A graph-based approach for text query expansion using pseudo relevance feedback and association rules mining

Siham Jabri, Azzeddine Dahbi, Taoufiq Gadi
Laboratory Informatics, Imaging and Modelling of Complex Systems, Faculty of Science and Technology,
Hassan 1st University Settat, Morocco

Article Info

Article history:

Received Nov 17, 2018
Revised Mei 31, 2019
Accepted Jun 27, 2019

Keywords:

Association rules
Dominance relations
Information retrieval
Pseudo-graph feedback
Query expansion
TREC

ABSTRACT

Pseudo-relevance feedback is a query expansion approach whose terms are selected from a set of top ranked retrieved documents in response to the original query. However, the selected terms will not be related to the query if the top retrieved documents are irrelevant. As a result, retrieval performance for the expanded query is not improved, compared to the original one. This paper suggests the use of documents selected using Pseudo Relevance Feedback for generating association rules. Thus, an algorithm based on dominance relations is applied. Then the strong correlations between query and other terms are detected, and an oriented and weighted graph called Pseudo-Graph Feedback is constructed. This graph serves for expanding original queries by terms related semantically and selected by the user. The results of the experiments on Text Retrieval Conference (TREC) collection are very significant, and best results are achieved by the proposed approach compared to both the baseline system and an existing technique.

Corresponding Author:

Siham Jabri,
Laboratory Informatics, Imaging and Modelling of Complex Systems,
Faculty of Science and Technology, Hassan 1st University,
577 Casablanca Road, Settat, Morocco.
Email: si.jabri@uhp.ac.ma

1. INTRODUCTION

The Information Retrieval (IR) domain is as old as the computers themselves, its systems are originally designed in order to automate the documents management by storing a collection of them as index, then retrieving information for mapping the user's query to a set of associated documents. With the advent of the Internet, the volume of documents and the number of people to manage have increased exponentially and valued at hundreds of millions. As result, the web search has become a standard source of information finding. This growth of data was and still is a big challenge for information retrieval systems.

Most queries are short and ambiguous for describing the relevant documents that meet the user information needs. This is the term mismatch problem in which the indexers and the users don't use the same words for describing the same idea. One of the successful techniques to handle the problem of term mismatch is to reformulate the original query by adding related terms that describe the user need and have not been mentioned, this process is called Query Expansion (QE).Query Expansion may be done in different ways: manual, interactive and automatic. Interactive query expansion process that involves both the system and user is better than the automatic process, but it is not feasible to involve the user in most of the time [1, 2]. The most popular technique in the literature is to define words in a vector space and giving weights to them. Rocchio *et al* [3] proposed a classical relevance feedback model to find text similarity and identifying relevant and non-relevant documents. Other methods for relevance feedback and ranking used contextual and word similarity modelled as co-occurrence [4-12].

Pseudo relevance feedback (PRF) is one of useful techniques to ameliorate retrieval performance. It obtains the expansion terms or phrases from the top ranked retrieved documents in response to a given query. However, if the documents used for this relevance feedback are irrelevant, the selected expansion terms impact the retrieval performance negatively [13]. Ariannezhad *et al* [14] proposed a new approach which consider that the documents containing more informative terms for PRF should have higher relevance scores. Moreover, an iterative algorithm is provided for ensuring the satisfaction of the proposed constraint for any PRF model. In this regard, the algorithm calculates the feedback weight of terms and the relevance score of feedback documents, simultaneously. Singh *et al* [15] presented a new fuzzy logic-based QE method for document retrieval based on PRF techniques. This approach combined the Borda, Condorcet and reciprocal weights of candidate expansion terms and produced a single fuzzy weight for every candidate expansion term. Then the degree of importance of a relevant term is calculated, and the higher this degree, the higher the chance to select relevant terms for query expansion. For filtering out irrelevant terms from candidates, the Fuzzy logic-based semantic similarity algorithms are used. Colace *et al* [16] introduced a new term extraction method for query expansion. The initial query is expanded with a structured representation made of weighted word pairs extracted from a set of training documents (relevance feedback). Bouziri et al [17] proposed a query expansion approach based on association rules between terms. The expansion is modelled as supervised classification problem and solved using a supervised learning algorithm. For this purpose, a training set is generated using a genetic algorithm-based approach that explores association rules space for retrieving the best expansion terms and generating a training instances that are used to build a classifier implementing decision tree algorithm. In our previous work [18], a query expansion approach based on an external structured knowledge resource namely Wikipedia, Explicit semantic analysis (ESA) and association rules technique has been proposed. The semantic interpretation ESA has been used for building the expansion graph. Then we calculated a new semantic relatedness measure that combines an association rules technique, semantic measure and the expansion graph avoiding the inclusion of irrelevant terms.

