Pondération des concepts en recherche d'information sémantique

Fatiha Boubekeur* — Wassila Azzoug**

**Université Mouloud Mammeri,*
Route de Hasnaoua, BP 17 RP 15000, Tizi-Ouzou, Algérie
amirouchefatiha@mail.ummto.dz

***Laboratoire Limose*
Université M'Hamed Bougara,
Avenue de l'Indépendance, 35000 Boumerdès, Algérie
azzoug_wassila@umbb.dz

RÉSUMÉ. *L'objectif principal de la pondération en recherche d'information (RI) est d'assigner aux termes d'index des poids sensés traduire leur importance dans les documents où ils apparaissent. En RI sémantique, les termes d'index représentent des concepts. L'importance d'un concept est généralement mesurée soit à travers sa fréquence d'occurrence, soit à travers sa centralité, définie comme son degré de relation avec les autres concepts du document. Dans ce papier, nous proposons et évaluons une approche de pondération des concepts basée sur une nouvelle définition de la centralité. La centralité d'un concept est une mesure combinée de sa fréquence relative et de sa proximité sémantique avec les autres concepts du document. Nous montrons en particulier que notre approche offre de meilleurs résultats que les approches de pondération classiques sur des concepts.*

ABSTRACT. *The main objective of term weighting in information retrieval (IR) is to assign index terms with weights assumed to reflect their importance in the documents where they appear. In semantic information retrieval, index terms are concepts. Concept importance is measured either through its occurrence frequency or its centrality, defined through the degree of its relations to other concepts in the document. In this paper, we propose and evaluate an approach to concept weighting that is based on a novel definition of centrality. Concept centrality is a combined measure of its relative frequency and its semantic relatedness to other concepts in the document. In particular we show that our approach provides better results than classical weighting approaches on concepts.*

MOTS-CLÉS : *Recherche d'information sémantique, concepts, pondération, centralité.*

KEYWORDS: *Information retrieval, Semantic indexing, weighting, centrality, concepts.*

1. Introduction

Un système de recherche d'information (SRI) a pour but de sélectionner, dans une collection de documents préalablement enregistrée, l'ensemble des documents pertinents pour une requête utilisateur. Pour ce faire, documents et requêtes sont d'abord indexés. L'indexation a pour but de construire une représentation simplifiée des documents et requêtes, à base de termes d'index. En indexation classique, les termes d'index sont des mots-clés identifiés à partir du texte du document (ou de la requête). En indexation sémantique, les termes d'index dénotent des concepts, éventuellement combinés à des mots-clés représentatifs du contenu sémantique du document (ou de la requête). Un concept est une abstraction généralisée de propriétés communes à plusieurs objets, faits ou évènements, … (Chevallet et al., 2009). Dans un texte, le concept est représenté par un ou plusieurs termes synonymes ayant une entrée dans une ressource terminologique (ontologie, thesaurus, dictionnaire…) (Chevallet, 2009). Les concepts sont identifiés par des techniques de *mapping* du texte sur la ressource utilisée (Dinh et al., 2010), et/ou des techniques de désambiguïsation des sens des mots (*WSD -Word Sense disambiguation-*) (Katz et al.,1999). A l'issue de l'indexation, les termes d'index sont pondérés par des poids sensés traduire leur importance dans le document (ou la requête). Les poids des termes du document et de la requête sont combinés dans un score de pertinence associé au document à l'issue de la recherche. De la qualité de la pondération dépend donc la qualité de la recherche.

De nombreux travaux en RI ont été consacrés à la définition de schémas de pondération performants, à l'exemple de *tf*idf* (Salton et al., 1988), Okapi-BM25 (Robertson et al., 1994), normalisation pivotée (Singhal et al., 1996), … qui malgré leurs différences, sont fondamentalement tous basés sur un même principe sous-jacent : quantifier *l'importance apparente* du terme principalement sur la base de sa fréquence d'occurrences et de sa fréquence documentaire. De telles approches, basées sur le comptage d'occurrences lexicales, ne permettent pas d'exprimer *l'importance latente* d'un concept liée à son apport sémantique au contenu du document.

