Journal of Computer Science

Original Research Paper

Improving Question Answering System based on a Hybrid Technique

[1]Ayatallah Gamal Abass, [1, 2]Sameh Abd El-Ghany and [1]Ahmed Abo Elfetoh

[1]Department of Information System, Faculty of Computers, Mansoura University, Egypt
[2]Department of Information System, College of Information and Computer Sciences, Al JOUF University, Saudi Arabia

Article history
Received: 09-09-2016
Revised: 31-03-2017
Accepted: 25-07-2017

Corresponding Author:
Ayatallah Gamal Abass
Department of Information System, Faculty of Computers, Mansoura University, Egypt
Email: ayagamal.011@gmail.com

Abstract*:* Question Answering (QA) is a specialized form of information retrieval characterized by information needs that are expressed as natural language statements or questions. Query expansion is an approach which is used to extend question key words with new related words that are not included in question. In this study, anew association rule based question processing model is introduced. This model is used for expanding question keywords with the most related terms using new hybrid association rule base. This hybrid rule base takes into account both the exact match of association rules and the hierarchical match of semantic similarity to overcome the mismatch problem between questions and answer words. Therefore, it contains association rules between document words at a semantic level. The main goal of this new model is to enrich the question by reformulating it into similar meaning queries with additional information and hence for improving question answering process.

Keywords: Association Rule Mining, Semantic Similarity, Query Expansion, Interesting Measure, Question Answering System

Introduction

Currently, the information available through the internet is gradually growing. Hence, accurate identification and extraction of a specific piece of information is becoming one of the most necessary functions for internet users. The most major way to obtain information is through information retrieval systems. This system accepts a user's query as input and gets back a set of documents sorted by their relevance to the query. Web search engines like (Google and yahoo) are one of the typical technologies which are used to perform the information retrieval task (Manning *et al*., 2009). Usually a search engine answers user's query expressed by a list of keywords with an ordered list of documents which are expected to contain the needed information. But most of them leave it to the user to extract the desired pieces of information from the retrieved ranked list of documents. As a result, users have to read a lot of returned pages to extract by themselves the information they need. This process usually is time consuming and the obtained information is not concentrative. Consequently, traditional information retrieval approaches became insufficient for finding and evaluating answers. The research of Question Answering (QA) intends to resolve this problem by allowing users to access knowledge resources and asking natural language questions then retrieving relevant answers in concise Words. The QA technology takes both information retrieval and information extraction a step further. It provides specific and brief answers to naturally formulated questions (Moldovan and Surdeanu, 2003).

Users always prefer to ask questions in their local language without being restricted to a certain query language, query formation rules, or even a particular knowledge domain. Moreover, they would like the discover answers to be short and precise. Depending on the user's capability in choosing the appropriate keywords, the result might be an empty list or a long list of documents and the user being supposed to look into these documents for getting the required correct information. This includes moving from more simple taxonomies of question to richer question analysis process. Thus, query expansion is one of the question enriching methods which have been used to get more relevant answers to the query.

Relationships among words can be a powerful tool to extract and get answers. Association rule mining

(Han *et al.*, 2011) is one of the most vital data mining techniques. Its objective is to extract hidden knowledge and correlations between words in data repositories; On the other hand, semantic similarity deals with computing similarity between conceptually similar but not necessarily lexically similar terms. It is computed by mapping terms (concepts) to an ontology and then use that ontology for examining their relationships. Usually, association rule mining considers only exact match between items in transactions. However, different terms can represent similar meanings like in the case of semantic similarity, where there is no an exact match, but a kind of similarity match.

In this study, association rule based query expansion model was introduced for expanding question keywords in question processing phase. It exploits the context and semantic relation between words to analyze and extract structure and meaning for both questions and candidate sentences. This QA model uses a hybrid semantic rule base for reformulating the asked question into more enhanced question that best match the answer. These association relations help to identify more relevant and precise answers for the asked questions from Frequently Asked Question (FAQ) database which improve the precision of QA systems.

This paper is organized as following: Section1 is an introduction. Section 2 is a survey about QA systems and their general architecture and main types. Section 3 discuss query expansion importance in QA system. Section 4 reviews related work about semantic similarity or association rule mining based query expansion in QA system. Section 5 introduces a new hybrid interesting measure and explains the reason for using of semantic similarity through association rule mining. Section 6 is about using this new measure in query expansion process for answering question. Experimental result and the used data set are described in section 7. Finally, conclusion and future work are presented in section 8.

Question Answering System

Generally, QA is an information retrieval task which is constrained by an expression of all or a part of the needed information as a set of natural language questions or statements. Question is defined as a natural language sentence, which usually begins with an enquiring word and expresses some users' information need (Kolomiyets and Moens, 2011). Natural language questions on one side specify a well-defined information need, but on the other side they include more information than a set of search terms, as they represent syntactic and semantic relationships between the search terms. Question type is defined as the certain semantic category of questions which is characterized by some mutual properties. QA research attempts to deal with a wide range of question types including: Fact, list, definition, how, why, hypothetical, semantically-constrained and cross-lingual questions. Generally, questions types can be classified into two kinds, i.e., factoid and non-factoid. The former type questions usually ask the names of people or places, where the latter type asks definitions, reasons, or methods.

