International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 6, December 2022, pp. 6350~6362
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i6.pp6350-6362

Knowledge graph-based method for solutions detection and evaluation in an online problem-solving community

Houda Sekkal[1], Naïla Amrous[2], Samir Bennani[1]
[1]Laboratory of Research in Computer Science and Education Mohammadia School of Engineers, Mohammed V University Rabat, Rabat, Morocco
[2]Laboratoire Mediation-Information-Knowledge-Society, School of Information Sciences Rabat, Rabat, Morocco

Article Info

Article history:

Received Jul 9, 2021
Revised Jul 14, 2022
Accepted Aug 10, 2022

Keywords:

Community mining
Information usefulness
Knowledge extraction
Knowledge graph
Knowledge management
Knowledge representation
Online communities

ABSTRACT

Online communities are a real medium for human experiences sharing. They contain rich knowledge of lived situations and experiences that can be used to support decision-making process and problem-solving. This work presents an approach for extracting, representing, and evaluating components of problem-solving knowledge shared in online communities. Few studies have tackled the issue of knowledge extraction and its usefulness evaluation in online communities. In this study, we propose a new approach to detect and evaluate best solutions to problems discussed by members of online communities. Our approach is based on knowledge graph technology and graphs theory enabling the representation of knowledge shared by the community and facilitating its reuse. Our process of problem-solving knowledge extraction in online communities (PSKEOC) consists of three phases: problems and solutions detection and classification, knowledge graph constitution and finally best solutions evaluation. The experimental results are compared to the World Health Organization (WHO) model chapter about Infant and young child feeding and show that our approach succeed to extract and reveal important problem-solving knowledge contained in online community's conversations. Our proposed approach leads to the construction of an experiential knowledge graph as a representation of the constructed knowledge base in the community studied in this paper.

Corresponding Author:

Houda Sekkal
Laboratory of Research in Computer Science and Education in Mohammadia School of Engineers, Mohammed V University Rabat
Avenue Ibn Sina B.P 765, Agdal Rabat 10090, Morocco
Email: houda.sekkal@gmail.com

1. INTRODUCTION

Online communities have become one of the principal sources of practical knowledge among users seeking for problem solving knowledge in the web [1]–[3]. Those communities are a gathering of members discussing and solving common interest problems and issues in different domains. This makes online communities room for the creation and gathering of special type of knowledge called problem solving knowledge. Indeed some researchers argue that online social interactions in online communities facilitate learning, expertise sharing, problem solving and innovation [4]. Many authors have defined problem solving knowledge as a way of knowing and understanding through direct engagement and being in the situation [5] [6] and as a truth learned from personal experience with a phenomenon [7]. Problem solving is a primary vehicle for better understanding of our environment, learning, and discovery of new opportunities [8]. This

type of knowledge is what distinguishes online communities as they contain discussions of users solving problems of other members through mutual efforts and the presence and testimony of persons who have solved similar problems [7]. The process of solving issues through online discussions leads to the accumulation of big amounts of valuable knowledge in the online community. Users need to seek, reuse and exploit this problem-solving knowledge to benefit from it in different contexts. In [9], Tanis focused on information seeking in online health communities and found that there are several reasons behind seeking information in online communities as to look for emotional support, inclusion, support others or pass time. According to [10], seeking information from peers in online health communities is a new way of pursuing health by banding together and sharing knowledge. Despite all those information seeking needs, actual platforms are limited in terms of extracting knowledge from the online communities. According to [11], existing methods for knowledge and information seeking in online communities lack of relevance and reliability as they are based only on full-text search or topics search which the process is shown in Figure 1. This is due to the limited matching terms in the query to the indexed post without consideration for the context introduced by these terms [11].

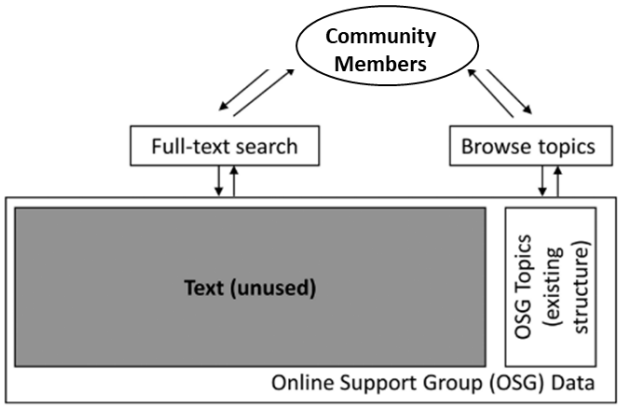

Figure 1. Existing methods of information retrieval from OSGs [11]

The full-text search does not enable seekers to rapidly get the answer for their questions or even to evaluate the usefulness and accuracy of the results. It only returns raw textual data as a result and push the seeker to spend more time and effort to look for the right answer to his question among all the online community conversations. Information usefulness as a topic is addressed in the literature from different perspectives. Generally, most studies discussed information usefulness in the context of business intelligence [12] and more specifically in business internet services [13]. Evaluating information usefulness shared in online communities is a critical issue specially in this area of big data and knowledge sharing in social media. In [14], the author proposed an approach for classifying the usefulness of solutions shared in knowledge communities. In his method, the author used a text analytic framework to extract important features from online forums in order to evaluate the information usefulness. In another work, Daradkeh [12] examines the usefulness of user generated content with large amounts of redundant information in order to analyze the digital voice of users in online open innovation communities. Generally, used approaches for information usefulness evaluation are based on extracting many features from online discussions and classify the usefulness of either threads or posts [14]. Such features concerns for example relevancy, objectivity, timelessness an completeness of each post [12], [14]. Despite the critical need for evaluating information usefulness in online communities it is noticeable that few studies have tackled this issue. In our work, we propose a new approach for evaluating information usefulness based on knowledge graph technology. According to [15] a knowledge graph is a structured representation of facts, consisting of entities, relationships, and semantic descriptions. Knowledge graphs enable the representation of knowledge and reasoning inspired by human problem solving [16]. Knowledge graphs are generally used in many real-world applications such as recommendation systems and question answering tasks with the ability of commonsense understanding and reasoning [15]. In our case, we built a process for extracting and evaluating best solutions to problems in an online community using knowledge graph technology and more specifically link prediction method used for knowledge graph completion. The goal is to reveal hidden semantic links between entities enabling to evaluate them. To the best of our knowledge, our work is the first to propose such an approach for information usefulness evaluation.