Dans ce papier, nous proposons un schéma de pondération sémantique basé sur la centralité d'un concept. La centralité est vue comme une mesure combinée de l'importance apparente du concept (basée sur sa fréquence d'occurrence) et de son importance latente (basée sur son degré de relation sémantique aux autres concepts) dans le document. Ce schéma de pondération est évalué dans le cadre d'une recherche sémantique puis comparé à des schémas de pondération classiques.

La suite du papier est structurée comme suit : La section 2 présente une synthèse des travaux en pondération de concepts en RI sémantique et situe notre contribution. La section 3 introduit brièvement l'approche d'indexation par les concepts utilisée. La section 4 présente notre approche de pondération des concepts. La section 5 est dédiée à son évaluation. Les résultats expérimentaux y sont présentés. La section 6 conclut le papier.

2. Travaux connexes

La pondération est un problème crucial en RI. De la qualité de la pondération dépend la qualité de la recherche. Dans le contexte de la RI sémantique, plusieurs approches de pondération de concepts ont été adoptées. Dans (Katz et al., 1999), les concepts identifiés à partir de la base lexicographique WordNet (Miller, 1995) sont pondérés par une mesure classique *tf*idf*. Dans (Harrathi, 2010), le schéma *tf*ief* proposé est une version de *tf*idf* adaptée à la pondération de concepts relativement à des éléments de document XML. Les auteurs dans (Baziz et al., 2005) proposent un schéma de pondération dit *Cf*idf*, qui étend la pondération *tf*idf* pour tenir compte des termes composés. Cette approche a été enrichie dans (Boughanem et al., 2010) avec les notions de centralité et de spécificité. La centralité d'un concept définit le nombre de ses relations avec les autres concepts du document (Kang et al., 2005). Sa spécificité définit son degré de «spécialité», estimé en fonction de sa "profondeur" dans la hiérarchie *is-a* de WordNet. Ces notions sont également utilisées en indexation sémantique de documents biomédicaux (Dinh et al., 2010). Dans (Blanco et al., 2012), la centralité d'un concept est autrement mesurée à travers le nombre de ses relations de cooccurrence avec les autres concepts du document. Dans notre approche (Boubekeur et al., 2010), l'importance d'un concept dans le document est exprimée non plus par le nombre de relations qu'il a avec les autres concepts du document mais par sa proximité sémantique avec les autres concepts du document. Cette mesure est combinée à la fréquence dans notre proposition (Boubekeur et al., 2011). Des approches similaires avaient été proposées dans le cadre de l'indexation des documents XML multimédias (Torjman et al., 2008) ou non (Zargayouna, 2005). Les auteurs considèrent les balises du document comme des entités sémantiques qui sont représentées par les termes qu'elles contiennent. Les termes sont pondérés en fonction de leurs fréquences d'occurrences dans la balise (Zargayouna, 2005) ou dans le nœud correspondant (Torjman et al., 2008) d'une part et de leur similarités sémantiques avec les autres concepts du contexte d'autre part..

Notre proposition dans ce papier concerne la pondération de concepts par la mesure de centralité. Nous revisitons la notion de centralité introduite dans (Boubekeur et al., 2011) selon le principe sous-jacent suivant : « *un concept est d'autant plus central dans un document qu'il est fortement présent et fortement corrélé aux autres concepts de son contexte dans le document* ». Ce principe suggère d'agréger fréquence et proximité sémantique cumulée au sein d'un même score sensé définir la centralité du concept dans le document. Nous proposons en particulier un nouveau schéma d'agrégation en vue de mesurer la centralité. Notre approche de pondération est évaluée expérimentalement sur une collection de documents indexée par des concepts. Ces concepts sont issus de notre approche d'indexation sémantique proposée dans (Boubekeur et al., 2011) et résumée en section suivante.

