الجامعة البريطانية في دبي

The British University in Dubai

Improving performance of collaborative question answering systems by using semantic resources

تحسين أداء النظم التعاونية للإجابة علي السؤال باستخدام مصادر دلالية

By

Muhammad Arshad Javed

(Student ID number 120166)

Dissertation submitted in partial fulfillment of the requirements for the degree of MSc in Informatics

Faculty of Engineering & IT

Dissertation Supervisor

Professor Dr. Khaled Shaalan

June-2015

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Muhammad Arshad Javed	120166	MSc. in Informatics	June, 2015

Improving Performance of Collaborative Question Answering Systems Using Semantic Resources

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

Electronic Submission Copyright Statement
Please choose one of the following two licenses and check appropriate box.

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation worldwide including the users of the repository, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Public access to my dissertation in the Repository shall become effective:
☒ Immediately ☐ 24 months after my submission
☐ 12 months after my submission ☐ 48 months after my submission

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation to students, faculty, staff and walk-in users of BUiD Library, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Signature

~~*Muhammad Arshad Javed*~~

ABSTRACT

In this modern age of technology, World Wide Web (WWW) provides us a platform to share the information with each other. People use different types of web applications for example online forums/blogs, portals for question answering, e-mail, and prompt messaging tools to collect and share their information and develop online communities. All these shared information on the web create a huge collection of data. This data is increasing day by day.

Online social networks gather data from individual users and offer them to create link with other users of mutual interests in the same network. In this fashion, the social networks evolved as platforms to launch and uphold the social relationships in addition to share their knowledge and information. To manage such a large information, we need to use Information Retrieval (IR) techniques in efficient way. An Information Retrieval (IR) system retrieves the text related to the query of the user from massive collection of documents in real time. A document may comprise a collection of text, like a web page or an article. Information Retrieval system efforts to gratify the user's requirements effectively. Usually, an IR system takes the user query in natural language and returns the documents containing information pertinent to the question. One typical example of an IR system is Question Answering System. Usually a question answering system contains three phases namely question analysis, document retrieval and answer analysis. The question analysis phase takes the user questions and applies several processes such as question classification, query expansion to increase the probability of finding the relevant documents. The document analysis phase takes the processed question and retrieves the documents containing possible answers. The answer analysis phase identifies the relevant passages or set of sentences containing the possible answers and presents it to users. Thus, Question Answering Systems are very useful for retrieving documents from a collection of documents.

In order to take full advantage of data generated by users over the social networks, a special class of Question Answering Systems was designed. These systems are called Collaborative Question Answering (CQA) Systems or Community Question Answering Systems. There are dozens of Collaborative Question Answering Systems available on the internet. The research proposed in this dissertation focuses mainly on CQA Systems and proposes methods to improve performances of these systems. One major problem with the existing CQAs is the mismatch between the user questions and the set of questions present in the CQAs. Though these CQAs contain the question, which is semantically similar to the user question, they fail to return the answers. The research in this dissertation proposes the methods to solve this issue. Thus, the scope of this dissertation is limited to the question analysis phase of the CQA systems. The overall performance of a CQA depends a lot on the question analysis phase. The question analysis phase in the proposed research attempts to improve the question matching in two steps. In the first step, called Question classification, questions are classified into several coarse grained and fine grained classes based on some rules. Based on predicted class of the question, the entity type (person, location, time etc.) expected to be present in the answers are determined. In question classification, we have used Wikipedia and WordNet tools. In the second step, called query expansion, irrelevant words are removed and semantically equivalent words are added. We have used a freely available open source thesaurus named Collaborative International Dictionary of English (CIDE) to find the semantically equivalent words. The methods proposed in this research are tested over a number of questions collected from existing CQA systems. The results are presented in the thesis.