Question Answering System Architecture and Component

The fundamental idea behind QA system is to assist human computer interaction. It gives the ability to answer natural language questions by extracting from a documents repository fragments of documents that contain relevant answer (Jurafsky and Martin, 2009). However, most of the QA researches are diverse in their scope, design, approaches and evaluation metrics etc. They follow the same basic architecture of QA system (Mathur and Haider, 2015). In general, a typical QA system consists of three modules (phases): Question processing, document processing and answer processing. Each of these main modules is also divided in some supplementary core sub-modules. For query processing module, its core sub module is question classification and for document processing module, its sub-module part is information retrieval. Finally, for answer processing module the core sub-module is answer extraction.

Question processing phase plays an important part in QA systems. It categorizes user questions and then derives expected answer types, extracts keywords to determine question focus and rephrase a question into semantically equivalent multiple questions. So, if this module does not work correctly, it will make problems for other sections. The goal of document processing module is to get a set of candidate documents that contains answers. It submits the reformulated questions from document processing module to information retrieval systems. Then, it returns a ranked list of relevant documents. Although the set of documents are generally ranked by their relevance to the query, the top ranked returned document may are not the appropriate answer to the question. Hence, documents are not adequate ranking unit with regard to the objectives of a QA system. Consequently, the next stage that extracts a set of potential answer passages from the retrieved documents is required. Answer processing and extraction module is the final phase of QA structural design. It is the tag of discrimination between QA systems. It has the responsibility to identify, extract and validate answers from set of ordered paragraphs which are received from document processing module. This phase are needed to rank and validate candidate answers which is classified into two general types is of factoid and non-factoid. Figure 1 displays the core component of the three modules of QA system (Mathur and Haider, 2015).

DOI: 10.3844/jcssp.2017.■■■.■■■

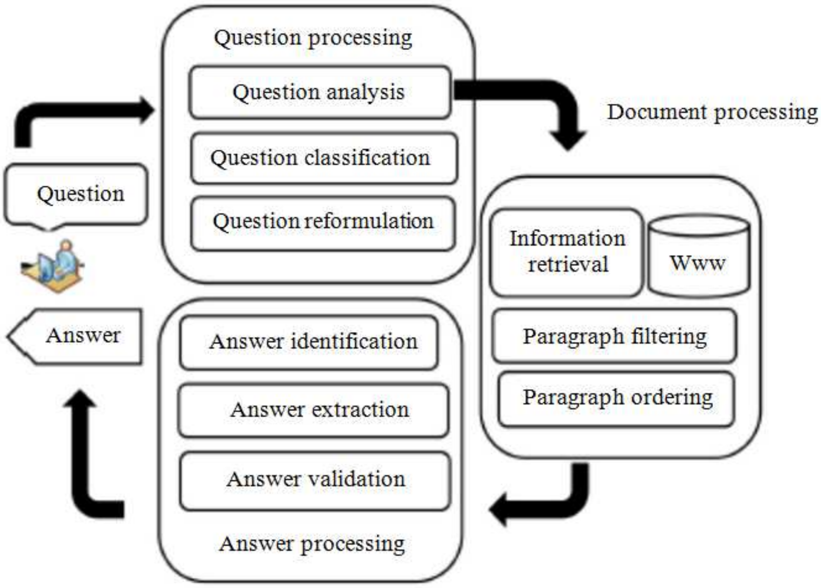

Fig. 1. QA system core component

Question Answering System Types

QA systems are categorized based on data content into two main types: Open domain and closed domain QA system. Open domain QA system concerns with questions which are approximately about everything and it can only depend on general domain ontology and world knowledge. Open-domain might refer to situations where the system answers unlimited questions type.

Therefore, these systems usually have to extract answer from much more data; On the other hand, closed-domain QA concerns with questions about a definite domain (such as medicine or weather forecasting and etc.).

Mishra and Jain (2016) identified eight criteria for sorting available large number of QA systems. These criteria are: Application domains for which QA systems are developed, users questions types, analyses types performed on users questions and source documents, used data sources types, characteristics of data sources, questions matching functions and their representations types, types of techniques used for retrieving answers and the forms of answers generated by QA systems.

Information Retrieval Role in Question Answering System

The Three technologies information retrieval, information extraction and QA are main technologies which are used to extract information from large document collections. QA systems are pretty much different from web search engines. Usually, web search engines response to the users query with reference and URL of related document but they fail when user needs a specific answer. But, QA system have to provide users with a short, comprehensible and accurate answer.

