Egyptian Informatics Journal (2014) 15, 211–220

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Enhancements to knowledge discovery framework of SOPHIA textual case-based reasoning

Islam Elhalwany [a,*], Ammar Mohammed [a], Khaled T. Wassif [b], Hesham A. Hefny [a]

[a] Institute of Statistical Studies & Researches, Cairo University, Egypt
[b] Faculty of Computers and Information, Cairo University, Egypt

Received 20 January 2014; revised 2 October 2014; accepted 12 October 2014
Available online 18 November 2014

KEYWORDS

Textual case based reasoning;
Questions answering systems;
Text classification;
Text clustering;
Knowledge discovery

Abstract Many approaches were presented recently for developing Textual Case-Based Reasoning (TCBR) applications. One of the successful approaches is SOPHisticated Information Analysis (SOPHIA), which is distinguished by its ability to work without prior knowledge engineering, without domain dependency and without language dependency. SOPHIA is based on the distributional document clustering approach, which facilitates an advanced and rich knowledge discovery framework for case-based retrieval.

This paper contributes to propose enhancements to SOPHIA approach that aims to enhance the retrieval efficiency and increase the precision degree. It also aimed to grantee that all results will have the same subject of the user query. The enhancements include performing an automatic classification to the case-base before the clustering step in the indexing stage, and include performing an automatic classification to the user query before the retrieval stage. Moreover, proofing that SOPHIA approach is a domain and language independent by applying it in the domain of Islamic jurisprudence in Arabic language.

1. Introduction

Traditional Case-Based Reasoning (CBR) is an Artificial Intelligence (AI) technique to support the capability of reasoning and learning in advanced decision support systems. It is specifically a reasoning paradigm that exploits the specific knowledge collected on previously encountered and solved situations, which are known as cases [1]. It is an approach for problem solving and learning in which expertise is embodied in a library of past cases. Each case typically contains a description of the problem, plus a solution and/or the outcome. The knowledge and reasoning process used by an expert to solve the problem is not recorded, but is implicit in the solution [2].

Textual Case-Based Reasoning (TCBR) – as with traditional CBR – aims to compare a problem description with a set of past cases maintained in a case base with the exception that descriptions are predominantly textual. The final aim is to reuse solutions of similar cases to solve the problem at hand.

* Corresponding author.
Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2014.10.002

Clearly, ability to compare text content is vital in order to identify the set of relevant cases for solution reuse. However, a key challenge in the text is variability in vocabulary that manifests as lexical ambiguities such as the polysemy and synonymy problems reasoning [3].

The fundamental idea of this problem-solving paradigm is to collect experiences from earlier problem solving episodes and to reuse them for dealing with new tasks explicitly. In real life, many of the most valuable experiences are stored as textual documents [4], such as:

- Reports by physicians.
- Documentations and manuals of technical equipments.
- Frequently Asked Question (FAQ) collections.
- Informal notes and comments on specific functions and observed behavior.

Many approaches were presented for developing TCBR, all of them apply the CBR cycle but they mainly differ in knowledge representation and Knowledge retrieval stages. The differences are in the techniques used to implement these important steps by researches such as:

- Information Retrieval and Word Net techniques.
- Pure statistical approaches.
- Shallow NLP techniques with manually constructed domain specific ontology and a generic thesaurus.
- Shallow NLP with a nearest neighbor algorithm in a text representation.
- Keywords and Rank Computation and Maintenance Algorithms.
- Employment of bag-of-concept (BOC) approach to extend case representations by utilizing is-a relationships captured in the taxonomy.

A typical application area of TCBR is a Question Answering (QA) system that may be inspired from the hotline support or help desks. In such systems, user submits his inquiry then gets the closest answer to his inquiry. As a scientific definition, Question Answering (QA) is an application area of Computer Science which attempts to build software systems that can provide accurate, useful answers to questions posed by human user in natural language (e.g., English) [5].

Patterson et al. [6] presented SOPHIA; a new approach for TCBR stands for *SOPHisticated Information Analysis* for TCBR (SOPHIA-TCBR), based on the distributional document clustering approach [7]. It facilitates an advanced and rich knowledge discovery framework for case-based retrieval. It is based on the conditional probability distributions of terms within documents. It then intelligently discovers important themes within the case-base and organizes cases into a large number of clusters, which have these themes as attractors. This process of forming clusters, allows both a very efficient and competent case retrieval process. In addition to the automated discovery and utilization of textual cases, it discovers similarity knowledge, which allows the semantic meaning of the cases to be considered, thus enabling more meaningful similarity comparisons among cases. (This means that cases that are on the same or similar subjects but use different terminology can be recognized as similar.)