Finding useful solutions proposed and discussed in online communities is the key goal behind our research. In our study context, we are interested on extracting useful solutions to issues discussed in an online health community about breastfeeding and to evaluate the best one according to the member's experiences. According to [14] there is a strong motivation to identify useful solutions automatically in online communities to better serve users' knowledge needs and bolster the success of these communities. The main goal of our study is to conceive a process capable of capturing, distilling and evaluating automatically problem-solving knowledge. We attempt to apply our process to an online community discussing breastfeeding issues. The main goal is to extract the problems encountered by new mothers and the proposed "working" solutions for each problem. The contribution this study is to demonstrate how a textual content analysis-based process and a graph analysis-based technique can be applied to textual online community's conversations to extract problem-solving knowledge and to evaluate it. The scope of this article is limited to self-help-group concerning breastfeeding issues. The purpose is to build a knowledge graph to link problem's type entities to solution's type entities and then to infer solutions of each problem in the graph. Finally, we deduce the best solution for each problem. Existing studies about information usefulness evaluation are based on complexes and time-consuming computational methods that do not take into consideration graph theory techniques. Our approach provides simple and accurate method for evaluating best solutions to problems with measurable score. To the best of our knowledge, our study is the first to provide processes to evaluate breastfeeding problem-solving information shared in online communities.

The reminder of this paper unfolds: we firstly present the main theoretical basis related to information and knowledge usefulness evaluation in online communities. We then introduce our proposed research method to extract and evaluate best solutions for specific issues related to breastfeeding from online community's discussions. We then describe the experimental results and the discussions. Finally, we present the conclusion and future work.

2. METHOD

2.1. Modelization of the problem-solving knowledge creation process in online communities

Modeling the process of problem-solving knowledge creation (PSKC) in an online community facilitates its understanding for extraction and capitalization. As we are focusing on problem solving knowledge, it is important first to understand the process of the creation of such content. As online support groups offer an environment of connection, friendship, information sharing, and an increase in confidence among their members [17], members expose unresolved problems and seek for solutions from experienced members. This leads to the creation of problem-solving knowledge based on problem solving process which can be represented as shown in Figure 2.

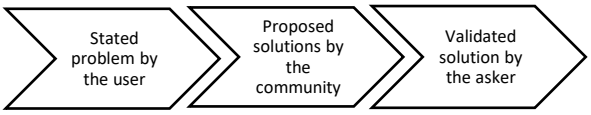

Figure 2. Modelization of the problem-solving knowledge creation (PSKC) process in online communities

According to Figure 2, PSKC process can be created starting by the definition of its steps: stated problem by the user, proposed solutions by the community members and the asker feedback of the best solution. For example, a member asks a question to describe a specific encountered problem. The community members start proposing solutions which are based on personnel experiences or heard from others. The asker is then in front of one or many proposed solutions to be tested. At this level, the asker can test the solution and replay to the community by giving feedback of the best tested solution. In many cases and situations, the asker may not give his feedback about the best tested solution that worked for him. In this case, the process of knowledge creation will be composed basically of the first two steps: stated problem and proposed solutions. In order to constitute a representation of problem-solving knowledge in online communities, we propose in the next chapter the extraction process of problem-solving knowledge components in online communities (OC).

2.2. Problem-solving knowledge extraction (visualization and evaluation) in online communities

Our goal is to automate the extraction of problem-solving knowledge. As shown in Figure 3, we conceived a process of problem-solving knowledge extraction composed of three phases: keywords problem

and solutions detection, keywords problem and solution mapping and finally best solution evaluation phase. Each step of this process will be detailed in the following sub-sections.

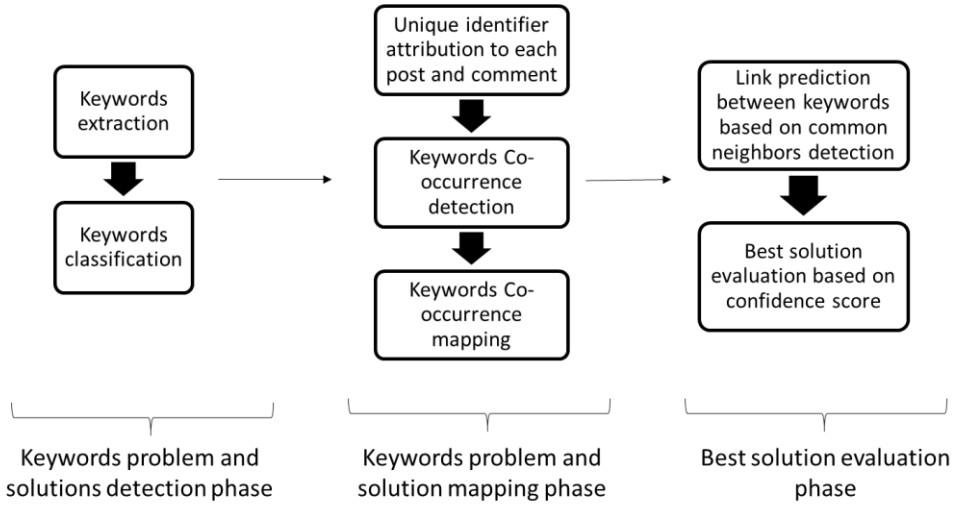

Figure 3. The process of problem-solving knowledge extraction in online communities (PSKEOC)