تحسين أداء النظم التعاونية للإجابة علي السؤال باستخدام مصادر دلالية

في هذا العصر الحديث من تكنولوجيا وشبكة ويب العالمية (شبكة الاتصالات العالمية) يوفر لنا منصة لتبادل المعلومات مع بعضها البعض . الناس استخدام أنواع مختلفة من التطبيقات على شبكة الإنترنت على سبيل المثال على الانترنت المنتديات / بلوق ، وبوابات ل مسألة الرد ، والبريد الإلكتروني ، و الرسائل الفورية أدوات لجمع وتبادل المعلومات ، وتطوير المجتمعات المحلية على الانترنت . كل هذه المعلومات المشتركة على شبكة الإنترنت إنشاء مجموعة هائلة من البيانات . هذه البيانات يزداد يوما بعد يوم.

الشبكات الاجتماعية على الانترنت جمع البيانات من المستخدمين الفرديين وتوفر لهم لخلق صلة مع مستخدمين آخرين المصالح المتبادلة في نفس الشبكة. على هذا النحو، تطورت الشبكات الاجتماعية كمنصات لإطلاق ودعم العلاقات الاجتماعية بالإضافة إلى تبادل المعارف والمعلومات. لإدارة مثل هذه المعلومات كبيرة، ونحن بحاجة إلى استخدام استرجاع المعلومات (الأشعة تحت الحمراء) تقنيات بطريقة فعالة. نظام استرجاع المعلومات (الأشعة تحت الحمراء) باسترداد النص المتعلق الاستعلام للمستخدم من مجموعة ضخمة من الوثائق في الوقت الحقيقي. يمكن أن تشمل وثيقة عبارة عن مجموعة من النصوص، مثل صفحة ويب أو مقال. الجهود الإعلامية نظام استرجاع لإرضاء متطلبات المستخدم بشكل فعال. عادة، وهو نظام الأشعة تحت الحمراء يأخذ الاستعلام المستخدم في اللغة الطبيعية وإرجاع الوثائق التي تتضمن المعلومات ذات الصلة لهذه المسألة. أحد الأمثلة النموذجية لنظام الأشعة تحت الحمراء هو السؤال نظام الرد. عادة ما يكون مسألة نظام رد يحتوي على ثلاثة مراحل وهي تتساءل تحليل وتوثيق واسترجاع وتحليل الجواب. مرحلة التحليل السؤال يأخذ الأسئلة المستخدم وتنطبق عدة عمليات مثل تصنيف السؤال، والتوسع الاستعلام إلى زيادة احتمال العثور على الوثائق ذات الصلة. مرحلة التحليل المستند تأخذ مسألة معالجة واسترداد الوثائق التي تتضمن الأجوبة المحتملة. تحدد مرحلة التحليل الجواب الممرات أو مجموعة من الجمل التي تحتوي على الإجابات المحتملة ذات الصلة ويقدمها للمستخدمين. وبالتالي، السؤال أنظمة الإجابة مفيدة جدا لاسترجاع الوثائق من مجموعة من الوثائق. من أجل قبل المستخدمين على الشبكات الاجتماعية، وقد تم تصميم فئة خاصة من سؤال أنظمة الرد. وتسمى هذه الأنظمة التعاونية سؤال الإجابة (CQA) نظم أو الجماعة سؤال أنظمة الرد. هناك العشرات من التعاونية سؤال الإجابة أنظمة المتاحة على شبكة الانترنت. البحث المقترح في هذه الأطروحة يركز أساسا على نظم CQA ويقترح أساليب لتحسين أداء هذه الأنظمة. مشكلة واحدة كبيرة مع CQAs الحالية هي عدم تطابق بين الأسئلة المستخدم ومجموعة من الأسئلة الموجودة في CQAs. ورغم أن هذه CQAs تحتوي على السؤال، الذي يشبه غويا على سؤال المستخدم، فإنها تفشل في العودة الأجوبة. البحث في هذه الأطروحة يقترح طرق لحل هذه القضية. وهكذاالاستفادة الكاملة من البيانات التي تم إنشاؤها من ، فإن نطاق هذه الأطروحة يقتصر على مرحلة التحليل مسألة النظم CQA. الأداء العام للCQA يعتمد كثيرا على مرحلة التحليل السؤال. مرحلة التحليل السؤال في البحث المقترح محاولات لتحسين مطابقة سؤال في خطوتين. في الخطوة الأولى، ودعا تصنيف سؤال، تصنف الأسئلة إلى عدة فئات الحبيبات الخشنة الحبيبات وغرامة استنادا إلى بعض القواعد. على أساس الطبقة وتوقع للقضية، ونوع كيان (شخص، والموقع، والوقت الخ) من المتوقع أن تكون موجودة في تتحدد الأجوبة. في تصنيف سؤال، واستخدمت ويكيبيديا و وردنت الأدوات. في الخطوة الثانية، ودعا التوسع الاستعلام، تتم إزالة كلمات غير ذات صلة وغويا تضاف عبارة مماثلة. لقد استخدمت متاحة بحرية المكنز