The contribution of this research is applying SOPHIA as one of the successful approaches in TCBR scientific area to Fatawa QA System and then modifying the approach to get an enhanced version of SOPHIA; this enhanced version will provide a better performance and a higher precision. The proposed Fatawa QA System inherits the main characteristics and advantages of SOPHIA-TCBR approach [6].

This is the first application of SOPHIA approach in the area of Arabic Islamic Fatawa; it could be proven that SOPHIA is a domain and language independent, scalable and therefore applicable for large case-bases.

In the next section, the domain of application will be discussed. Moreover, it will clarify challenges and restrictions for implementing systems that respond to Fatwa requester and automatically provide predefined answers. In Section 3, SOPHIA approach will be presented and the proposed Fatawa Question Answering System based on textual case-based reasoning will be introduced then the enhanced version of the proposed approach will be described. Section 4 will show the experiments, results and evaluation. Finally Section 5 will address the future work.

2. Domain of application

The domain of application is the *Islamic Fatawa*, the term Fatawa (Religious verdict) refers to seeking a legal ruling for religious issues that Muslims all over the globe pose on a daily basis. Fatwa is a couple of a question and an answer, which can be called as "a case", so Fatwa and a case have the same meaning. Although most of religious questions have multiple answers (opinions), there is an organization that can provide us with certified opinions on any question and can do researches to answer new issues that recently appeared due to development and change of time, place, people and circumstances. This organization is official religious organization for Fatwa in Egypt named *Egypt's Dar al-Ifta* (Fatwa House) which is considered among the pioneering foundations for Fatwa in the Islamic world. It was established in 1895 by high command of khedive Abbas Hilmi.

Due to the limitation of human resources, this organization could not handle a huge amount of questions daily, and that pushed people to get their answers from non-qualified or non-specialized scholars. So, the target is to help people to get the right answers from the right place by helping this organization to achieve its huge duty and increase its ability to answer inquiries by the proposed Fatawa QA System.

The proposed *Fatwa QA System* is an intelligent system that can respond to questions with the closest answers that already recorded before. It's fully automated system. The proposed system can reduce the need for submitting questions, if there are similar answers recorded. Over the passage of time, the need to pose questions will be limited to the new issues only. As a result, the organization can handle more and more users without human resources interference.

After many personal interviews with the scholars who work in the field of jurisprudence and Fatwa, talking about the specialty of this field, the following challenges were detected:

- Scholars working on this field believe that questions have to be manually handled one by one. This is because any misunderstanding may result in a wrong or imperfect answer; which is not accepted. They see that a human can do that job perfectly than the machine system.

- Some Fatawa cannot be generalized or be publically browsed, where the verdict is for a certain special case or it may cause instability in society.
- Some Fatawa have names, public figures, or private situations where reader can know who is the original owner of the case base question.
- The sensitivity of the field, where mistakes are not allowed and may lead to a public rejection from the Islamic scholars and all Muslims as well.
- Linguistic mistakes or poor linguistic expressions are not accepted.

To overcome the previous challenges, the following conditions and restrictions are imposed.

- Answers can never be automatically generated or modified or depend on syntax combinations.
- Language accuracy must be at the highest level.
- Case question and answer are always coupled and answers should not be automatically exchanged between cases.
- Remove Fatawa that cannot be publicly browsed from the case base.
- Remove names or any indication to the original inquirer in case base.
- The new system teaches user how to perfectly ask his question, and offer him all relate cases. Additionally, if the user doubts about the result, he can submit his questions. That's beside, users already do that by nature, when they browse Fatawa websites, or watch religious TV channels, they always apply what they read or watch on their own cases.

3. Related work

Based on analyzing research contributions, there are many approaches were developed to implement TCBR. The differences among TCBR approaches may exist in the two main steps of CBR cycle:

- *Retain:* Knowledge representation (indexing and clustering).
- *Retrieval:* (similarity assessment).

The differences reside in the techniques used to implement these important steps.

Some researchers use Information Retrieval and WordNet techniques, while others use pure statistical approaches. Researchers likewise may use Shallow NLP techniques, a manually constructed domain specific ontology and a generic thesaurus. Others prefer Shallow NLP with a nearest neighbor algorithm in a text representation. Some researchers use Keywords, Rank Computation, and Maintenance Algorithm. Moreover, others employ the bag-of-concept (BOC) approach to extend case representations by utilizing "is-a" relationships captured from the taxonomy. All of these approaches are domain and language dependent and require a lot of knowledge on engineering work. Following are the summaries of some researchers' contributions:

Burke et al. [8] developed FAQ-Finder, a question–answering system, they uses techniques that combine statistical and semantic knowledge. This system starts with a standard information retrieval approach based on the vector space model. Cases are compared as term vectors with weights based on a term's frequency in the case versus in the corpus.

