour	aire jardin
cuisine	piscine
salle de bain	salle de sport
grange	salle d'activité sportive
sas	
salle de conférence	
galerie	
hall	
halle	
salle d'activité	
chambre	
sanitaire	
cage d'escalier	
bureaux	
boutiques	

Figure 8 : Composante de la rubrique Espace

2.4- Bilan

➢ Difficultés :

La grande difficulté rencontrée au cour de l'élaboration du travail préalable réside du fait que toutes les étapes fut réalisées manuellement. Les outils dont nous disposons pour réalisation de nos travaux à l'aide du matériel informatique n'ont pu satisfaire nos attentes. Dans le cas de la normalisation des termes qui devait être effectuée par le logiciel Tree Tagger pour l'avancement rapide du travail. Cet outil qui permet, la normalisation des mots automatiquement (c'est-à-dire mettre les termes au pluriel au singulier, mais aussi de pouvoir retrouver la forme canonique des termes). Au vue du temps qui nous été imparti, étant un outil assez complexe nécessitant l'intervention d'un informaticien, nous n'avons pu l'utiliser pour notre étude.

➢ Suggestions :

Dans l'élaboration de votre travail nous avons jugé utile de retenir notre attention que sur la légende d'images. Pourquoi ce choix arbitraire ? Nous partons du principe que celle-ci doit pouvoir décrire au mieux le contenu d'une image plutôt que le commentaire du projet. Il est certes vrai que nous avons rencontré des handicaps dans ce choix, c'est-à-dire :

- Il nous est arrivé d'être confronté à des images dont le contexte immédiat ne décrivait pas le contenu de celle-ci avec précision.
- Ou plus encore que le contexte ne décrivait pas du tout le contenu de l'image.

Dans de telles situations, nous nous referons au commentaire du projet, en supposant que la probabilité de trouver des renseignements relatifs à la description de ces images sera grande.

De manière générale, le commentaire du projet décrit de façon globale l'ensemble des images qui le compose. Il est vrai qu'il d'écrit des images, mais pas de manière individuelle, à cela se pose le problème de la sélection des éléments textuels pouvant permettre de retenir notre attention, afin d'être utilisé lors de l'indexation. La présence des phrases décrivant de manière sucinte une image, mais en faisant référence à des concepts qui ne s'y retrouvent pas dans l'image.

Il est vrai que le commentaire de projet nous apporte des données intéressantes dans la description générale du projet, mais dans la description des images de manière spécifique cela reste à satisfaire. Dans une telle situation, il paraît judicieux d'orienter notre indexation de deux manières, une indexation se limitant uniquement à la légende de l'image et une autre se référant à la fois à la légende de l'image et au commentaire de projet. Seule l'expérimentation nous permettra de juger de la pertinence de chacune.

Dans l'ensemble, nous pouvons dire que l'une des grandes difficultés rencontrées réside dans le fait que l'ensemble des différentes étapes de notre travail de recherche s'est effectué manuellement, ce qui fut en grande partie un handicap majeur, en matière de temps. Au-delà de ce handicap, ce travail manuel nous a permis de mieux élaborer certains problèmes, par exemple dans le cadre des choix des termes retenus pour l'indexation.

CHAPITRE 3 : Analyse de l'indexation manuelle

3.1- Présentation de l'indexation manuelle

L'indexation manuelle fut effectuée par une thésarde du CRAI, elle s'est fait sur la base d'un thésaurus préalablement réalisé dans le domaine de la construction bois. Ce thésaurus est regroupé en cinq grandes familles à savoir : Produit, Type de réalisation, Matière, Ouvrage architectural, Espace. Tous ces éléments sont subdivisés en trois sous unité hiérarchique (Fig: 9 & 10). Cette indexation fut réalisée par l'intermédiaire d'une interface : 4D (améliorée par Mr. Humbert) permettent la visualisation des images et de leurs contextes.

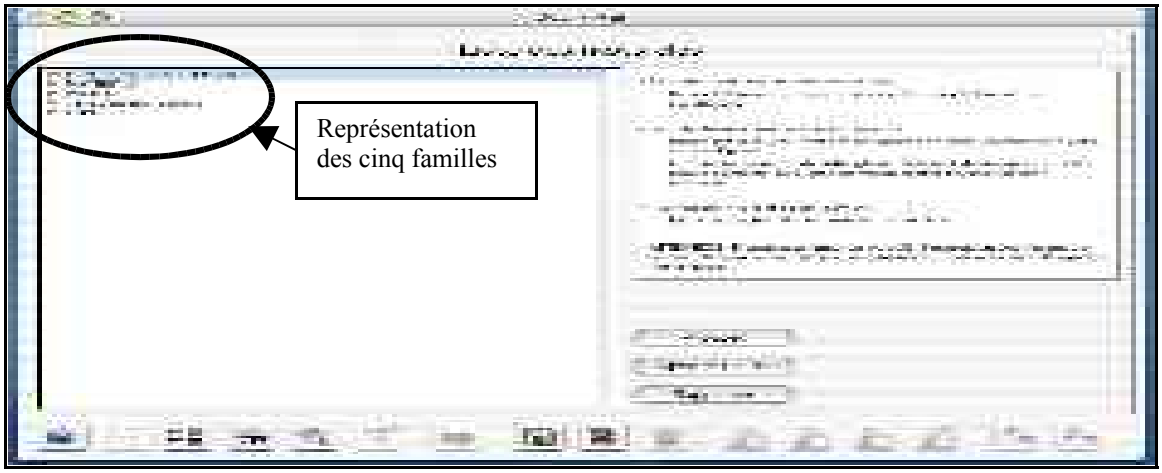


Figure 9 : Classification des familles

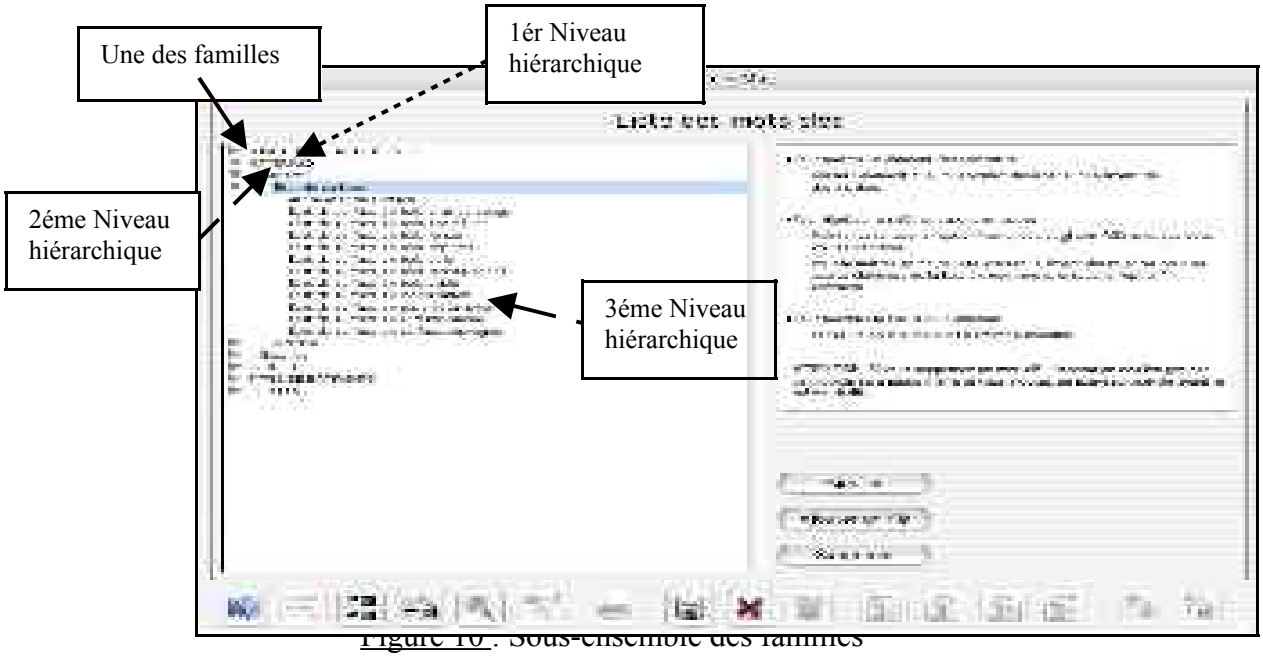


Figure 10 : Sous-ensemble des familles

3.2- Procédé d'indexation

3.2.1- Choix des images

Dans un premier temps les images sont sélectionnées, à travers le logiciel de récupération des images sous le format photo, afin d'être utilisé dans l'interface 4D. Les images de type dessin technique, n'ayant pas été retenue, jugé non pertinent (EngineSk ne les prend pas en

considération) dans le cadre de notre étude, nous ne retenons que les images photographiées (Fig: 11).

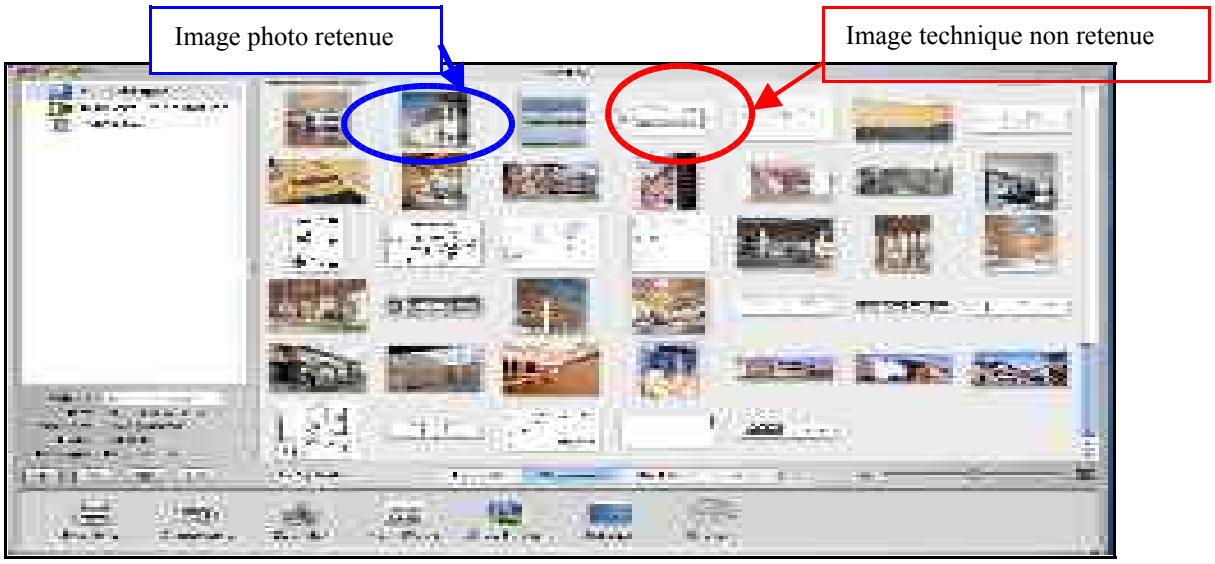


Figure 11 : Sélection des images

3.2.2- Création de la base de donnée

Dans un premier temps on insère l'image retenue, puis on lui attribut des renseignements qui permettront de retrouver l'image par les mots contenus dans son contexte. La création de cette base de donnée s'effectue comme suit (Fig: 12) :

- Insertion d'une image en lui attribut un numéro pour sa reconnaissance (00110002.jpg)
- Un commentaire de l'image permettant de décrire le contenu de l'image à l'aide de mots clé (sujet)
- On donne le thème général décrivant l'image (sous sujet)
- On détermine la source de l'image (CNDB)
- Le chemin par lequel l'image a été extraite (ANGO :imagesprojetbois : 00110002.jpeg)
- La taille de l'image (effectuée automatiquement)

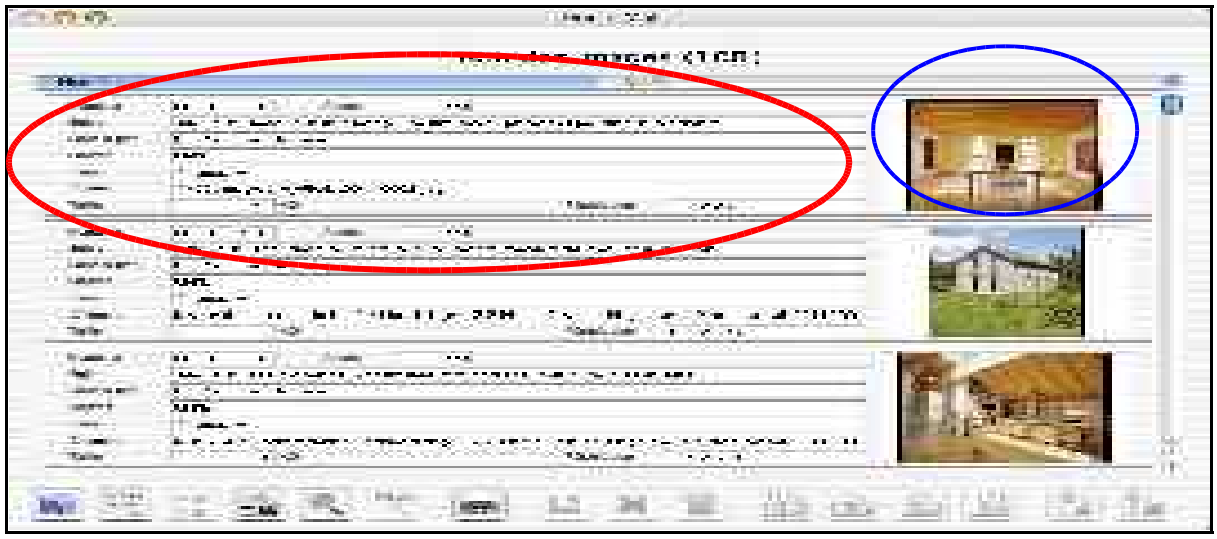


Figure 12 : Création de la base de donnée

Sous la rubrique « sujet » qui se rattache au contexte de l'image, le commentaire de ce dernier ne peut comporter que quatre-vingt (80) caractères en prenant en compte aussi les espaces. Cette limite de caractère permet d'effectuer une sélection assez restrictive des mots qui permettra une meilleure description de l'image, qui facilitera de retrouver plus aisément une image. Mais cette restriction peut avoir comme limite de ne pas pouvoir décrire de façon plus explicite une image qui contient une variété d'information importante.