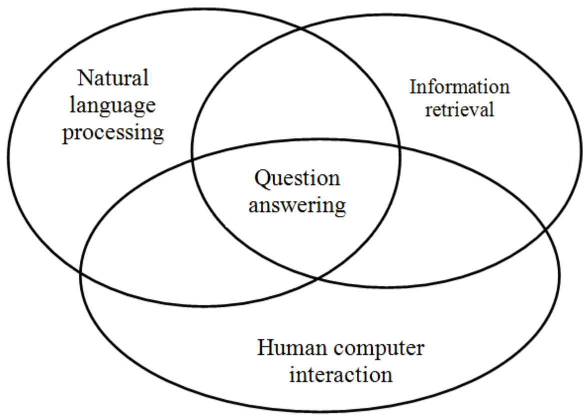

Fig 2. The interdisciplinary nature of the QA system

After facing these problems, information extraction techniques comes to be used in QA**.** Information extraction techniques are essential for natural language questions analyzing. These techniques often depend on Natural Language Processing (NLP) tools such as part-of-speech taggers, which determine the syntactic category of a word. QA places at the intersection of various scientific technologies including NLP, information retrieval and information extraction. Figure 2 shows the interdisciplinary nature of the QA field (Maybury, 2008).

Frequently Asked Question Answering System

Several FAQ retrieval systems have been introduced in the literature (Sneiders, 2009; Karan and Šnajder, 2015). The major task of FAQ retrieval system is to return the most relevant question-answer pairs for a specific users query. Its position is between classic information retrieval and QA. The main problem of retrieving FAQ is that query terms are short and domain specific, which increase the probability of a lexical gap. For example; the question "Can't connect to the net" should be mapped to the question "Why is my internet down?" though the two queries has no common words.

Query Expansion and Question Answering System

Natural language questions are a natural way to express a user information need. Two cases that are concerned with query terms lead to information retrieval failure. On one hand, the term mismatch problem in which users employ query terms which may be different from terms in particular internet resources; On the other hand, understanding the natural language questions correctly for forming queries and deducing their correct meaning to retrieve exact answers is not an easy task. It is usually difficult for a user to describe his or her need precisely (intention gap) according to just a few query terms. So, users' terms may not be discriminating enough. Other web challenges were mentioned in (Kwok *et al*., 2001). Usually QA differs from traditional keyword based information retrieval in the increased information that a question can convey over a set of keywords and in the target of answers. Finding answers to a natural language questions involves not only the awareness of what to look for, but also where to look for the answer. These differences make QA queries to be more than a bag of keywords. Thus, query expansion methods are needed to expand query terms in hopes of matching the exact form of the answer as it appears. This might include all morphological variants or synonyms of the words contented in the question.

Moldovan and Surdeanu (2003) did an experiment which displayed the distribution of the errors per system module in QA system. In this system there are ten modules; the first five modules correspond to question processing, the next two modules perform document and passage processing and the last three modules perform answer processing. The goal in this experiment is to identify the earliest module in the chain that prevents the system to find the right answer, i.e., causes the error. Two of the ten modules are responsible for more than half of the errors. These modules are derivation of the expected answer (M3) and keyword expansion (M5) which are a part of question processing. In keyword expansion module, if the question keywords which are used for passage retrieval are not expanded with the semantically related terms occurring in the answers, the relevant passages are not retrieved.

Related Work

Semantic Query Expansion in Question Answering System

Beginner users may lack enough knowledge in the domain of search. As a result, the query framed by them may not meet the information requirements. To overcome this drawback, a querying approach can be used based on domain specific ontologies and some NLP techniques for better results. Based on ontology semantic relations, keywords extracted in question processing module are expanded to semantically similar words. Many QA systems use various linguistic resources, such as WordNet (Miller, 1995), to identify relation between terms and hence improve query construction (Pasca and Harabagiu, 2001).

Bo and Yunqing (2008) introduced intelligence automatic QA system based on restricted domain ontology. This specific area ontology benefited from the accurate description of concept and relation for expanding question keywords. Consequently, the system accuracy and recall rates were improved.

Abouenour *et al*. (2012) presented the basic components of a new Arabic QA system IDRAAQ. In IDRAAQ, the passage retrieval phase was based on multi-level processing for improving the quality of retrieved passage and then the performances of the whole system. It was relied on keyword-based and structure-based levels that respectively consist of Arabic WordNet based query expansion process and a Distance Density N-gram Model.

Athira *et al*. (2013) used ontology and domain knowledge for identifying the relations and reformulating queries. This system enhanced the abilities of the current QA with the capability of processing complex questions. The aim of the system was to generate short and specific answer in the natural language asked question for a specific domain which was a step towards semantic web QA.

Association Rule based Query Expansion in Question Answering System

For a user's question, the codified query fails to find hidden knowledge and relationships. As a result, many irrelevant documents are returned. Thanks to the availability of large document collections on the web combined with in information retrieval improvements and NLP techniques, a new class of query expansion based QA systems has been appeared. These QA systems answer users' natural language questions with the help of a repository of documents (Hawking *et al*., 1999).