In addition, FAQ-Finder includes a semantic definition of similarity between words, which is based on the concept hierarchy in WORDNET; a semantic network of English words provides a system of relations between words and synonym sets and between synonym sets themselves.

Lenz et al. [4] presented the CBR-Answers Project, another question–answering system that compares textual cases through the meanings of terms. The program processed the free text components to identify Information Entities (IEs), which are indexing concepts that may occur in text in different forms, mapping from texts to sets of IEs where an IE may be more than just a keyword.

This approach requires some domain-specific knowledge engineering to identify task-specific terms, which may include product names or physical units.

FAllQ's similarity assessment checks word similarity using two lexical sources: a manually constructed domain specific ontology and a generic thesaurus. Case Retrieval Nets, which support FAllQ's retrieval strategy, represent the case base as a network of IE nodes where similarity arcs connect nodes with similar meaning. Retrieval is performed by propagating activation through this network. FAllQ differs from FAQFinder because FAllQ is a domain specific, while FAQFinder is a domain independent.

Brüninghaus and Ashley [9] present methods that support automatically finding abstract indexing concepts in textual cases and demonstrate how these cases can be used in an interpretive CBR system to carry out case-based argumentation and prediction from text cases. They implemented these methods, which predict the outcome of legal cases using the given textual summary. Their approach uses classification-based methods for assigning indices. They compare different methods for representing text cases, and consider multiple learning algorithms. They also show that combination of some background knowledge and Shallow NLP with a nearest neighbor algorithm in a text representation leads to the best performance for TCBR task. They introduced a program called SMILE + IBP; it stands for "*SMart Index LEarner + Issue-Based Prediction*". It uses CBR to predict the outcomes of legal disputes inputted directly as text and to explain those predictions.

Han et al. [10], introduced a Q&A system. They put forward an interactive and introspective Q&A engine, which uses keywords of the question to trigger the case and sorts the results by the relationship. The engine can also modify the weights of the keywords dynamically based on the feedbacks of the user. Inside the engine, they use *feature-weight maintenance algorithm* to increase the accuracy. They also extend the 2-layer architecture of CBR to a 3-layer structure to make the system more scalable and maintainable. They split this representation into three levels: a feature level corresponding to feature values F, a problem description level corresponding to P and an answer level corresponding to S.

Recio-García and Wiratunga [11] presented a novel approach to acquiring knowledge from Web pages. It focuses

on the primary knowledge structure that is a dynamically generated taxonomy. Once taxonomy was created can be used during the retrieve and reuse stages of the CBR cycle. They are interested in gathering taxonomy to capture the semantic knowledge in textual cases that cannot be obtained through statistical methods alone. Firstly, they propose to guide the taxonomy generation process using a novel CBR specific disambiguation algorithm.

Secondly, case comparison is improved by means of Taxonomic Semantic Indexing, a novel indexing algorithm that utilizes the pruned taxonomy.

They employ the bag-of-concept (BOC) approach to extend case representations by utilizing is-a relationships captured in the taxonomy. Results suggested significant performance improvements with the BOC representation and best results were obtained when taxonomies are pruned using their disambiguation algorithm.

Patterson et al. [6] presented SOPHIA; a new approach for TCBR stands for *SOPHisticated Information Analysis* for TCBR (SOPHIA-TCBR), based on the distributional document clustering [7] approach of SOPHIA, which facilitates an advanced and rich knowledge discovery framework for case-based retrieval. It is based on the conditional probability distributions of terms within documents. It then intelligently discovers important themes within the case-base and organizes cases into a large number of clusters, which have these themes as attractors. This process of forming clusters, allows both a very efficient and competent case retrieval process. In addition to the automated discovery and utilization of textual cases, it discovers similarity knowledge, which allows the semantic meaning of the cases to be considered, thus enabling more meaningful similarity comparisons among cases. (This means that cases that are on the same or similar subjects but use different terminology can be recognized as similar.)

Vattam and Goel [12] were interested in cross-domain TCBR. They have encountered this problem in the context of biologically inspired design (BID) – the invention of new technological products, processes and systems by analogy to biological systems. The needed biological knowledge typically is found in the form of unstructured textual documents, typically on the Web. Due to its growing importance, they posit that BID presents a great opportunity for exploiting and exploring cross-domain TCBR. They have developed a technique for semantic tagging of biology articles based on Structure–Behavior–Function models of the biological systems. They have also implemented the technique in an interactive system called Biologue; Controlled experiments with Biologue indicate improvements in both findability and recognizability of useful biology articles. Their work suggests that task-specific but domain-general model-based tagging might be useful for TCBR in support of complex reasoning tasks engaging cross-domain analogies.

4. Applying "SOPHIA-TCBR approach

SOPHIA-TCBR approach has several advantages such as; it is domain independent and does not require any user intervention to acquire domain knowledge. As such, all knowledge can be discovered automatically. It is also a language independent, scalable and therefore applicable for large case-bases, that's beside it can work with non-structured documents.