In this paper, another a query expansion technique is introduced using pseudo relevance feedback and association rules for building a Pseudo-Graph Feedback in order to expand queries by semantically related terms selected by the user. The contributions of this work are organized as follows:

a. A set of retrieved documents in response to the original query is selected and judged to be relevant for generating association rules using a technique based on dominance relations [19].
b. The extracted rules allow to discover the strength correlations between query terms and the candidate ones, to then construct an oriented and weighted graph called Pseudo-Graph Feedback.
c. To avoid the integration of non-similar terms in the expanded queries, the user is invited to select from the built graph the most related terms describing his information need.

The remainder of this paper consists of the proposed approach analysis presented in Section 2, results and discusson reported in section 3 and the conclution is given in the last part.

2. PROPOSED METHOD ANALYSIS

In this section, the proposed approach for query expansion based on pseudo relevance feedback and association rules is described. The approach consists of building, from the retrieved documents in response to a given a query, the semantic graph, called Pseudo-Graph Feedback, which represents the candidate expansion terms. Roughly, three main steps are carried out. The system architecture of the query expansion is illustrated in Figure 1. The first step concerns association rules generation where the vector space model is used for ranking text documents according to the given query [20]. For then applying an association rules algorithm based on dominance relation that will be detailed later. This phase allows to discover the strength correlations between document terms and original query. The second phase used the generated association rules as data source for building a graph called Pseudo-Graph Feedback. As third step the best expansion terms are extracted from the generated graph by the user avoiding the inclusion of inadequate terms.

2.1. Associaiton rules generation

The idea is to use the TF-IDF of vector space model to find an initial set of most relevant documents for a given query, to then estimate that the top k ranked documents are relevant without any user interaction. This process is called Pseudo Relevance Feedback, it allows to automate the manual part of relevance feedback. The selected documents are used to generate association rules using an algorithm based on dominance relations. It allows to rank association rules according to a real value and to find the most relevant rules among very large datasets. This algorithm uses a combination of a set of measures and not only one [19]. An illustrative example of association rules algorithm principle is presented in Table 1. Supposing that Measures = {Support, Confidence, Lift, Jaccard, GI}. The rule “R1” strictly dominates the second rule “R2” because R1(Support) = 240, R1(Confidence) = 0.84, R1(Lift) = 18,35, R1(Jaccard) = 0,73 and

R1(GI) = 10.35 which are all (pair by pair) bigger than R2(Support) = 70, R2(Confidence) = 0.72, R2 (Lift) = 10.24, R2(Jaccard) = 0,56 and R2(GI) = 2.90. Similarly, R2 dominates R3.