2.2.1. Keywords problem and solutions detection phase

Keywords are technical terms agreed and understood by the community that describe special issues, situations or solutions given by the members. Those terms are used and mentioned in online discussions by the members. The goal behind this phase is to automate the extraction and the classification of keywords representing problems and solutions.

a. Keyword's extraction

The objective is to extract those technical terms from user-generated content in online discussions. Our approach for keywords extraction is based on the occurrence frequency of the candidate terms that may represent keywords. Those keywords may be unigram terms or noun phrases. Keywords should be semantically rich and encapsulate the maximum of information in order to be representative. A domain-independent automatic term recognizer (ATR) is used for this step. Automated term extraction or recognition (ATE) from textual documents is a sub-field in text mining. It can be used for different important purposes, for example as inputs in ontology learning [18]. ATE aims to identify terminological units in specific domain corpora [19]. Section 3 present the details of the conducted experiment.

b. Keyword's classification

After keyword's extraction step comes keyword's classification. As our objective is the extraction of problem-solving knowledge composed of the stated problems and proposed solution, the extracted keywords are of two types: keywords describing encountered problems and keywords describing proposed solutions. Keyword's classification step aims to classify the extracted keywords into these two categories. We choose to build a machine learning classifier to classify our keywords based on a supervised learning approach as detailed in section 3.

2.2.2. Keywords problem and solution mapping phase

In this phase, we aim to visualize all the extracted keywords in order to infer the relationship between them. The relation between a keywords of type "problem" and a keywords of type "solution" that we want to infer is: "is a solution for". In the following, we describe our approach for keywords mapping phase.

a. Unique identifier attribution (UIA)

As shown in Figure 3, the first step of this phase is to assign to each post and comment created in the online community a unique identifier. This step will allow to our process the mapping of all the content shared in the online community and to detected the exact position (post or comment) where a keyword is mentioned in the online discussions. Figure 4 describe our UIA approach. In Figure 4, Pi represents a content of type "Post" whereas AjPi represents a content of type "Comment". For each post Pi there is many comments AjPi assigned which means that for a published post, there are many comments assigned.

P1	A1P1	A2P1	...	AmPn
P2	A1P2	A2P2	...	AmP2
P3	A1P3	A2P3	...	AmP3
...	...	...	...	...
Pn	A1Pn	A2Pn	...	AmPn

Figure 4. Unique identifier attribution to posts and answers of the online community

b. Keywords co-occurrence detection

This step consists of assigning to each keyword all the positions (unique identifier UI) of posts or comments where it is cited in the online community discussions. The goal is to link keywords with the UI previously created. To this end we developed an algorithm as shown in the following code (algorithm 1: mentions creation). According to the algorithm, the idea is to fetch in all posts and answers in our database each keyword found in step "Keyword's extraction" and then assign the cell address of the post or answer where it is found to the keyword. To do so, the algorithm is composed of three loops: the first one (line 1) browse every keyword to look if it is mentioned in posts (the second loop, line 2) and in answers (the third loop, line 3). If the first keyword is mentioned in a post or an answer (line 4) then the algorithm returns the cell address of the post or the answer (line 5). The algorithm generates a double column table where the first one represents all the keywords and the second one the UI (cell address) where each keyword was mentioned.

Algorithm 1. Mentions creation

```
For all keywords
For all posts (Pi) in the first column
For all answers (AjPi) from the second column to the last column
    if a keyword "Ki" is mentioned in a Post "Pi" or in an answer "AjPi"0
    Then return the cell address where the keyword "Ki" is found and assign it to the
keyword "Ki"
End if
```

c. Keywords co-occurrence mapping

Keyword's co-occurrence mapping step consists of building a keyword's graph to infer the relation between keywords of type problems with keywords of type solutions. The keyword's graph we are willing to construct is described: i) G (V, E) is an undirected graph where V is the set of nodes and E the set of edges; ii) the set of nodes V is composed of two types of nodes: a set of keyword nodes {Ki} and the set of posts and comments nodes {AjPi}; and iii) E is the set of edges between nodes that represents the occurrence of a keyword in a specific content (post or comment). As illustrated in Figure 5, i) K_p is a keyword node of type P (Problem) mentioned in three answers: A_iP_i and A_sP_i and AnPi. So, K_p has edges with nodes of type A_tP_i related to the post P_i and ii) K_s is a keyword node of class S (solution) mentioned in three answers: A_iP_i and A_sP_i and A_nP_i. So, K_s has edges with nodes of type A_tP_i related to the post P_i. The resulted graph is presented in the result section.

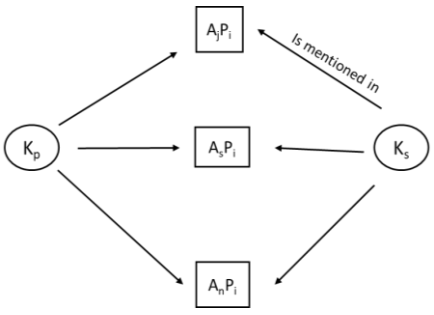

Figure 5. Illustration example of our keyword's graph

2.2.3. Best solution evaluation phase

In the previous phase, we detected the problems discussed and solutions proposed by the online community. As cited previously, for each post describing a particular problem, many solutions may be

proposed by the community. We need then to recommend to the community users the best solution for each problem. In this last phase, we present our method for detecting and evaluating the best solution for each problem mapped in the graph. This phase is composed of link prediction and best solution evaluation steps. Each step is detailed bellow.

d. Link prediction

After mapping keywords co-occurrences through the construction of a graph of keywords and posts and comments, the present step consists of inferring the relationship between keywords. This type of relation is not represented directly in the graph. This relationship between keywords we are aiming to infer from the graph will enable to detect the solutions proposed for a specific problem discussed in the online community. This approach will help users and members to find relevant solutions to encountered problems related to the subject discussed by the community as shown in Figure 6. For link prediction between keywords nodes, we use common neighbor's method from graph theory.