Knowledge is automatically discovered within the following five stages of the SOPHIA-TCBR framework:

- Case knowledge discovery.
- Narrow theme discovery.
- Similarity knowledge discovery.
- Case assignment discovery.
- Internal cluster structure discovery.

The following section will show how this approach can be useful in the proposed Fatwa QA System.

4.1. Fatawa Question Answering System

This research mainly contributes to proposing Fatawa QA System based on TCBR. Fatawa QA System is an intelligent system that can respond to a user's question with the semantically closest questions that already answered before. The proposed system can deduct when a new question is asked. The proposed system inherits the main characteristics and advantages of SOPHIA-TCBR approach [6].

Fatawa QA System cycle contains indexing, retrieval, and learning stages. Indexing includes many steps such as selecting Fatawa (cases) according to specific criteria to be used as *a case-base*, putting Fatwa in the form of *terms vector*, calculating term frequency, forming *terms contexts* by calculating the probability of randomly selecting the term y in a randomly selected case within which the term z co-occurs. After that, calculating the degree to which the term z represents its context to determine the terms that can be defined as *narrow themes*, then, *theme context* of narrow themes will be selected, and finally, every case in the case-base is assigned to the nearest theme context (cluster) to get the *clustered case-base*. Retrieval includes the same indexing steps but on the individual incoming question to retrieve its closest questions. Learning stage includes adding new questions to case-base and rebuilding indexes as well.

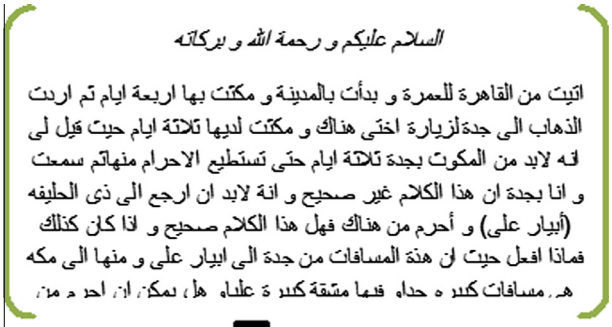
The following steps will describe in details how the proposed system was implemented:

1. Indexing and clustering processes started with applying stemming process as follows:
   - Removing diacritic characters, numbers, symbols and any non-Arabic characters.
   - Convert the multiform characters to a single form (ex $\acute{\imath}, \grave{\imath}, \hat{\imath}, \imath => \imath$).
   - Split question text into words.
   - Remove stop words; which is extremely common words such as pronouns and prepositions.

2. Calculate the term frequency and probability distribution for all terms of each case in the case-base table using the following formula.

$$P(y|X) = \frac{\mathrm{tf}(X,y)}{\sum_{t \in Y} \mathrm{tf}(X,t)} \tag{1}$$

where $tf(X, y)$ is the term frequency of the term y in document X, and Y denotes the set of all terms in All Cases. In Fig. 1, a sample question is displayed and a table of terms and its frequency.


السلام عليكم و رحمة الله و بركاته

اتيت من القاهرة للعمرة و بدأت بالمدينة و مكثت بها اربعة ايام ثم اردت الذهاب الى جدةلزيارة اختى هناك و مكثت لديها ثلاثة ايام حيث قيل لى انه لابد من المكوث بجدة ثلاثة ايام حتى تستطيع الاحرام منهاثم سمعت و انا بجدة ان هذا الكلام غير صحيح و انة لابد ان ارجع الى ذى الحليفه (أبيار على) و أحرم من هناك فهل هذا الكلام صحيح و اذا كان كذلك فماذا افعل حيث ان هذة المسافات من جدة الى ابيار على و منها الى مكه هى مسافات كبيرة جدا فهما مشقة كبيرة علياه هل يمكن ان احرم من

ID	FatwaID	Term	Freq
13154	518559	القاهره	1
13155	518559	للعمره	1
13156	518559	بدات	1
13159	518559	بها	1
13160	518559	اربعه	1
13161	518559	ايام	3
13163	518559	اردت	1
13164	518559	الذهاب	1
13166	518559	جده	4

Figure 1 Sample question and frequency table.

3. Calculating the probability distribution "$P(y|z)$" which equals to the probability of randomly selecting the term y in a randomly selected case within which the term z co-occurs, i.e. grouping of semantically related cases bound together by the specific theme.

This distribution can be approximated as you can see in the following equation:

$$P(y|z) = \frac{\sum_{x \in X(z)} \mathrm{tf}(X,y)}{\sum_{x \in X(z), t \in Y} \mathrm{tf}(X,t)} \tag{2}$$

where $X(z)$ is the set of all cases from the corpus which contain the term z.

