Méthodologie de construction automatique du thésaurus pour l'indexation et la recherche des images

Inga Gheorghita

Méthodologie de construction automatique du thésaurus pour l'indexation et la recherche 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 de construction automatique du thésaurus à l'aide du Trésor de la Langue Française informatisé (TLFi). Nous utilisons les définitions du TLFi pour désambiguïser et enrichir les mots-clés présents dans les descriptions textuelles associées aux images, en construisant un arbre hiérarchique. L'approche proposée peut être utilisée pour la catégorisation très précise d'images, pour l'indexation de grandes quantités d'images et pour la recherche.

This article presents a methodology for automatic thesaurus construction using the "Trésor de la Langue Française informatisé" (TLFi). We use the definitions of TLFi to disambiguate and expand the keywords of image's textual descriptions, by building a hierarchical tree. The proposed approach can be used for accurate categorization of images, the indexation of large amounts of images and search.

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

1 Introduction

Avec l'arrivée de l'Internet le marché de l'image numérique a progressé de manière exponentielle. L'offre d'illustrations n'a jamais été aussi grande. Sachant qu'une agence de photo gère habituellement entre un et vingt millions d'images, qu'un satellite météo envoie plusieurs giga-octets de données chaque jour, qu'un possesseur d'appareil photo numérique actif prendra de l'ordre de cents mille photos en trente ans (Gros, 2007), l'accès et la recherche dans cette masse d'informations énorme posent de nouveaux défis. L'organisation non structurée des images provoque un très grand désordre et une extraordinaire confusion lorsque l'on cherche à les identifier ou les repérer. L'une des causes est que les descriptions textuelles associées aux images ne sont souvent pas suffisantes pour permettre leur structuration. Afin de gérer et d'utiliser efficacement ces bases d'images, un système d'indexation et de recherche est donc nécessaire. C'est pour cette raison que la recherche d'images est devenue un sujet très actif dans la communauté internationale et a connu un véritable engouement au cours des deux dernières décennies.

Le but d'un système d'indexation et de recherche d'images est d'organiser l'information selon certains critères et de permettre aux utilisateurs un accès rapide et une recherche des images qui correspondent mieux à leurs besoins d'information. 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 formes, les textures, les couleurs. Ainsi suivant les objectifs recherchés, certains systèmes de ce type sont utilisés pour reconnaître des formes, des objets, des scènes dans une image et même des images similaires à une « image requête » comme dans le cas du système QBIC (Query By Image Content) (Niblack, et al., 1993). Le problème majeur des systèmes de recherche par le contenu est la prise en compte insuffisante de l'aspect sémantique des images, ce qui par conséquent réduit leur efficacité.
La plupart des systèmes d'indexation et de recherche d'images, notamment ceux existant sur le Web tels que Google Image, Flickr, Photolia etc., est basé sur l'utilisation de 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 la requête et les termes associés aux images.
Toutefois les images proposées à l'utilisateur ne semblent pas toujours correspondre à sa requête initiale. Ces écarts sont essentiellement liés à l'opacité des systèmes de recherches par rapport à la sémantique : aucune analyse du contenu de la requête pour déterminer sa signification n'est en fait réalisée. D'autre part, la description textuelle de l'image est souvent trop courte pour décrire son contenu. Enfin, une difficulté réside dans les 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 description d'une image.

Afin de rendre à l'utilisateur des résultats répondant au mieux à sa requête, il convient d'utiliser des ressources sémantiques. Ainsi (Popescu, Grefenstette, & Moëllic, 2007) exploitent les connaissances ontologiques contenues dans le WordNet (Miller, 1990) et les techniques de traitement des images afin d'améliorer la recherche des images. WordNet est aussi utilisé pour trouver la correlation entre les mots-clés et les régions visuelles de l'image afin de réaliser l'annotation des images (Li & Sun, 2006), (Jin, Wang, Khan, & Awad, 2005). Mais étant donné que la construction manuelle des ressources linguistiques (WordNet etc.) est un processus assez couteux en temps et en argent, actuellement plusieurs systèmes réutilisent l'information disponible sur Internet afin de produire des informations sémantiques structurées. (Wang, Jiang, Chia, & Tan, 2010) utilisent Wikipedia pour la construction automatique d'une ontologie et obtiennent des résultats encourageants en l'utilisant pour rechercher des images sur le web. (Nakayama, Hara, & Nishio, 2007) proposent l'exploitation des hyperliens de Wikipedia afin de construire un thésaurus associatif. Dans (Wang, Ma, & Li, 2004) les auteurs créent un thésaurus en utilisant les images du Web. A partir du texte situé autour de l'image, ils extraient les termes qui ont une pondération élévée et essaient d'associer ces mots-clés avec les régions visuelles de l'image. Pour obtenir une structure hiérachique ils font appel au WordNet.
Dans nos recherches, nous nous focalisons sur l'élaboration d'un moteur d'indexation et de recherche d'images utilisant les ressources lexicales de l'ATILF, plus particulièrement le Trésor de la Langue Française Informatisé (TLFi : www.atilf.fr/tlfi). Pour ce faire, nous proposons la construction automatique du thésaurus à partir du texte brut associé aux images : la liste des mots-clés, le titre et la légende. Plus précisément, il s'agit d'associer à chaque mot de la description textuelle de l'image une liste de mots extraite des définitions TLFi (pour chaque domaine), en construisant ainsi un arbre hiérarchique, et ce de façon automatique. Le thésaurus sera enrichi au fur et à mesure que les nouvelles images, avec ses descriptions textuelles, seront ajoutées à la base de données images (XILOPIX : www.xilopix.com). L'approche proposée permettra de réaliser une indexation automatique des images par domaine et une recherche plus précise.