3. Approche d'indexation sémantique utilisée

L'approche d'indexation sémantique utilisée (Boubekeur et al., 2011) permet d'indexer les documents et requêtes par une combinaison de concepts et de mots-clés orphelins. Deux types de concepts sont utilisés : les collocations et les sens des mots. Les collocations sont identifiées par une technique de *mapping* du texte sur une liste prédéfinie de collocations de WordNet (Miller, 1995), tandis que les sens (synsets de WordNet) sont identifiés par une technique de désambiguïsation des sens des mots, basée sur l'utilisation conjointe de WordNet et de son extensions aux domaines WordNetDomains (Magnini et al., 2000). Les mots-clés orphelins sont des mots non vides qui n'ont pas d'entrée dans WordNet.

4. Approche de pondération proposée

Partant de l'idée qu'un concept est d'autant plus représentatif du contenu du document qu'il est *central* localement dans ce document et *central* globalement dans la collection, nous définissons :

— La centralité locale d'un concept C^i dans un document d, notée $cc(C^i,d)$, sur la base de sa *pertinence* dans le document d'une part et sa fréquence d'occurrence d'autre part. La pertinence d'un concept est mesurée à travers sa proximité sémantique aux autres concepts du document. Sa fréquence égale la fréquence cumulée de tous ses termes représentatifs dans le document.

Formellement :

$$cc(C^i, d) = \alpha * tf(C^i, d) + (1-\alpha)\sum_{i \neq l} Sim(C^i, C^l) \quad \textbf{[1]}$$

où α est un facteur de pondération qui permet de balancer la fréquence par rapport à la pertinence (ce facteur est fixé expérimentalement), $Sim(C^i, C^l)$ mesure la similarité sémantique entre les concepts C^i et C^l, et $tf(C^i, d)$ est la fréquence d'occurrence du concept C^i dans le document d.

— La centralité globale d'un concept comme son pouvoir de discrimination dans la collection. L'idée est qu'un concept qui est central dans trop de documents n'est pas discriminant. En considérant qu'un concept C^i est central dans un document d si sa centralité est supérieure à un seuil s fixé, la centralité documentaire du concept est définie par :

$$dc(C^i) = \frac{n}{N} \quad \textbf{[2]}$$

où N est le nombre total de documents de la collection, et n est le nombre de documents d de la collection où le concept C^i est central (ie. $cc(C^i, d) > s$).

Le pouvoir de discrimination d'un concept est alors vu comme une mesure de sa *centralité documentaire inverse* (notée *idc*). Formellement :

$$idc\left(C^i\right) = \frac{1}{dc\left(C^i\right)} \quad \textbf{[3]}$$

REMARQUE. — Dans le cas où s = 0, la centralité documentaire d'un concept est assimilée à sa fréquence documentaire, et sa centralité documentaire inverse est assimilée à sa fréquence documentaire inverse *idf*.

Le poids $W\left(C_d^i\right)$ d'un concept C^i dans un document *d* est alors défini comme la mesure combinée de sa centralité locale et de sa centralité globale. Formellement:

$$W\left(C_d^i\right) = cc\left(C^i, d\right) * idc\left(C^i\right) \quad \textbf{[4]}$$

Le schéma de pondération proposé, dit schéma *cc-idc*, permet outre la pondération des concepts, la pondération des mots orphelins. Dans ce dernier cas, seul le facteur *tf***idf* de la centralité locale est considéré puisque les distances sémantiques inexistantes dans ce cas, sont initialisées à zéro.

5. Evaluation expérimentale

5.1. *Collection de test*

Nous avons évalué notre approche de pondération des concepts sur la collection TIME. La collection TIME est une petite collection composée 423 documents issus d'articles de presse du magazine TIME de 1963. Elle propose outre cette collection de documents, un nombre important de requêtes (83) et des jugements de pertinence.

Le sous-ensemble de la collection TIME que nous avons utilisé se compose des 423 documents de la collection et des 15 requêtes, parmi les 83 de la collection, qui donnent les résultats les plus significatifs dans une recherche classique à base de mots clés pondérés par *tf***idf*. Des jugements de pertinence sont en outre associés aux requêtes.