4. Calculating the degree to which the term z represents its context or its theme where the entropy for every word context is calculated. The entropy is found on the word context conditional probability distribution by the following formula:

$$H(Y|z) = -\sum_{y} P(y|z)\log(P(y|z)) \tag{3}$$

"$H(Y|z)$" is used as a criteria for narrow contexts selection of a theme.

5. The whole set of words was divided into disjoint subsets according to the case frequency cf:

$$Y = \bigcup_i Y_i$$
$$Y_i = \{z : z \in Y, cf_i \leqslant cf(z) \leqslant cf_{i+1}\}$$
$$i = 1 \ldots r \tag{4}$$

where $Y(z)$ denotes the set of all different terms from cases in $X(z)$ and the case frequency $cf(z) = |X(z)|$ of the term z.

Here, the thresholds cf_i satisfies the condition $cf_{i+1} = \alpha cf_i$ where $\alpha > 1$ is a constant.

Choosing narrow word themes is based on the assumption that in total there are N narrow word themes and r case frequency intervals. For every $i = 1 \ldots r$ a set $Z_i \subset Y_i$, is selected such that:

$$|Z_i| = \frac{N \cdot |Y_i|}{\sum_{i=1 \ldots r} |Y_i|} \tag{5}$$

And

$$z_1 \in Z_i, z_2 \in Y_i - Z_i \rightarrow H(Y|z_1) \leqslant H(Y|z_2), \quad then\ Z = \bigcup_i Z_i \tag{6}$$

where Z is set of selected narrow themes. And N is set to 1000. Fig. 3 shows a table of some of the narrow themes and their case frequency.

6. Until this step, N groups of terms were created where every group or cluster has semantically related terms. Now there is a need to measure the similarity between cases in the case base and every cluster context that can be achieved by Jensen–Shannon divergence [13] between the probability distributions P_1 and P_2 representing the case and the theme, respectively as in Eq. (7):

$$\mathrm{JS}_{\{0.5,0.5\}}[P_1,P_2] = H[\overline{P}] - 0.5H[P_1] - 0.5H[P_2] \tag{7}$$

where $H[P]$ denotes the entropy of the probability distribution P and $\overline{P}$ denotes the average probability distribution $= 0.5P_1 + 0.5P_2$, where the lower value of JS divergence, the higher semantic similarity between case and its theme. JS is non-negative bounded function of P_1 and P_2, which is equal to zero if and only if $P_1 = P_2$.

7. In this step, cases are assigned to a cluster based on their semantic similarity to a theme. In Eq. (8), a case is assigned to a cluster $C(z)$ with attractor z if:

	Question	Answer	ZID	JSDist
1	الحلف بالله كذبا	الاستغفار والتوبة النصوح وإخراج كفارة ال...	594	0.50210970464135
2	بسم الله الرحمن الرحيم السلام عليكم ورحم...	تنتف العانة للمرأة أولى من حلقها، وليس ...	580	0.367765283206789
3	ما المقصود بان الدين ليس بالتعقل هل ١- الطا...	العقل فى الشريعة الإسلامية هو مناط ا...	974	0.412730293053437
4	السلام عليكم هل حرام الاحتفال بالمولد النبو...	يجوز الاحتفال بذكرى المولد النبوي الشر...	594	0.50210970464135
5	هل يوجد وصيلة أخرى لفك الإحرام فى العمرة ...	التحلل من الإحرام في العمرة لا يكون الا ...	594	0.50210970464135
6	هل يقضى الصلاة الشخص الذى اغمى عليه ا...	ذهب المالكية والشافعية ، وهو قول عند ا...	594	0.502109704641351
7	هل اذا أمت المرأه النساءفى الصلاة ولم تعلم ...	أخرج البخارى عن أبى هريرة أن النبى ص...	440	0.493996598501483

Figure 2 Case assignment to narrow themes "ZID" with the distance "JSDist".

ID	Term	CaseFreq	CaseInterval
63	زوجي	88	6
53	صيام	87	6
55	قمت	83	6
40	واحده	82	6
46	الاحرام	81	6

Figure 3 Narrow themes table.

ID	CaseAID	CaseBID	ZID	JDist
118638	341425	520333	698	0.745484265886677
118639	341425	520334	698	0.745484265886677
118640	341425	520537	698	0.745484265886677
118641	341425	520682	698	0.86999551567033
118643	348902	505252	856	0.788581588002531
118823	352589	479866	594	0.891530900168394
118907	352589	495012	594	0.893995028690807
119015	352589	504170	594	0.882239282720313

Figure 4 Distance between cases.

$$z = \text{argmin}_{t \in Z}\ JS_{\{0.5,0.5\}}[P(Y|x), P(Y|t)] \quad (8)$$

i.e., a case is assigned to the cluster whose attractor it has the highest semantic similarity to.