مفتوحة المصدر اسمه التعاونية قاموس اللغة الإنجليزية الدولي (CIDE) للعثور على كلمات تعادل غويا. ويتم اختبار الأساليب المقترحة في هذا البحث على عدد من الأسئلة التي تم جمعها من أنظمة CQA القائمة. يتم عرض النتائج في الأطروحة.

DEDICATION

Dedicated to my sweet family and friends.

ACKNOWLEDGMENT

First of all, I would like to take this opportunity to express my sincere gratitude towards my supervisors Professor Dr. Khaled Shaalan and Dr. Santosh Ray for their generous guidance, support, and motivation given to me during my studies.

I wish to express my cordial gratitude to my teacher Dr. Sherief Abdallah as well.
I would like to thank all staff members and administration departments in BUiD for their kind support during my studies.

Finally, I would like to thank my colleagues, particularly the higher management of Khawarizmi International College (KIC) for their continuous support to complete my studies successfully.

.

Table of Contents

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The World Wide Web (WWW) is a huge repository of information containing large number documents in form of text, images, video and audio. . Due to increasing popularity of social networking websites, this collection of data is getting larger day by day. Example of social network includes MySpace, Facebook, YouTube, Twitter and many others. All such online social networks are very well established around their users. All these online social networks collect the personal data from the users and provide them to establish link with each other in term of their common interests. In this way, the social networks became the place to establish and maintain the social relationships as well as to share their knowledge and information (Mislove et al, 2007).

The objective of the Information Retrieval (IR) process is to find the text related to the query of the user from enormous collection of documents within minimum time. An IR system works with the storage and collecting of information from a huge collection of documents. A document may contain a collection of text, like a web page or an article. A very known example of IR system is Question Answering System. A Question Answering System tries to satisfy the user's information requirements successfully. Conventionally in a Question Answering System, the user enters a query in natural language and the documents having such information related to the question are provided by the system.

1.2 Question Answering Systems

Beyond the search engines, Question Answering Systems have emerged as a new technology to provide the correct answer instead of whole document. A Question Answering System must be capable to find the answer to a question written in natural language. To understand the significance of Question Answering Systems, consider the question, how can a user search the right answer of a given question? When someone asks some question, the person attempts to understand the question completely. Then he tries to find the source of knowledge where he can find the correct answer. It depends on the type of question, he can search a book or an expert or he can use online sources in the form of documents that contain the required answer of the question. In the document, he finds some additional clues to get the answer of the question. Question Answering System efforts to manage an extensive type of questions including, fact, description, why, how, list, theoretical, semantically controlled and cross lingual type of questions. The search collections vary from lesser local document collections, to in-house organization documents, to accumulated newswire reports, to the World Wide Web. Accordingly, the Question Answering System can be categorized as closed domain and open domain questions. Closed domain type of question answering works with questions related to a particular domain, for example, education, business etc. Close domain question answering can be considered as relatively easy job because they can employ NLP techniques to manipulate specific domain knowledge regularly. Open domain question answering works with questions related to everything and only depend on common ontologies and knowledge. Usually such systems have more data accessible from which to retrieve the answer.

To improve the accuracy, the Question Answering Systems implement a number of language resources and software procedures, which contain progressively, compound natural language process (NLP) methods (Shaalan, 2014). From the existing or available collection of documents to extract a correct answer, refined syntactic, semantic and related processing techniques of text must be implemented. There are some questions that need more than extraction of concisely specified answer, such questions must be fragmented into modest questions and answers from the available question answering procedures should be collected in a smart way, perhaps with cognitive and implication proficiencies.