5.2. *Protocole d'évaluation*

Nous avons évalué notre approche de pondération sur un système de RI sémantique basé sur le modèle vectoriel. Dans ce système que nous avons implémenté, les index des documents et requêtes, assimilés à des vecteurs de concepts et mots-clés pondérés, sont comparés à travers la mesure du cosinus classiquement utilisée dans le modèle vectoriel. Les concepts sont issus de notre approche d'indexation sémantique (résumée en section 3). Les poids des termes

d'index sont les poids *cc-idc* (formule [1]), dans lesquels la similarité entre concepts est calculée sur la base de la mesure de Resnik (Resnik, 1999). L'index sémantique pondéré ainsi construit est désigné par Sem-CC-IDC.

L'évaluation est faite selon le protocole TREC en se basant sur deux métriques d'évaluation : les précisions à différents points *x*, *P@x* (*x* = 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 100), et la précision moyenne *MAP* (*Mean Average Precision*). La précision au point *x*, *P@x*, est le ratio des documents pertinents parmi les *x* premiers documents restitués. La *MAP* est la moyenne arithmétique des précisons moyennes *AP* calculées pour l'ensemble des requêtes utilisées (rappelons que pour une requête donnée, la précision moyenne *AP* correspond à la moyenne des précisions calculées aux différents rangs *p* où un document pertinent est restitué). Pratiquement, pour chaque requête, les 100 premiers documents restitués par le système sont examinés, et les précisions *P@x* (*x*= 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 100) ainsi que la *MAP* sont calculées. Le programme trec_eval[1] a été utilisé pour ces calculs. Nous avons ensuite comparé ces résultats issus de notre index sémantique Sem-CC-IDC à ceux restitués par des systèmes de référence (ou baselines). Dans nos expérimentations, nous avons considéré quatre baselines: Classic-TF_IDF, Classic-BM25, Sem-TF_IDF et Sem-BM25. Les deux premières correspondent à un index classique à base de mots-clés pondérés respectivement par *tf*idf* et BM25. Les deux dernières correspondant à un index sémantique à base de concepts pondérés respectivement par *tf*idf* et BM25.

La comparaison des résultats de la recherche issus respectivement des baselines Classic-TF_IDF et Classic-BM25 avec ceux issus de Sem-CC-IDC permet de montrer globalement l'apport de notre indexation sémantique par rapport à une indexation classique. Dans cette première comparaison, nous ne pouvons pas distinguer l'apport de la pondération proposée de l'apport des concepts. Pour ce faire, il faudra considérer un même index sémantique avec différents schémas de pondération. C'est l'objet des index Sem-CC-IDC, Sem-TF_IDF et Sem-BM25, dans lesquels le même index sémantique (celui issu de notre approche d'indexation résumée en section 3) est utilisé avec les pondérations respectives *cc-idc*, *tf*idf* et *BM25*. La comparaison des résultats de la recherche issus de Sem-TF_IDF et Sem-BM25 avec ceux de Sem-CC-IDC permet de montrer l'apport de la pondération des concepts par *cc-idc* par rapport à leur pondération par *tf*idf* ou *BM25*.

5.3. *Evaluation de l'approche de pondération*

5.3.1. *Choix du paramètre* α

Le paramètre α (formule [1]) permet de balancer entre la fréquence et la pertinence d'un concept. L'objectif de cette étape préalable est d'évaluer l'impact de α sur les résultats de la recherche. Pour cela, différentes valeurs sont affectées à α

[1] http://trec.nist.gov

qui sont utilisées pour pondérer les termes de notre index sémantique. Les index pondérés ainsi obtenus sont ensuite évalués à travers les résultats de recherche qu'ils retournent. L'évaluation est réalisée selon le protocole décrit ci-dessus (section 5.2). Le tableau 1 présente les résultats de cette évaluation (où *MP@x* est la moyenne arithmétique des différentes précisions *P@x* (*x*=1,2,3,4,5,10,15,20,30,50,100).