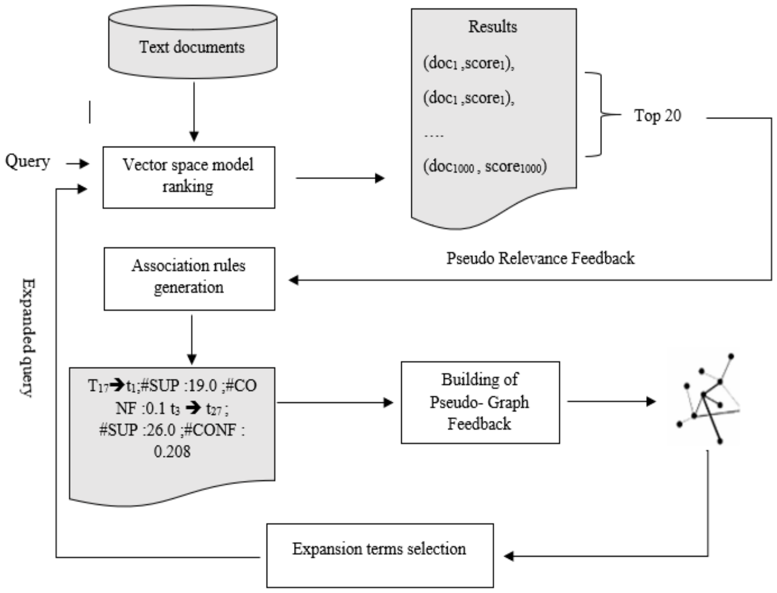
Text documents
Query
Vector space model ranking
Results
(doc₁ ,score₁),
(doc₁ ,score₁),
....
(doc₁₀₀₀ , score₁₀₀₀)
Top 20
Pseudo Relevance Feedback
Association rules generation
Expanded query
T₁₇➔t₁;#SUP :19.0 ;#CONF :0.1 t₃ ➔ t₂₇ ; #SUP :26.0 ;#CONF : 0.208
Building of Pseudo- Graph Feedback
Expansion terms selection

Figure 1. The proposed query expansion process

Table 1. Association rules examples

R	Supp	Conf	Lift	Jaccard	GI
R1: colombia ➔cocaine	240	0.84	18.35	0.73	10.35
R2: acid➔rain	70	0.72	10.24	0.56	2.90
R3: aids➔virus	84	0.71	1.52	0.55	0.42

Association rules generation process for a given query is executed as in the following steps:

a. Step 1: Preprocessing is an essential phase in text mining process. This step transforms the data source contents into a format that will be more effectively processed by subsequent steps. So, the document's contents are tokenized and only text is kept. After that stop words such as common words, prepositions and illegal characters are filtered, and the sentences are identified. Then the algorithm of Porter [21] for English text is used for stemming inflected or derivational words to their root form.
b. Step 2: For constructing the transactional dataset each keyword is considered as item, the transactions are represented by the sentences and the document in which the occurred sentence represents transaction elements.
c. Step 3: Transactional dataset is imported, and the referenced algorithm [19] is applied, it executes Apriori algorithm [22] to find the frequency of itemsets and generates all association rules. Finally, significant measures to evaluate and rank the obtained rules are calculated.
d. Step 4: Ranking of irredundant association rules.

2.2. Building of pseudo-graph feedback

The proposed graph called Pseudo-Graph Feedback is based on the generated rules in the first phase. This graph determines the candidate expansion terms, and the relations between them and the original query. The aim of the Pseudo-Graph Feedback is to transform the user query into a structured query that can be mapped to known terms. So, in this second phase the influence among association rules items is considered to find the adequate terms. The algorithm to build this oriented and weighted graph Gpgf = (V,E,w) takes as input a set of rules R={R_1,R_2,R_3,..,R_m},which are selected among the generated rules in the first phase. In these rules, terms are correlated. Logically, when a term t_i is related to the initial query, the term t_k which is correlated to t_i in some rules, should also be related to the query. Thus, any association rule R_j from R must contain at least one query term or a correlated term with the query term, and its confidence must be greater than a certain threshold. The set of nodes V in the Pseudo-Graph Feedback is the set of distinct terms t in R. Each term represents a graph node, and the relatedness between two terms represents an edge. Given two terms t,t' ∈ V ,they are connected with a directed edge if there is at least one rule R_j from R in which t and t' are located in the premise and the conclusion respectively. In other words, the set of edges E is formed as:

$$E=\{\ (t,t')\ |\exists R_j \in R\ /\ t\in R_j(premise)\ \wedge\ t'\in R_j(conclusion)\ \} \qquad (1)$$

The key aspect of the construction of Pseudo-Graph Feedback is to define the weighting function w : E → [0 ,1] as the maximum of the confidence of any association rule R_j from R, which contains the two vertices t and t' in the premise and the conclusion respectively.

$$w\ (t,t') = \max_{t,t'\in E,R_j\in R} \text{Confidence } (\mathrm{R}_j(\mathrm{t,t'})) \qquad (2)$$

For example, Figure 2 illustrates a possible Pseudo-Graph Feedback resulting from the generated association rules for the query "Water pollution".

