DOI: https://doi.org/10.31341/jios.42.2.5 UDC 004.659
Open Access Preliminary Communication

Improving the Results of Google Scholar Engine through Automatic Query Expansion Mechanism and Pseudo Re-ranking using MVRA

Mawloud Mosbah *mos_nasa@hotmail.fr*
Informatics Department
Faculty of Sciences
University 20 Août 1955 of Skikda, Algeria

Abstract

In this paper, we address the enhancing of Google Scholar engine, in the context of text retrieval, through two mechanisms related to the interrogation protocol namely: query expansion and reformulation. Besides the both mechanisms, we adopt re-ranking scheme using a pseudo relevance feedback algorithm we have proposed previously in the context of Content based Image Retrieval (CBIR) namely Majority Voting Re-ranking Algorithm (MVRA). The experiments conducted using ten queries reveal very promising results in terms of effectiveness.

Keywords: Information Retrieval, Google engine, Query Expansion, Query Reformulation, Re-ranking, Pseudo Relevance Feedback, MVRA.

1. Introduction

Information retrieval system aims at extracting, from a large information repository, a subset of information satisfying the user requirement expressed as a query [1]. As common users, we feel unsatisfied with the returned results. From our modest point of view, the reasons behind this non satisfaction are: (1) the less significance of encoding methods considered into indexing stage, (2) the less expressivity of languages considered into interrogation protocol and (3) the considering of both document and query just as vectors belonging to features space vector over the matching process.

Owing to the pre-cited non-satisfaction reasons, it is mandatory to enhance results answered by information retrieval system. Re-ranking and automatic query reformulation/expansion are some schemes for materializing this enhancement.

In this paper, we address the improvement of Google Scholar engine [2] through combining re-ranking and query expansion/reformulation schemes. The concepts to be added are extracted from the first returned documents. For re-ranking, we utilize Majority Voting Re-ranking Algorithm (MVRA) we have previously proposed in the context of CBIR.

The rest of the paper is arranged then as follows: Section 2 is devoted to present MVRA. In Section 3, we present query expansion and reformulation schemes. Section 4 talks about Google Scholar engine. We present, in Section 5, the architecture and the execution scenario of the proposed approach. Section 6 shows the experiments conducted and we conclude the paper with a conclusion.

2. Majority Voting Re-ranking Algorithm (MVRA)

The retrospect of what have been achieved into CBIR field reveals that CBIR is a mixture of two very interesting research domains namely information retrieval and machine vision. Indeed, in its first steps, CBIR [3] has adopted documentary information retrieval techniques [4]. Even with ontology approach [5], CBIR is still borrowing from text retrieval field.

Both of CBIR and text retrieval are qualified as problems difficult to be defined formally. Such kind of problems requires involving user for reaching solution. Relevance feedback is the scheme materializing user involvement through exploiting results judgment. This mechanism of relevance feedback has been firstly adopted within text retrieval field [6] and applied thereafter within CBIR [7].

The pseudo code of MVRA, introduced in [11], [12], [13], is given in Figure 1.

```
MVRA Algorithm
Let N : the number of the first returned documents.
Initialization:
Re-ranking_set=Φ.
Documents={$doc_1$, $doc_2$,.., $doc_N$} (the first returned documents)
Candidates= {$doc_{(N-4)}$, $doc_{(N-3)}$, $doc_{(N-2)}$, $doc_{(N-1)}$, $doc_N$}
Electors = Documents minus Candidates={$doc_1$, $doc_2$, .., $doc_{(N-5)}$}
        1.  Calculate the distance matrix : Electors*Candidates
WHILE (Electors≠Φ) do
        2.  Organizing a vote: Each elector from Electors gives one point to one document
            of Candidates whose the distance is the smallest.
        3.  Re-ranking the Cadidates based on the points collected during the vote
            operation
        4.
            Re-ranking_set=Re-ranking_set plus the last two documents within the
            Candidates.
            Candidates=the last two documents within the Electors plus Candidates.
            Electors= Electors minus the last two documents within the Electors.
            END WHILE
            Inversing the Re-ranking_set before visualizing it to the user.
            END.
```

Figure 1. MVRA pseudo Code.