3.2.3- Indexation manuelle

L'indexation des images s'effectue uniquement avec des mots du thésaurus, et particulièrement avec des mots du troisième niveau hiérarchique, qui sont moins générique et permette d'éviter des ambiguïtés. Les mots choisis dans celui-ci doivent décrire avec plus ou moins d'exactitude le contenu de l'image, afin d'obtenir une réponse plausible lors d'une requête. Les mots retenus sont pondérés manuelle par une classification qui varie de 1 (moins pertinent) à 5 (plus pertinent) et qui permet d'ordonnancer par ordre d'importance. Cette classification permet de mettre en avant l'élément saillant de l'image en valeur, afin de permettre une bonne indexation manuelle (Fig: 13).

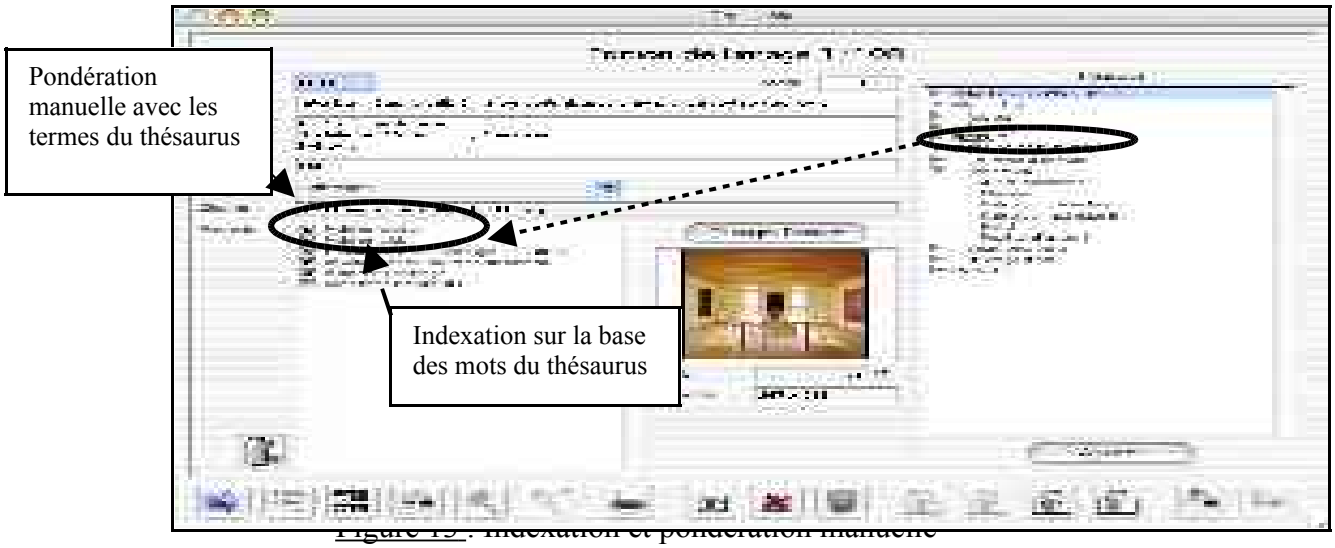


Figure 13 : Indexation et pondération manuelle

3.2.4- Requête

Dans le cas précis, lors de notre requête nous recherchons toutes les images contenant des poteaux. Pour juger pertinent la réponse à cette requête il faudrait qu'à la fois l'élément poteau soit représenter sur l'image, mais aussi que le mot « poteau » soit dans le commentaire d'image (Fig: 14).

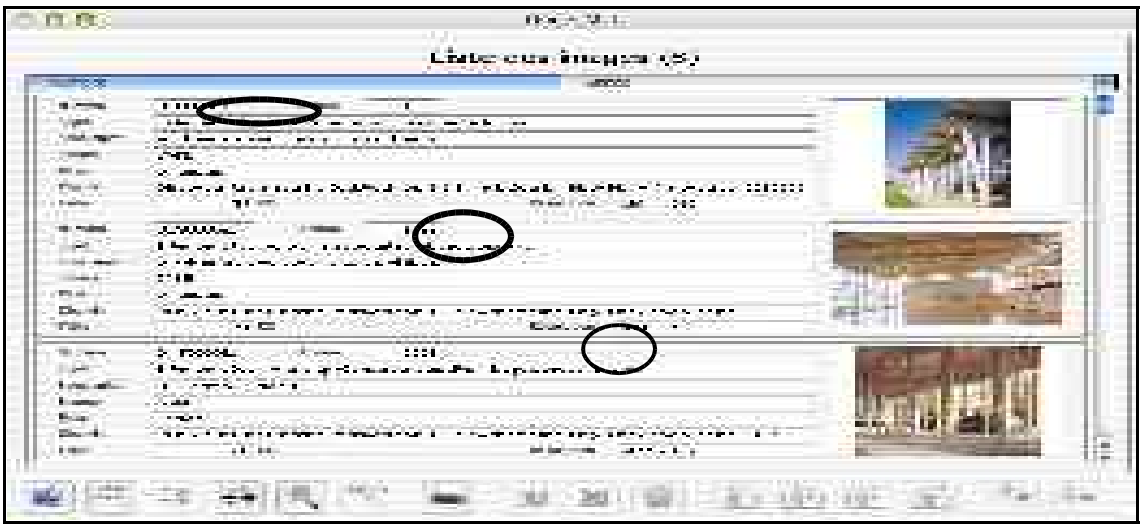


Figure 14: Réponse d'une requête

CHAPITRE 4 : Analyse du robot d'indexation EngineSK

4.1- L'Architecture de EnginSK

Etant au stade embryonnaire, il est amélioré par Mr. Pascal Humbert au laboratoire du CRAI. L'utilisation de celui, avait pour but de réaliser des indexations d'images par leur contenu textuel. N'étant qu'à l'étape de prototype, il est confronté à certains handicaps : la non reconnaissance des mots vides, l'absence d'un normalisateur des mots, la non reconnaissance des mots composés. Ce dernier ne pouvant résoudre ces problèmes, nous passons par l'intermédiaire d'un filtre capable d'enrayer ces problèmes.

4.2- Travail préalable effectué par le filtre

4.2.1- Suppression des mots vides :

Le filtre devant être utilisé pour la suppression des mots vides, s'effectue automatique à l'aide d'une liste des termes vides (Fig: 4) « il s'agit : des déterminants, pronoms, prépositions, conjonctions (le, ou, s', votre, etc.) » qui est préalablement établi. A cet effet nous obtenons un texte dépourvu de ceux-ci (Fig :15).

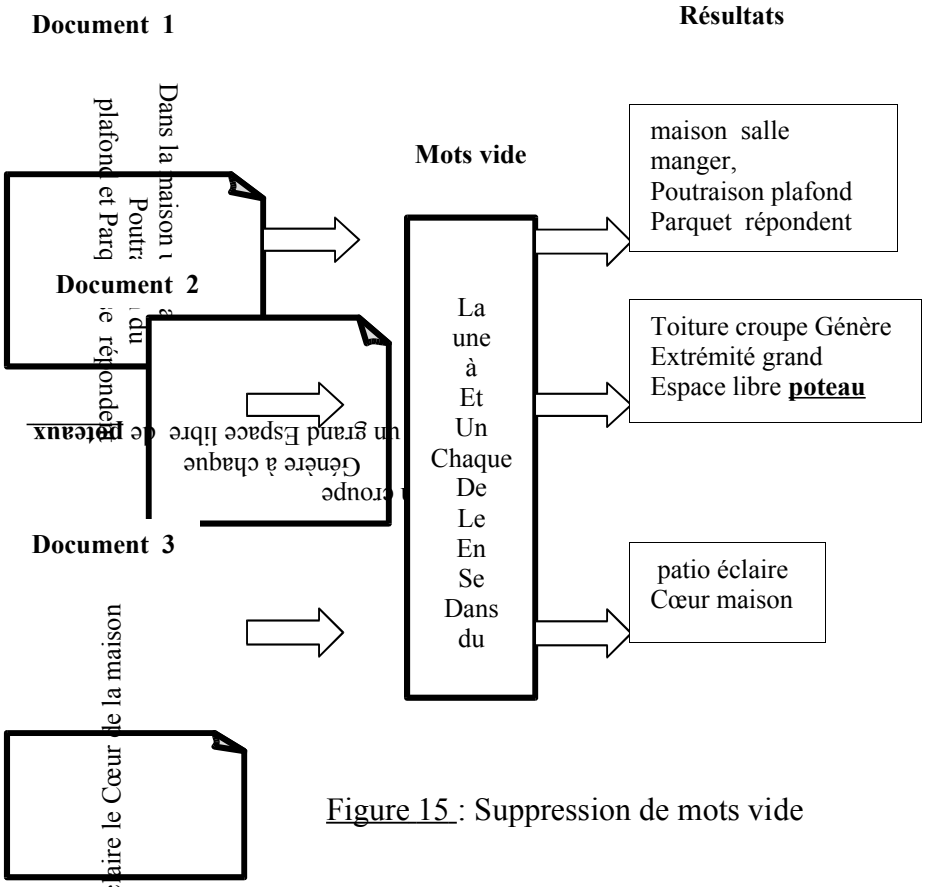


Figure 15 : Suppression de mots vide

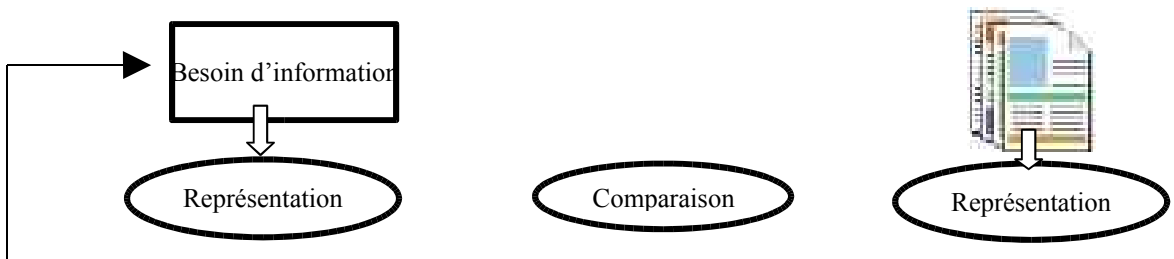
4.2.2- Normalisation des termes :

Elle consiste à mettre tous les mots pluriels au singulier, car EngineSk considère un même terme singulier et pluriel comme étant deux termes différents (exemple: *poteaux* et *poteau*). Pour palier à ce problème une normalisation y est nécessaire à ce titre nous passons toujours par l'intermédiaire du filtre. A cela, ce dernier fut aussi utilisé pour la reconnaissance des

mots composés qui n'étaient pas pris en compte comme étant un seul mot, mais plutôt comme étant des termes à part entière.

4.3- Fonctionnement d'EngineSK

Le travail préalablement établi par le filtre, nous a permis l'utilisation de EnginSk dans de bonne condition. Les différents textes étant traités, ils furent insérés dans le robot afin d'être indexés. EngineSk est un robot qui indexe uniquement du texte en utilisant la méthode statistique (Fig: 16) : C'est-à-dire qu'à chaque mot indexé il lui affecte un poids qui représente son importance vis-à-vis du contenu du document (exemple le mot « *poteau* » pris dans une phrase aura un poids plus important que pris dans un paragraphe). Le poids attribué à un terme est proportionnel à sa fréquence d'apparition dans le texte. Plus le poids est important, plus le terme est jugé apte à décrire mieux le contenu (dans le cadre de notre étude le contenu textuel est en relation avec le contenu de l'image).
Le robot ayant indexé les textes, pour la visualisation des images auxquels les textes se réfèrent on utilise une interface dont le logiciel 4D(Cf : chapitre 3) déjà utilisé lors de l'indexation manuelle.