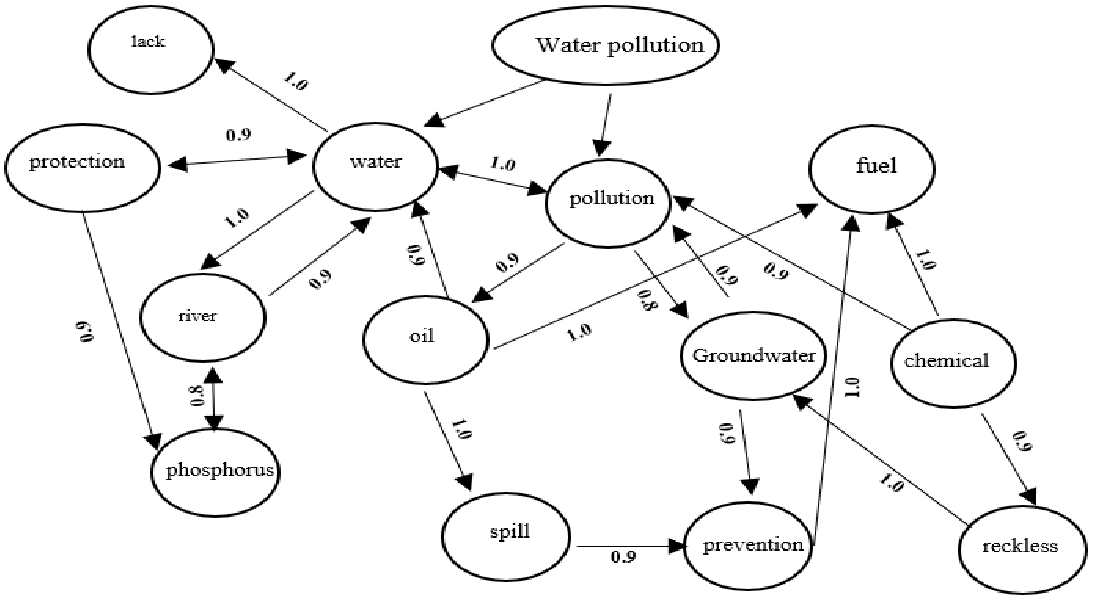

Figure 2. A portion of the pseudo-graph feedback using the association rules for the query "water pollution"

2.3. From pseudo graph feedback to the expanded query

Once a Pseudo-Graph Feedback is built, the number of candidate terms generated still is too large for expanding the short user query. The user can influence the expanded query by selecting adequate terms and ignoring bad ones. So, to avoid the inclusion of large number of terms which can negatively influence the information retrieval performance, the user is asked to provide feedback information by selecting the terms that better satisfy his information need. It is simple for him to determine which of the available terms better describes his interest. The Pseudo-Graph Feedback is labelled by sets of terms extracted from the association rules generated from the documents in the answer set. it allows the system to manage ambiguities. Once the user has selected the related terms to the query from the graph, the terms are added to the original query and the reformulated query is processed.

3. RESULTS AND DISCUSSION

In this last section, the experimental studies to test the retrieval effectiveness is presented. The dataset on which the runs are conducted and the evaluation metrics used to test the approach are described, then the obtained results are discussed.

3.1. Test collection and evaluation metrics

The collection TREC AP8890 chosen to apply the proposed approach is a set of english news articles published by Associated Press (1988-1990).The collection contains 242 918 documents with 150 topics wich represent the queries and a relevance judgments file made by domain experts. Only 50 titles of the TREC topics are used as queries for simulating search scenarios where users tend to submit short queries. In this work, the collection of documents is indexed using Lucene [23], which is an open-source Java full-text search library. The same library is then used for retrieving the top 1,000 documents, for each query using the TF-IDF of the vector space model [24]. The following metrics are used for evaluating the information retrieval performance of the proposed approach by comparing the responses of a system according to a query with a relevance judgement [25]:

a. Precision: measures the proportion of relevant documents among all documents retrieved by the system.
b. Recall: measures the proportion of relevant documents among all relevant documents in the database.
c. MAP: Mean average precision, which measures the area underneath the entire recall precision.
d. recip_rank: the rank of the first relevant document.