Fig. 2 displays a table of Fatawa subjects, questions, answers, narrow themes IDs "ZID", and Jensen divergence "JSDist".

8. Calculating the similarity between two cases from the same cluster can be achieved using the JS divergence. Such that, the lower JS divergence, the higher similarity and this is similarity knowledge which forms the key to discover semantically related cases.

In Fig. 4, a list of couples of cases from the same cluster "ZID" and JS divergence "JDist" between them.

9. Retrieval process can be initiated from Fatwa requester form depicts a typical interaction with Fatawa QA System.
10. Suppose the user enters his question. He has the option to submit his question directly to *Dar Aliftaa* backend team, or to check the case base first.

If the user preferred to check the case base first he will get two different search mechanisms, he can use both or either of them.

First, one is a textual search, which searches sequentially the whole case-base for the closest cases; it does not depend on any text clustering. This option matches only terms of question with terms of cases in the case base after applying stemming step on both of them. The matching result will be ordered by number of matched words and difference in terms count between them.

Second option is the similar cases search, which takes the benefit of clustered case-base. JS Divergence function is used to determine to which cluster the question will be assigned, and then check the most similar cases. Fig. 5 shows the user form, which allows the user to enter his question and to select between the two options.

11. Questions that marked as new will be added to case-base through the learning process and indexes will be periodically rebuilt.

4.2. Enhanced SOPHIA approach

In this section, the enhancement will be applied to SOPHIA approach. As discussed in previous sections, SOPHIA is unsu-

Figure 5 User form.

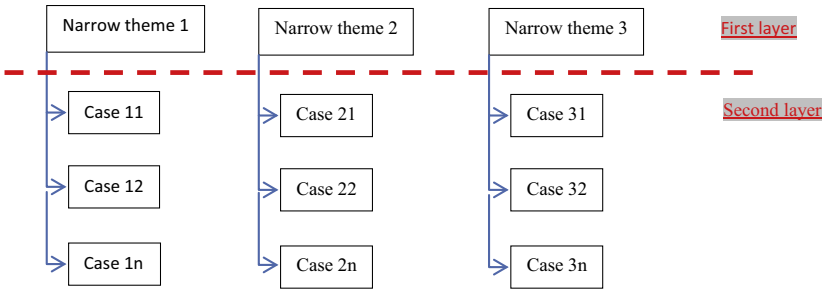


Figure 6 SOPHIA two layers model.

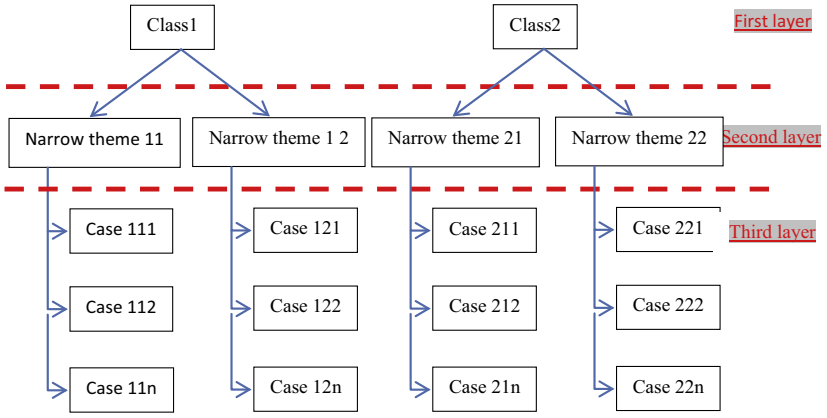


Figure 7 Three layers of the enhanced model.

pervised approach that does not require any knowledge engineering where it has a full automatic process for knowledge discovery. It is clear now that SOPHIA model consists of two layers; first layer is the narrow themes (attractors), and the second is the cases that assigned to narrow themes to form the clusters.

The two layers model will be extended to three layers model, where the first layer will be proposed as a new layer to contain the classes of cases, while the second layer will be the narrow themes, and the third layer will be the cases as before. Figs. 6 and 7 show the difference between the original SOPHIA as two layers model and the proposed enhanced SOPHIA three layers model. In Fig. 7 you will see a new layer contains classification of case-base then each class of cases is clustered.

The modification will be achieved by feeding the knowledge discovery process by a little predefined knowledge, which acts as a training set of cases, and let the system discover the rest of knowledge. This piece of knowledge will enhance the clustering process and will grantee balanced clusters which will cause a better performance for the system.

Luckily, there already exist classified Fatawa, which saved the classification effort and made it easy to extract the narrow themes and themes' contexts of the case-base regarding to their classification. The following steps will describe how this modification affects the original algorithm.