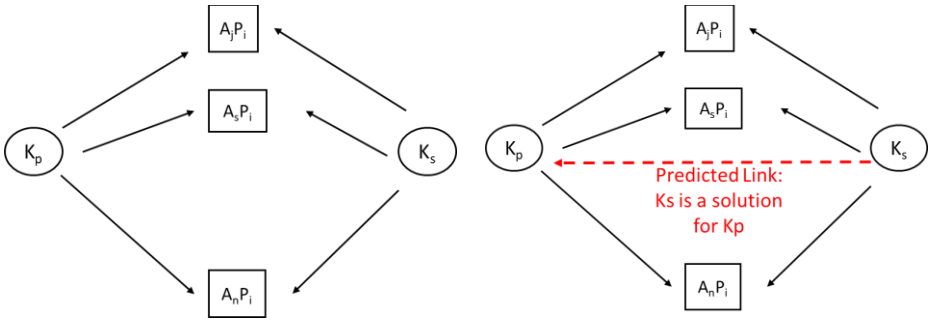

Figure 6. Illustration example of keyword's graph, dashed red line depicts possible link between keyword's nodes

Considering an undirected simple network *G (V, E)*, where *V* is the set of nodes and *E* is the set of links. For a pair of nodes *x*, *y*∈*V*, let *Γ(x)* denote the set of neighbors of *x*. According to common neighbors theory [20], [21], in common sense, two nodes, *x* and *y*, are more likely to have a link if they have many common neighbors. The simplest measure of this neighborhood overlap is the directed count, namely:

$$\mathrm{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (1)$$

In Figure 6, nodes K_p and K_s has many common neighbors which are A_jP_i, A_sP_i and A_nP_i, then according to link prediction approach based on common neighbors, there is an inferred link between K_s and K_p. The inferred link between the two nodes of different type namely "solution" and "problem" indicate that the keyword K_s is a solution for keyword K_p. Concretely speaking, if many members recommend the same solution many times (keyword node K_s) for the same problem (keyword node K_p) in a community, then necessary that solution is working. Our approach supports the fact that if the same problem keyword has many common neighbors with the same solution keyword, then it is highly the right solution for that problem. Now, as many solutions might be recommended for the same problem in the community, one may ask which is the best and working one? In the next sub-section, we present an approach for solutions evaluation based on a confidence score.

e. Best solution evaluation

In order to rank the best solutions proposed by the online community to a given problem, we propose to process a score for each pair of nodes of different type (problem/solution) based on common neighbors' theory. We say that the keyword K_s is a solution to keyword K_p to a certain degree of confidence called confidence degree (*Cd*). We propose (1) to calculate the confidence degree of K_s being a solution to K_p:

$$\mathrm{Cd}\left(\left(K_s|K_p\right)\right) = \frac{\mathrm{CN}(K_s, K_p)}{\mathrm{D}(K_p)} \quad (2)$$

where: $\mathrm{Cd}\left(\left(K_s|K_p\right)\right)$ the confidence degree of K_s being a solution for K_p, with K_p is a keyword node of type problem and K_s a keyword node of type solution; $\mathrm{CN}(K_s, K_p)$ is the number of common neighbors between nodes K_s and K_p (1); $\mathrm{D}(K_p)$ is the degree of node K_p meaning the number of all connections K_p has with all

nodes in the graph. In the case where a node of type problem K_p is not linked to any common nodes with a node of type solution K_s, then Cd=0. In the following section we expose the conducted experiment and test of our approach.

3. RESULTS AND DISCUSSION

In this section, we study the effectiveness of our proposed approach on a real online community forum of breastfeeding mothers. The goal behind this section is to test if our method succeeded at extraction valuable knowledge from raw instructed forum text. For this purpose, we compare the obtained results with the World Health Organization Publication about breastfeeding knowledge.

3.1. Dataset

Data was extracted from an online discussion's forum about breastfeeding publicly available using Octoparse tool. Octoparse is a powerful web scraping tool that can help extracting open data from almost all the websites and turn it to a well-structured file [22], [23]. In our case, we used Octoparse software to extract all posts and comments created in the online forum until the date of the data collection. The tool generated a cell surface vimentin (CSV) file containing two types of columns about online discussions: posts and answers (each column contains a comment to a specific post). In the generated CSV file, each line represents a post. As mentioned previously, the constituted dataset contains posts and answers generated by mothers sharing their problems and issues about breastfeeding. The generated CSV file contains 2,557 posts and 20,456 answers. All the data's characteristics are detailed in Table 1. In the following we give results about the conducted experiment.

Table 1. Dataset characteristics

Characteristic	Description
Text Language	English
Type of text	User-generated text in an online forum community
Number of posts	2,557 posts
Number of answers	20,456 answers

3.2. Results

3.2.1. Keywords problem and solutions extraction

After constructing our dataset composed of posts and their related answers, we start the keyword extraction step. Our goal is to extract meaningful keywords characterizing each problem and its solutions. In our context, we choose TerMine [24], a multi-word automatic term recognizer (ATR) that annotates the input text with candidate multiword terms recognized by the C-value method and acronyms. C-value is a domain-independent method for multiword ATR which aims to improve the extraction of nested terms as detailed in [24]. Nested terms are those that appear within other longer terms, and may or may not appear by themselves in the corpus [24]. This method takes as input a corpus and produces a list of candidates multi-word terms. The C-value approach combines linguistic and statistical information. The linguistic information consists of the part-of speech tagging of the corpus, the linguistic filter constraining the type of terms extracted, and the stop-list [24]. The part-of-speech tagger used for our dataset is the TreeTagger [25]. The statistical information consists of measuring the above characteristics: the total frequency of occurrence of the candidate string in the corpus, the frequency of the candidate string as part of other longer candidate terms, the number of these longer candidate terms and the length of the candidate string (in number of words). The resulting keywords database is composed of 69 keywords extracted from posts (questions) and 135 keywords extracted from answers as presented in Table 2.