Making use of the insight gain of Association Rule data mining, association rule text mining was applied to query reformulation problem (Feldman, 1996) by converting the textual document into transaction format.

In this mining technique if particular terms appear in a document, there is a high probability that certain other term will also appear in that document. Generally, Association Rules (Han *et al*., 2011) are used to discover the correlations between terms that happened concurrently in the database or other data repositories. It is usually represented as a directed relation (X Y) between two sets of items from the antecedent to the consequent. They are regularly evaluated with interest measures such as support and confidence metrics.

Association rules based query expansion was applied into intelligent QA system to reformulate the question key terms to more enhanced query terms that best match the document content (Voorhees, 2001).

Yang *et al*. (2003) defined QUestion Answering by LexIcal FabrIc and External Resources (QUALIFIER) which is an event-based question answering system that answers definition, factoid and list questions in the TREC12. This system performs event mining to find out and then include the knowledge of event structure systematically for more effective QA. During the knowledge acquisition stage, it integrates the knowledge of the pre-retrieved TREC documents, web, WordNet and manually constructed ontology to extract additional terms which are used to expand the original query term. So, the new query contains terms that are related to the local context in the web and the lexical context in WordNet. Answer candidate sentences are selected from the top returned documents and are ranked based on association rules obtained from QA event analysis.

Yunjuan *et al*. (2011) introduced intelligent QA system that applied the association rules algorithm to discover the potential rules between keywords which users use to sort the result. in this intelligent system and from the generated rules, a keyword associated table is generated and used to calculate correlation between the keywords. By finding out keywords which often appear together, correlation and relationship between knowledge points and the students frequently questions asked are calculated.

Qu and Wang (2012) introduced intelligent QA system for online teaching to help students search problems. This system was based on a database that contains questions and their corresponding answers. It used improved association rules based searching answer algorithm to get the similarity value of user question with each question in database and then sum up these similarities. The answer of the user input question is the answer of the question with the biggest corresponding similarity. This searching answer algorithm used frequent item sets other than the whole Q&A to search answers.

Shortcoming

Previously stated work used either semantic similarity or association rules for expanding query in QA system. These approaches consider only the word co-occurrence and the exact match of association rules or the hierarchical match of semantic similarity. But, no one of them has utilized involving of ontologies in association rule mining process. Consequently, anew query expansion model is needed to solve these limitations. This model benefits from coupling the semantic relation of WordNet and context relation of association rule for creating a new hybrid interesting measure.

New Interesting Association Rules Measure

The need of Semantic Similarity in Data Mining Process

Typical text mining techniques usually transform text into flat bags of words representation that does not make use of the semantic information which illustrates the conceptual roles of the text. Based on such simple representations, text mining techniques can only discover shallow patterns, such as term associations. Using semantic similarity measure in conjunction with association rule mining process comes from two perspectives: First, most of association rule mining methods depend on statistical measurements and rarely take into account the semantic knowledge behind the statistical numbers. It can only annotates frequent pattern with non-semantically information (e.g., support, confidence and so on,); which cannot help users for the complete understanding of the patterns. This statistical information is not enough for measuring the interestingness. Generally, in the case of semantic similarity, there is no an exact match, but there is a kind of similarity match which can be useful to discover more relevant association rules and therefore important information. From the other perspective, though semantic similarity measure gives good results in discovering equivalences and hierarchical based relationships between terms, other relationships may stay hidden. Moreover, semantic similarity between terms changes over the time and across domains.

New Hybrid Measure

In this study, a new association rule interesting measure is introduced to overcome the previously stated shortcoming. This hybrid interesting measure results from involving the semantic similarity of WordNet ontology in association rule mining process. These two kinds of measures are combined by annotating frequent pattern with more structured information that can better indicate the hidden meanings of the pattern.

Semantic similarity is used at two phases: The first, in the preprocessing phase and the second, after generating the most frequent words. Considering semantic similarity in the preprocessing phase by using concepts instead of words before the frequent item generation process. Making use of the underlying hierarchical structure of WordNet ontology, generalized association rule mining principle (Han and Yongjian, 1995) was applied by mapping every word in the transactional text database to its concept. The main motivation for using word concept mapping is to find meaningful association between concepts and give rules not generated when not considering semantic mapping. This is typically in the case of words that each has a small frequency value but when represented as a common concept it could be important especially when different variation of the same concept occur in the document.

The second phase is after calculating the most frequent words or pattern. In this phase, semantic similarity between every two frequent word pairs is calculated using the general domain ontology WordNet (Miller, 1995). As a result, a new measure which is called semantic support for each of the most frequent word group is introduced:

$$\begin{aligned}&Semantic\ support\ (t1,\ t\ ..t_k) = (average\ semantic\ similarity(ti,tj) + \\ &\quad average\ support(ti,tj)) * \ support\ (t1,t2,..t_k)\end{aligned} \qquad (1)$$

where, *i*,*j* are two frequent words or concept and k is the frequent words.

