Ressources lexicales au service de recherche et d'indexation des images

Inga Gheorghita[1,2]

(1) ATILF-CNRS, Nancy-Université (UMR 7118), France
(2) XILOPIX, 37 rue de la Plaine, 75020 Paris, France
inga.gheorghita@atilf.fr

Cet article présente une méthodologie d'utilisation du Trésor de la Langue Française informatisée (TLFi) pour l'indexation et la recherche des images fondée sur l'annotation textuelle. Nous utilisons les définitions du TLFi pour la création automatique et l'enrichissement d'un thésaurus à partir des mots-clés de la requête de recherche et des mots-clés attribués à l'image lors de l'indexation. Plus précisement il s'agit d'associer, de façon automatisé, à chaque mot-clé de l'image une liste des mots extraits de ses définitions TLFi pour un domaine donné, en construisant ainsi un arbre hiérarchique. L'approche proposée permet une catégorisation très précise des images, selon les domaines, une indexation de grandes quantités d'images et une recherche rapide.

This article presents a methodology for using the "Trésor de la Langue Française informatisée" (TLFi) for indexing and searching images based on textual annotation. We use the definitions of TLFi for automatic creation and enrichment of a thesaurus based on keywords from the search query and the keywords assigned to the image during indexing. More specifically it is automatically to associate, to each keyword of the image a list of words from their TLFi's definitions for a given area, thus building a hierarchical tree. The proposed approach allows a very accurate categorization of images, depending on the fields, a indexing of large amounts of images and a quick search.

Mots-clés : TLFi, indexation, recherche, images, thésaurus
Keywords: TLFi, indexing, search, images, thesaurus

1 Introduction

Avec l'arrivée de l'Internet le marché de l'image numérique a progressé de manière exponentielle. L'offre d'illustration n'a jamais été aussi grande. En tenant compte des faits qu'une agence de photo gère habituellement entre un et vingt millions d'images, qu'un satellite météo envoie plusieurs giga-octets de donnés chaque jour, qu'un possesseur d'appareil photo numérique actif prendra de l'ordre de cents mille photos en trente ans (Gross, 2007), nous voyons bien que le problème d'accès et de recherche dans cette masse d'information énorme pose de nouveaux défis. La grande quantité des images provoque un très grand désordre et l'extraordinaire confusion qui en résultent concernant leur identification. La cause principale est le fait que les principes d'annotation des images par mots clés restent souvent très approximatifs et ne garantissent ni la concordance, ni la pertinence des images.

Pour gérer et utiliser efficacement ces bases d'images, un système d'indexation et de recherche d'images est nécessaire. C'est pour cette raison que la recherche d'images est devenue un sujet très actif dans la communauté internationale depuis plus d'une dizaine d'années. Dans nos recherches nous nous focalisons sur l'élaboration d'un moteur de recherche et d'indexation d'images utilisant les ressources lexicales de l'ATILF, plus particulièrement le Trésor de la Langue Française Informatisé (TLFi : www.atilf.fr/tlfi). Nous articulons notre analyse autour de l'insensibilité des systèmes de recherche à la sémantique en nous appuyant sur une recherche des images basée sur l'exploitation du texte brut associé aux images. Notre objectif, afin de pouvoir réaliser l'indexation et la recherche des images, est de trouver une intersection suffisante entre le vocabulaire utilisé pour la description d'images par son auteur et la définition hiérarchique des descripteurs dans le thésaurus de la base de données images (XILOPIX : www.xilopix.com).

2 Généralité et problématique

L'indexation des images a comme but de faciliter l'organisation d'information selon certains critères comme la date de prise de photo, le lieu, la marque de l'appareil photo numérique etc. L'objectif de la recherche est, par contre, de mieux répondre aux besoins d'information de l'utilisateur, c'est-à-dire de retrouver l'information pertinente.

Il existe deux approches concernant l'indexation et la recherche d'images : soit *par le contenu visuel,* soit par *le contenu textuel* de leur description. L'indexation par le contenu visuel se réalise en utilisant les caractéristiques symboliques de l'image comme les histogrammes, les formes, les textures, les couleurs. Ce type d'indexation est utilisé dans les domaines spécifiques (médicine, astronomie, design, publicité) où la modélisation de l'image est possible. Par exemple, à partir des images médicales des patients on peut construire une modélisation informatique de ses organes et de leurs pathologies.