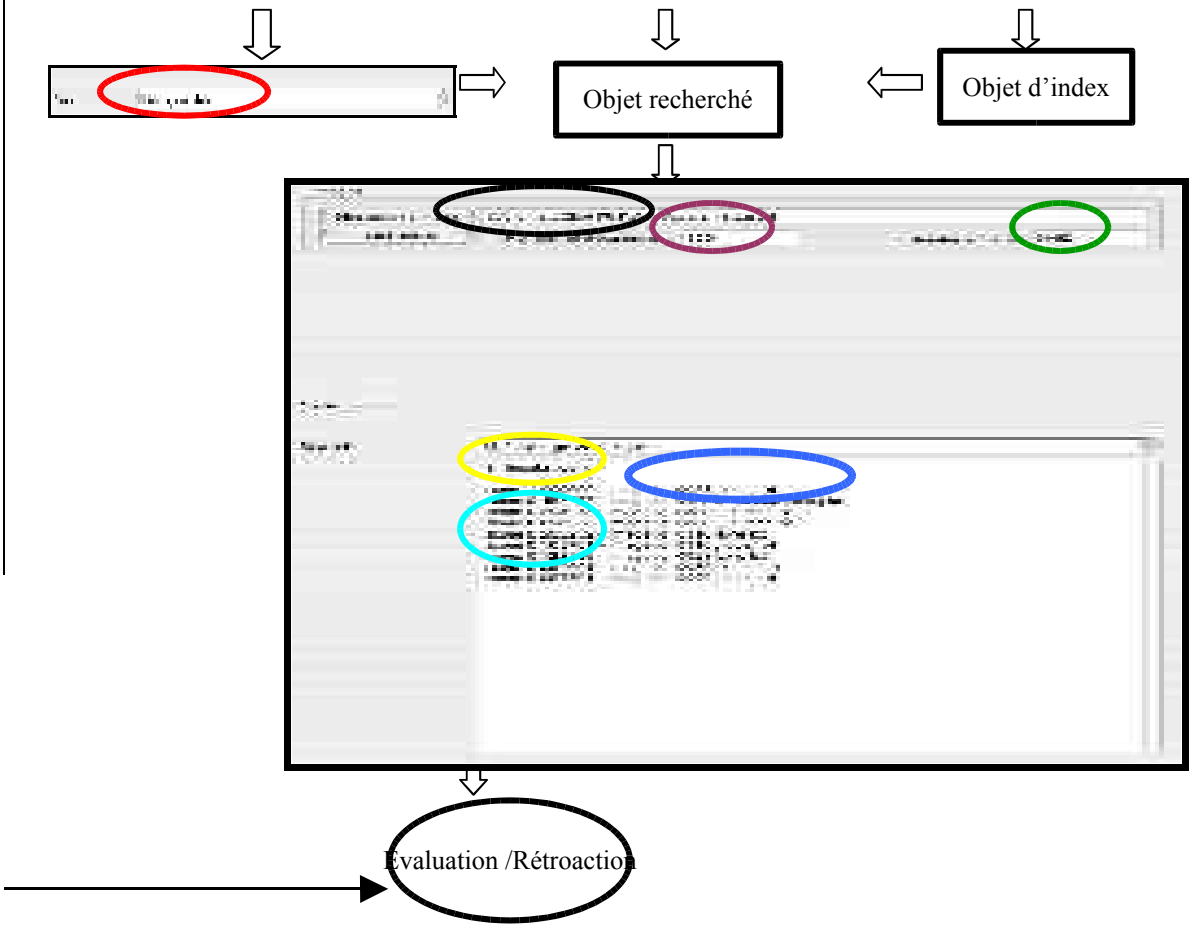


Figure 16 : Représentation de la requête

Formulation d'une question

Site des documents sources

Nombre de fichier

Nombre total de résultat

N° du document appartement le terme

Nombre de termes

Score du terme dans le document

CHAPITRE 5 : Analyse du robot d'indexation et d'extraction d'image WimexBot

5.1- Présentation de l'application WimexBot

5.1.1- Introduction

WimexBot (Web Image Extractor robot) [NAK 03], est un prototype écrit en JAVA, spécialisé dans l'extraction et l'indexation d'images à partir de site français de produits du bâtiment. Il s'apparente aux moteurs de recherche sur le Web. Il scrute le site; récupère des images et extrait de leur environnement textuel proche, des termes en s'appuyant sur le vocabulaire du thésaurus, dans le cadre de notre étude le thésaurus utilisé est relatif à la construction bois.

5.1.2- Vocabulaire/Définition

Pour comprendre l'application de ce logiciel, il est important de prendre connaissance des différentes étapes nécessaires pour la configuration de cette application :

- *WimexBot* : est une application permettant de trouver, de mémoriser et indexer des images à partir du Web,
- *Site* : est une représentation un ensemble de documents liés entre eux et dont chaque URL(18) a la même racine,
- *Page* : c'est un document texte dans lequel on peut trouver des URL. L'application considère comme page que les documents de type « texte/HTML »,
- *Image* : c'est un document de type « image/jpeg »
- *Paragraphe* : c'est un groupe de mots limité par des séparateurs. Les balises HTML suivantes sont considérées comme des séparateurs de paragraphe :
  - <TABLE> : tableau,
  - <TD> : cellule d'un tableau,
  - <BODY> : corps de la page HTML,
  - <H1> à <H7> : titre,
  - <HR> : ligne de séparation horizontale,
  - <IMG> : image.
- *Contexte* : un paragraphe peut être plus ou moins proche d'une URL (lien hypertexte, image, …). Il représente alors un contexte pour l'URL. L'application s'intéresse qu'aux contextes proches des URL afin de déterminer les termes entourant une page ou une image,

(18) Uniform Resource Locator est un format de nommage universel pour désigner une ressource sur Internet

- *Thésaurus*, il est élaboré sur la base d'une liste de termes descriptifs relatifs à la construction bois et de termes descripteurs de cette construction, qui se classifient en plusieurs thèmes comme : matériaux, espaces, produits, ouvrages, réalisation.

5.2- Principe de fonctionnement du WimexBot

Le principe de fonctionnement du WimexBot) [CANE 04] s'effectue par le robot en parcourant un site donné en paramètre et extrait toutes les images répondant positivement à certains critères de sélection (critères que l'utilisateur peut éventuellement paramétrer:

- ***Parcours du site spécifié, et analyse de leur contenu (images, textes, lien hypertextes).***

Un premier parcourt du site est effectué afin de trouver des URL de pages et des URL d'images intéressantes. Parmi les images trouvées, seulement celles considérées comme graphiquement pertinentes, seront sélectionnées pour être analysées ultérieurement

- ***Sélection des images en fonction de leur formes et de leur contexte textuel***

Ainsi une première sélection des images est réalisée en considérant les critères de formes (hauteur, largeur, proportion et taille de l'image). Afin d'obtenir des images pertinentes, une analyse du texte autour de l'image a été réalisée. Cette analyse est faite en utilisant le thesaurus de la construction bois.

- ***Recherche des termes présents à la fois dans le thésaurus et dans les textes proches de l'image***

Lors d'un deuxième parcours des pages qui contiennent les images graphiquement pertinentes, nous recherchons les paragraphes afin de réaliser la deuxième étape de sélection de ces images. Seulement les images qui ont au moins un contexte, qui contient un thème du thésaurus seront retenues.

Dans le principe de fonctionnement de WimexBot, nous distinguons trois étapes principales :

- *recherche de documents intéressants,*
- *recherche des contextes d'une image,*
- indexer les images par des thèmes du thesaurus

5.2.1- Recherche de documents intéressants

L'étape consiste à l'application d'un premier parcours du site afin de rechercher les URL de pages et d'images, qui vont représenter les documents intéressants. Lors du parcours, l'application récupère toutes les URL et teste si elles sont correctes. Si c'est le cas, nous effectuons une analyse spécifique adaptée au type de document que ces URL représentent.

5.2.2- Recherche des contextes d'une image

La deuxième étape est la recherche des paragraphes des pages qui ont été retenues comme intéressantes lors de la première étape et qui contiennent des images. Rechercher les paragraphes d'une page permet d'extraire le texte qui accompagne les images de cette page, afin d'avoir le plus de renseignements possibles sur le contenu de ces images.

Le robot parcourt une deuxième fois la page HTML afin de construire les paragraphes, calculer les chemins des paragraphes et des images par rapport à la racine de la page HTML,

ainsi que de calculer la distance entre chaque image et chaque paragraphe de la page. Le but est d'identifier les contextes proches des images.

5.2.3- Recherche des contextes d'images pertinentes

Les contextes proches des images étant repérés, il faut rechercher parmi eux, ceux qui contiennent au moins un des thèmes du thesaurus. Pour savoir si un contexte contient un thème du thesaurus, il faut comparer les deux textes. Pour pouvoir réaliser cette analyse de comparaison entre un contexte et un thème du thesaurus, la stratégie de recherche optimale par groupe nominal. Les thèmes trouvés dans les contextes proches d'une image seront utilisés pour indexer cette image. Avec l'indexation des images, l'application fournit l'ensemble des résultats nécessaires à être utilisés ultérieurement. Il n'en reste plus qu'à mettre ces résultats à disposition de l'utilisateur.

NB :Pour plus de compréhension confère annexe n°3

CHAPITRE 6 : Expérimentations et résultats

L'expérimentation que nous avons mené est un moyen de juger de la pertinence des résultats obtenus par chacun des modes d'indexation. Dans le cadre de notre étude, la pertinence est relative au fait que les résultats obtenus doivent être le plus proche possible des réponses obtenues par l'indexation effectuée par l'expert. Pourquoi avoir pris l'indexation manuelle comme étant le point de comparaison des deux robots ? Les projets étant élaborés par le raisonnement humain, l'indexation de l'expert sera plus pertinente. A cela, elle a la possibilité

d'effectuer un retour sur les mots et les images retenues, ce qui n'est pas forcément le cas automatiquement.

6.1- Protocole de l'expérimentation :

6.1.1- Choix de l'échantillon :

Dans le cadre de notre expérimentation, les utilisateurs que nous ciblons appartiennent au domaine de l'architecture, comme les architectes. Nous envisageons d'effectuer notre étude sur deux principaux types de besoins auxquels répond notre expérience :

- La première approche s'effectue sur des critères se rattachant à des thématiques bien précise qui sont de quatre ordres à savoir: poteau, charpente, bardage, fenêtre. Cette rubrique s'applique dans le cas ou l'utilisateur connaît le nom de l'élément qu'il recherche. Dans ce cas, une recherche par mots-clés est souvent plus appropriée (Fig:17).
- La deuxième consiste à une approche aléatoire. Cette rubrique s'applique, dans le cas ou l'utilisateur n'a qu'une idée plutôt vague de ce qu'il recherche. L'utilisateur recherche des idées, mais ne sait pas quoi précisément. En premier lieu, il visualise au hasard puis au fur et à mesure de son parcours, il affine sa recherche jusqu'au moment ou il amorce l'acceptation ou le rejet des images présentées.

Chacune de ces rubriques (thématique et aléatoire) est représentée par neuf images se rattachant un thème choisi pour l'expérimentation (Fig :18).

Figure 17: Présentation des cinq rubriques pour l'expérimentation sur la page d'accueil de 4D

Figure 18 : Représentation des neuf images retenues dans la rubrique charpente

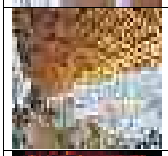
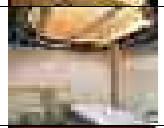
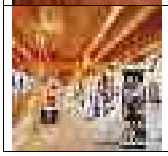
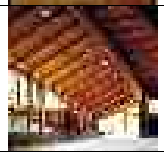
Rubrique Charpente					
Image	**Code**	**Poids**	**Image**	**Code**	**Poids**
	**Image requête 00800006**			**Image requête 00530006**	
	3200002 Résultat	5		01100006 Résultat	5
	00530003 Résultat	5		00110004 Résultat	4
	03220002 Résultat	4		02890008 Résultat	4
	01100005 Résultat	4		00110002 Résultat	3
	02300002 Résultat	3		00530005 Résultat	2

Figure 19 : Résultat possible lors d'une requête de la rubrique « charpente »

Dans un second temps, nous avons effectué manuellement une classification des différentes images devant apparaître au cours d'une requête en fonction des différentes rubriques (Fig :19). Cette classification fut établie par ordre d'importance, sous forme de graduation allant de 1 (moins pertinent) à 5 (plus pertinent). Plus une image décrira avec précision une des rubriques plus sa pondération tendra vers 5, moins il l'exprimera plus sa pondération tendra vers 1.

Cette attitude nous permettra d'affirmer ou d'infirmer. S'il est possible de retrouver une image par le seul critère, que son contexte proche décrit son contenu de manière explicite par des mots clés le constituant. Dans un autre coté cette attitude permet de voir les limites de chacun des robots sur la base de leur mode d'indexation réciproque.

6.2- Essai d'expérimentation :

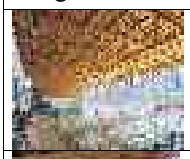
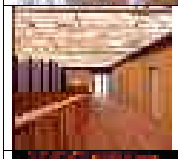
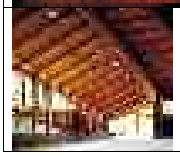
Rubrique aléatoire : thème charpente		
Image	N°	Poids
	3200002 Requête	
	00800006 Résultat	5
	01100005 Résultat	5
	00530003 Résultat	4
	00530005 Résultat	3

Figure 20 : classification manuelle des images attendues en fonction d'une requête

Au cour de notre expérimentation nous avons sélectionné une image de la rubrique aléatoire. Cette dernière représente le concept de *charpente*. Mais au préalable nous avons établi

manuellement les images devant répondre à cette requête (Fig : 20). Les critères permettant de classifier les images devant apparaître sont de trois ordres :

- Identifier les éléments présents dans l'image de manière précise
- Ces éléments doivent renvoyer à un concept du domaine de l'architecture
- Ces concepts doivent-être présent dans le commentaire de l'image

Pour chacune des méthodes d'indexation, nous allons essayer de comparer leurs résultats avec celui établi manuellement, mais aussi entre eux.

6.2.1- Résultats de l'indexation manuelle spécialisée (utilisation d'un thesaurus) :

Dans ce cas nous constatons que les résultats obtenus (Fig: 21) ne sont pas conformés avec les réponses prédéfinies (Fig :20). Quel peuvent bien être les causes de cette divergence de résultat ? C'est ce dont nous essayerons de comprendre ci-après.