For every most frequent n word: The average semantic similarity or average support is the summation of semantic similarities or support for every two word pairs divided by the number of terms and the support is the support of this frequent words. The main reason for choosing this equation design and its parameters is to involve the semantic measure in the mining process not only just to use it in optimizing items or the generated rules. Moreover, to automatically derive the items support at a semantic level.

Furthermore, using the average values will always make the new semantic support value between 0 and 1. As a result, a new hybrid measure or score that indicates the support of item set in knowledge base at a semantic level is automatically derived.

By replacing traditional support with semantic support, semantic interesting measure (confidence, left, etc.) is then calculated. For an association rule($x\ y$):

$$Semantic\,confidence(t1 \rightarrow t2) = \frac{semantic\,support(t1,t2)}{semantic\,support(t1)} \qquad (2)$$

Finally, a new knowledgebase which contains a list of semantically annotated rules between words is constructed.

For calculating semantic similarity Wu and Palmer (1994) measure was suggested to be used. This measure considers the depth of concepts in the taxonomy as a measure of their similarity. The depth is calculated by counting edges that separate terms from their Least Common Subsumer (LCS):

$$Sim(t_1,t_2) = (2 * depth\ \ (lcs)) / (depth(t_1) + depth(t_2)) \qquad (3)$$

In which depth (ICS) denotes the depth of the LCS of two concepts, LCS means the least common subsumer of concepts. According to the formula, this means that 0 < similarity value <= 1.

Hybrid Query Expansion Model for Question Answering System

This new measure exploits association rule relations and the semantic relation of WordNet for making dynamic patterns. This patterns could be used to expand the asked question keywords by working on query reformulation in question processing module. It is based on semantic rule base for expanding question keywords which contain association rules between terms at a semantic level. With the aid of this hybrid rule base, the system can expand the keywords to increase the search area for the question. Then, these expanded words are used to query the FAQ base by locating and then extracting the correct answers to users.

Proposed Method

This association rule based questions answering model is based on two knowledge sources. The first component is the semantic rule base which contains rules result from using WordNet semantic similarity in association rule mining. The second data source is the FAQ base which is a library of questions and answer pairs. With the aid of this hybrid rule base, the system can expand the keywords to increase the search area for the question. Then, the expanded words are used to query the FAQ base by locating and then extracting the correct answers to users.

For building the new semantic rule base from a collection of document data set: As association rule text mining requires plain text data to be converted to structured data format. All sets of documents are considered as a transaction database, where each document (questions or answers) is regarded as a transaction, the words in document are a collection of item sets and the document keyword are regarded as a set of transactions. Then the transaction can be expressed as: {document id, keyword 1, keyword 2, keyword 3, ..., keyword n}. Before text mining, NLP techniques are used to preprocess each document by select the most important term. In doing so: Words are extracted (tokenization) from each document. Stop words and the most common words are excluded in order to reduce the vocabulary number and increase the quality of the contributing terms. Stemming is performed using

DOI: 10.3844/jcssp.2017.■■■.■■■

(Porter, 1980) stemming algorithm. Stemming is the process of reducing related words to their stem, base or root form through affix removal. Its aim is to adapt various derivational alternatives of the same word to a single indexing form. After keywords have converted to their base form they are indexed into inverted list. Then, every word is mapped to its concept using WordNet ontology structure. Finally, Apriori-association rule mining algorithm (Agrawal *et al*., 1993) is used to get the most frequent words with minimum predefined support = 0.1. For each most n-frequent group of words semantic similarity between every word pairs is calculate using Equation 3 for similarity measure using WordNet. Using Equation 1, semantic similarity is involved with the support value to get the new semantic support. Using the new semantic support value and Equation 2 and according to a predefined minimum confidence value = 0.8, the confidence values of the rules are then calculated. Consequently, a new semantic rule base that contains semantic rules between words and their corresponding semantic confidence is generated. Figure 3 displays the process of building the new semantic rule base.

The second data source for this intelligent QA system is a frequently asked question (FAQ) database. This knowledge source stores question and answer pairs which can save answering time. In the context of question answering system, for a given collection of documents (such as a local collection or the World Wide Web), Users' query interface is used to retrieve the question posted by the user. Firstly, in question processing module, the key terms which are the term for which information is being sought are identified from questions. Linguistic techniques such as tokenization, stemming and part of speech (POS) tagging are implemented to user's question for formulating it into a precise query. For increasing system chances of finding pages with the desired answer and by utilizing the new hybrid semantic rule base, query vocabulary are extended via query expansion by adding the most contextually and semantically related secondary terms.