La recherche des images par le contenu visuel reste encore à l'étape de travaux de recherche et d'élaboration de prototypes, tandis qu'aujourd'hui les moteurs de recherche d'images existants sont basés sur la recherche par mots clés. Les procédures d'accès aux données photographiques peuvent alors être schématisées ainsi : l'utilisateur formule une requête composée de termes langagiers et le système lui propose en réponse des images qu'il considère comme proches des termes de sa requête. Pour ce faire, le système, le plus souvent à l'aide de calculs statistiques, détermine la similarité entre les termes de requête et les termes associés aux images.

Toutefois les images proposées à l'utilisateur ne semblent pas toujours correspondre à sa requête de recherche. Ces écarts sont liés essentiellement à l'opacité des systèmes de recherches par rapport à la signification sémantique : aucune analyse du contenu de la requête pour déterminer sa signification n'est pas réalisée. Une autre problématique de l'indexation textuelle des images est le fait que la description textuelle de l'image est souvent trop courte pour décrire son contenu. Il y a aussi des aspects subjectifs du contenu d'une image, qui dépendent du domaine de connaissances et de la perception de celui qui la regarde, et qui déterminent la diversité de la description d'une image.

Afin de rendre à l'utilisateur des résultats répondant mieux à sa requête, il convient d'utiliser des ressources sémantiques[1].

3 Le TLFi

Le Trésor de la langue française (TLF) est le plus grand dictionnaire de langue française rédigé en 16 volumes par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS). Son apparition en 2000 sur l'Internet a été un vrai succès, 30 000 à 40 000 demandes quotidiennes provenant de tout le monde (Dendien & Pierrel, 2003). Il est utilisé dans le milieu d'enseignement (les écoles, les lycées, les universités etc.) avec le même engouement que dans le milieu de recherche. Ainsi le TLFi devient la base

[1] Les ressources sémantiques sont représentées sous forme de bases de connaissances structurées comme les taxonomies, les thésaurus, les ontologies.

de plusieurs projets de recherche comme par exemple le projet Definiens (Barque & Polguère, 2009) ou le projet RELIEF[2].

Quelles sont alors les facteurs qui déterminent l'utilisation du TLFi dans ces projets ? Tout d'abord TLFi est une ressource normée en XML qui est accessible à la toute communauté scientifique. C'est aussi une source de données lexicales très riche. On y retrouve toute l'information synchronique et diachronique d'un mot. Sa structure assez normalisée et structurée permet une extraction des connaissances, par exemple pour le domaine de Traitement Automatique des Langues. Le TLFi présent aussi un grand intérêt dans la structure de ses définitions lexicographiques. Les définitions du TLFi sont des définitions logiques (hyperonymiques) constituées d'*une classe* ou *genre prochain* à laquelle appartient le mot défini et des *propriétés* qui le particularisent à l'intérieur de cette classe.

Si les deux projets précédemment cités visent l'utilisation du TLFi pour la création des nouvelles ressources linguistiques, nous nous proposons ici comme but l'utilisation du TLFi pour la recherche et l'indexation des images fondée sur l'annotation textuelle. Nous utilisons les définitions du TLFi pour la création automatique et l'enrichissement d'un thésaurus à partir des mots-clés de la requête de recherche et des mots-clés attribués à l'image lors l'indexation. Plus précisement il s'agit d'associer, de façon automatisé, à chaque mot-clé de l'image une liste des mots extraits de ses définitions TLFi pour un domaine donné, en construisant ainsi un arbre hiérarchique.

4 Représentation et analyse de corpus de travail

Le corpus de travail que nous utilisons est constitué des données de la ressource lexicale SEMEME, construite à partir du TLFi dans le cadre du projet DIXEME. SEMEME contient 78 476 fichiers XML pour 93 697 entrées. L'intérêt que nous portons pour SEMEME est qu'il contient les définitions TLFi lemmatisées[3] et filtrées, en ne gardant que les mots à sémantisme pleine (substantifs, verbes, adjectives, adverbes). A chaque définition du TLFi sont aussi attribuées des statistiques d'occurrences de chaque lemme, le domaine d'emploi et des informations de contraintes structurelles liés au lexème. Par rapport au TLFi initiale, dans SEMEME ne sont pas gardées les informations relatives à l'organisation hiérarchique (marques de plan, marques de niveau hiérarchique), l'information lexicographique sur l'indicateur d'emploi, les synonymes et les antonymes. L'information sur les données de corpus de travail est présentée dans le tableau 1 ci-dessous :

	Total	Vedette	Syntagme
Lemmes différents	38 617	35 468 (91.84%)	22 779 (58.98%)
Définitions	265 475	206 052 (77.61%)	59 423 (22.38%)

Tableau 1 : L'information sur les données de corpus de travail

Pour pouvoir faire une analyse profonde de notre corpus de travail, toutes les données contenues dans les fichiers XML ont été mises dans une base de données.