Unfortunately, users are not always ready to give their relevance feedback. This leads the system adopting pseudo relevance feedback relying on the assumption that the first results of the initial search are relevant. Pseudo relevance feedback mechanism attempts then extracting correlation from the first returned results. There are many approaches for pseudo relevance feedback automatic learning [8]: clustering approach using algorithms such as K-means [9] and Hierarchical Agglomerative Clustering Method (HACM) [10], query reformulation, and parameterization of matching measures approach. In previous works [11], [12], [13], we have proposed a new algorithm, for re-ranking results into CBIR field, known as Majority Voting Re-ranking Algorithm (MVRA).

Using MVRA, as textual information retrieval enhancement, is then an attempt to say that it is time that algorithms introduced within CBIR contribute for improving the ancestor field that of information retrieval. It is time then that information retrieval field takes advantage of MVRA based on the collective relevance judgment and the new comparison way that of comparing between two documents rather than the traditional comparison manner, considered in documentary information retrieval field, that of comparing between a query and a document.

3. Query Expansion and Query Reformulation

The main purpose of an information retrieval system is satisfying the information user requirement through allowing access to the relevant documents [14], [15]. This requirement is commonly expressed as a query according to the adopted interrogation protocol. The major problem within the information retrieval field is the gap between the following couples: (information need, query), (query, index) and (corpus, index). Moreover, there are surely other problems of other necessary aspects such as: making the passage from data to information [16], the appropriate matching measure taking into account the semantic aspect [17], [18]. An additional issue to point out here is that there is a strong relation between the all these cited problems.

In this subsection, we address (information need, query) problem as a pre-processing operation in order to well deal with the (query, index) problem. In other words, we are interested here in query expansion and reformulation [19], [20].

Contrary to query expansion process which adds some other concepts to the original query, query reformulation proceeds to reformulate completely the query discarding the initial query concepts. Provided that the retrieval processes is based on data statistics rather than information where concepts are replaced simply by words or terms, the theoretical comparison between the both notions (expansion and reformulation) does not enable us to say anything conclusive. Indeed, relying just on new terms, keeping the same concepts, seems to be a good direction for accessing new documents which is good in terms of recall. In the other hand, keeping the initial query terms, through expansion mechanism, may lead to better results refinement which is good in terms of precision.

As done within *PhraseFinder* system [21], we add terms co-occurred, in first documents, with the initial query terms. Otherwise, we appeal to rely on occurrences number as term selection criteria.

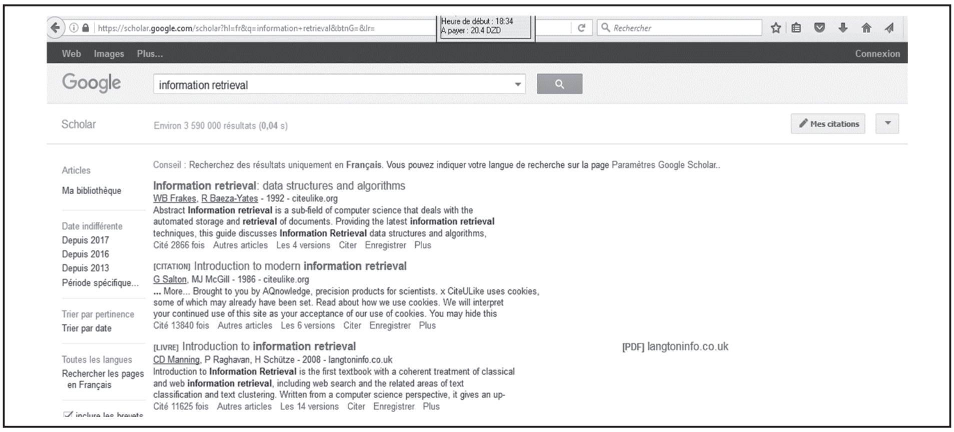
We aim then to improve the performance of Google Scholar through adopting pseudo relevance feedback mechanism and combing two schemes: query expansion coming from information retrieval field and MVRA proposed in the context of CBIR. For doing so, we proceed to:

- Utilize Local Context Analysis (LCA) [22] as an expansion method.
- Rely seldom on abstract terms rather than all document terms for the reason that documents are research scholars. This is efficient in the both cases: expansion and MVRA.
- Besides discarding the stop list (articles, prepositions, etc) as any basic information retrieval field do, we proceed to discard other research terminology such as Experimental Results and Conclusion.

4. Google Scholar Engine

Google Scholar [2] is an information retrieval engine handling, since its appearance in November 2004, scientific searching and allowing to scientific researches being visible to the academic community [23], [24], [25], [26], [27].