Each query in TREC collection is expanded with the expansion terms selected by the user from the Pseudo-Graph Feedback. The expanded queries are answered by the information retrieval system based on Lucene [23]. For the baseline method, the original queries are interrogated without any expansion. The following runs are conducted and the generated responses are evaluated:

a. Baseline: The original queries without any expansion.
b. PGF-approach: Query expansion based on the Pseudo Graph Feedback and user interaction for terms selection.
c. 0-Filtering: Query expansion based on the Pseudo Graph Feedback without any user interaction.
d. PRF: The classical Pseudo-Relevance Feedback technique implemented using Lucene.

The parameters for the experiments have been set experimentally as follows:

a. The number of text documents used in association rules is fixed to 20 documents retrieved at the top of results.
b. The values of measures used in association rules generation are determined by taking minimal values for not excluding any important rule: minSupport =1; minConfidence=0.1; minLift=0.1; minJacard=0.1; minGI=0.1;
c. The confidence threshold for terms selection from the generated association rules in Pseudo-Graph Feedback building is empirically set to 0.7.
d. The number of expansion terms selected by the user have been set to 3 to 5 terms at most, that allows to get the best results, because in this case the queries are short.

The aim principal of this work is to present simple form of information to the user in order to select the adequate terms for query expansion. For this reason, we proposed the Pseudo Graph Feedback based on association rules which describes the vocabulary terms related to a given query and the relations between them. In order to evaluate the performance of this proposed approach, it is recommended to compare it with recent query expansion approaches based on association rules. But, despite of using the same data collection and the same approaches, contradictions in results are detected which prevent a fair comparison due to use of large variety of configuration parameters like stemming algorithms, stop words filtering, ranking models, etc. Therefore, for comparing the proposed approach, the same search engine Lucene is used for implementing a method proposed by authors [17] and detailed in introduction section, using the same parameters values and data set TREC AP8890.

3.2. Results and discussion

Table 2 shows the different values of the Mean Average Precision (MAP), and the rank of the first relevant document (recip_rank) obtained by the system without and with using the proposed expansion techniques. For each query the MAP, recip_rank and the rate of improvement compared to the baseline (MAP-Gain) are calculated. Regarding the results obtained and summarized in Table 2, it can be seen that the proposed query expansion technique achieves a significant improvement in terms of MAP and recip_rank compared to the baseline and other runs (0-Filtering, PRF and AG-approach).

Table 2. Comparison of the runs with respect to the baseline and an existing algorithm

Run	MAP	recip_rank	MAP-Gain
PGF-approach	0,2004	0,5109	86%
AG-approach [10]	0.184	0.4454	71%
PRF	0,134	0,4046	25%
0-Filtering	0,1365	0,4071	27%
Baseline	0,1076	0,3685	-

The increase of MAP means that whenever the query contains more relevant query expansion terms the number of relevant documents is increasing. This seems clear in PGF-approach, the rate of improvement is +86% than the baseline. While the AG-approach achieved 71% of improvement. So, the use of the Pseudo Graph Feedback with the user interaction for expansion terms filtering improves the retrieval effectiveness. In the other hand, the MAP of 0-Filtering looks better than the baseline and PRF, they have approximately 27% and 25% improvement over the baseline respectively. This means that the terms composed the Pseudo-Graph Feedback even without filtering process are more relevant for reformulating the queries than PRF terms retrieved by the classical Pseudo Relevance Feedback process.

Figure 3 presents the precision when X documents are retrieved (P@X). X denotes the proportion of relevant documents in the top X documents in the returned list for a given queries. X is set to 5, 10, 15, 20 and 30 respectively. It is observed that using the Pseudo-Graph Feedback and the user interaction for selecting the expansion terms, leads to the improvement of the retrieval effectiveness when compared to the baseline and other approaches. The user interaction in this approach ensured that the extended queries containing adequate terms.