[2] Ressource Lexicale Informatisée d'Envergure. Il s'agit d'un nouveau projet lexicographe qui vise la construction d'une nouvelle ressource lexicale du français à large couverture, appelée le *Réseau Lexical du Français (RLF),* à partir de TLFi.

[3] Les définitions ont été lemmatisées et annotées avec un outil réalisé par l'ATILF. La lemmatisation n'a pas été sans erreurs, le fait qui doit être pris en compte lors l'exploitation des résultats.

4.1 La structure des définitions TLFi

Les définitions d'un dictionnaire sont rédigées selon certaines règles qui déterminent le type et la structure des définitions. En lexicographie il y a plusieurs types de définitions : logique (hyperonymique), par équivalence synonymique, morphosémantique, méronymique, par approximation (Touratier, 2000).
En général, les définitions du TLFi sont construites selon schéma suivante « classificateur+spécifications » où le classificateur représente la classe à la quelle appartient le mot défini et les spécifications désignent les caractéristiques spécifiques du mot au sein de cette classe. D'habitude le classificateur représente la classe à laquelle appartient le mot défini et il varie en fonction du domaine de définition. Par exemple le mot « crime » dans le domaine juridique a comme classificateur «infraction » et dans une définition par hyperbole son classificateur est « action ». Donc suivant le domaine, le mot « crime » peut être groupé dans deux classes « infraction » et « action ».
Les définitions nominales du TLFi suivent le mieux la structure d'une définition hyperonymique. Nous avons remarqué que, dans ces définitions, le classificateur correspond dans la plupart des cas au premier substantif de la définition, en indiquant ainsi le concept le plus voisin du mot à définir.

4.2 Les connaissances lexicales pour le regroupement des définitions TLFi

Pour une entrée du TLFi, les définitions sont structurées selon une hiérarchie de niveaux. Cette structuration n'est pas toujours similaire. Pour certaines entrées, les définitions précédées des indicateurs d'emploi, des domaines, apparaissent aux niveaux plus hauts (I, II) de la hiérarchie, mais pour d'autres entrées les définitions apparaissent aux niveaux inférieurs. Donc le fait de considérer la définition de haut niveau comme la plus appropriée pour un mot de l'entrée du TLFi n'est pas tout à fait juste.
Les indicateurs d'emploi, les domaines, les informations mises entre crochets semblent être plus pertinentes pour la détermination des définitions appropriées pour un mot.
Dans les fichiers XML de SEMEME, les domaines sont les seules connaissances lexicales qui permettent d'associer à un mot une définition particulière. Cela nous permet de regrouper les définitions d'une entrée du TLFi selon les domaines.
Par contre seulement 30.87% définitions vedettes du TLFi ont des domaines. Les définitions sans domaines sont groupées dans nouveau domaine nommé « generic ».

4.3 Les expressions métalinguistiques des définitions

Chaque type de définition (nominale, verbale, adjectivale, adverbiale) est lié à des expressions métalinguistiques caractéristiques de son type. Dans les définitions nominales on rencontre souvent les expressions métalinguistiques de type : « action de », « manière de », « fait de », « partie de », « manque de » etc. Pour les définitions verbales sont caractéristiques des locutions verbales et des locutions comprenant les verbes fonctionnels tels « faire » ou « laisser ». Les définitions adjectivales, quant à elles, ont plus souvent la structure « Qui+être+Adj.», « Qui+verbe », « propre à », « relatif à », « se dit de » etc.
Dans nos recherches nous avons porté plus d'attention aux définitions nominales car lors de la construction du thésaurus seuls seront utilisés seulement les substantifs.
Si la classe des événements ou la classe des transports est définissable, la classe des actions, des manières est plus difficile à définir. Les hyperonymes de type « action, manière, caractère etc. » sont trop génériques pour représenter des concepts. C'est pour cette raison que nous avons créé une classe, nommée la classe M, qui contient toutes les expressions métalinguistiques trouvées dans les définitions du TLFi. Les mots qui appartiennent à cette classe ne sont pas utilisés. Toutefois pour pouvoir les garder[4], ils sont ajoutés de manière suivante au substantif qui les précède :

[4] Nous considérons important de garder les expressions métalinguistiques afin de pouvoir spécifier le mot qui les précède. Par exemple le mot « continent » ne peut pas être considéré comme « monde » quand en réalité il représente une partie de ce monde.