Google Scholar, as a Google service, adopts semi automatic query expansion scheme. Indeed, the system proposes, for the user, other keywords to be added for each submitted concept. Moreover, Google Scholar uses the feedback for improving performances and allowing refinement results.



information retrieval

Environ 3 590 000 résultats (0,04 s)

Conseil : Recherchez des résultats uniquement en **Français**. Vous pouvez indiquer votre langue de recherche sur la page Paramètres Google Scholar..

Information retrieval: data structures and algorithms
WB Frakes, R Baeza-Yates - 1992 - citeulike.org
Abstract **Information retrieval** is a sub-field of computer science that deals with the automated storage and **retrieval** of documents. Providing the latest **information retrieval** techniques, this guide discusses **Information Retrieval** data structures and algorithms,
Cité 2866 fois Autres articles Les 4 versions Citer Enregistrer Plus

[CITATION] Introduction to modern **information retrieval**
G Salton, MJ McGill - 1986 - citeulike.org
... More... Brought to you by AQnowledge, precision products for scientists. x CiteULike uses cookies, some of which may already have been set. Read about how we use cookies. We will interpret your continued use of this site as your acceptance of our use of cookies. You may hide this
Cité 13840 fois Autres articles Les 6 versions Citer Enregistrer Plus

[LIVRE] Introduction to **information retrieval** [PDF] langtoninfo.co.uk
CD Manning, P Raghavan, H Schütze - 2008 - langtoninfo.co.uk
Introduction to **Information Retrieval** is the first textbook with a coherent treatment of classical and web **information retrieval**, including web search and the related areas of text classification and text clustering. Written from a computer science perspective, it gives an up-
Cité 11625 fois Autres articles Les 14 versions Citer Enregistrer Plus

Figure 2. Snapshot screen of Google Scholar Graphical User Interface.

5. The Architecture of our approach

The execution scenario of our approach is as follows: the user submits his/her information requirement, as query keywords, to the system. Google Scholar will then be communicated using the query submitted by the user. Two scenarios are expected:

- The first results set answered by Google Scholar should be re-ranked employing MVRA. The new keywords being added to the original query are extracted just from the first subset of MVRA output. The expanded query is communicated again to Google Scholar and the corresponding results will be visualized to the user.

- The second scenario is that the new keywords being added to the original query are extracted just from the first subset of returned documents. The new reformulated query is communicated again to Google Scholar. The returned results are re-ranked using MVRA to be visualized to the user.

Figure 3 presents the general architecture of the proposed approach.

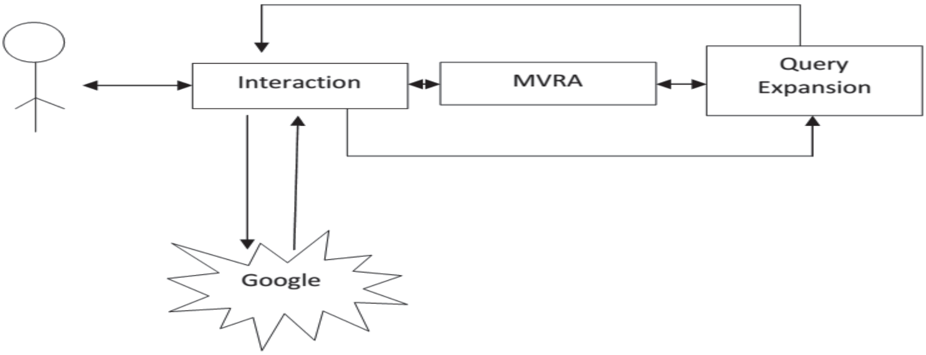


Figure 3. The General Architecture of the Proposed Approach.

6. Experimental Results

For testing our approach, we have built a system to which we submitted 10 queries of different disciplines. For each query, Google Scholar proposes some keywords being added. Table 1 presents the expanded queries proposed by Google Scholar and those proposed by our system for the different considered queries.