These secondary terms are coupled with the initial query key term to be used to select the most related answer that best match the question. Consequently, the new structured query contains terms that are related to the local context in the document and the lexical context in WordNet. In document processing module, the query terms that are result from the question processing phase is next used to query frequently asked questions database. In this process query expansion is utilized to reformulate questions into equivalent multiple questions. These reformulated query terms are used to select and retrieve question answer pairs that best match the reformulated question.

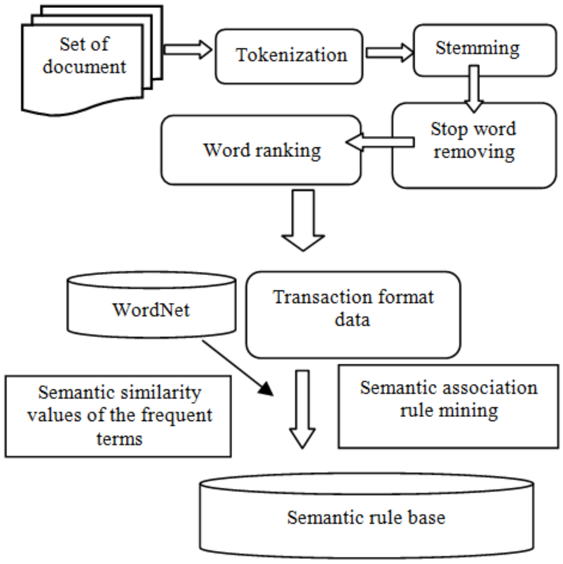

Fig. 3. Semantic association rule mining process

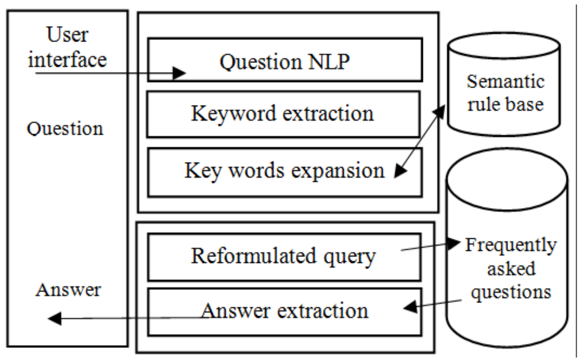

Fig. 4. Proposed system architecture

As the new model rules are used in query refinement, if the terms are not good discriminators, the expanded query may result worst performance than the original one due to the poor discriminatory ability of the added terms. The most suitable weighting scheme is TF-IDF. Because the IDF weight helps in decreasing the significance of high frequent terms which are poor content discriminator. Figure 4 displays the process of answering user`s question by utilizing the new hybrid rule base and the FAQ database.

Experimental Result

Dataset

In order to observe the improvement using this new semantic interesting measure in query expansion process, a document set of 479 webpage are used. These pages were crawled from Webopedia website. Webopedia is an online dictionary and internet search engine for information technology and computing definitions. It provides definitions to words and abbreviations related to computing and information technology. The number of words after pages preprocessing is 2670 word and after WordNet mapping process is 2825 word. After deleting useless words, the final semantic rule base contain 2692 rule between every two words or concept. For building FAQ data source every crawled page is set in the form of question and its related answer (question-answer pairs).

Result

There are many evaluation measures which are vary from one QA researcher to another. Some of the commonly used evaluation metrics are precision, recall and f-measure which are used for testing system efficiency. Recall and precision are traditional metrics used for information retrieval where f-measure is the harmonic mean of the precision and recall. Query expansion refers to a family of recall-boosting techniques particularly suitable for Boolean keyword or phrase document retrieval engines. Information retrieval system recall is very important for QA. If no correct answers are existing in a document, no further processing could be carried out to find an answer. Precision and ranking of candidate passages can also affect QA performance in the information retrieval phase. Recall for a QA system is defined as the fraction of number of correct answers to the number of the questions to be answered. While, precision is the fraction of the number of relevant retrieved answers to the total number of the answered questions. F-measure is the measure that combines precision and recall.

$$F - measure = 2.\frac{Precision.Recall}{Precision. + Recall} \tag{4}$$

Table 1 displays the difference between the values of traditional confidence, WordNet semantic similarity and the new semantic confidence for word1 to word 2 rule. From the table, it could be concluded that: The value of new semantic confidence depend not only on support value but also on the semantic similarity. For example, the rule (email communication) its confidence is 0.8 and due to the high value of the two words similarity = 0.823 the value of the new confidence increase too and so on for the other words rules.

The values of precision, recall and F-measure after using the new measure in expansion process are displayed in Table 2-4 while Table 5 displays the average values of them. It is obvious that the values of tables for the three expansion models vary from one measure to another but generally the average value of the new hybrid measure is larger than the average values of the other two measure. The superiority of this new measure comes from that semantic similarity query expansion model usually is only based on the hierarchical match and association rule model based on the exact match between terms; On the other hand, the new measure combines the two previous measures to annotate a frequent pattern with more structured information that can better indicate the buried meanings of the pattern. Moreover, it involves the semantic measure in the mining process not just to use it in optimizing items or the generated rules.