Classe M + Subst. => Subst. (Classe M)

Dans les définitions du TLFi sont utilisées aussi des expressions métalinguistiques d'oppositions de type « sans, pas avec, non, ni, par opposition au ». Pour ce type des expressions nous avons créé une autre classe N. Les substantifs situés dans la définition après les mots qui appartiennent à cette classe, sont éliminés :

Classe N + Subst. => ~~Subst.~~

5 Pondération des mots dans une définition TLFi

Dans une définition les mots ont des statuts différents et n'apportent pas la même quantité d'information. Afin de pouvoir attribuer à un mot-clé une liste des noms qui représente le classificateur et les spécifications dans un domaine donné, nous regroupons toutes les définitions du TLFi selon les domaines et calculons la pondération de chaque lemme.

Par rapport aux autres formules de pondération comme TF.IDF (Sparck Jones, 1972) qui privilège les termes discriminants et rares, notre but est de donner plus de poids aux termes situés au début de la définition, considérés comme des représentants des classes, et aux termes discriminants dans la collection des définitions pour un domaine donné, considérés comme des caractéristiques spécifiques. Les 4 mesures qui composent notre formule de pondération ont été obtenues grâce aux analyses effectuées sur notre corpus de travail.

6 Hiérarchisation des termes

Afin de pouvoir faire une hiérarchisation des termes, d'autres critères doivent être pris en compte car la pondération n'est pas tout à fait suffisante. En total nous avons analysé 10 critères qui pourraient être utilisés lors de la hiérarchisation des termes, mais seulement 2 ont été retenus :

1. La présence du mot vedette dans les définitions de ses lemmes.
2. La présence des lemmes dans les définitions des autres lemmes pour le même mot vedette.

Quand le lemme L1 s'inclut dans les définitions de lemme L2 et inversement, on peut parler de phénomènes d'hypéronymie ou de synonymie. Ainsi dans le cas quand *herbe* $\subset$ *pâturage* et *pâturage* $\subset$ *herbe* on peut dire qu'ils sont des synonymes mais il est assez difficile de déterminer la relation de l'hypéronymie. Afin de pouvoir déterminer quel lemme représente un hypéronyme pour l'autre il faut vérifier sa position dans les définitions. Le lemme avec la position la plus initiale est considéré comme l'hyperonyme de l'autre lemme.

Ainsi dans notre exemple l'herbe est un hyperonyme de pâturage, car il apparaît en troisième position dans la définition du lemme « pâturage » alors qu'à l'inverse « pâturage » n'apparait qu'en 17^{e} position dans une des définitions du lemme « herbe ».

7 Les principes de construction du thésaurus

Les termes du thésaurus sont les mots-clés utilisés par utilisateur lors de la recherche ou de l'indexation des images. Afin de pouvoir réaliser la hiérarchisation, à chaque mot-clé est attribuée une liste de lemmes avec les pondérations dans chaque domaine auquel peut appartenir le mot-clé. En s'appuyant sur les critères décrits ci-dessus nous avons pu définir un processus automatique de hiérarchisation des termes. Le thésaurus est en constante évolution car il est construit au fur et à mesure que de nouveaux mots-clés apparaissent dans la base d'indexation des images. Le thésaurus a une structure hiérarchique à arborescence simple où les fils du nœud père représentent des éléments de celui-là liés par la relation *« est un »* (cf. fig.1).

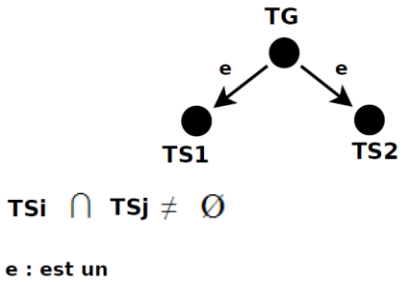

Figure 1 : Modèle de thésaurus hiérarchique à arborescence simple

L'arbre de la hiérarchie est peu profond et se développe avant tout en largeur.

8 Analyse des résultats

L'analyse des premiers résultats a montré que les pondérations des termes et les critères utilisés ne sont pas toujours suffisants pour la construction autonome de la hiérarchie. Le problème principal est la capacité à extraire à partir d'une définition les lemmes qui représentent la classe à laquelle appartient le mot-clé et ses spécifications. Ci-dessous (cf. tableau 2) nous présentons les résultats obtenus pour le mot-clé « tanaisie ».