1. In the original algorithm, the whole set of words was divided into disjoint subsets as in Eq. (4). However, this equation will be modified to divide the whole set of words in the same way but according to their class instead of their case frequency.
2. In the original algorithm, there was a different entropy threshold for every interval addressed in Eq. (6). However, that will be modified to have the same entropy threshold where dividing words according to their class instead of their case frequency.
3. The third modification is to limit the size of narrow themes context to be 200 terms for each context instead of taking the whole context in the original algorithm. That will enhance the performance as well. Fig. 8 shows a sample of the classes and a sample of its narrow themes. The retrieval algorithm also was modified to take the benefits of the added knowledge (classification), where the class of user inquiry will be detected first, and then it will be compared to only clusters of this class. That modification will clearly affect the performance of retrieval process.

5. Experiments

Two experiments were run; first one used SOPHIA approach without any modifications, while the second used the enhanced version of SOPHIA approach.

5.1. Dataset

Dar Alifta Fatawa archive was used; this archive contains a great amount of Fatawa in nine different languages, these Fatawa have different statuses which define if Fatwa is new

Figure 8 Samples classes and their narrow themes.

(not answered yet), answered, under revision, under translation, refused, or completed (finished); These Fatawa also are classified in different classes such as Marriage, Prayers, Divorce, Purification, Banks, …. There were also many other details recorded in Fatawa archive.

Arabic Fatawa were selected, that were answered and classified to any class except Inheritance "Merath" were avoided, because it depends on calculations that make this kind of Fatawa inappropriate for TCBR. The total number of selected cases (Fatawa) was (7337) cases. These cases will be used as a training set where all steps described in the previous section will be applied on them. Table 1 shows the statistics of the dataset that were obtained from Egypt's Dar alifta archive.

5.2. Experiment output

The output of the experiments can be initiated from Fatwa requester form depicts a typical interaction with Fatawa QA System. The user has the option to submit his question directly to *Dar Aliftaa* backend team, or to check the case base first. If the user preferred to check the case base first he will get two different search mechanisms, he can use both or any of them.

First option is a keyword search, which does not depend on TCBR. This option matches only terms of question with terms of cases in the case base after applying stemming step on both of them. The matching result will be ordered by number of matched words and difference in terms count between them. Second option is the similar cases search which depends on TCBR. In the same manner, JS Divergence function is used to determine to which cluster the question will be assigned, and then check the most similar cases.

The output of the second experiment is faster than the first experiment because clusters are smaller, and here it grantees that all results will be from the same class which increases the precision due to the classification step.

5.3. Experiment results

The experiment environment included a notebook with Intel Core 2 Due 2.5 GHz, 4 GB Ram. SQL Server 2012 Enterprise Edition was used for the database and C# 2010 for the forms. For the first experiment; execution time ranges from 19 to 30 s depends on the size of cluster that contains the result. However in the second experiment; execution time ranges from 2 to 3 s depends on the size of cluster that contains the result.

5.4. Assessment of the classification process

In this section the classification process will be assessed to measure to what degree the automatic classification will match the manual expert classification; steps were achieved as follows:

Table 1 Cases count in each class.

Class English Name	Class Arabic Name	Cases Count
Marriage	الزواج	1214
Prayers	الصلاة	1073
Divorce	الطلاق	868
Purification	الطهارة	775
Banks	البنوك	736
Obligatory charity	الزكاة	682
Morals	الأخلاق	517
Hajj and minor pil	الحج والعمرة	407
Fasting	الصوم	287
Crimes, capital pu	الجنايات والحدود والكبائر	237
Vows and oaths	الأيمان والنذور	192
Funeral prayers	الجنائز	128
Stocks and marke	البورصة ومايتعلق بها	79
Insurance	التأمين	65
Maintenance	النفقة	52
Lineage	النسب	25

Figure 9 Clusters' sizes in applied SOPHIA approach.

Figure 10 Clusters' sizes in the enhanced SOPHIA approach.

The classification process starts with detecting narrow context for each class; this context has the highest frequency in the class and the lowest entropy values.

Jensen–Shannon divergence [13] technique is used to know the closest class to each case.

The test was run for the whole case-base cases (7337 case) that were already classified to compare between the manual classification of experts and the automatic classification.

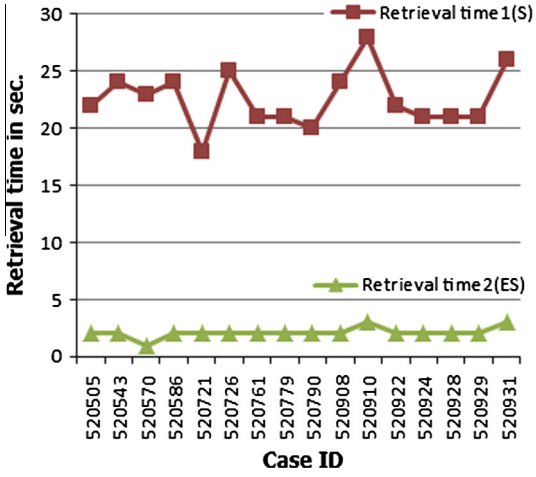


Figure 11 A comparison between SOPHI(S) and enhanced SOPHIA version according to retrieval time.