The 10 queries submitted to the System	The Expanded Queries proposed by Google Scholar	The Expanded Queries proposed by our System
Query Expansion	1. Retrieval query expansion 2. Automatic query expansion 3. Query expansion ontology 4. Relevance feedback query expansion 5. Latent semantic query expansion 6. Query expansion Wordnet 7. Concept based query expansion 8. Query expansion term 9. Interactive query expansion 10. Web query expansion	1. Query expansion technique 2. Query expansion method 3. Probabilistic query expansion 4. Probabilistic query expansion model 5. Document query expansion
Information Retrieval	1. Modern information retrieval 2. Introduction to modern information retrieval 3. Cross language information retrieval	1. Information retrieval model 2. Information retrieval document 3. Web information retrieval

Information Retrieval	4. Semantic information retrieval 5. Information retrieval Rijsbergen 6. Relevance information retrieval 7. Information retrieval query 8. Indexing information retrieval 9. Multimedia information retrieval 10. Rank information retrieval	4. Information retrieval text 5. Information retrieval interaction 6. Information retrieval dataset 7. Information retrieval collection
Object Design	1. Object design and sequencing theory 2. Object design and implementation 3. Object designs 4. Object designation 5. Object design phrase	1. Learning object design 2. Object design performance 3. Object design model 4. Object design reusability 5. Learning object design by contract
Software Maintenance	1. Software maintenance cost 2. Software maintenance management 3. Software management and evolution 4. Software maintenance process 5. Software maintenance environment 6. Software maintenance concepts and practice 7. Software maintenance metrics 8. Software maintenance activities 9. Software maintenance models 10. Software maintenance tasks	1. Software maintenance process 2. Software maintenance and evolution 3. Software maintenance management 4. Software maintenance activities 5. Software maintenance work 6. Software maintenance project 7. Software maintenance task
Validation and Verification	1. Validation and verification tools	1. Code verification and validation 2. Model verification and validation 3. Verification and validation of simulation models 4. Validation and verification techniques 5. Validation and verification tools

Functional Programming	1. Functional programming language 2. Functional programming matters 3. Functional programming style 4. Functional programming paradigm 5. Functional programming approach 6. Functional programming game engine PHD 7. Functional programming community 8. Functional programming systems 9. Functional programming environment 10. Functional programming in scala	1. Functional programming matters 2. Functional programming languages 3. Functional programming paradigm 4. Functional programming techniques 5. Functional programming methodology 6. Functional programming community
Remote Object Access	///	1. Remote sensing 2. Remote sensing data
Code Performance Optimization	///	1. Code optimization technique
Object Modelling with UML	///	1. Object modelling notations 2. Object modelling techniques 3. Object modelling language
Business Component Resource	///	1. Business process 2. Business model 3. Business process management

Table 1. The expanded Queries suggested by Google Scholar and by our System.

As Google engine, Google scholar is a black box whose the function way is still unknown. This fact is proved in Table2 where the rank presented by Google Scholar, for “Query Expansion” query is different of our developed system based on vector model and Cosine similarity as a matching measure. Table3 shows the rank of the 9 first documents without and with MVRA (two cases are adopted: separation of terms, non separation of terms) in the case of “Query Expansion” query. Contrary to Google, the results returned, by Google Scholar, have the PDF format. This format, contrary to HTML format, is rich in information content and simple to be processed automatically. Our system is developed using Java language. We have utilized *iTextpdf.jar* as an API allowing making the passage from PDF to Text.

Rank of Google Scholar	Rank of our developed system where the query terms are separated	Rank of our developed system where the query terms are non separated
1	6	6
2	5	3
3	3	5
4	4	9
5	7	4
6	1	7
7	9	1
8	2	2
9	8	8

Table 2. Some queries as ranked by Google Scholar and by our system.

Rank without MVRA	*Rank with MVRA where the query terms are separated*	*Rank with MVRA where the query terms are non separated*
1	1	1
2	3	2
3	2	4
4	4	5
5	5	3
6	6	7
7	8	8
8	7	6
9	9	9

Table 3. Some queries as ranked with and without MVRA.

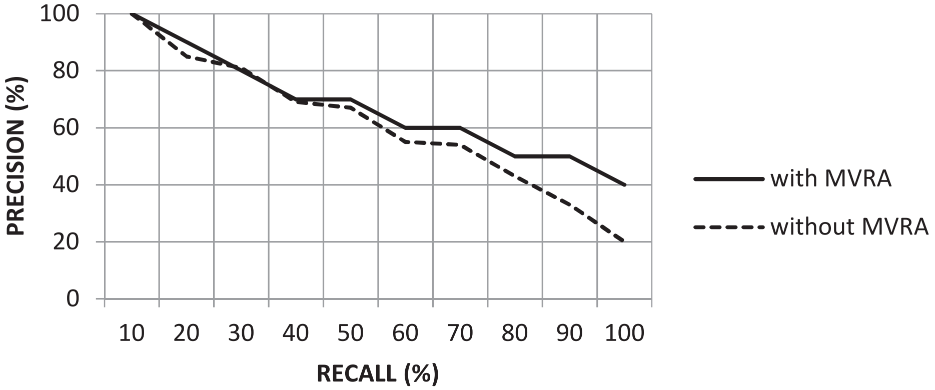


Figure 4. The average Precision\Recall with and without applying MVRA.