2 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). Il est utilisé dans le

milieu de l'enseignement avec le même engouement que dans le milieu de la recherche. Il est aujourd'hui la base de plusieurs projets de recherche comme le projet Definiens (Barque & Polguère, 2009) ou le projet RELIEF[1]. Le TLFi a aussi été utilisé pour l'enrichissement du thésaurus Thesaulangue[2] afin d'obtenir une onto-terminologie destinée à l'annotation sémantique de textes de spécialités (Kister, Jacquey, & Gaiffe, 2009).

Si les dictionnaires et encyclopédies en ligne sont utilisés pour la production d'informations sémantiques structurées pourquoi le TLFi, produit de 30 ans de travail de lexicographes, ne pourrait-il pas être utilisé dans le même but ? Quels sont les facteurs qui nous conduisent à utiliser le TLFi dans notre projet? Tout d'abord le TLFi est une ressource codée en XML, 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 présente aussi un grand intérêt dans la structure de ses définitions lexicographiques. Ce sont des définitions logiques (hyperonymiques) constituées à la fois d'*une classe* ou *genre prochain* à laquelle appartient le mot défini et des *propriétés* qui le particularisent à l'intérieur de cette classe.

2.1 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 de SEMEME est qu'il contient les définitions TLFi lemmatisées et filtrées, en ne gardant que les mots à sémantisme plein (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 initial, SEMEME n'a pas conservé 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 du corpus de travail est présentée dans le tableau 1 ci-dessous :

	Total	Vedette	Syntagme
Lemmes différents	38 617	35 468	22 779
Définitions	265 475	206 052	59 423
Domaines	7 786[3]	6 851	3 341

Tableau 1 : Information sur les données du corpus de travail

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

[1] 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.

[2] Le thésaurus de linguistique française du laboratoire ATILF.

[3] Le grand nombre de domaines s'explique par le fait que certains domaines apparaissent en combinaison avec d'autres comme « Math., Arithm. », « Math., Géom. », « Math.mod » etc. L'homogénéisation des domaines et leur hiérarchisation restent à faire.

2.2 La structure des définitions du TLFi

Les définitions d'un dictionnaire sont rédigées selon certaines règles qui déterminent le type et la structure des définitions. En lexicographie, il existe plusieurs types de définitions : logique (hyperonymique), par équivalence synonymique, morphosémantique, méronymique, par approximation (Touratier, 2000).
En général, les définitions du TLFi, surtout pour les noms, sont construites selon le schéma suivant « classificateur+spécifications » où le classificateur représente la classe à laquelle 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 » ; 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 ». 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.
Dans la version en ligne du TLFi, les définitions sont parfois structurées selon une hiérarchie (I, Ia, II, IIa etc.). Or les fichiers XML de SEMEME ne contiennent pas d'indicateurs de hiérarchie. Toutefois les domaines qui leur sont associés représentent une connaissance lexicale importante et nous l'utilisons pour regrouper les définitions par domaine. Par exemple le mot « lion » appartient à 8 domaines et en fonction du domaine lui sont associés les lemmes des définitions correspondantes. Ainsi pour le domaine « zoologie » au mot « lion » sont associés les lemmes « mammifère, Félidés, crinière » etc. Par contre dans le TLFi seulement 30.87% des définitions ont des domaines. Les définitions sans domaines sont groupées dans un nouveau domaine nommé « générique ».

3 Pondération des mots dans une définition du TLFi

Dans une définition les mots ont des statuts différents et n'apportent pas la même quantité d'informations. Afin de pouvoir attribuer à un mot-clé une liste de noms qui représentent 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. Contrairement aux autres formules de pondération comme TF.IDF (Sparck Jones, 1972) qui privilégient 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. Ainsi pour calculer la pondération finale nous prenons en compte l'importance du mot dans la définition (pondération locale), l'importance du mot dans la collection des définitions pour un domaine donné (pondération globale) et la position du mot dans la chaine de caractères de la définition. Les 3 mesures qui composent notre formule de pondération finale ont été obtenues grâce aux analyses effectuées sur notre corpus de travail.