Table 2. Keyword's extraction results

	Post's/question's text	Answer's text
Number of extracted keywords	69	135

The extracted keywords are bigram composed of two terms. Bi-gram keywords play a very important role in our knowledge extraction process because they allow us to capture more meaning and semantics than uni-gram terms. In order to obtain meaningful results, we decide to filter the extracted keywords and only keep those who's scores are greater than or equal to 2 [26]. The following step in our

approach concerns keywords classification. To this end, we conceived an implementation process based on machine learning classifiers presented in Figure 7 using Weka [27], an open-source machine learning software.

The first step of the process is about the dataset constitution. Our task consists of classifying keywords into two categories as mentioned previously. As we adopt a supervised machine learning approach, our test dataset of keywords should be annotated. The annotation process consisting of manually labelling the test data set with right classes was done by a breastfeeding experimented mother who manually classified the keywords according to solutions or problems. The second step consist of applying a *StringToWordVector* filter to convert text attributes into numerical attributes representing word occurrences.

Figure 7. Keywords classification process

The final step consists of applying the machine learning algorithms to train the classifiers. In our case we choose to apply and compare three machine learning classifiers: support vector machine (SVM) algorithm, multilayer perceptron (MLP) algorithm and DL4jMlp deep learning algorithm in WEKA platform. The results of this step are detailed in subsection 3.3.1. The performance of the applied classifiers was estimated based on a k-fold cross-validation. In our case, we tried many values for k and choose the value 10 because it generated the best result.

3.2.2. Keyword mapping, graph construction, and solution evaluation

After assigning a unique identifier (UI) to each post and comment in our dataset, we run the algorithm 1 presented in sub-section 2.2.2. The result is an excel file that we imported in Gephi [28], an open software for graph visualization. The resulted graph contains nodes representing keywords, posts and answers and edges representing mentions of co-occurrences. Figure 8 shows part of our resulted knowledge graph. The next and final step of the graph construction process aims to answer the following question: given a problem, what is, according to the community the solution for it? Best solution evaluation step is performed after applying (2) to each keyword pairs (problem/solution) in the generated graph. In order to give example of some results generated by the knowledge graph, we apply (2) to some keywords represented in the constructed graph as shown in Figure 8. The result is shown in Table 3.

As exposed in Table 3, the problem “clogged duct” has three possible solutions according to the community: “warm compress”, “hot shower” and “hot water”. This means that a breastfeeding mother may suffer from clogged duct which is an area of the breast where milk is blocked. According to the community shared experiences, this issue can be resolved by taking a hot shower or by putting a warm compress on the blocked area of the breast. To decide which of the solutions is effective, we applied our proposed confidence degree measure in (2). According to Table 3, “hot shower” is more likely to be a good solution for clogged duct issue with a confidence degree of 22.7% compared to “warm compress” where its confidence degree is 4.5%. In the same Table 3, solutions “nipple cream” and “Lactation consultant” are proposed by the community to the problem tongue tie. Tongue tie can cause serious nipple pain for the mother. According to calculated Cd, the best solution for this issue is to consult a lactation consultant which the Cd is 27.2%. For the problem “Nipple confusion” the only given solution by the community members is to use “Nipple Shield” which is a flexible silicone nipple that is worn over the mom’s nipple during a feeding. In the next sub-section, we present the evaluation method to assess the effectiveness of our approach in extracting and evaluating solutions to problems discussed in online communities about breastfeeding.

3.3. Evaluation and discussion

In order to evaluate our process of problem-solving knowledge extraction in online communities (PSKEOC) applied to breastfeeding knowledge, we choose to compare the obtained results with the Infant and young child feeding: model chapter [29] by the World Health Organization (WHO). It describes essential knowledge and basic skills that every health professional who works with new mothers and young children health care should master [29]. Most importantly, it contains in his 7th session details and solutions about how to manage breastfeeding difficulties. Our evaluation process aims to compare solutions to problems provided by this chapter to those extracted by our system and test the credibility of the obtained results. Our evaluation

process is composed of two parts: the first one concerns evaluating the extraction and classification step and the second concerns evaluating the attribution of correct solutions to extracted problems.