For evaluation, we utilize Precision and Recall metrics adapted for the case of the web [28]. These metrics are given respectively as follows:

$$Precision(C_q) = \frac{1}{|c_Q|} \sum_{p \in C_q} sim(r_q, p) \quad (1)$$

Where C_q is the number of considered documents and r_q is a description of relevant documents computed here via a binary classification using Hierarchical

Agglomerative Clustering Algorithm (HACM). The class containing the highest number of documents is qualified as the relevant class. $sim(\ )$ is Cosine function given as follows:

$$sim(q,p) = \frac{\sum_{k \in q \cap p} f_{k_q} f_{k_p}}{\sqrt{\left(\sum_{k \in p} {f_{k_p}}^2\right)\left(\sum_{k \in q} {f_{k_q}}^2\right)}} \qquad (2)$$

$$Recall(C_q) = Precision(C_q).\left|C_q\right| = \sum_{p \in C_q} sim(r_q, p) \qquad (3)$$

As shown in Figure 4, the good effect of using MVRA is evident. According to Figure 5, experiments show the clear superiority of enhancing scheme using query expansion and MVRA especially (expansion + MVRA) scenario which yields the best performances over the considered queries.

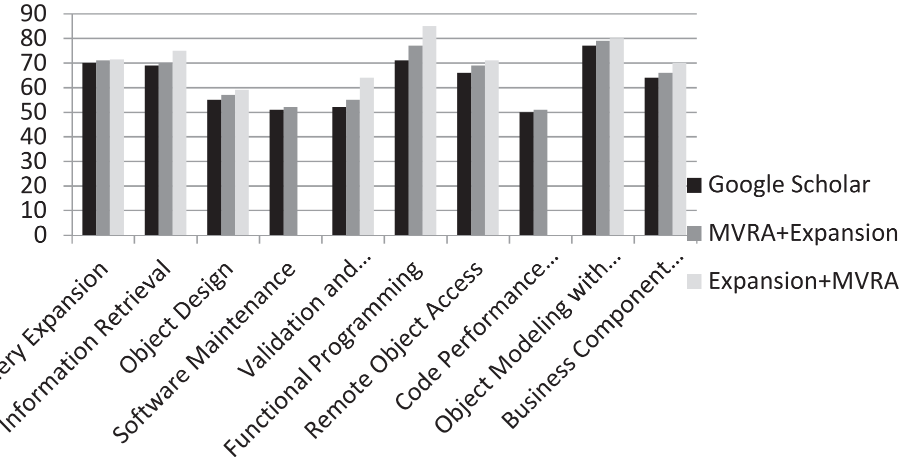


Figure 5. The Average Precision over the considered 10 queries: without improvement, with MVRA after Expansion, with Expansion after MVRA.

7. Conclusion

In this paper, we have addressed the enhancement of Google Scholar engine through using query expansion and Majority voting re-ranking Algorithm. The experiments show the clear superiority of enhancing scheme especially (expansion + MVRA) scenario which yields the best performances over the considered queries.

References

[1] R. Baeza-Yates, B. de A. N. Ribeiro, and others, Modern information retrieval. New York: ACM Press; Harlow, England: Addison-Wesley, 2011.

[2] Google, [Online]. Available: http://scholar.google.com/.

[3] F. Menczer, “Complementing search engines with online web mining agents,” Decis. Support Syst., vol. 35, no. 2, pp. 195–212, 2003.

[4] P. Jacsó, “Google Scholar: the pros and the cons,” Online Inf. Rev., vol. 29, no. 2, pp. 208–214, 2005.

[5] A. Noruzi, “Google Scholar: The new generation of citation indexes,” Libri, vol. 55, no. 4, pp. 170–180, 2005.

[6] J. Pomerantz, “Google Scholar and 100 percent availability of information,” Inf. Technol. Libr., vol. 25, no. 2, pp. 52–56, 2006.