Domaine	Lemme	Pondération
Botanique	plante	0.09659559704474198
Botanique	famille	0.0298607629748796
Botanique	fleur	0.01445838603979702
Botanique	bouquet	9.579406148522748E-5
Botanique	tige	0.0018571130644637402
Botanique	herboristerie	7.80560689215469E-6
Botanique	propriété	2.508329670387084E-4

Tableau 2 : Pondérations des lemmes du mot-clé « tanaisie »

Le lemme « plante » a une pondération la plus grande et est considéré comme l'hyperonyme du mot « tanaisie ».
Afin de pouvoir construire la hiérarchie nous vérifions quels lemmes sont inclus dans les définitions des autres lemmes (cf. tableau 3).

Lemmes 1	Lemmes 2
famille	tige
fleur	tige
fleur	bouquet
tige	plante
plante	fleur
plante	tige
plante	herboristerie
plante	bouquet

Tableau 3 : La liste des Lemmes 1qui se trouvent dans les définitions des Lemmes 2

Puisque le lemme « plante » a été choisi comme l'hyperonyme du mot « tanaisie » (cf. tableau 3), nous vérifions seulement pour ce lemme sa position dans les définitions des autres lemmes et choisissons le lemme dans la quelle le lemme « plante » a la position la plus initiale. Dans le cas présent, c'est dans la définition de lemme « fleur » que le lemme « plante » a la position la plus initiale. La structure hiérarchique construite est donc de la forme suivante (cf. image 1) :

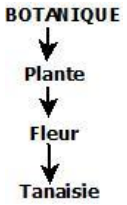

Image 1 : Exemple de construction d'une hiérarchie pour le mot-clé « tanaisie »

Les autres lemmes « tige, herboristerie, bouquet » sont considérés comme des caractéristiques de terme « plante » et dans le thésaurus seront situés au niveau le plus bas.
Toutefois les règles proposées, lors de la hiérarchisation ne sont pas suffisantes pour satisfaire tous les cas. Comme un système informatique ne possède pas d'intuition épilinguistique, l'intervention de l'homme est indispensable.
La présence des domaines comme des termes génériques du thésaurus est très importante et permet de résoudre les problèmes d'ambiguïté. Lorsqu'un mot-clé appartient plusieurs domaines, l'utilisateur peut préciser le domaine concret qu'il souhaite pour indexer son image. L'interaction homme-machine permet ainsi d'améliorer les résultats de recherches en proposant à l'utilisateur seulement les images d'un domaine exact.
La validation des résultats pourra être réalisée à l'aide de projet Definiens, où la segmentation des définitions TLFi, en composante centrale et composante périphériques, a été réalisée manuellement[5].

[5] Les premiers résultats de ce projet Definiens en cours devaient être disponibles dans quelques mois.

9 Conclusion

L'objectif poursuivi dans nos recherches est la création d'un système automatique d'indexation et de recherche d'images exploitant des sémantiques du TLFi. La solution proposée dans cet article consiste en création d'un thésaurus construit à l'aide des ressources linguistiques qui permettent d'enrichir les descriptions textuelles de l'image et d'interpréter son contenu. Grâce à sa forte capacité d'apprentissage, le thésaurus permet une catégorisation très précisément des images, selon les domaines, une indexation de grandes quantités d'images et une recherche rapide.
A l'étape actuelle les recherches ont été menées sur l'analyse des définitions du TLFi et la détermination des règles de hiérarchisation. Prochainement les règles seront implémentées dans le système afin de réaliser les premières expérimentes sur un corpus d'entraînement.

Remerciements

Je tiens à exprimer mes remerciements à Monsieur Jean-Marie Pierrel, mon directeur de recherche, pour sa disponibilité et ses conseils, et messieurs Eric Mathieu et Cyril March, responsables de XILOPIX, pour m'offrir la possibilité de mener mes recherches dans le cadre de leur entreprise.

Références

BARQUE L., & POLGUERE A. (2009). Structuration et balisage sémantique des définitions du Trésor de la Langue Française informatisé (TLFi). *Fourth International Conference on Meaning-Text Theory*. Montréal.

DENDIEN J., & PIERREL J.-M. (2003). Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL (Traitement Automatique des Langues),* Vol. 44 - n°2. Hermes Sciences.

GROSS P. (2007). *L'indexation multimédia : description et recherche automatique.* Lavoisier.

SPARCK JONES K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation , 28(1)*, 11-20.

TOURATIER C. (2000). *La sémantique.* Paris: Armand Colin.