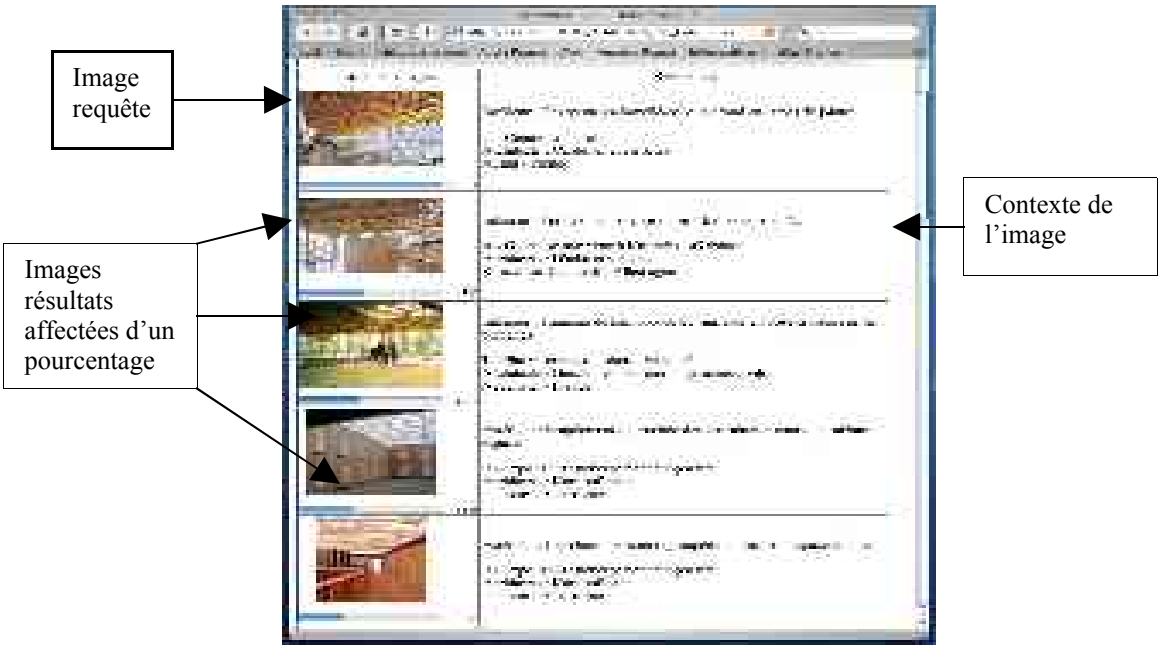


Figure 21 : Résultat obtenu par l'indexation spécialisée

- Etude de l'image requête et ses résultats :

Pour comprendre les réponses obtenues, nous analysons les mots contenus dans la légende d'image ; les termes utilisés pour l'indexation de celle-ci et la pondération qui leur est attribuée. Nous observons que l'image requête est représentée en grande partie par le concept « charpente », et que ce terme est utilisé pour l'indexation de l'image avec une forte pondération de cinq (Fig : 22).

Par ce constat peut on dire que les résultats obtenus sont relatifs uniquement au commentaire du texte ? ,Ou simplement à la pondération des termes indexeur ? Ou prend-il en compte les deux éléments précédents ?.

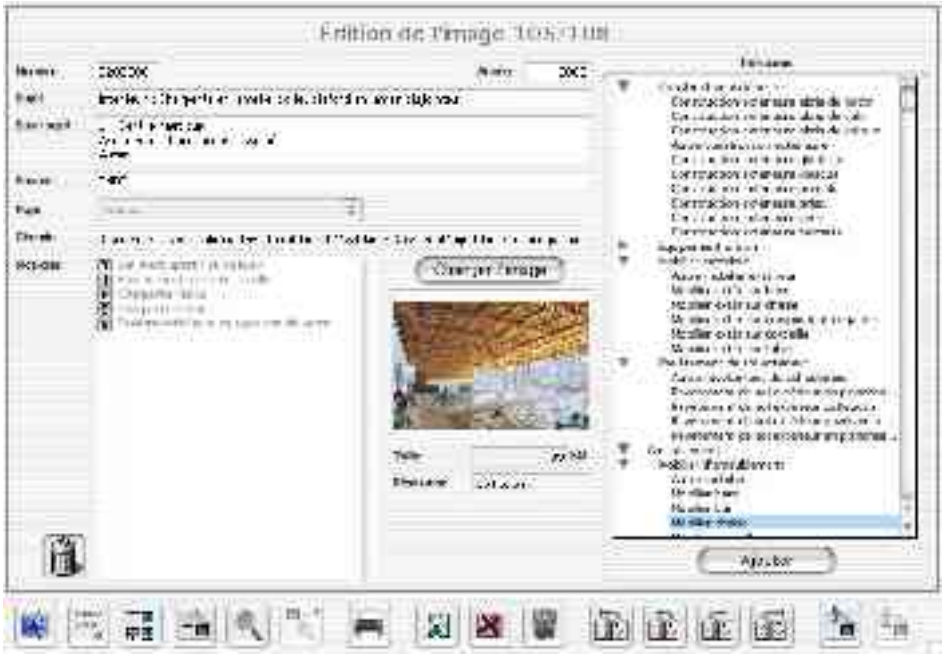


Figure 22 : Indexation effectuée sur l'image requête

➢ Etude des images réponses :

Tout d'abord, nous comparons la légende de l'image « requête » avec celle des trois premiers résultats. Nous remarquons qu'il n'existe aucun lien, du fait que les termes utilisés dans le commentaire d'image requête sont totalement différents de ceux des résultats.

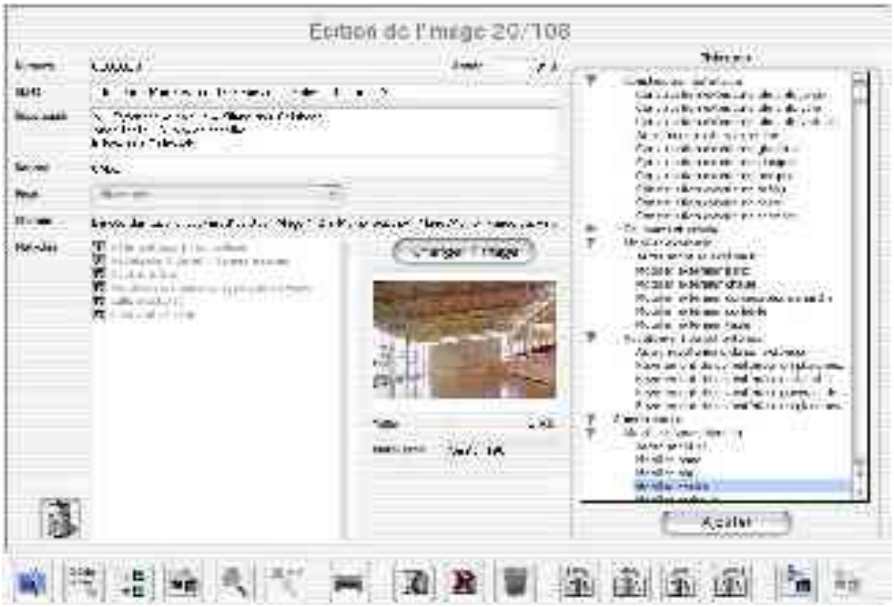


Figure 23 : Premier résultat requête

A première vue, l'image requête est indexée par deux termes dont leur pondération est cinq à savoir « *charpente treillis et charpente résille* (Fig 22) ». A cela les trois premiers images (résultats) sont indexées et pondérées par un même terme « charpente treillis » au poids 5, (Fig : 23 ; 24 ; 25). Cette classification se justifie par le fait que ; la première image « réponse (Fig :23) » est reliée au premier plan avec l'image requête du fait qu'en plus du terme « charpente treillis », elle est indexée par d'autres termes étant en commun avec l'image source tel que « *Bâtiment sportif, fenêtre extérieure* » (Fig :22 ;23). Ces liens en plus font en sorte qu'elle se trouve à la première place.

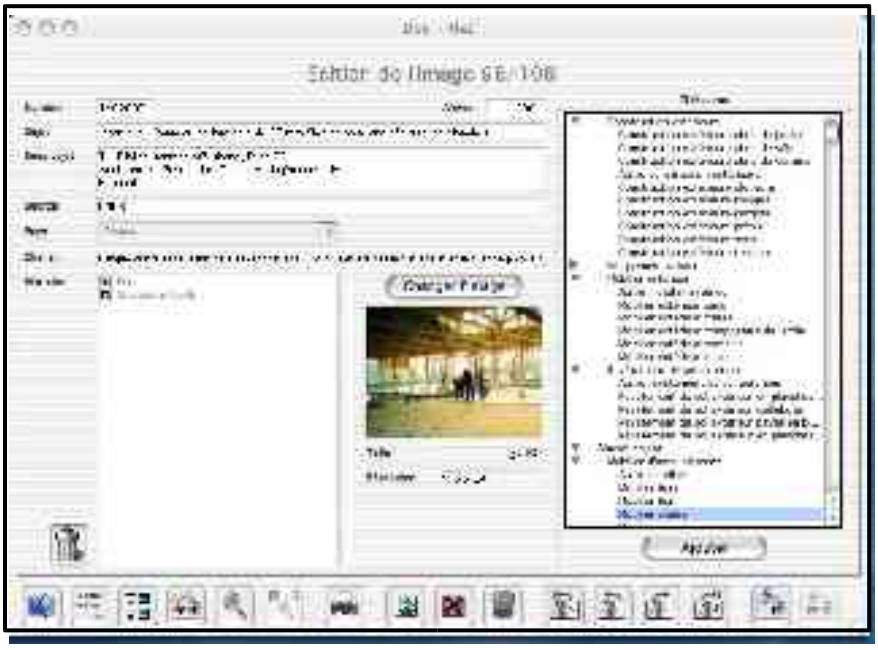


Figure 24 : Deuxième résultat

Figure 25 :Troisième résultat

Par rapport à la deuxième et la troisième place, elle s'explique du fait que la deuxième image n'est indexée que par deux termes et dont le terme ayant une forte pondération y est représenté (Fig :24) dans l'image requête. Dans la troisième réponse (Fig :25), celle-ci est indexée par quatre termes dont trois d'entre eux ont une forte pondération, dont l'un des trois est « *charpente treillis* ». Cela est dû au fait que la seconde image est indexée par deux termes uniquement et que l'un des deux termes indexeur à une faible pondération (donc elle n'est pas prit en compte par le robot qui agit comme si l'image n'était qu'indexée par un seul terme). Or la troisième image étant indexée par trois termes de même poids le robot ne la considère la deuxième réponse plus pertinente.

L'analyse des résultats (dix premières images) obtenus au cour de la requête de rubrique charpente (Cf :annexe 5) montre que 80% des images obtenus sont pertinentes contre 20% non pertinentes (bruit).

6.2.2- Résultats de l'indexation sur texte intégral (EngineSK)

Les résultats obtenus dans cette méthode d'indexation, sont différents de ceux souhaités. Il faut rappeler que l'indexation effectuer par EngineSk sur le texte intégrale, s'effectue en indexant tous les termes contenus dans le texte en leur affectant un poids relatif à leur fréquence dans le texte. Nous faisons ce rappel du fait que les résultats obtenus au cour de l'image requête dépendent fortement de ce critère précité.

- Indexation sur la légende de l'image

Dans le contenue textuel de l'image requête, nous constatons qu'il est composé de cinq termes décrivant l'image. En vérifiant le contenu textuel des images obtenues au cours de la requête : la première image contient le mot « *charpente* », la deuxième image contient les mots « *plafond et lames* » termes que l'ont retrouve dans l'image requête (Fig: 26). La classification établie par le robot s'explique par le fait que, l'image qui occupe la première place à son contexte qui contient le mot « charpente » et qu'il effectue une correspondance entre le poids qu'il est attribué sur le commentaire de l'image requête et celui de la première réponse. Or la troisième image contient-elle aussi deux termes important, mais il les considère au second plan du fait que ces dernier s on un poids faible par rapport à leur apparition dans leur contexte comparé à ces derniers.

Figure 26 : Résultat obtenu par l'indexation sur texte intégral sur la légende

Nous pourrions dire que dans ce cas précis la classification des réponses s'effectue sur le critère du poids qui est affecté à un terme par rapport à sa fréquence d'apparition dans le contexte. Les résultats obtenus montrent que 60% des images sont pertinentes contre 40% qui sont du bruit.

- Indexation sur la légende de l'image et au commentaire du projet

Figure 27 : Résultat obtenu par l'indexation sur texte intégral sur la légende et le commentaire du projet

Dans ce cas précis nous constatons que les résultats obtenus n'ont aucun point commun avec la requête, c'est-à-dire que les images obtenues ne représente pas la rubrique charpente. Comparé aux résultats précédents nous pouvons affirmer que cette méthode d'indexation est moins pertinente, du fait que 10% des images sont pertinentes contre 90% qui sont du bruit.

6.2.3- Résultats de l'indexation sur Wimexbot (utilisation d'un thésaurus)

WimexBot devant permettre la comparaison sur les modes d'indexation n'a pu se faire; dû à certains défauts d'ordre technique. Au vue du temps qui nous est impartie ce problème n'a pu être résolu. Mais par rapport au résultat obtenu au cours d'une requête, on observe des résultats non satisfaisant du fait que les images obtenues au cour d'une requête ne son pas associé à leur contexte réel. Ce qui mette en doute la crédibilité de notre expérience, à cela le robot pose un handicap majeur aux utilisateurs du fait que la requête s'effectuée en langage SQL (Structured Query Language est un langage de définition et de manipulation de données relationnelles normalisé ANSI) maîtrisé par les informaticiens .