	α=0,1	**α=0,2**	**α=0,3**	**α=0,4**	**α=0,5**	**α=0,6**	**α=0,7**	**α=0,8**	**α=0,9**
P @ 1	**0,1333**	**0,1333**	**0,1333**	**0,1333**	**0,1333**	**0,1333**	**0,1333**	**0,1333**	**0,1333**
P @ 2	**0,1333**	**0,1333**	0,1000	0,0667	0,1000	0,1000	0,0667	0,0667	0,0667
P @ 3	**0,0889**	**0,0889**	**0,0889**	0,0667	**0,0889**	**0,0889**	0,0667	0,0444	0,0444
P @ 4	**0,0667**	**0,0667**	**0,0667**	0,0500	**0,0667**	**0,0667**	**0,0667**	**0,0667**	**0,0667**
P @ 5	**0,0800**	**0,0800**	**0,0800**	**0,0800**	**0,0800**	**0,0800**	**0,0800**	0,0667	0,0667
P @ 10	0,0667	**0,0733**	**0,0733**	**0,0733**	**0,0733**	0,0667	0,0667	0,0667	0,0467
P @ 15	0,0534	0,0534	0,0534	0,0534	0,0534	**0,0578**	**0,0578**	0,0534	0,0489
P @ 20	0,0400	0,0400	0,0400	**0,0433**	**0,0433**	**0,0433**	**0,0433**	**0,0433**	0,0400
P @ 30	**0,0378**	**0,0378**	**0,0378**	**0,0378**	**0,0378**	**0,0378**	0,0355	0,0333	0,0333
P @ 50	**0,0253**	**0,0253**	**0,0253**	**0,0253**	**0,0253**	**0,0253**	0,0240	0,0240	**0,0253**
P @100	0,0160	0,0160	0,0160	0,0160	0,0167	0,0173	0,0173	**0,0180**	**0,0180**
MP@x	0,0674	**0,0680**	0,0650	0,0587	0,0653	0,0652	0,0598	0,0560	0,0536
MAP	0,0968	**0,0987**	0,0968	0,0968	0,0970	0,0970	0,0929	0,0861	0,0787

Tableau 1. *Impact du facteur α sur les résultats de la recherche.*

Les résultats expérimentaux montrent que l'index issu de la valeur α= 0.2 donne les meilleurs précisions aux points *x* (*x* <=10). Aux rangs 15 et 20, les performances de cette pondération sont très proches des performances maximales (0,0534 contre un maximum de 0,0578 et 0,0400 contre un maximum de 0,0433 respectivement pour P@15 et P@20). Puis aux rangs 30 et 50, les meilleures performances sont atteintes. Une comparaison globale des performances de la recherche pour les différentes valeurs de α est réalisée alors sur la base de la moyenne des précisions à *x* (*MP@x*) et de la *MAP*. Les résultats montrent que la valeur α= 0.2 retourne les meilleurs valeurs. Dans la suite, nous retenons α= 0.2 comme facteur de pondération des concepts de la collection TIME. Ce choix permet de favoriser, dans la formule de pondération proposée, la pertinence sémantique du concept par rapport à sa fréquence dans le document. Il reste cependant à déterminer dans quelle mesure ce choix de la valeur de α est dépendant ou non de la collection utilisée.

5.3.2. *Etude comparative*

Les premières comparaisons réalisées entre les baselines Classic-TF-IDF et Classic-BM25 d'une part et notre index sémantique Sem-CC-IDC (pour α= 0.2) d'autre part, sont représentés à travers le graphique de la Figure 1. Il y apparait clairement que l'indexation sémantique apporte de meilleurs résultats à tous les points de précision par rapport à l'indexation classique basée mots-clés. A ce niveau de l'évaluation, il apparait donc que notre indexation sémantique est plus performante que l'indexation classique.