1.2.1 Collaborative Question Answering Systems

When more than one people try to learn together and share their knowledge and ideas about some topics, this is called collaborative learning. It is completely different from personal alone learning, in collaborative learning the people may utilize and distribute their resources and knowledge by inquiring for information, assessing, observing the information of each other. This approach of learning act as a model that information can be produced by sharing understanding and experience of different member of community where they energetically interact with each other (Arai & Handayani, 2012). The main objective of the collaborative Question Answering Systems (CQA) is to provide the opportunities to its users to share their knowledge. Yahoo Question Answering System lets its users to submit their questions to get the answers, and reply the answer of the questions asked by another user.

In CQA System comparatively the answer selected by the users as the best one is not certainly the answer of excellent quality. The decision taken by the user about the best answer of the question can be influenced by individual thinking, for example the relationship between the users, individuals own opinion, lack of information and understanding about the topic. The users who were very well familiar with author of the answer are a number of causes because of that the client might be considered the given answer as the best, like as the answer is related to the subject requirements completely or it. The priority might be given to that author of the answer whose view is nearly matched to the client of the question even a number of other similarly correct answer available as well. Consequently, a good system that recommends the best answer automatically is required to improve this state because it will select the right answer accurately, fairly than just choosing according to the likeness of the asker of the question.

In contrast to CQA systems, automated Question Answering System aims to deliver brief information, which covers answers to the asked questions. There are mainly three important research techniques utilized in automated Question Answering System. First one is Natural language Processing (NLP) matches the questions asked by the user into an official world model and makes sure the maximum consistent answers. Second information retrieval (IR) strengthens the QA, along with NLP. It focuses on real information extraction from huge collection of text. Third type is used template based technique of QA compare the user questions to various templates that conceal often-inquired parts of information or facts domain (Andrenucci & Sneiders, 2005).

1.3 History of Question Answering Systems

In the beginning of 60's some of the Artificial Intelligence systems were considered as Question Answering Systems. The most famous Question Answering Systems of that time are Baseball and LUNAR. These systems were created in 1960s. Baseball system was developed to provide the answers of the questions related to the competitions of baseball league of USA for last one year. Whereas the LUNAR system answered, the questions related to the geological study of various types of rocks that were fluttered from the surface of lunar sent by Apollo expeditions. Both of these systems were very efficient in their own domains. LUNAR system was presented in a conference during 1971 to demonstrate the issues related to lunar studies. That system was capable to answer about 90% of the questions related to its domain asked by the people. Those systems were consisted of some common features like fundamental database / information system written by the professionals of the selected domain. In 1960s, some systems were developed and question answering modules were attached along with them as subsystems. SHRDLU and ELIZA are very famous. SHRDLU was implemented in the toys to simulate the operation of robot in the toys and provided the option of asking the questions related to the world state. This system was related to a particular domain like rules of physics and it was not much efficient. However, on other hand ELIZA was used to simulate the talk with psychologist. ELIZA was capable to communicate on different topics by choosing simple rules, which identified the significant words in the given inputs. There was an elementary level approach to answer the questions.

During 1970s and 1980s, some important and extensive ideas and concepts on computational linguistics were developed and presented which directed to the improvement of determined the tasks in text understanding and Question Answering System. Unix Consultant system was known as a good example of such kind of system, this system was answering the questions relating to the Unix operating system. This system contained all-inclusive hand constructed information related to its field and it was designed at jargon the answer to serve different types of users. There was another system known as LILOG, it was able to understand the text related to the tourism information about the cities in Germany. The systems used in both UC and LILOG projects helped greatly to develop ideas related to reasoning and computational linguistics.

During 1990s in Text Retrieval Conference presented the idea of Question Answering System that worked very successfully. The systems exhibited in that context were supposed to answer all questions related to any subject by finding corpus of the text that changed time to time. This context promoted the research and progression in the field of open domain Question Answering System of text bases. In the beginning, only clean English text was used in the corpus but later noisy text contained poor form of English also included. With addition of noisy text stimulated the Question Answering Systems towards credible setting. In real life, the text is very noisy as people become very careless while writing in unstructured media such as blogs. In the beginning, the corpus of TREC data comprised of just news wire text that was precisely clean.