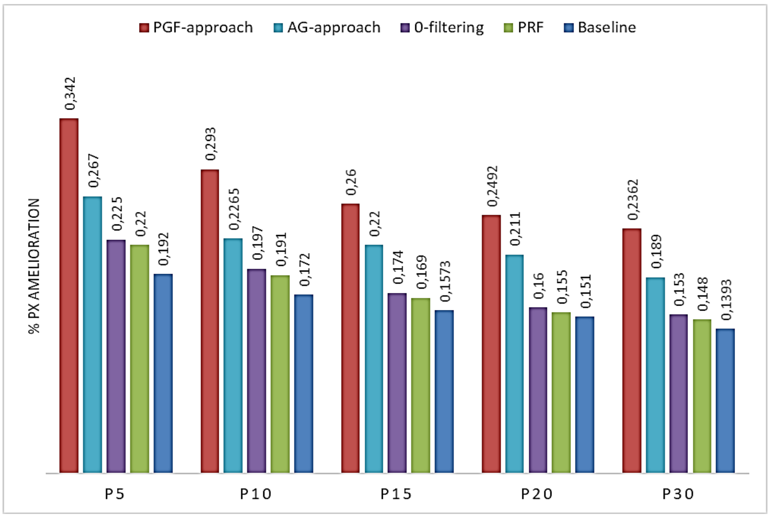

Figure 3. Improvement percentage in P@X

For example, the PGF-approach precisions are 0,342 and 0.293 for the top five and top ten retrieved documents respectively, while the baseline brings only 0,192 (+78%) and 0.172 (+70%). For the AG-approach 0.267(+39) and 0.2265(+32), and for 0-filtering approach 0.225 (+17%) and 0.197(+15%). These experiments offer the advantage to rank the relevant documents according to queries in the top of results.This performance could be explained by the use of the association rules technique based on multiple criteria for building the Pseudo Graph Feedback. This algorithm is efficient to rank and and keep only important rules by considering multiple measures and dominance relations. The valued graph is a simple and structured form of informations representing the correlations between query terms and the candidate's ones, and the edges represent the semantic relations between them. The number of expansion terms can be too large for short queries engendering low performance. For ensuring that the expanded queries will contain the adequate terms, the generated graph is presented to the user for selecting the best expansion terms. This phase has a positive impact for expanding the queries with adequate terms and improving the retrieval effectiveness. For AG-approach, presented by bouziri *et al* [10] adds terms extracted from the association rules to the original queries. The rules generated by Charm algorithm from the whole documents collection are modelled as classification problem and resolved by the decision tree algorithm for detecting the best terms for query expansion. Despite of the precision in the process of selecting relevant terms, irrelevant ones can be added to the original query.

4. CONCLUSION

The proposed query expansion approach expands queries with terms selected by the user from Pseudo-Graph Feedback. This graph is built using the association rules generated by a technique using multiple criteria and dominance relations. The experimental study was conducted on TREC AP8890 collection. Our method leads to significantly improved retrieval performance, and exceeds the baseline significantly. In terms of Mean Average Precision (MAP) the proposed approach has approximately 86% over the baseline, although the improvements attained by the comparison method don’t outperform 71%. This confirms that the association rules techniques is a significant way to present a simple and structured form of information’s to the user in the form of graph for selecting adequate terms for expansion. As perspectives, other data sources and text mining algorithms will be used for selecting and ranking query expansion terms.

REFERENCES

[1] D. Pal, *et al*., “Exploring query categorisation for query expansion: A study,” *arXiv preprint* arXiv:1509.05567, 2015.

[2] C. Buckley, *et al*., “The effect of adding relevance information in a relevance feedback environment,” *SIGIR’94*, Springer, London, pp. 292-300, 1994.

[3] J. Rocchio and G. Salton, “The SMART retrieval system,” *Relevance feedback in information retrieval*, pp. 313-323, 1971.

[4] A. Ilgarriff, *et al*., “Itri-04-08 the sketch engine,” *Information Technology*, vol. 105, pp. 116, 2004.

[5] Y. Matsuo and M. Ishizuka, “Keyword extraction from a single document using word co-occurrence statistical information,” *International Journal on Artificial Intelligence Tools*, vol. 13, pp. 157-169, 2004.

[6] E. Terra and C. L. Clarke, “Frequency estimates for statistical word similarity measures,” *Proc. The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistic, vol. 1, pp. 165-172, 2003.