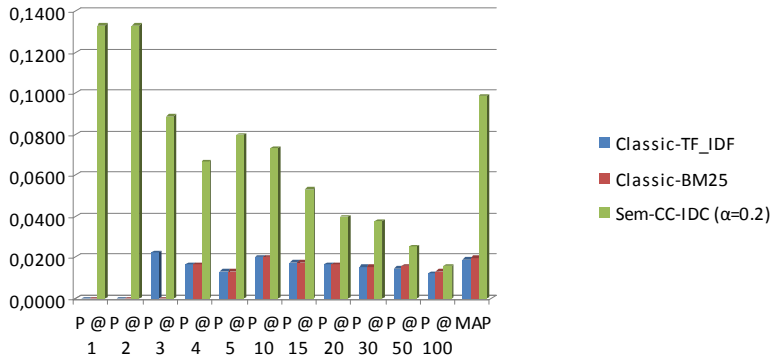

Figure 1. *Apport de notre index sémantique par rapport aux index classiques*

Les comparaisons suivantes réalisées entre les baselines Sem-TF-IDF et Sem-BM25 d'une part et Sem-CC-IDC (pour α= 0.2) d'autre part, sont représentées à travers le graphique de la Figure 2.

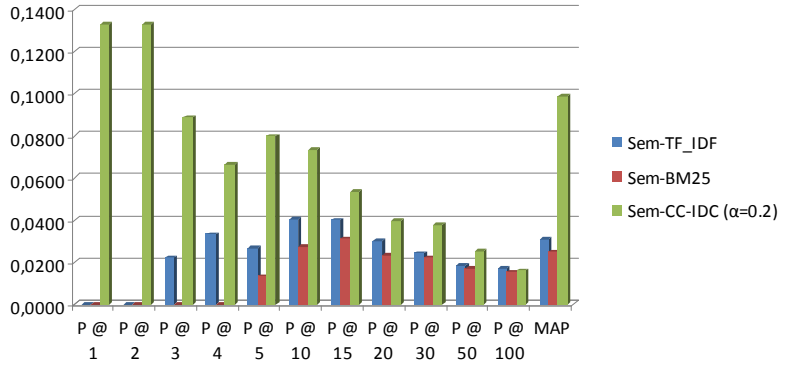

Figure 2. *Apport de notre pondération sémantique*

De cette évaluation, il ressort que l'index Sem-CC-IDC présente de meilleurs résultats à tous les points de précision par rapport aux index sémantiques Sem_TF_IDF et Sem-BM25. Les gains de performances significatifs (supérieurs à 50%) par rapport à Sem-TF-IDF sont de 300%, 100%, 200%, 80%, 54,47% et 218,38% respectivement pour les précisions P@3, P@4, P@5, P@10, P@30 et MAP respectivement. Les précisions *P@1* et *P@2* obtenues par l'index sémantique *Sem_CC-IDC* atteignent la valeur 0.133 tandis qu'elles sont nulles pour les baselines considérées impliquant des gains de performances certains. Notons cependant qu'au rang 100, un décroissement de la performance de l'ordre de -7% est observé. Les gains de performances significatifs par rapport à Sem-BM25 sont de 500%, 168,29%, 71,42%, 71,42%, 69,91% et 293% pour les précisions P@10, P@15,

P@20,P@30, P@50 et MAP respectivement. Le gain de performance au point de précision 100, bien que non significatif, est néanmoins positif de l'ordre de 4%. Les précisions P@1, P@2, P@ 3 et P@4 dans le cas de l'index sémantique varient entre 0,133 et 0.0889 tandis qu'elles sont nulles dans le cas de la baseline Sem-BM25.

A cette échelle de l'évaluation, nous concluons que notre approche de pondération est plus performante, pour les concepts, qu'une pondération classique.