6.2.4- Synthèse des méthodes d'indexations

L'analyse effectuée sur les résultats obtenus dans chacun des modes d'indexation montrent que, les résultats de la méthode d'indexation spécialisée sont plus proche de nos attentes. Du fait qu'il fournit en grande partie des images pertinentes, mais aussi l'on constate moins le phénomène de bruit (c'est-à-dire qu'au cour d'une requête nous aurons en plus des images pertinentes, celles qui ne répondent pas à la demande).

Ce qui n'est pas le cas de l'indexation sur le texte intégral. On pourra expliquer ce bruit par le fait que dans l'indexation à texte intégral, le robot indexe tous les mots contenus dans le texte. Par ce fait, il peut affecter à certain terme inadéquat (ne décrit pas forcement un concept relatif au contenu de l'image par exemple : abrite, couverte, etc.) une pondération importante. A cela elle présente un autre inconvénient : le phénomène du silence (c'est-à-dire au cour d'une requête une image n'apparaît pas du fait que son contenue textuelle ne décrie pas le contenue d'une image avec des termes adéquats). Or dans le cas de l'indexation spécialisé le simple fait d'indexer une image avec des termes précis (tiré d'un thésaurus) et de lui attribué un poids réduit assez bien le phénomène de « bruit et de silence ».

D'autre expérience fut réalisée en choisissant deux images requêtes. Cette méthode a permis d'obtenir des résultats assez satisfaisants du fait qu'elle se rapproche des résultats attendus. Mais cette démarche nécessite du temps, car la recherche s'obtient par plusieurs étape de recherche et qui du premier coup n'est pas pertinente.

Si nous devons classifier les modes d'indexation par ordre d'importance en tenant compte des résultats obtenus par chacun, il se présentera comme suit (Fig: 28) :

Méthode d'indexation	**Classement**	**Silence**	**Bruit**	**Observation**
Indexation manuelle spécialisée (thésaurus)	première	Moyen 40%	Faible 60%	Dû à la pondération des termes indexeur
Indexation automatique sur texte intégral (légendes)	deuxième	Fort 25%	Moyen 75%	Dû à un manque de pondération des termes clés, mais aussi du à une prise en compte de termes inadéquat
Indexation automatique sur texte intégral (légendes + commentaire projet)	troisième	Fort 50%	Fort 50%	Dû à un manque de pondération des termes clés,mais aussi à une surcharge d'informations non intéressantes

Figure 28 : Classement des différentes méthodes d'indexation

6.3- Comparaison des résultats

La vérification des résultats est une étude expérimentale ; qui consiste à prendre comme échantillon une dizaine de personnes. Celles-ci devront effectuer des requêtes sur les différentes méthodes d'indexations, les classer par ordre de pertinence et tirer des conclusions sur ces dernières. Le but de cette seconde expérimentation est de voir ; si les remarques et observations énumérées ci-dessus (Cf : 6.2.4) rejoignent celle de la seconde expérimentation.

6.3.1- Analyse des résultats

Les résultats obtenus montre que sur dix personnes (Cf annexe n°4) ayant effectués cette étude ; sept (70%) d'entre eux pense que l'indexation spécialisée est la plus pertinente puis vient l'indexation sur la légende et la dernière et l'indexation sur l'ensemble du projet. Les trois (30%) autres personnes pensent que l'indexation sur la légende est plus pertinente de celle spécialisée.

Par rapport au premier pronostique, nous constatons que les résultats obtenus se joignent à notre analyse. La classification de cette dernière s'est effectuée sur le fait qu'une indexation est plus pertinente qu'une autre, du fait qu'elle produit moins de bruit ou de silence(Fig :29), mais aussi par rapport à l'ordonnancement des images auquel l'utilisateur espère.

Causes	**Silence**	**Bruit**
Prise en compte d'un concept inadéquat		*
Non prise en compte d'un concept informatif	*	
Prise en compte d'un concept non informatif		*
Niveau de spécificité mal compris	*	*
Mauvaise traduction d'un concept	*	*

Figure 29 : Les causes des problèmes de bruit et de silence

Nous constatons que les causes de bruit et silence que nous avons relevé vient en partie de certaines termes de la rubrique Espace. Si au départ ces termes nous paraissaient important pour décrire le contenu d'une image, elles ont pour conséquence d'enduire en erreur les résultats au cour d'une requête face à cela, il nous semble important de ne pas les prendre en considération.

6.3.2- Propositions :

Etant en tendus de la faible représentation de notre corpus, qui ne dispose pas assez d'images représentatives dans notre base de donnée (108 images). Cela a joué en notre défaveur, du fait qu'au cours de notre requête on note la présence importante de « bruit et du silence » dans les

résultats obtenus dans les différentes méthodes d'indexation. La classification obtenue par rapport à la pertinence des différents modes d'indexation effectuée par l'utilisateur, peut être dû aussi à une faible représentation de notre échantillon, mais aussi du fait que nous nous sommes limités à une seule indexation experte. Par rapport à la deuxième perspective du classement (Cf 6.3.1) qui place l'indexation intégrale (sur le texte de la légende) au second rang n'est pas à négliger. Nous pensons qu'il existe une possibilité d'améliorer cette méthode.

- Proposition 1

Dans le cas de l'indexation effectuée uniquement sur la légende de l'image, nous envisageons supprimer à l'aide d'un filtre tous les termes qui ne se trouvent pas dans le thésaurus (relatif à la construction bois). Cette suppression permettra l'élimination des concepts inadéquats (c'est-à-dire des termes qui ne décrivent en rien un concept relavant du domaine de l'architecture, par exemple des termes tels que : abrite, trois, haut, niveau etc.).

Dans le cas de l'indexation effectuée à la fois sur la légende et le commentaire de projet, nous envisageons la même méthode cité ci-dessus. A cela on envisage d'attribuer des poids différents pour les termes provenant de la légende et du commentaire. Ceux qui proviennent de la légende aurons un poids plus important que ceux provenant du commentaire de projet. Nous faisons ce choix du fait que le commentaire de projet décrit l'ensemble d'un projet, mais ne décrit pas de manière spécifique le contenu d'une image.

- Proposition 2

Dans l'indexation manuelle spécialisée, nous envisageons ajouter un ou deux termes (terme utilisé au cours de l'indexation de l'image avec une forte pondération de 5) à chaque image ; pour permettre à l'utilisateur de mieux cibler les images qu'il retient au cours d'une requête. Cela peut avoir un avantage celui d'effectuer un choix judicieux, dans la mesure ou ces termes nous orientent dans le concept qui est le plus mis en valeur. Mais cela peut avoir un handicap : surcharger l'interface.

L'objectif de ces hypothèses est de voir comment nous pourrions réduire les effets du « bruit et du silence » observer dans les différentes méthodes d'indexation. Chacune de ces méthodes, ayant des avantages et des inconvénients, seule l'expérimentation peut nous départager dans la crédibilité de chacune d'elles.

6.3.3- Expérimentation de la proposition 1 :

L'expérimentation de l'hypothèse 1 consiste à supprimer du commentaire de projet et de la légende, tous les termes ne figurant pas dans le thésaurus. Cette étude à pour but de voir si le phénomène de bruit peut être enrayer ou atténué dans le cas de l'indexation sur texte intégral par l'analyse statistique. D'autre part nous avons effectué une autre analyse qui consiste à

supprimer les pondérations attribuées aux termes descripteurs ceci étant un moyen de voir si cette dernière joue un rôle important dans la réduction du phénomène de bruit (Cf : annexe 4). les résultats obtenus sont présentés comme suit ci-dessous :

N°	Types	Bruit	pertinent	Observations
1ère	**Indexation spécialisée avec pondération**	20%	80%	Nous constatons l'importance de la pondération au cours de l'indexation. Sans cette dernière le phénomène de *bruit* est considérable.
2ème	**Indexation spécialisée sans pondération**	40%	60%	
1ère	**Indexation automatique sur texte intégral (légende sans mots vides)**	40%	60%	Nous constatons que la 2ème méthode d'indexation sur texte intégral se rapproche plus de la 1ère méthode d'indexation spécialisée avec pondération
2ème	**Indexation automatique sur texte intégral (légende contenant uniquement les termes du thésaurus)**	30%	70%	
1ère	**Indexation automatique sur texte intégral (légende + commentaire projet, sans les mots vides)**	90%	10%	Nous constatons que le phénomène de *bruit* n'est réduit que de 10% ce qui vient confirmer que la surcharge d'information peut nuire à une bonne indexation
2ème	**Indexation automatique sur texte intégral (légende+commentaire projet contenant uniquement les termes du thésaurus)**	80%	20%	

Pour Les autres propositions ,nous n'avons pu les réaliser faute d'outils le permettant, mais aussi du temps qui nous est imparti. Ces propositions pourraient-être expérimentées ultérieurement dans la poursuite d'un travail de recherche beaucoup plus poussé.

CONCLUSION

L'objectif de ce projet était de comparer des méthodes d'indexations dans le cadre de la recherche d'image par le contenu sémantique. Nous pouvons affirmer qu'il est possible de retrouver une image pertinente que sur la base de certains termes contenus dans le contexte auquel elle est rattachée. Mais pour ce faire, il faudrait que le contexte auquel il est lié contient des concepts pertinents qui décrire avec clarté le contenu d'une image. Dans un autre temps la possibilité de retrouver une image pertinente dépend principalement du mode d'indexation. La méthode d'indexation choisie devra être capable d'extraire du contexte de l'image les termes clé permettant de retrouver les images les plus proche de l'image requête.

Au cours de notre étude nous avons constaté que l'indexation manuelle à l'aide d'un thésaurus donnait des résultats assez satisfaisant, du fait que les termes utilisés était fait sur la base d'un thésaurus et qu'à chacun des termes un poids leur a été attribué par rapport à leur pertinence, sur l'image. Il est vrai que l'indexation manuelle a donné des résultats satisfaisant, mais il n'en demeure pas moins que l'indexation sur la méthode statistique peut-être fiable à certain niveau. Cette dernière peut apporter des résultats satisfaisants ; si on supprime les termes dont les concepts sont inadéquats. Mais aussi s'il est possible d'affecter des poids aux termes jugés pertinents non pas en fonction de leur fréquence d'apparition dans le contexte auquel il appartient ; mais par rapport à leur importance à décrire le contenu de l'image.

ANNEXES

Annexe 1

mot	type	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	TOTAL
Accolée	ouvrage								1													1
acée	espace							1														1
acier	ouvrage		1		1	1						4						2				9
aggloméré	matière													1								1
aire	espace								1		1							2				4
Aire-jardin	espace										2											2
arcs	ouvrage																1					1
arêtiers	ouvrage									1												
armé	matière																		1			1
arrondis	ouvrage						2															2
articulation	ouvrage						1												1			2
aspect	ouvrage						2						1				1	1				5
assemblage	ouvrage						2										1			1		4
autoclave,	ouvrage						1															1
Autostable,	ouvrage			1																2		1
autrichienne.	pays											2										2
auvent	ouvrage				1																	1
avant-corps	ouvrage																1					1
baies	ouvrage								1													1
bains.	espace	1														1						2