[7] J. Beel, B. Gipp, and E. Wilde, “Academic Search Engine Optimization (aseo) Optimizing Scholarly Literature for Google Scholar & Co.,” J. Sch. Publ., vol. 41, no. 2, pp. 176–190, 2009.

[8] P. Mayr and A.-K. Walter, “An exploratory study of Google Scholar,” Online Inf. Rev., vol. 31, no. 6, pp. 814–830, 2007.

[9] J. Xu and W. B. Croft, “Improving the effectiveness of information retrieval with local context analysis,” ACM Trans. Inf. Syst., vol. 18, no. 1, pp. 79–112, 2000.

[10] Y. Jing and W. B. Croft, “An association thesaurus for information retrieval,” in Intelligent Multimedia Information Retrieval Systems and Management-Volume 1, 1994, pp. 146–160.

[11] B. Audeh, P. Beaune, and M. Beigbeder, “SMERA: Semantic Mixed Approach for Web Query Expansion and Reformulation,” in Advances in Knowledge Discovery and Management, Springer, 2017, pp. 159–180.

[12] C. Carpineto and G. Romano, “A survey of automatic query expansion in information retrieval,” ACM Comput. Surv., vol. 44, no. 1, p. 1, 2012.

[13] M. Mosbah and B. Boucheham, “Distance selection based on relevance feedback in the context of CBIR using the SFS meta-heuristic with one round,” Egypt. Informatics J., vol. 18, no. 1, pp. 1–9, 2017.

[14] M. Mosbah and B. Boucheham, “Matching Measures in the Context of CBIR: A Comparative Study in Terms of Effectiveness and Efficiency,” in World Conference on Information Systems and Technologies, 2017, pp. 245–258.

[15] R. L. Ackoff, “From data to wisdom,” J. Appl. Syst. Anal., vol. 16, no. 1, pp. 3–9, 1989.

[16] S. Krishnamurthy and V. Akila, “Information Retrieval Models: Trends and Techniques,” in Web Semantics for Textual and Visual Information Retrieval, IGI Global, 2017, pp. 17–42.

[17] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to information retrieval, vol. 39. Cambridge University Press, 2008.

[18] M. Mosbah and B. Boucheham, “Pseudo relevance feedback based on majority voting mechanism,” Int. J. Web Sci., vol. 3, no. 1, pp. 58–81, 2017.

[19] M. Mosbah and B. Boucheham, “New algorithm for re-ranking based on the vote operation in the context of CBIR,” in Computer Applications & Research (WSCAR), 2014 World Symposium on, 2014, pp. 1–7.

[20] M. Mosbah and B. Boucheham, “Majority voting re-ranking algorithm for content based-image retrieval,” in Research Conference on Metadata and Semantics Research, 2015, pp. 121–131.

[21] M. A. Hearst and J. O. Pedersen, “Reexamining the cluster hypothesis: scatter/gather on retrieval results,” in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996, pp. 76–84.

[22] J. M. Pena, J. A. Lozano, and P. Larranaga, “An empirical comparison of four initialization methods for the k-means algorithm,” Pattern Recognit. Lett., vol. 20, no. 10, pp. 1027–1040, 1999.

[23] M. Mosbah and B. Boucheham, “Re-ranking in the Context of CBIR: A Comparative Study,” in World Conference on Information Systems and Technologies, 2017, pp. 297–307.

[24] M. Mosbah and B. Boucheham, “Relevance feedback within CBIR systems,” Int. J. Comput. Inf. Sci. Eng., vol. 8, no. 4, pp. 19–23, 2014.

[25] J. J. Rocchio, “Relevance feedback in information retrieval,” SMART Retr. Syst. Exp. Autom. Doc. Process., pp. 313–323, 1971.

[26] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, “Ontology-based photo annotation,” IEEE Intell. Syst., vol. 16, no. 3, pp. 66–74, 2001.

[27] H. Elghazel, K. Idrissi, C. Ben Amar, and A. Baskurt, “Approches textuelles pour la recherche d'images,” in 3ème conférence internationale sur les Sciences Electroniques, Technologies de l’Information et des Télécommunications" SETIT2005, 2005, pp. 27–31.

[28] H. Abed and L. Zaoui, “Système D’Indexation et de Recherche d’Images par le Contenu.,” in CIIA, 2009.