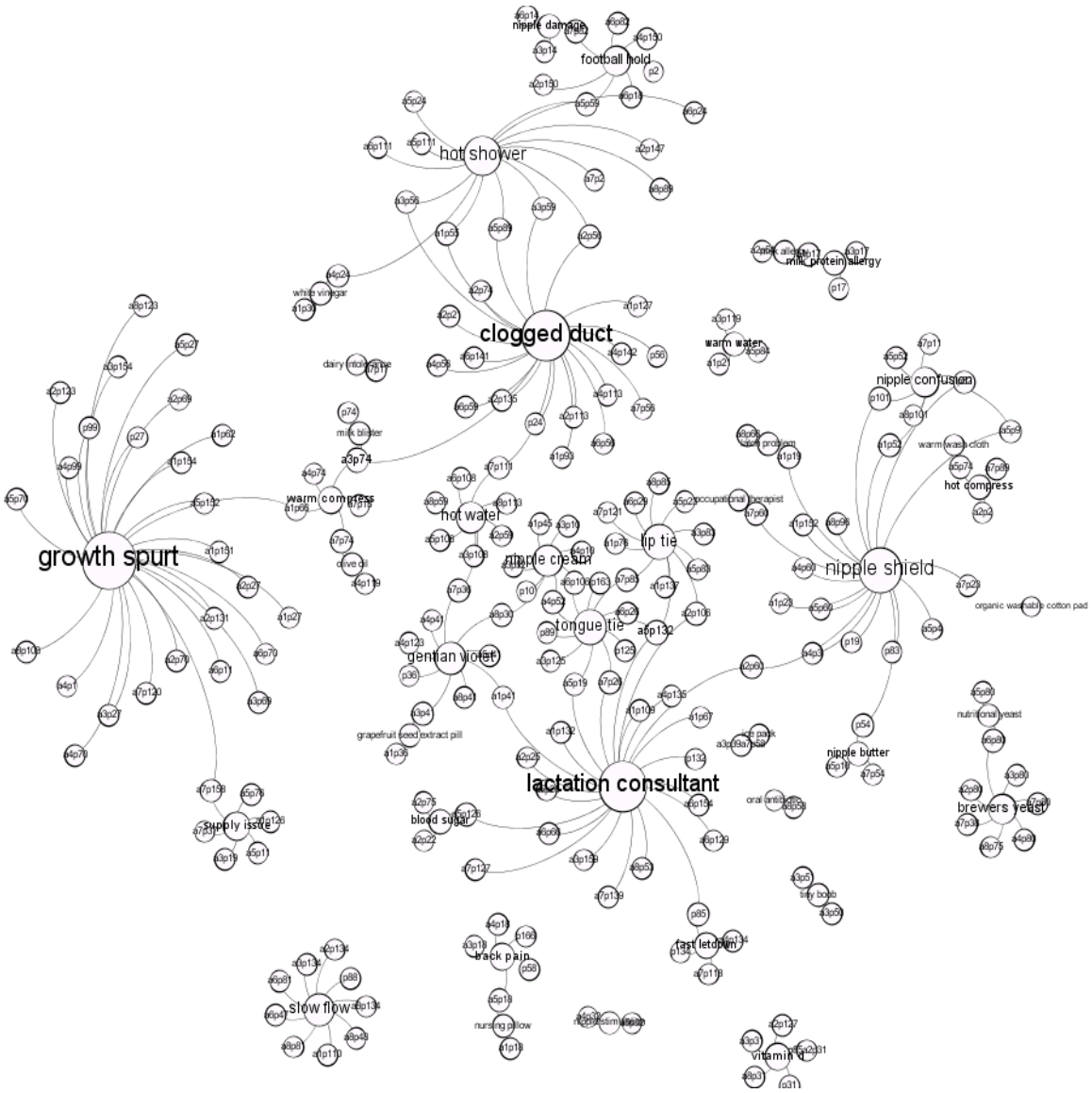

Figure 8. Problem solving knowledge graph

Table 3. Comparison of confidence degree value between solutions according to the obtained graph in Figure 8

Keywords of type Problem	Common neighbor nodes of type keywords (Solutions)	Cd*100 (confidence score*100)	Best deduced solution (with higher Cd)
"Clogged duct"	"Warm compress"	=1/22*100=4,5%	"Hot Shower"
	"Hot Shower"	=5/22*100 =22.7%	
	"Hot water"	=1/22*100 =4.5%	
"Tongue tie"	"Nipple cream"	=2/11*100 =18.2%	"Lactation consultant"
	"Lactation consultant"	=3/11*100 =27.2%	
"Nipple confusion"	"Nipple shield"	=3/18*100 =16.6%	"Nipple shield"

3.3.1. Evaluation of keywords problem and solutions extraction and classification

To evaluate our classification task, we use three metrics: precision, recall and F-measure. Precision measures the percentage of items that the system detected as in fact positive (according to human labels). Recall measures the percentage of items (posts) present in the input that were correctly identified by the system. F-measure incorporates aspects of both precision and recall [30]. As detailed in as detailed in sub-section 3.2.1, we trained three machine learning classifiers which are: SVM, MLP and deep learning-based classifier. The results are presented in Tables 4 to 6 which expose the tree evaluation metrics of the classifiers for each category.

Table 4. Classification evaluation metrics using SVM classifier

	Precision	Recall	F-measure
Category "Problem"	0.745	0.897	0.814
Category "Solution"	0.877	0.704	0.781

Table 5. Classification evaluation metrics using MLP

	Precision	Recall	F-measure
Category "Problem"	0.762	0.821	0.790
Category "Solution"	0.813	0.753	0.782

Table 6. Classification evaluation metrics using deep learning classifier

	Precision	Recall	F-measure
Category "Problem"	0.750	0.808	0.778
Category "Solution"	0.800	0.741	0.769

To compare the three classifiers' results, we processed the weighted average of each classifier as presented in Figure 9. According to Figure 9, SVM achieved the best results with 79.874% of correctly classified instances. This means that SVM classifier can reach good results when it comes to classify natural language text related to breastfeeding difficulties.

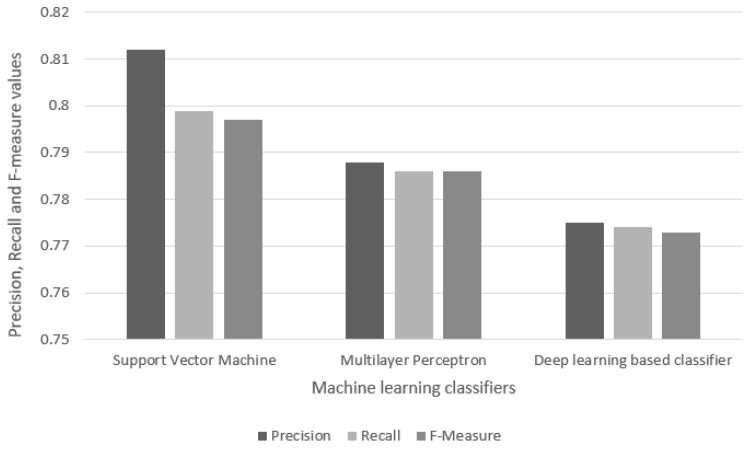

Figure 9. Evaluation metrics weighted average of the three classifiers

3.3.2. Evaluation of best solution evaluation phase

Table 7 presents an overview of the difficulties mentioned in the WHO reports about breastfeeding with the proposed solutions for managing them. Table 7 presents also extracted problems by our system with their extracted solutions as well. The goal behind this table is to compare both difficulties and solutions to conclude if our system succeeded in extracting solutions to each breastfeeding problem.