Annexe 2

#Base de données wimexbot - Table dictionnaire sur le serveur localhost

 # phpMyAdmin SQL Dump

 # version 2.5.3

 # http://www.phpmyadmin.net

```
# Serveur: localhost
# GÈnÈrÈ le : Vendredi 16 Juillet 2004 ‡ 11:32
# Version du serveur: 3.23.58
# Version de PHP: 4.3.3
# Base de données: `wimexbot
#--------------------------------------------------------


# Structure de la table `dictionnaire
DROP TABLE IF EXISTS `dictionnaire`;
CREATE TABLE `dictionnaire` (
  `noMotDico` int(11) NOT NULL auto_increment,
  `motDico` varchar(100) NOT NULL default '',
  PRIMARY KEY  (`noMotDico`)
) TYPE=MyISAM AUTO_INCREMENT=490 ;
# Contenu de la table `dictionnaire`
INSERT INTO `dictionnaire` VALUES (1, 'AmÈnagement');
INSERT INTO `dictionnaire` VALUES (2, 'extÈrieur');
INSERT INTO `dictionnaire` VALUES (3, 'Construction');
INSERT INTO `dictionnaire` VALUES (4, 'extÈrieure');
INSERT INTO `dictionnaire` VALUES (5, 'abris');
INSERT INTO `dictionnaire` VALUES (6, 'de');
INSERT INTO `dictionnaire` VALUES (7, 'jardin');
INSERT INTO `dictionnaire` VALUES (8, 'vÈlo');
INSERT INTO `dictionnaire` VALUES (9, 'voiture');
INSERT INTO `dictionnaire` VALUES (10, 'Autre');
INSERT INTO `dictionnaire` VALUES (11, 'gloriette');
INSERT INTO `dictionnaire` VALUES (12, 'kiosque');
INSERT INTO `dictionnaire` VALUES (13, 'pergola');
INSERT INTO `dictionnaire` VALUES (14, 'prÈau');
INSERT INTO `dictionnaire` VALUES (15, 'serre');
INSERT INTO `dictionnaire` VALUES (16, 'terrasse')


Base de donnÈes wimexbot - Table mot_thesaurus sur le serveur localhost
# phpMyAdmin SQL Dump
# version 2.5.3
# http://www.phpmyadmin.net
# Serveur: localhost
# GÈnÈrÈ le : Vendredi 16 Juillet 2004 ‡ 11:30
# Version du serveur: 3.23.58
# Version de PHP: 4.3.3
# Base de donnÈes: `wimexbot`
```

```
# --------------------------------------------------------

# Structure de la table `mot_thesaurus`

DROP TABLE IF EXISTS `mot_thesaurus`;

CREATE TABLE `mot_thesaurus` (

  `noThesaurus` int(11) NOT NULL auto_increment,

  `texte` varchar(250) NOT NULL default '',

        `type` enum('FONCTIONS CONSTRUCTIVES','MATERIAUX','PRODUIT','TYPES DE BATIMENTS','NON DESCRIPTEUR') NOT NULL default 'FONCTIONS CONSTRUCTIVES',

  `noDocMat` int(11) NOT NULL default '0',

  PRIMARY KEY  (`noThesaurus`)

) TYPE=MyISAM AUTO_INCREMENT=607 ;

# Contenu de la table `mot_thesaurus`

INSERT INTO `mot_thesaurus` VALUES (71, 'AmÈnagement extÈrieur', 'FONCTIONS CONSTRUCTIVES', 1);

INSERT INTO `mot_thesaurus` VALUES (302, 'Construction extÈrieure', 'FONCTIONS CONSTRUCTIVES', 2);

INSERT INTO `mot_thesaurus` VALUES (303, 'abris de jardin', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (304, 'abris de vÈlo', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (305, 'abris de voiture', 'FONCTIONS CONSTRUCTIVES',3);

INSERT INTO `mot_thesaurus` VALUES (310, 'Autre construction extÈrieure', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (306, 'gloriette', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (307, 'kiosque', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (308, 'pergola', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (534, 'prÈau', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (309, 'serre', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (527, 'terrasse', 'FONCTIONS CONSTRUCTIVES', 3);

INSERT INTO `mot_thesaurus` VALUES (311, 'Equipement urbain', 'FONCTIONS CONSTRUCTIVES', 2);

, 3)



# phpMyAdmin SQL Dump

# version 2.5.3

# http://www.phpmyadmin.net

# Serveur: localhost

# GÈnÈrÈ le : Vendredi 16 Juillet 2004 ‡ 11:35

# Version du serveur: 3.23.58

# Version de PHP: 4.3.3

# Base de donnÈes: `wimexbot`

# --------------------------------------------------------
```

```
# Structure de la table `se_trouve_dans`
DROP TABLE IF EXISTS `se_trouve_dans`;
CREATE TABLE `se_trouve_dans` (
  `noTheme` int(11) NOT NULL default '0',
  `noMotDico` int(11) NOT NULL default '0',
  PRIMARY KEY  (`noMotDico`,`noTheme`)
) TYPE=MyISAM;
# Contenu de la table `se_trouve_dans`
INSERT INTO `se_trouve_dans` VALUES (48, 1);
INSERT INTO `se_trouve_dans` VALUES (49, 1);
INSERT INTO `se_trouve_dans` VALUES (71, 1);
INSERT INTO `se_trouve_dans` VALUES (71, 2);
INSERT INTO `se_trouve_dans` VALUES (322, 2);
INSERT INTO `se_trouve_dans` VALUES (328, 2);
INSERT INTO `se_trouve_dans` VALUES (329, 2);
INSERT INTO `se_trouve_dans` VALUES (334, 2);
INSERT INTO `se_trouve_dans` VALUES (335, 2);
INSERT INTO `se_trouve_dans` VALUES (344, 2);
INSERT INTO `se_trouve_dans` VALUES (302, 3);
INSERT INTO `se_trouve_dans` VALUES (310, 3);
INSERT INTO `se_trouve_dans` VALUES (65, 4);
INSERT INTO `se_trouve_dans` VALUES (66, 4);
INSERT INTO `se_trouve_dans` VALUES (176, 4);
INSERT INTO `se_trouve_dans` VALUES (302, 4);
INSERT INTO `se_trouve_dans` VALUES (310, 4);
INSERT INTO `se_trouve_dans` VALUES (496, 4);
INSERT INTO `se_trouve_dans` VALUES (303, 5);
INSERT INTO `se_trouve_dans` VALUES (304, 5);
INSERT INTO `se_trouve_dans` VALUES (305, 5);
```

Annexe 3

1- Conception de l'application

1.1- Page d'accueil

La page d'accueil du site. Elle permet à l'utilisateur de :

- consulter la liste des sites entièrement ou partiellement analysés, trouvés
- ou prêts à être analysés,
- choisir un site à analyser,

- saisir l'URL d'un site à analyser,
- pouvoir supprimer un site apparaissant dans la liste des sites et les informations le concernant,
- lancer l'exécution des trois étapes de recherche de pages et d'images, de recherche de contextes des images et d'indexation des images,
- visualiser les résultats obtenus après exécution des trois étapes,
- configurer des paramètres concernant l'image ou les sites, les pages et les mots interdits.

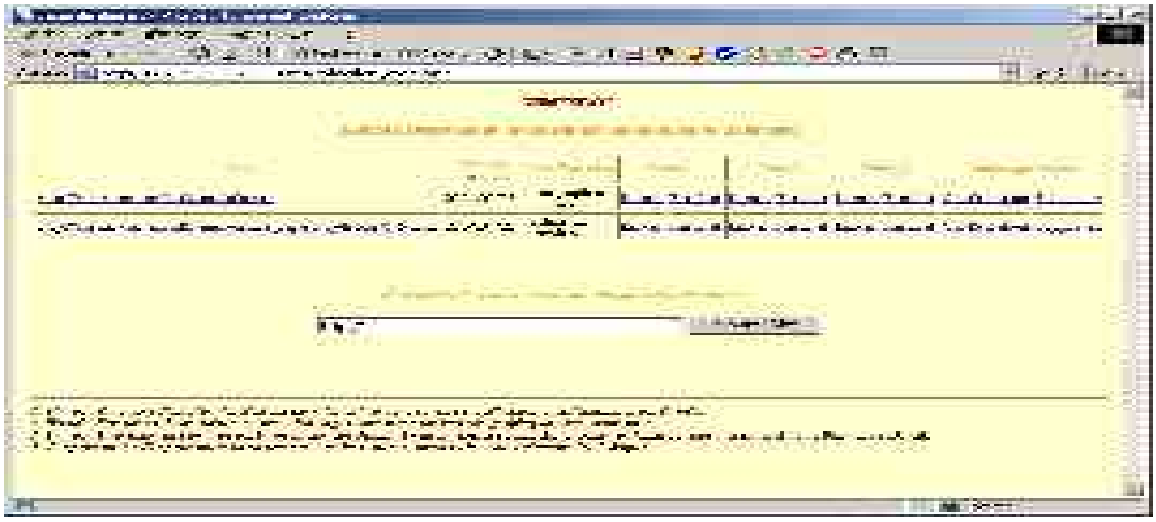


Figure 1 : Page d'accueil

Dans la figure ci-dessus, apparaît le site dans lequel se trouvent les documents que nous voulons indexer. Le robot indique, la dernière fois qu'il a été parcouru, c'était lors de l'étape 1 (recherche des pages et des images) qui a été terminée avec succès (état parcours = « Documents trouvés »).

Les sites qui apparaissent dans la liste de la page « Depart.jsp » sont enregistrés dans la table "site" de la base de données (Fig: 20).

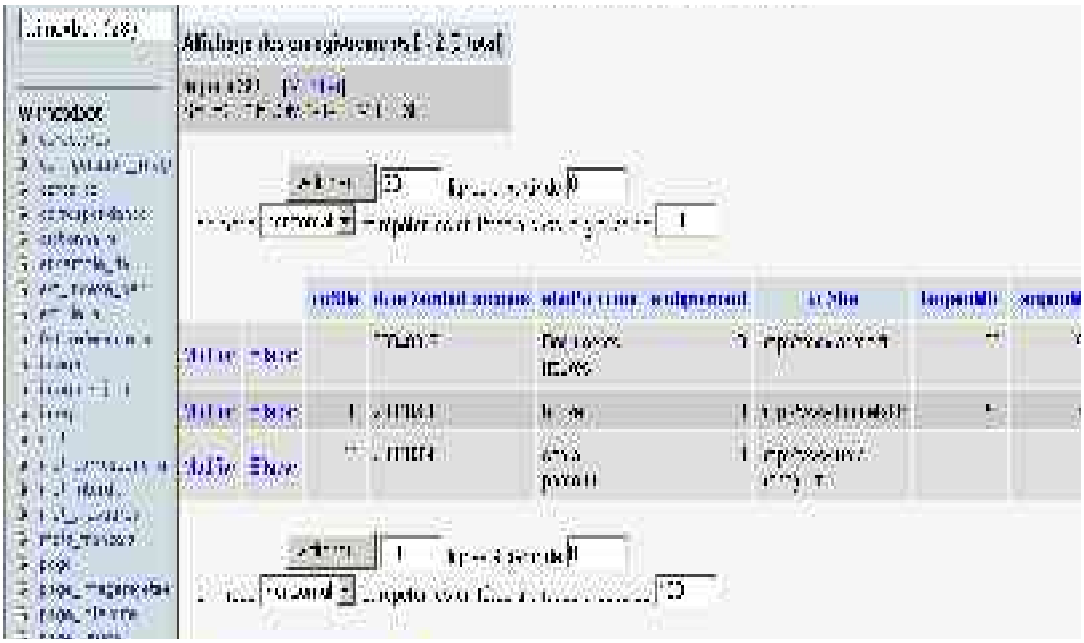


Figure 2 : Table « Site »

- **Page « ConfigImage.jsp »**

Cette page JSP(*) cette page permet d'établire certains paramètres pour un site choisi par l'utilisateur. Elle permet de :

- La configuration de paramètres des images (Fig: 21) qui seront considérés lors la recherche et de la sélection des images graphiquement pertinentes d'un site. Les paramètres peuvent être personnalisés au gré de l'utilisateur
- Ajouter ou retirer des URL de sites ou pages interdits ou des mots interdits ((Fig: 22) Ces éléments sont utilisés dans le but de restreindre le parcours d'un site lors de la recherche des pages ou des images.
- Relancer l'exécution de chacune des étapes après un certain nombre de jours. Nombre de jour c'est le paramètre que l'utilisateur pourra définir.

(*)Java Server Pages est un standard permettant de développer des applications Web interactives.

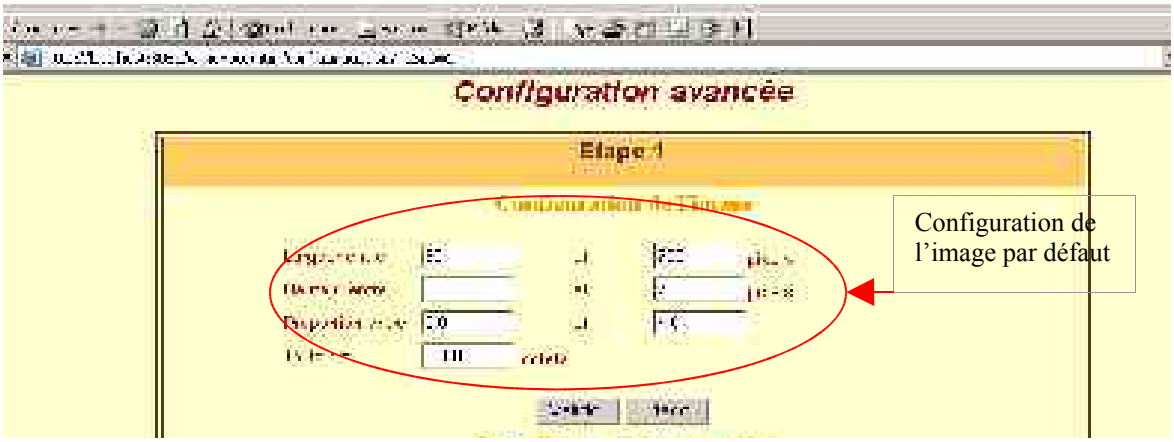


Figure 3 : Configuration d'image par défaut

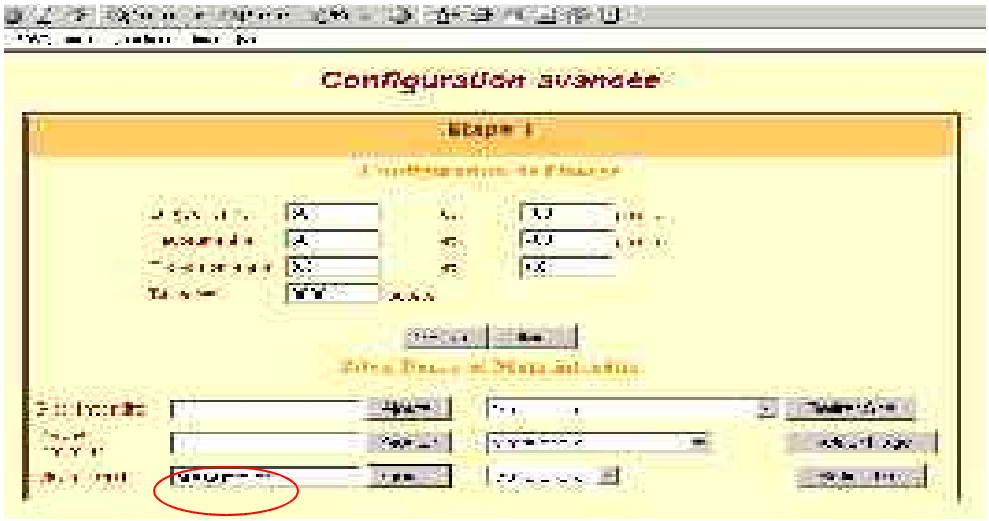


Figure 4 : Ajout d'un mot interdit

La figure ci-dessus montre l'ajout d'un mot interdit à la liste des mots interdits d'un site. Lors du parcours du site les URL de pages contenant ce mot seront rejetées.