3.1 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. Au total nous avons analysé 10 critères qui pourraient être utilisés lors de la hiérarchisation des termes comme la présence du mot dans les formes des syntagmes, les catégories grammaticales que peut avoir le mot etc., mais seulement 2 ont été retenus :

1. La présence du lemme avec la pondération maximale dans les définitions des autres lemmes pour le même mot vedette.
   Si le lemme *L1* a une pondération maximale dans les définitions d'un mot vedette *L2* pour un domaine donné *D*, alors *L1* est considéré comme l'hyperonyme de *L2*, c'est-à-dire $L2 \subset L1$. Dans la hiérarchie *L1* devient le nœud-père pour *L2* et s'inclut dans le domaine *D* qui représente le nœud générique du thésaurus. Ainsi si le mot vedette « lion » appartient au domaine « zoologie » et $lion \subset mammifère$, nous obtenons que $lion \subset mammifère \subset zoologie$. La hiérarchie ainsi formée n'est pas tout à fait complète. C'est pour cette raison que nous vérifions le cas où le lemme « mammifère » a

une position initiale dans les définitions des autres lemmes pour le même mot vedette « lion », c'est-à-dire s'il ne représente pas l'hyperonyme pour un autre lemme. Si c'est le cas, alors nous obtenons la structure hiérarchique suivante *lion ⊂ félidés ⊂ mammifère ⊂ zoologie*.

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. Toutefois dans le cas où *herbe ⊂ pâturage* et *pâturage ⊂ herbe* 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 ».

3.2 Les principes de construction du thésaurus

Le thésaurus est un lexique hiérarchisé pour un domaine donné. Les techniques utilisées pour la construction automatique des thésaurus sont en grande majorité de nature statistique. Elles consistent à déterminer une relation entre deux termes soit sur la base de leur présence dans un même document, soit en analysant les termes qui co-occurrent avec eux (Bruandet & Chevallet, 2003). L'information sur la co-occurrence peut aussi être utilisée pour l'indentification des relations sémantiques entre les termes (Schütze & Pedersen, 1997). Ainsi un terme X est un synonyme de Y si les termes avec lesquels ils co-occurent sont les mêmes. Toutefois les méthodes linguistiques basées sur des relations linguistiques entre termes sont aussi utilisées. Ce sont des approches qui exploitent par exemple des patrons syntaxiques (Hearst, 1992), (Morin, 1999) ou bien les marqueurs de causalité (Nazarenko, 1994).

L'approche que nous proposons est plutôt une approche hybride (Le & Chevallet, 2006), car nous exploitons en même temps les données statistiques et l'information linguistique représentée sous forme de relations sémantiques entre les termes. Ainsi pour la construction du thésaurus nous utilisons les mots-clés qui apparaissent dans les descriptions textuelles associées aux images. Afin de pouvoir réaliser la hiérarchisation, en fonction du domaine, une liste de lemmes avec les pondérations est attribuée à chaque 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 des images. Le thésaurus a une structure hiérarchique à arborescence simple où les nœuds-fils représentent des éléments du nœud-père liés par la relation *« est un »* (cf. fig.1).

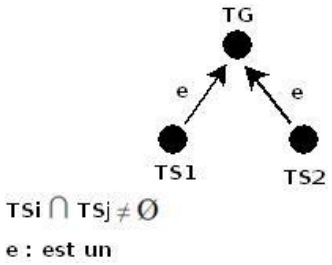

Figure 1 : Modèle de thésaurus hiérarchique à arborescence simple

L'arbre de la hiérarchie se développe avant tout en profondeur plutôt qu'en largeur.

4 Analyse des résultats

Les résultats préliminaires que nous obtenons sont assez satisfaisants, mais ils nous ne permettent pas encore d'évaluer notre modèle. 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 maximale et représente l'hyperonyme du mot « tanaisie ». Ainsi selon les règles de hiérarchisation présentées ci-dessus, nous obtenons que $tanaisie \subset plante \subset botanique$.
Afin de pouvoir construire une hiérarchie complète, 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 : Liste des Lemmes 1 qui se trouvent dans les définitions des Lemmes 2

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

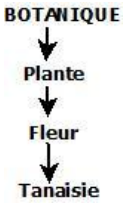

Image 1 : Exemple de construction d'une hiérarchie pour le mot-clé « tanaisie »

Les autres lemmes « tige, herboristerie, bouquet » sont considérés comme des caractéristiques du terme « plante » et seront situés dans le thésaurus à un plus bas niveau.