According to the Table 7, there are many difficulties extracted by our system that are discussed in the WHO chapter. For instance, our system succeeds to extract keyword difficulties like: irritated nipples, clogged duct, breast abscess, and slow flow. However, the same difficulties are not managed the same way in the community as by the WHO chapter. For example, according to the WHO model chapter, the issue "large nipples" should be resolved by "expressing breast milk" or by "taking a different position for breastfeeding". However, with the same issue, the online community recommend to apply a "nipple shield" to breastfeed the baby as shown in Table 7. Additionally, there are some breastfeeding difficulties discussed in the community that are not in the WHO Chapter like "growth spurt" issue.

Breastfeeding is a domain where new mothers are highly concerned as they need a lot of support in order to succeed breastfeeding their babies. Lacking of support while breastfeeding is considered as a serious problem that can interrupt breastfeeding. We believe that our process can help mothers and breastfeeding professionals as well. The obtained knowledge graph can be considered as a real knowledge base for the community as it can help both professionals and members of the online community to understand encountered problems and find most working solutions to specific issues related to breastfeeding domain. Our approach can also be helpful in the decision-making process as it provides a confidence score that characterizes the confidence degree of each possible solution based on lived experiential knowledge of members.

Table 7. comparison between difficulties and solutions mentioned in the WHO report and our PSKEOC process

Some difficulties mentioned in the WHO report	Proposed solutions by the WHO report	Some problem keywords extracted by our PSKEOC process	Proposed solutions to problems extracted by our PSKEOC process
Full breasts	Breastfeed the baby frequently to remove the milk	Huge breasts	-
Breast engorgement	remove the breast milk or remove the breast milk	-	-
Blocked duct: localized lump in one breast	– Feed from the affected breast frequently and gently massage the breast over the lump while her baby is suckling. – Apply warm com-presses, and to vary the position of the baby. – Gentle massage over the lump	Clogged duct	– Hot Shower – Hot compress
Mastitis	– Improve the removal of milk. – Apply warm compresses – Vary the position of the baby	Milk blister	Lactation nurse
Breast abscess	– Penicillinase-resistant antibiotics – Express milk	Breast abscess	-
Sore or fissured nipple	– Improve baby's position and attachment.	Irritated nipple	Coconut oil
Mastitis, abscess and nipple fissure in an HIV- infected woman	– Avoid breastfeeding on the affected side – Remove the milk from the affected breast by expression – Give antibiotics for 10–14 days		
– Candida infection (thrush) in mother and baby – White spots inside the cheeks or over the tongue	– Gentian violet or nystatin	Supply issue	– Pump session – Lactation consultant – Milk tea – Brewers yeast
Inverted, flat, large and long nipples	– The mother takes a different position for breastfeeding. – Express breast milk and feed it by cup	Flat nipple	Nipple shield
-	-	Growth spurt	– Hot shower – Licithine capsule – Birth control pill – Lactation consultant – Warm compress – Football hold and – Ice Pack.

4. CONCLUSION

Extracting, evaluating and facilitating the reuse of knowledge encapsulated in user generated content is an important and challenging task. Online communities are a medium for sharing knowledge components between users through asking and answering questions and resolving problems. Our work proposes an approach for solutions extracting and evaluation in an online community of breastfeeding mothers. Few studies have tacked this important question of capitalizing shared knowledge in online communities. Our presented method is based on knowledge graphs technology and graph theory approaches which enable representing extracted knowledge and reasoning on it. We developed a process for problem-solving knowledge extraction in online communities (PSKEOC) composed of three steps. The first one enables the extraction and classification of knowledge components from the online community conversations based on machines learning algorithms. The second step concerns the representation of knowledge components in a knowledge graph. The third step aims to evaluate and return best solutions to problems discussed in the

community using link prediction method. In order to evaluate the obtained results, we compared them to the WHO chapter on managing breastfeeding issues. The results show that our approach succeeded at extracting breastfeeding difficulties and their right evaluated solutions. Our process even extracted some difficulties that are not discussed in the WHO report but shared by the online community members with their corresponding solutions. The resulted knowledge graph about breastfeeding difficulties management can be considered as a real way for improving problem solving process in online communities and capitalizing its results. We aim in our future work to apply our process to many other datasets in different domain in order to evaluate the performance of our process in extracting problem solving-knowledge components in online communities. We aim also to enrich the obtained knowledge graph with more semantics in order to maximize the knowledge extraction, representation and capitalization.

REFERENCES

[1] Y. Tausczik and X. Huang, "Knowledge generation and sharing in online communities: current trends and future directions," *Current Opinion in Psychology*, vol. 36, pp. 60–64, Dec. 2020, doi: 10.1016/j.copsyc.2020.04.009.

[2] A. Zagalsky, D. M. German, M.-A. Storey, C. G. Teshima, and G. Poo-Caamaño, "How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists," *Empirical Software Engineering*, vol. 23, no. 2, pp. 953–986, Apr. 2018, doi: 10.1007/s10664-017-9536-y.

[3] J. Cole, C. Watkins, and D. Kleine, "Health advice from internet discussion forums: how bad is dangerous?," *Journal of Medical Internet Research*, vol. 18, no. 1, Jan. 2016, doi: 10.2196/jmir.5051.

[4] I. Buunk, C. F. Smith, and H. Hall, "Tacit knowledge sharing in online environments: Locating 'Ba' within a platform for public sector professionals," *Journal of Librarianship and Information Science*, vol. 51, no. 4, pp. 1134–1145, Dec. 2019, doi: 10.1177/0961000618769982.

[5] N. Nimkulrat, C. Groth, O. Tomico, and J. Valle-Noronha, "Knowing together-experiential knowledge and collaboration," *CoDesign*, vol. 16, no. 4, pp. 267–273, Oct. 2020, doi: 10.1080/15710882.2020.1823995.