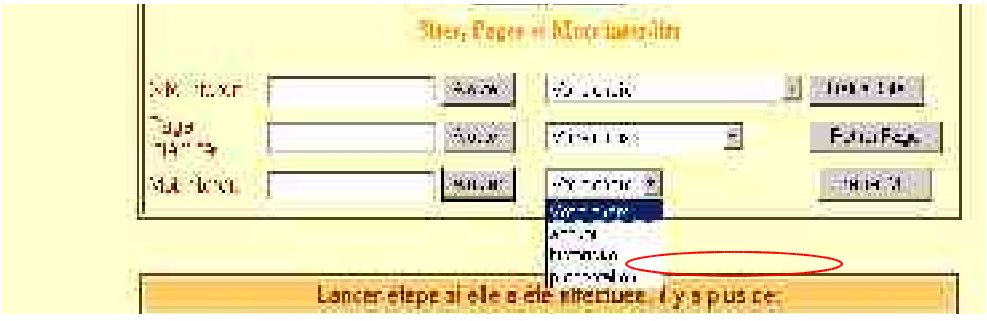


Figure 5 : Liste des mots interdits

1.2- Première étape

- **Page « Etape1.jsp »**

C'est la page qui permet de tester si la première étape correspondant à la recherche de pages et d'images peut être exécutée pour le site choisi par l'utilisateur. Dans le cas où cette étape aurait été exécutée récemment, la page représentée par la figure suivante apparaît.

- **Page « Resultat1.jsp »**

Cette page permet de visualiser les résultats obtenus lors de l'exécution de la première étape. L'utilisateur peut choisir de visualiser les images pertinentes (Fig: 24 & 25) ou les images rejetées. L'ensemble des images sera affiché avec pour chacune ses caractéristiques graphiques et la liste des pages où l'image a été trouvée. Pour les images rejetées, nous affichons en plus la cause du rejet. Les informations nécessaires pour cette visualisation sont stockées dans les tables « image », « image_rejete » et « page » de la base de données.

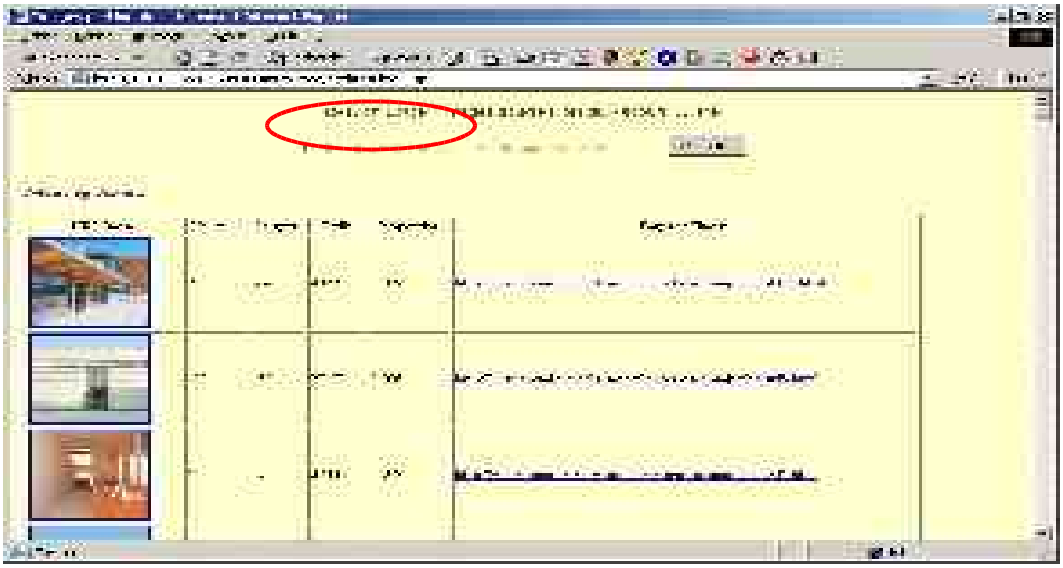


Figure 6 : Images pertinentes

La première image trouvée ci dessus, a été pris dans le page du site http://mistral.crai.archi.fr/mariefrance/projetsbois/d11.html.

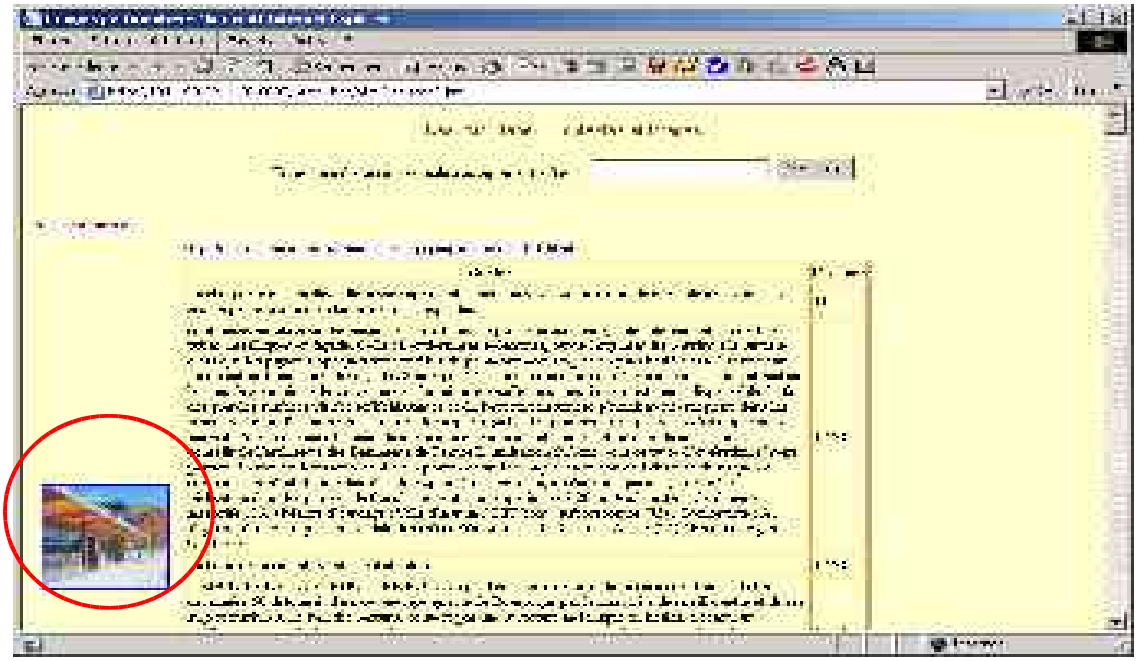


Figure 7 : Page contenant des images pertinentes

Les images dans la figure suivante ont été rejetées (Fig: 26)car leurs caractéristiques graphiques ne correspondent pas aux valeurs configurées au début. Ainsi, la première image a été rejetée à cause de sa larguer (702>700)

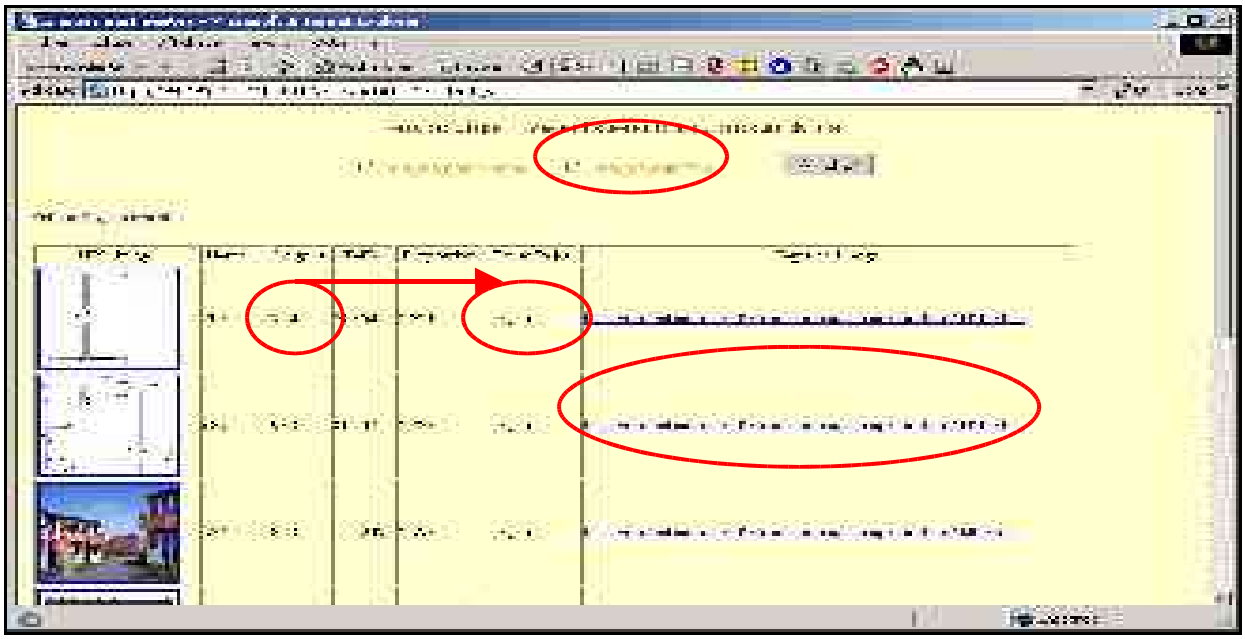


Figure 8 : Images rejetées

1.3- Deuxième étape

- **Page « Etape2.jsp »**

Dans cette page, nous testons si la deuxième étape de recherche des contextes, pour les images, peut avoir lieu. Si c'est le cas, l'exécution de cette étape est lancée, sinon, l'utilisateur pourra consulter les résultats de cette étape.

Après exécution de cette étape, les contextes de chaque page contenants des images graphiquement pertinentes sont enregistrés dans la table « contexte » (Fig: 27) de la base de données.

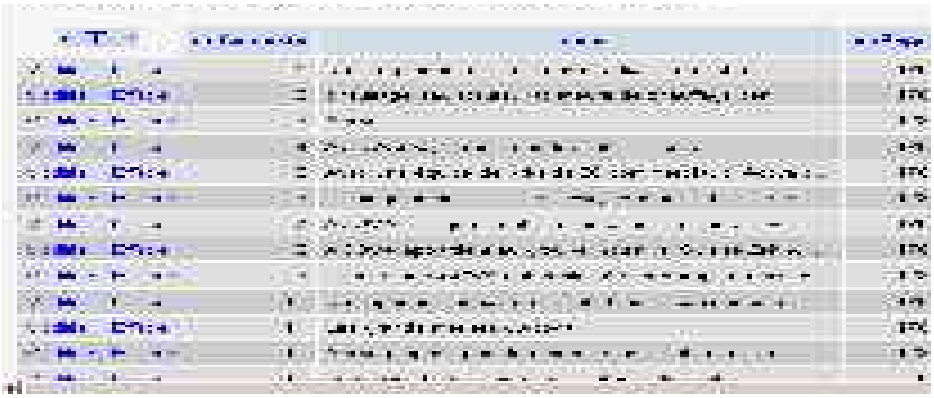


Figure 9 : Table « contexte »

- **Page « Resultat2.jsp »**

Cette page permet d'afficher, pour chaque image graphiquement pertinente, l'ensemble des contextes se trouvant dans les pages où l'image a été trouvée(Fig: 28). Elle affiche également la distance entre une image et un contexte.

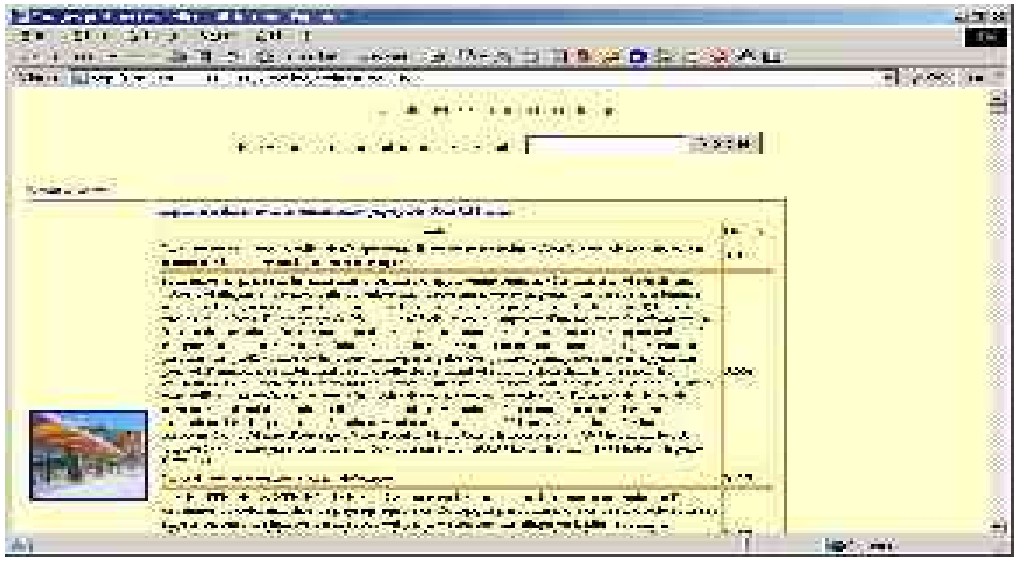


Figure 10 : Les contextes d'une image et choix d'une distance

L'utilisateur peut restreindre l'ensemble des contextes qu'il souhaite afficher. Pour cela, il doit préciser une distance maximale entre 0 et 1 (les valeurs des distances varient entre 0 et 1). Dans cette même figure, l'utilisateur a définit une distance maximale de 0.5.