Table 1. (Word1-word2) semantic confidence

Word1	Word2	Semantic similarity	Confidence	New confidence
Disk	Device	0.8009	0.500	0.800
Disk	Drive	0.8409	0.600	0.800
Browser	Url	0.7279	0.600	0.650
Interface	Software	0.9009	0.600	0.800
Software	Interface	0.9009	0.645	0.750
Software	Program	0.9479	0.486	1.000
Software	Application	0.9009	0.532	0.600
Google	Browser	0.8339	0.750	0.801
Windows	Browser	0.7829	0.532	0.600
Email	Communication	0.8239	0.878	0.900
Windows	Software	0.9999	0.468	0.500
Url	Address	0.9479	0.750	1.000
Network	Application	0.4629	0.390	0.235
System	Software	0.4286	0.448	0.305
Internet	Application	0.1818	0.390	0.223
Lecture	Device	0.2353	0.317	0.315
Phone	Call	0.6667	0.624	0.504
Program	File	0.5333	0.321	0.143
Program	Computer	0.2857	1.000	0.893

Table 2. Precision values for query expansion using new semantic association rule -association rule and WordNet based semantic similarity

Query term	New semantic association rule	WordNet-semantic similarity	Association rule
Disk	0.079	0.077	0.065
Google	0.033	0.033	0.038
Email	0.008	0.004	0.008
Windows	0.056	0.044	0.033
Interface	0.013	0.015	0.010
Memory	0.042	0.015	0.035

Table 3. Recall values of query expansion using new semantic association rule -association rule and WordNet based semantic similarity

Query term	New semantic association rule	WordNet-semantic similarity	Association rule
Disk	0.974	0.949	0.795
Google	0.696	0.696	0.783
Email	0.800	0.400	0.800
Windows	0.964	0.750	0.571
Interface	0.667	0.778	0.556
Memory	0.513	0.436	0.179

Table 4. F-Measure values of query expansion using new semantic association rule -association rule and WordNet based semantic similarity

Query term	New semantic association rule	WordNet-semantic similarity	Association rule
Disk	0.147	0.143	0.120
Google	0.064	0.064	0.072
Email	0.017	0.008	0.017
Windows	0.107	0.083	0.063
Interface	0.025	0.029	0.020
Memory	0.077	0.028	0.059

Table 5. Average precision-recall and F-measure values of query expansion based on new measure and the other two measures

Average value	New semantic association rule	WordNet- Semantic similarity	Association rule
Precision	0.039	0.031	0.032
Recall	0.770	0.670	0.610
F-measure	0.073	0.059	0.058

Conclusion

Question answering is one step ahead of information retrieval. However "bag-of-words" representation has been used as an effective tool for retrieving large number of relevant documents in information retrieval system. It is not effective for QA where users need precise answers. Moreover, one of the main QA problems is that the question keywords are conveyed in natural language text in various ways. This is known as the semantic gap between the query and document. In order to bridge this gap, a new association rule based query expansion model was introduced for QA system. This QA model is based on a new knowledgebase which contain a list of semantically annotated rules between words. This hybrid knowledge base which is built from both semantic and contextual resources is used to bridge the surface shallow differences between questions and their correct answers. The main goal of expanding question key terms is to rewriting and reformulation question into similar enhanced questions. By asking question with another new enhanced query terms that best match the document content, less time and sources are used for search. According to the calculated result, this new measure show its superiority over the other traditional measures. It gives more valuable rules that can better indicate the hidden meanings of the pattern. Thus, precision and recall of the search in the QA system is improved.

Future Work

Future research will focus on the mining process to generate more enhanced rules by examining the usefulness of using another interesting measure and reducing the number of generated rules by using more enhanced association rule algorithm.

Author's Contributions

Ayatallah Gamal Abass: Participated in all experiments such as data preparation, selection and testing, coordinated the data-analysis and contributed to the writing of the manuscript.

Sameh Abd El-Ghany: Designed the research plan, organized the study, participated in all experiments, coordinated the data-analysis and contributed to the review of the manuscript.

Ahmed Abo Elfetoh: Designed the research plan, organized the study, contribute in revision of manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

Reference

Abouenour, L., K.M. Bouzouba and P. Rosso, 2012. IDRAAQ: New Arabic question answering system based on query expansion and passage retrieval. Proceedings of the Conference and Labs of the Evaluation Forum, (LEF' 12), Italy.