[7] G. Cao, *et al*., “Selecting good expansion terms for pseudo-relevance feedback,” *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 243-250, 2008.

[8] S. E. Robertson and K. S. Jones, “Relevance weighting of search terms,” *Journal of the American Society for Information science*, vol. 27, pp. 129-146, 1976.

[9] Y. Lv and C. Zhai, “Positional relevance model for pseudo-relevance feedback,” *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 579-586, 2010.

[10] J. Allan, “Relevance feedback with too much data,” *SIGIR*, vol. 95, pp. 337-343, 1995.

[11] S. Yu, *et al*., “Improving pseudo-relevance feedback in web information retrieval using web page segmentation,” *Proceedings of the 12th international conference on World Wide Web*, pp. 11-18, 2003.

[12] S. Jabri, *et al*., “Ranking of text documents using TF-IDF weighting and association rules mining,” *2018 4th International Conference on Optimization and Applications (ICOA)*, pp. 1-6, 2018.

[13] C. Macdonald and I. Ounis, “Expertise drift and query expansion in expert search,” *The sixteenth ACM conference on Conference on information and knowledge management, ACM*, pp. 341-350, 2007.

[14] M. Ariannezhad, *et al*., “Iterative Estimation of Document Relevance Score for Pseudo-Relevance Feedback,” *European Conference on Information Retrieval. Springer*, Cham, pp. 676-683, 2017.

[15] J. Singh, *et al*., “Fuzzy logic hybrid model with semantic filtering approach for pseudo relevance feedback-based query expansion,” *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on. IEEE*, pp. 1-7, 2017.

[16] F. Colace, *et al.*, “Improving relevance feedback-based query expansion by the use of a weighted word pairs approach,” *Journal of the Association for Information Science and Technology*, vol. 66, pp. 2223-2234, 2015.

[17] A. Bouziri, *et al*., “Learning query expansion from association rules between terms,” *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on, IEEE*, pp. 525-530, 2015.

[18] S. Jabri, *et al*., “Improving Retrieval Performance Based on Query Expansion with Wikipedia and Text Mining Technique,” *Int. J. Intell. Eng. Syst*, vol. 11, pp. 283-292, 2018.

[19] A. Dahbi, *et al*., “A new method for ranking association rules with multiple criteria based on dominance relation,” *Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of. IEEE*, pp. 1-7, 2016.

[20] G. Salton, *et al*., “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, pp. 613-620, 1975.

[21] M. Porter, “PorterStemmer (java version) [Software],” 1980.. Available: https: //tartarus.org/ martin/PorterStemmer/index-old.html.

[22] R. Agrawal, “Fast algorithms for mining association rules,” *20th int. conf. very large data bases, VLDB*, pp. 487-499, 1994.

[23] Lucene. Available: http:lucene.apache.org/core.

[24] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, pp. 513-523, 1988.

[25] A. Baccini, *et al*., “Analyse des critères d'évaluation des systèmes de recherche d'information. Technique et Science Informatiques,” vol. 29, pp. 289-308, 2010.

BIOGRAPHIES OF AUTHORS

Siham Jabri is Business Intelligence Engineer, graduted from the faculty of science and technologies (Hassan First University of Settat Morocco) in 2014. Since 2015, she is preparing her Ph.D in the Laboratory of Informatics, Imaging and Modeling of Complex Systems (LIIMCS). She is working on Natural Langage Processing and Datamining.

Azzeddine Dahbi got his Bachelor degree in computer science in 2010 from the faculty of science and techniques university Hassan 1st Settat, Morocco. Followed by a Master degree in mathematics and application from the same faculty. Now preparing his Ph.D degree in the Laboratory of Informatics, Imaging and Modeling of Complex Systems (LIIMCS). His research interests include knowledge discovery from database.

Taoufiq Gadi is a Professor on computer science at the faculty of science and technologies (Hassan First University of Settat Morocco). Since 2014, he is the Director of the Informatics, Imaging and Modeling of Complex Systems Laboratory. He has conducted more than tens PhD theses and written a fifty of scientific papers in the domain of 3D models analysis, models Driving Architecture, Datamining and Database Analysis, Modeling of Complex Systems.