L'analyse des premiers résultats a montré que les pondérations des termes et les règles de hiérarchisation utilisées ne sont pas toujours suffisantes pour la construction autonome de la hiérarchie. Ainsi par exemple les lemmes avec la pondération maximale ne représentent pas toujours des hyperonymes des mots-clés. Le mot « parterre » n'est pas l'hyperonyme du mot « fauteuil » pour le domaine « spectacle ». Il peut arriver aussi qu'un nœud-fils ait deux nœuds-pères. Par exemple pour le domaine « générique» le mot « bâtiment » a comme nœud-père les termes « construction » et « administration ».

Actuellement le système de construction du thésaurus est en cours de développement. Son implémentation permettra d'améliorer les règles de hiérarchisation et d'assurer sa croissance en profondeur. L'évaluation des résultats pourra être réalisée à l'aide du thésaurus déjà existant chez XILOPIX, qui est construit de manière manuelle. Dans un deuxième temps, il faudra résoudre les cas où le mot-clé n'est pas présent dans le TLFi (noms propres, etc.).

5 Conclusion

L'objectif poursuivi dans nos recherches est la création d'un système automatique d'indexation et de recherche d'images exploitant l'information sémantique du TLFi. L'approche proposée dans cet article consiste à construire un thésaurus à partir de ressources linguistiques (ici les définitions du TLFi) et des descriptions textuelles des images. Le thésaurus construit pourra être utilisé pour une catégorisation très précise des images selon les domaines, pour 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 expériences 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'avoir offert 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.

BOUJEMAA N., FAUQUEUR J., FERECATU, M., FLEURET F., GOUET V., SAUX, B. L., ET AL. (2001). IKONA: Interactive Generic ans Specific Image Rerieval. *International workshop on Multimedia Content-Based Indexing and Retrieval (MMCBIR)* .

BRUANDET M. F., & CHEVALLET J. P. (2003). Utilisation et construction de bases de connaissances pour la Recherche d'Informations. Dans M. -H. Stefanini, & E. Gaussier, *Assistance Intelligente à la Recherche d'Information* (pp. 85-118). Hermes.

CHANG S.-K., & LIU S.-H. (1984, Juillet). Picture indexing and abstraction techniques for pictorial databases. *Pattern Analysis and Machine Intelligence, IEEE Transactions , 6* (4), pp. 475-484.

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

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*, (pp. 539-545). Nantes, France.

JIN Y., WANG L., KHAN L., & AWAD M. (2005). Image Annotations By Combining Multiple Evidence & WordNet. *13th Annual ACM International Conference on Multimedia*, (pp. 706-715).

KISTER L., JACQUEY E., & GAIFFE B. (2009). Fusion d'un thesaurus et d'une terminologie : utilisation de ressources existantes pour amorcer une onto-terminologie. *TIA'39.* Toulouse.

LE T. H., & CHEVALLET J. -P. (2006). Extraction et structuration des relations multi-types à partir de texte. *RIVF'06*, (pp. 53-58). Ho Chi Minh, Vietnam.

LI W., & SUN M. (2006). Automatic Image Annotation Based on WordNet and Hierarchical Ensembles. *Computational Linguistics and Intelligent Text Processing.* Mexico City.

MILLER G. A. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography , 3* (4), 245-264.

Morin E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Thèse de doctorat.

NAKAYAMA K., HARA T., & NISHIO S. (2007). Wikipedia Mining for an Association Web Thesaurus Construction. *Proceedings of IEEE International Conference on Web Information Systems Engineering*, (pp. 322-334).

NAZARENKO A. (1994). *Compréhension du langage naturel: le problème de la causalité.* Thèse de doctorat.

NIBLACK W., BARBER W., EQUITZ M., FLICKNER M., GLASMAN E., PETKOVIC D., ET AL. (1993). The QBIC project : quering images by content using color, texture and shape. *Proceedings SPIE: Storage and Retrieval for Image and Video Database* , pp. 173-181.

POPESCU A., GREFENSTETTE G., & MOËLLIC P. -A. (2007). Improving Image Retrieval Using Semantic Resources. *Springer SCI series* .

SCHÜTZE H., & PEDERSEN J. O. (1997). A cooccurrence-based thesaurus and two applications to Information Retrieval. *Information Processing and Management , 33* (3), pp. 307-318.

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.

WANG H., JIANG X., CHIA L. -T., & TAN A. H. (2010). Wikipedia2Onto - Building Concept Ontology Automatically, Experimenting with Web Image Retrieval. *Informatica , 34* (3), 297-306.

WANG X. J., MA W. Y., & LI X. (2004). Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval. *Proceedings of ICME*, (pp. 2231-2234). Taipei, Taiwan.