[6] R. Julian, B. Bliesemann de Guevara, and R. Redhead, "From expert to experiential knowledge: exploring the inclusion of local experiences in understanding violence in conflict," *Peacebuilding*, vol. 7, no. 2, pp. 210–225, May 2019, doi: 10.1080/21647259.2019.1594572.

[7] T. Borkman, "Experiential knowledge: a new concept for the analysis of self-help groups," *Social Service Review*, vol. 50, no. 3, 1976

[8] P. H. Gray, "A problem-solving perspective on knowledge management practices," *Decision Support Systems*, vol. 31, no. 1, pp. 87–102, 2001, doi: 10.1016/S0167-9236(00)00121-4.

[9] M. Tanis, "Health-related on-line forums: what's the big attraction?," *Journal of Health Communication*, vol. 13, no. 7, pp. 698–714, Oct. 2008, doi: 10.1080/10810730802415316.

[10] Y. Zhao and J. Zhang, "Consumer health information seeking in social media: a literature review," *Health Information and Libraries Journal*, vol. 34, no. 4, pp. 268–283, Dec. 2017, doi: 10.1111/hir.12192.

[11] T. R. Bandaragoda, D. De Silva, D. Alahakoon, W. Ranasinghe, and D. Bolton, "Text mining for personalized knowledge extraction from online support groups," *Journal of the Association for Information Science and Technology*, vol. 69, no. 12, pp. 1446–1459, Dec. 2018, doi: 10.1002/asi.24063.

[12] M. K. Daradkeh, "Exploring the usefulness of user-generated content for business intelligence in innovation," *International Journal of Enterprise Information Systems*, vol. 17, no. 2, pp. 44–70, Apr. 2021, doi: 10.4018/IJEIS.2021040103.

[13] M. Nowakowski, "Comparative analysis of information usefulness evaluation methods on business internet services," *European Research Studies Journal*, vol. 23, no. 2, pp. 292–306, Nov. 2020, doi: 10.35808/ersj/1825.

[14] X. Liu, G. A. Wang, W. Fan, and Z. Zhang, "Finding useful solutions in online knowledge communities: a theory-driven design and multilevel analysis," *Information Systems Research*, vol. 31, no. 3, pp. 731–752, Sep. 2020, doi: 10.1287/isre.2019.0911.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.

[16] A. Newell, J. C. Shaw, and H. A. Simon, "Report on a general problem-solving program," 1969.

[17] T. Noorani, M. Karlsson, and T. Borkman, "Deep experiential knowledge: reflections from mutual aid groups for evidence-based practice," *Evidence and Policy*, vol. 15, no. 2, pp. 217–234, May 2019, doi: 10.1332/174426419X15468575283765.

[18] V. Kosa *et al.*, "Cross-evaluation of automated term extraction tools," 2017.

[19] M. da Silva Conrado, T. A. Salgueiro Pardo, and S. O. Rezende, "A machine learning approach to automatic term extraction using a rich feature set," in *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2013-Student Research Workshop*, 2013, pp. 16–23.

[20] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, Oct. 2009, doi: 10.1140/epjb/e2009-00335-8.

[21] X. Fan *et al.*, "Local core members aided community structure detection," *Mobile Networks and Applications*, vol. 24, no. 4, pp. 1373–1381, Aug. 2019, doi: 10.1007/s11036-018-0994-2.

[22] A. J. Arumugham, D. Ahamad, A. Mahmoud, M. Mahmoud, and M. Akhtar, "Strategy and implementation of web mining tools," *International Journal of Innovative Research in Advanced Engineering*, vol. 12, no. 12, pp. 1–7, 2017

[23] M. F. Shadiqin Thirafi and F. Rahutomo, "Implementation of naïve bayes classifier algorithm to categorize indonesian song lyrics based on age," in *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, Nov. 2018, pp. 106–109., doi: 10.1109/SIET.2018.8693201.

[24] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the C-value/NC-value method," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130, Aug. 2000, doi: 10.1007/s007999900023.

[25] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," 1994.

[26] H. Sekkal, N. Amrous, and S. Bennani, "Knowledge components detection in user-generated content," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Jun. 2020, pp. 1–6., doi: 10.1109/ISCV49265.2020.9204188.

[27] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining, fourth edition: practical machine learning tools and techniques*, 4th ed. Morgan Kaufmann, 2016.

[28] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *International AAAI Conference on Weblogs and Social Media*, 2009, vol. 3, no. 1, pp. 361–362.
[29] W. health organization, "Infant and young child feeding: Model Chapter," *World Health Organization*. 2009.
[30] S. Houda, A. Naila, and B. Samir, "Machine learning based knowledge organization in online communities," in *Proceedings of the 4th International Conference on Big Data and Internet of Things*, Oct. 2019, pp. 1–8., doi: 10.1145/3372938.3372965.

BIOGRAPHIES OF AUTHORS

Houda Sekkal is Data and Knowledge Engineer and a Ph.D. Candidate at the Laboratory of Research in Computer Science and Education, Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco. Her research project is about the use of knowledge management and data mining methods to improve the process of learning in online learning communities. She can be contacted at email: houda.sekkal@gmail.com.

Naïla Amrous is a full professor at School of Information Sciences. She has a PhD in Information and Communication Sciences. Ongoing research: information literacy (IL), knowledge management (KM); innovative teaching practices. She can be contacted at email: amrousnaila@yahoo.fr.

Samir Bennani is a Full Professor and Deputy Director of students and academic affairs at Mohammadia School of Engineers. Engineer degree in Computer Science in 1982; PhD in Computer Science in 2005; Professor at the Computer Science Department-EMI; 34 recent publications papers between 2014 and 2017; Ongoing research interests: SI, modeling in software engineering, information system, e-learning content engineering, tutoring, assessment and tracking. He can be contacted at email: sbennani07@gmail.com.