The result was around 6000 cases automatic classification matches the manual classification, the others were not; which means that automatic classification works correctly by 80%.

5.5. Evaluation

In this paper, two applications of TCBR were presented; one of them inspired of SOPHIA approach and the other was an enhanced version of SOPHIA approach. Based on results of the previous experiments, it was clear that both of the two applications can work, however the second one has a better result in the domain of application. Figs. 9 and 10 show a comparison between the clusters' sizes where it is clear that the modified version made clusters more balanced.

This difference in the clustering process affects the performance of retrieval process where the modified version provided a better performance that's beside adding classification guarantees that the results will be always related to user question which increases the precision.

Table 2 A comparison between SOPHI(S) and enhanced SOPHIA version according to retrieval time.

Case ID	Class	Retrieval time 1 (S) (s)	Retrieval time 2 (ES) (s)
520505	Vows and oaths	22	2
520543	Crimes, capital punishment, and major sins	24	2
520570	Stocks and markets	23	1
520586	Insurance	24	2
520721	Banks	18	2
520726	Morals	25	2
520761	Maintenance	21	2
520779	Hajj and minor pilgrimage	21	2
520790	Marriage	20	2
520908	Purification	24	2
520910	Divorce	28	3
520922	Prayers	22	2
520924	Funeral prayers	21	2
520928	Lineage	21	2
520929	Fasting	21	2
520931	Obligatory charity	26	3

Another performance test was run; a question from each class was randomly selected as a test set. The difference in retrieval time is clear as you can see in Table 2 and Fig. 11 as well.

6. Future work

Enhancing the clustering algorithm will be in consideration so that the retrieval process could be executed fast in case the case base is extended to include 100,000 cases. Also stemming process can be enhanced by using root extraction algorithms and removing words' prefixes and suffixes. Spelling checker may be used to help Fatwa requester to write his questions without spelling mistakes. Classification process also may be enhanced to be more accurate. In addition, synonyms problem will be in consideration.

References

[1] Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations and system approaches. AI Commun 1994;7:39–59.
[2] Abdrabou E, Salem A. Case-based reasoning tools from shells to object-oriented frameworks. In: Proceedings of international conference knowledge-dialogue-solution KDS, Varna, Bulgaria, June–July 2008.
[3] Recio-García J, Wiratunga N. Taxonomic semantic indexing for textual case-based reasoning. In: Proceedings of ICCBR; 2010. p. 302–16.
[4] Lenz M, Hubner A, Kunze M. Textual CBR. In: Case-based reasoning technology, from foundations to applications. London (UK): Springer-Verlag; 1998. p. 115–38.
[5] Freeucci D et al. IBM research report towards the open advancement of QA systems. RC24789 (W0904-093). Computer Science; 2009.
[6] Patterson D, Rooney N, Galushka M, Dobrynin V, Smirnova E. SOPHIA-TCBR: a knowledge discovery framework for textual case-based reasoning. Know-Based Syst 2008;21(5):404–14.
[7] Dobrynin V, Patterson D, Rooney N. Contextual document clustering. Proceedings of 26th European conference on information retrieval research, LNCS, vol. 2297. Springer; 2004. p. 167–80.
[8] Burke RD, Hammond KJ, Kulyukin V, Lytinen SL, Tomuro N, Schoenberg S. Question answering from frequently-asked questions files: experiences with the FAQ finder system. AI Mag 1997;18(1):57–66.
[9] Brüninghaus S, Ashley K. Reasoning with textual cases. In: Muñoz-Avila H, Ricci F, editors. Case-based reasoning research and development (LNAI 3620). Berlin: Springer; 2005. p. 137–51.
[10] Han P, Shen R, Yang F, Yang Q. The application of case based reasoning on a Q&A system. In: Proceedings of Australian joint conference on artificial intelligence AI-02; 2002. p. 704–13.
[11] Recio-García JA, Wiratunga N. Taxonomic semantic indexing for textual case-based reasoning. In: Proceedings of ICCBR; 2010. p. 302–16.
[12] Vattam S, Goel A. Biological solutions for engineering problems: a study in cross-domain textual case-based reasoning. In: Proceedings of ICCBR 2013, LNAI 7969; 2013. p. 343–57.
[13] Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theor 1991;37(1):145–51.