1.4- Troisième étape

- **Page « Etape3.jsp »**

C'est la page JSP permettant l'exécution de la troisième étape, correspondant à l'indexation des images, qui pourraient être lancée. Si c'est le cas, l'application commence l'analyse par des thèmes du thésaurus. Ces thèmes ont été trouvés dans les contextes proches des images.

- **Page « Resultat3.jsp »**

Cette page permet d'afficher les résultats de la troisième étape de l'analyse d'un site. Elle est constituée de deux frames. Le frame se trouvant à gauche (« liste_images.jsp ») se charge d'afficher la liste des images pertinentes qui ont été extraites lors de l'analyse d'un site. Nous pouvons choisir une des images. Les informations se rattachant à cette image seront affichées dans le frame à droite (« image_page.jsp »). Nous retrouvons la figure de l'image, ses caractéristiques graphiques, ainsi que les pages où elle a été trouvée. De plus, nous pouvons consulter une liste de thèmes qui indexent l'image et qui sont censés la décrire. Cette liste représente le résultat fourni après exécution de la troisième étape de l'application.

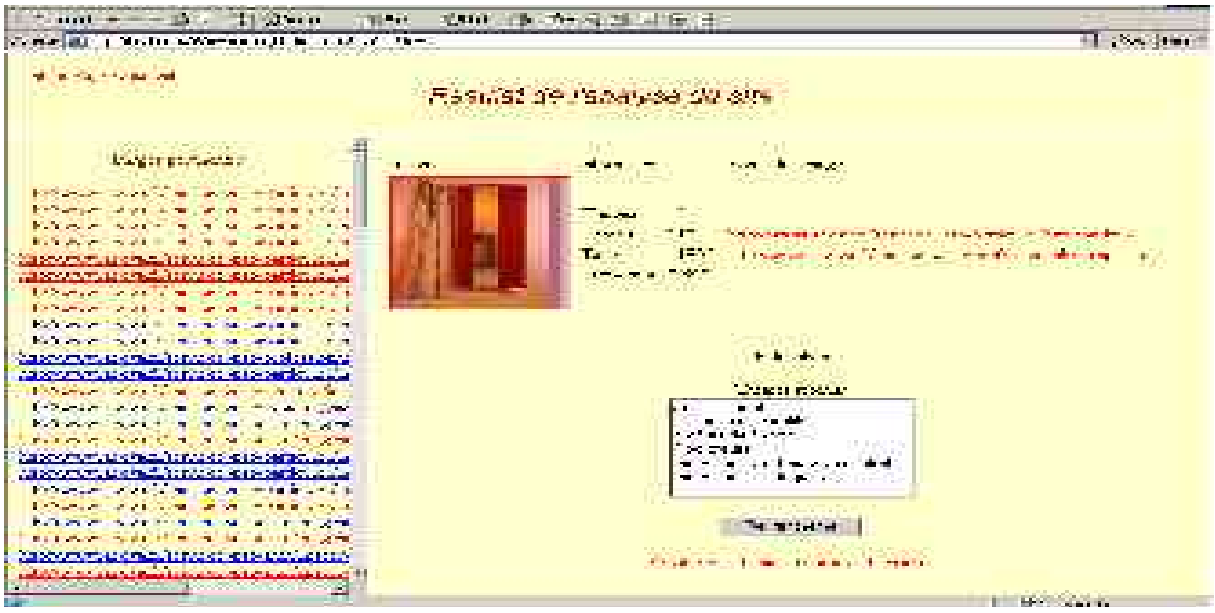
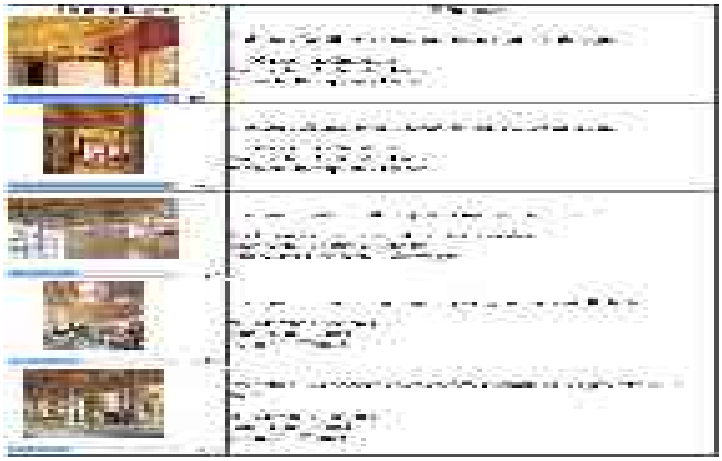


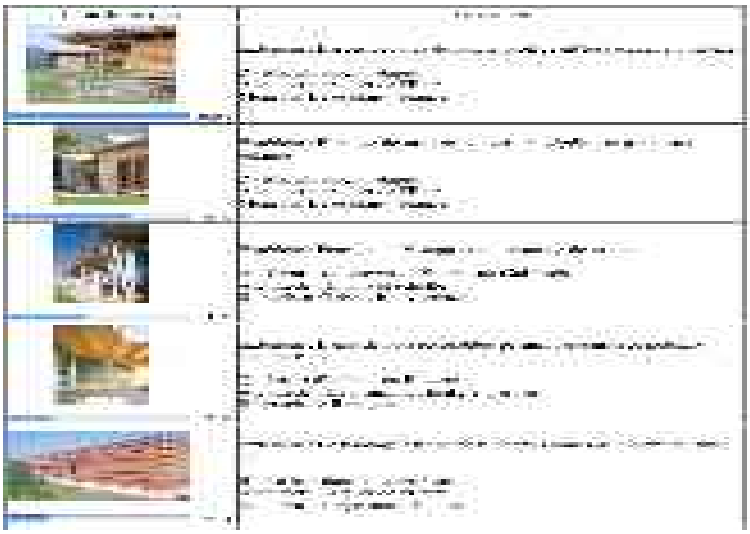
Figure 11 : Indexation de l'image

Annexe 4

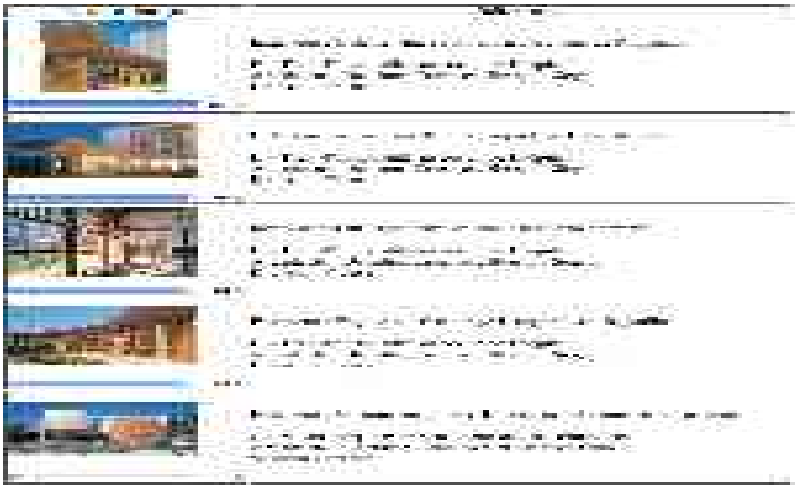
1 erè place : Résultats obtenus par le thesaurus



2 ème place : Résultats obtenus par la légende



3 ème place : Résultats obtenus par la légende + texte



Bibliographie sur Internet

Catherine BERRUT, *Modèles de documents :images fixes*, Laboratoire du MRIM de Grenoble, http://www-mrim.imag.fr/presentation/images.php

J.C BIGNON, *Assistance à la conception architecturale et technique par des méthodes et outils de recherché d'informations par l'image*; CRAI, 2001 ; http:/www.crai.archi.fr

J.C BIGNON ; *Extraction et indexation d'images appliquées au domaine de la conception architecturale et technique* ; CRAI, 2001,http: //www.crai.archi.fr

Matthieu CORD, *Interprétation d'image et indexation par le contenu* ; 1999 ;http://www.etis.ensea.fr/image/Activites/Equipe_Image_Indexation.html

Matthieu CORD, Christophe AMBROIS, *Techniques d'apprentissage pour l'indexation et la recherche d'image par le contenu* ; 2003 ; http://gdr-isis.org/rillk/bib/archive/00000586/

Pierre CHAPERON, *Indexation des images en mouvement : un tour d'horizon* ; 2001 ; Périodique électronique étudiant de l'Ecole de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal ; http://www.ebi.umontreal.ca/curcus/vol6no1/chaperon.html

Walaiporn NAKAPAN ;*Recherche d'informations techniques par l'image* ; CRAI, 2000 ; http://www.crai.archi.fr

C.BOUDRY ; *Principe de la recherche d'images sur Internet* ; URFIST de Paris/Ecole des Chartres, 2002, http://www.ccr.jussieu.fr/urfist/image_numerique/home_image.htm

KACHER.C ; *Méthode d'indexation appliquée à la recherche d'ouvrages architecturaux par l'image* ; CRAI 2003 ; http://www.crai.archi.fr

Sabrina TOLLARI ; *Mise en relation et fusion d'indices textuels et visuels pour une recherche d'images par le contenu* ; Laboratoire SIS-Equipe informatique, Université de Toulon et du Var, 2003, http: //sis.univ-tln.fr/tollari/ARTICLES/BDA2003/node1.html

Sabrina TOLLARI ,Hervé GLOTIN ,Jacques LE MAITRE ,*Mise en relation et fusion d'indices textuels et visuels pour une recherche d'images par le contenu*, Laboratoire SIS-Equipe informatique, Université de Toulon et du Var,2003, http://sis.univ-tln.fr

James M.TUENR ; *Le traitement de l'image en environnement numérique* ; Ecole de bibliothéconomie et des sciences de l'information, Université de Montréal, 1996,
http: //www.mapageweb.umontreal.ca/turner/francais/textes/asted96.htm

Bibliographie

[AMA 1998] AMAR Muriel, *Fondements théoriques de l'indexation*, Thèse de l'Université Louis -lumière, Lyon 2, Lyon 1998

[AOU 2003] AOUED.S, *Proposition d'une structure amont d'aide à l'indexation d'images*. Mémoire de DEA. Université Henri Poincaré, Nancy1, octobre 2003

[ART et GIL 1987] AITCHISON.J, GILCHRIST.A ,*Construire un thésaurus*, Edition ADBS, 1987

[BOU 1992] BOUDON, P, *Introduction à l'architecturologie*. Dunod, Paris, 1992

[CAN 2004] CANE Mirela, *Réalisation d'une interface pour un robot extracteur et* indexation d'images provenant de sites Internet du bâtiment. Mémoire de DESS S.I.D, UFR mathématique et informatique, université de Nancy 2,2004

[CHAL & VER 2000] CHALENET.M, VERDIER.H, *Thesaurus de l'architecture* : *document et méthodes, n° 7*, Edition du Patrimoine, 2000

[DEG & MEN 01] Danièle DEGEZ , Dominique MENILLET,*Thesaurolossaire des langages documentaires : un outils de contrôle sémantique*, éditions ADBS,Paris 2001

[DUP & ERM 2000] Gérard DUPOIRIER & Jean-louis ERMINE, *Gestion des documents et gestion des connaissances : document numérique volume 3 n°3-4*, Hermès science publication, Paris 2000

[GAUS & StEF 2003] Eric GAUSSIER &Marie-Hélène STEFANINI, *Assistance intelligente à la recherche d'information*, Hermès science et publication Lavoisier, Paris 2003

[HAL 1989] HALIN, G, *Apprentissage pour la recherche interactive et progressive d'images : processus EXPRIM et prototype RIVAGE*, Thèse, spécialité Informatique, l'Université de Nancy 2, 1989

[JOL 2001] Jean-Michel JOLION, *L'indexation : document numérique, volume 4-n°1-2*, Hermès science publication, Paris 2001

[PRO 1992] PROST, R. *Conception architecturale : une investigation méthodologique*, Edition l'Harmattan, Paris 1992

[NAK 2003] NAKAPAN, W, *Recherche d'information par l'image : application à la recherche interactive des produits du bâtiment*. Thèse de l'Institue National Polytechnique de Lorraine, discipline : Science pour l'architecture, Nancy 2003

[LEF 2000] Philippe LEFEVRE, *La recherche d'information : du texte intégral au thésaurus*, hermès science publication, Paris 2000

[VIG 1990] VIGAN.J. DICO BAT ; *Dictionnaire général du bâtiment*. Édition Arcature, 1990

[ZER 2001] N.ZERROUK, *Analyse et expérimentation d'un robot d'indexation et de recherche d'images WimexBot*. Mémoire de DEA. Université Henri Poincaré, Nancy1, octobre 2001

[ZID 2002] ZIDANE, C. *Conception assistée par l'image appliquée à la reconnaissance des textures de matériaux*. Mémoire de DEA. Université Henri Poincaré, Nancy1, octobre 2002