Agrawal, R. T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, ACM, USA, pp: 207-216. DOI: 10.1145/170035.170072

Athira, P.M., M. Sreeja and P.C. Reghuraj, 2013. Architecture of an ontology-based domain-specific natural language question answering system. Int. Jo. Web Semantic Technol., 4: 31-39. DOI: 10.5121/ijwest.2013.4403

Bo, W. and L. Yunqing, 2008. Research on the design of the ontology-based automatic question answering system. Proceedings of the International Conference on Computer Science and Software Engineering, Dec. 12-14, IEEE Xplore Press, Hubei, pp: 871-874. DOI: 10.1109/CSSE.2008.233

Feldman, R., 1996. Mining associations in text in the presence of background knowledge. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Aug. 02-04, AAAI Press, Portland, pp: 343-346.

Han, J. and F.Yongjian, 1995. Discovery of multiple-level association rules from large databases. Proceedings of the 21th International Conference on Very Large Data Bases, Sept. 11-15, Morgan Kaufmann Publishers Inc., USA, pp: 420-431.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier, Burlington, ISBN-10: 0123814804, pp: 744.

Hawking, D., E. Voorhees, P. Bailey and N. Craswell, 1999. Overview of TREC-8 question answering track. Proceedings of the 8th Text Retrieval Conference, (TRC' 08).

Jurafsky, D. and J.H. Martin, 2009. Question Answering: Speech and Language Processing, 2nd Edn., ISBN-13: 978-0131873216.

Karan, M. and J. Šnajder, 2015. Evaluation of manual query expansion rules on a domain specific FAQ collection. Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality and Interaction, Sept. 08-11, Springer, Toulouse, France, pp: 248-253. DOI: 10.1007/978-3-319-24027-5_24

Kolomiyets, O. and M.F. Moens, 2011. A survey on question answering technology from an information retrieval perspective. Inf. Sci., 181: 5412-5434. DOI: 10.1016/j.ins.2011.07.047

Kwok, C., O. Etzioni and D.S. Weld, 2001. Scaling question answering to the web. Trans. Inform. Syst., 19: 242-262. DOI: 10.1145/502115.502117

Manning, C.D., P. Raghavan and H. Schütze, 2009. Introduction to Information Retrieval. 1st Edn., Cambridge University Press, Cambridge, England, ISBN-13: 9780521865715

Mathur, A. and M.T.U. Haider, 2015. Question answering system: A survey. Proceedings of the International Conference on Smart Technologies and Management for Computing Communication Controls Energy and Materials, May 6-8, IEEE Xplore Press, Chennai, India, pp: 47-57. DOI: 10.1109/ICSTM.2015.7225389

Maybury, M., 2008. New Directions in Question Answering. In: Advances in Open Domain Question Answering, Strzalkowski, T. and S.M. Harabagiu (Eds.), Springer, pp: 533-558.

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38: 39-41. DOI: 10.1145/219717.219748

Mishra, A. and S.K. Jain, 2016. A survey on question answering systems with classification. J. King Saud Univ.-Comput. Inform. Sci., 28: 345-361. DOI: 10.1016/j.jksuci.2014.10.007

Moldovan, D. and M. Surdeanu, 2003. On the Role of Information Retrieval and Information Extraction in Question Answering System. In: Information Extraction in the Web Era, Pazienza, M.T. (Ed.), Springer, pp: 129-147.

Pasca, M. and S. Harabagiu, 2001. The Informative Role of WordNet in Open-Domain Question Answering. Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, (NAACL' 01).

Porter, M.F., 1980. An algorithm for suffix stripping. Program, 14: 130-137. DOI: 10.1108/eb046814

Qu, Z. and Q. Wang, 2012. Design and implementation of an intelligent automatic question answering system based on data mining. Proceedings of the 3rd International Conference on Advances in Swarm Intelligence, Jun. 17-20, China, pp: 137-146. DOI: 10.1007/978-3-642-31020-1_17

Sneiders, E., 2009. Automated FAQ answering with question-specific knowledge representation for web self-service. In Proceeding of the 2nd Conference on Human System Interactions, May 21-23, IEEE Xplore Press, Catania, Italy, pp: 298-305. DOI: 10.1109/HSI.2009.5090996

Voorhees, E.M., 2001. Overview of the TREC 2002 question answering track. Proceedings of the 11th Text Retrieval Conference (TREC' 02), pp: 115-123.

Wu, Z. and M. Palmer, 1994. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Jun 27-30, Association for Computational Linguistics, USA, pp: 133-138. DOI: 10.3115/981732.981751

Yang, H., H. Cui, M. Maslennikov, L. Qiu and M. Kan *et al*., 2003. QUALIFIER in TREC-12 QA main task. Proceedings of the 12th Text Retrieval Conference (TREC' 12), pp: 480-488.

Yunjuan, L., K. Dechuan, Z. Lijun and Y. Liqiang, 2011. The application of association rules algorithm in the optimization of information processing in intelligent question-answering system. Proceedings of the International Symposium on Computer Science and Society, Jul. 16-17, IEEE Xplore Press, Kota Kinabalu, Malaysia, pp: 296-298. DOI: 10.1109/ISCCS.2011.86