6. Conclusion

Nous avons présenté dans ce papier, une approche de pondération sémantique des concepts dans un contexte d'indexation sémantique en RI. Notre contribution a porté sur la proposition et l'évaluation d'un nouveau schéma de pondération des concepts basé sur une nouvelle définition de la centralité d'un concept. Dans notre proposition, la centralité d'un concept tient compte d'une part de son importance apparente (mesurée à travers sa fréquence d'occurrences) dans le document et d'autre part de son importance latente (mesurée à travers ses relations sémantiques avec les autres concepts) du document. Un facteur α, dont la valeur est déterminée expérimentalement, permet de balancer entre importance latente et importance apparente d'un concept. L'approche a été évaluée sur la collection TIME en considérant un sous ensemble de 15 requêtes présélectionnées. Les résultats de l'évaluation ont montré d'une part que notre indexation sémantique est plus performante qu'une indexation classique basée mots-clés, et d'autre part que notre approche de pondération des concepts est plus performante que la pondération des concepts basée sur *tf*idf* ou *BM25*. Pour valider nos résultats à grande échelle, il reste à tester notre approche sur des collections de tests de type TREC.

7. Bibliographie

Baziz M., Boughanem M., Aussenac-Gilles N., «*A Conceptual Indexing Approach based on Document Content Representation*», *CoLIS5: Fifth International Conference on Conceptions of Libraries and Information Science*, Glasgow, UK, 4 juin 8 juin 2005.

Blanco R, Lioma C., «*Graph-based term weighting for information retrieval*», *Information Retrieval*, Vol.15, Issue 1, p. 54-92, 2012.

Boubekeur F., Boughanem M., Tamine L., Daoud M., «*Using WordNet for Concept-based document indexing in information retrieval*», *Fourth International Conference on Semantic Processing (SEMAPRO)*, Florence, Italy, October 2010.

Boubekeur F., Azzoug W., Chiout S., Boughanem M., «*Indexation sémantique de documents textuels*», *14e Colloque International sur le Document Electronique (CIDE14)*, Rabat, Maroc, Décembre 2011.

Boughanem M., Mallak I., Prade H., «*A new factor for computing the relevance of a document to a query*», *IEEE World Congress on Computational Intelligence (WCCI 2010)*, Barcelone, 18/07/2010-23/07/2010.

Chevallet J.-P., Ressources endogènes et exogènes pour une indexation conceptuelle intermédia, Mémoire d'Habilitation à Diriger des Recherches, 2009.

Dinh D., Tamine L., «*Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients*», *Conférence francophone en Recherche d'Information et Applications, CORIA 2010 (2010)*. P. 325-336.

Harrathi R., Recherche d'information conceptuelle dans les documents semi-structurés, Thèse de Doctorat de L'Institut Nationale des Sciences Appliquées de Lyon. Septembre 2010.

Kang B. Y., Lee S., «*Document indexing: a concept-based approach to term weight estimation*, Information Processing and Management», Vol. 41, Issue 5, p. 1065–1080, 2005.

Koopman B., Zuccon G., Bruza P., Sitbon L., Lawley M., «Graph-based concept weighting for medical information retrieval», Proceedings of the Seventeenth Australasian Document Computing Symposium, p. 80-87, 2012.

Magnini B. and Cavaglià G. *"Integrating Subject Field Codes into WordNet"*, *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May - 2 June, 2000.

Miller G. «*WordNet: A Lexical database for English*», *Actes de ACM 38*, pp. 39-41.

Resnik P., «Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language», Journal of Artificial Intelligence Research (JAIR), 11, 1999, p. 95-130.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M., and Gatford M., "*Okapi at TREC-3", Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA, November 1994.

Singhal A., Buckley C., Mitra M., « *Pivoted Document Length Normalization*», *ACM SIGIR*, 1996, p. 21-29.

Salton G., Buckley C., «*Term weighting approaches in automatic text retrieval*», *Information Processing and Management*, 24(5), 513–523.

Torjmen M., Sauvagnat P., Boughanem M.. «*Towards a structurebased multimedia retrieval model*», In *Proceedings of the 1st ACM, SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada*, p. 350–357, 2008.

Uzuner O., Katz B., Yuret D., "*Word Sense Disambiguation for Information Retrieval*", *AAAI/IAAI* 1999:985.

Zargayouna H. Indexation sémantique de documents XML, Thèse de doctorat en sciences de l'université Paris XI Orsay, Décembre 2005.