With the development of World Wide Web, it is being used as a corpus of the text. Interest is increased to integrate the Question Answering Systems with web exploration. Ask.com is an example of one of the primary systems, later on Microsoft and Google began to assimilate Question Answering System and services in their own search engines.

1.3.1 History of Collaborative Question Answering (CQA) System

In the web there are many collaborative Question Answering Systems exist, for example the portal of Yahoo! answers[1]. PC World considers Yahoo! Answers as the best question answering portal. It was launched in 2005. In the beginning, it attracted a number of visitors and became very famous and popular website in its first six month in the group of education after Wikipedia and education.com (Sullivan, 2006). The clients of the Yahoo answers achieve their reward points in the following way: First login the system, second to provide the answers of submitted questions, and finally they do their vote for the best answer provided by other users. However, there are some limitations or drawback to the Yahoo! answers. The main issue is about the quality of the answers submitted by the users. The users are motivated to answer as many questions as they like, without considering the quality and accuracy of the answers they submit. Another important issue is about the size

[1] www.answers.yahoo.com

of data / information, extra overload by the users. A number of questions (in thousands) are provided to the system. Many users can provide their answers about the same question. All such answers are not graded based on their worth or quality but recorded in the mannered as provided by the users. In the system the author of the question is allowed to select the best answer as well as other users can do their vote for good answer. Another important and well-known question answering portal is known as "Google answers"[2] which launched in 2002 by Google. It was considered as online knowledge sharing market offered by Google. However, it was closed for any new activity in 2006 whereas the collection of data in its archives remains available (wikipedia.org).

Suryanto et al (2009) introduced various methods over CQA portal for example Yahoo! Answers. Those methods used for answer quality and answer importance as well. Through experiments on Yahoo! Answers, they showed that quality aware approaches experienced improved performance than non-quality aware methods. Cao et al (2010) proposed a method to exploiting class information associated with questions in CQA archives for enhancing the performance of question retrieval process. As Ray et al (2010) stated that true classification of questions supports to increase the general performance of question answering process in CQA systems. Liu & Agichtein (2011) explored the related aspects that influence the user's behavior in a huge, well-liked Collaborative Question Answering (CQA) system. Evaluation approaches were proposed to help in the development of question recommender systems. This is very important function in CQA systems to assign the questions from question inquirer to prospective answerer. Sakai et al (2011) used Graded Relevance Metrics methods to evaluate the answer ranking for a question and explained the benefits over zero cost assessment using the best answer data. Arai & Handayani (2012) developed a question answering system for a dynamic cooperative learning process, where an online system automatically answers the questions of the students in collaborative learning situation. System functioned upon knowledge base of question answering. Li et al (2012) analyzed the question quality in CQA and presented a joint reinforcement based tag proliferation algorithm to predict the quality of the question using its text features and profile of requester. Different types of flat classifiers such as NB,

[2] www.answers.google.com

SVM and ME applied to evaluate the classification models for question subject category. They combined them with hierarchical models of question classification (Qu et al, 2012).

1.4 Structural design of Question Answering System

Generally, an open domain Question Answering System is comprised of three main components, which are:

- **Question Processing:** A system examines a question completely and assigns labels to the question according to its predictable answer type through a process. This process is called question classification. In addition, new semantically equivalent keywords are added in the question to enhance the probability of retrieving relevant documents. This is called query expansion.

- **Document Retrieval:** Open domain Question Answering Systems use search engines to retrieve the documents related to user's questions. Close domain systems use information retrieval systems developed specifically to retrieve the documents. The usually compute the similarity between the expanded query and the documents to determine the relevancy of the documents.

- **Answer Processing:** It is an important component in question answering system that creates the correct answer from the passages of text. At first, it extracts and produces candidate answers from the paragraphs and then assign**s them some ranks according to some functions.** It is a process to select an appropriate answer from the available collection of answers for given question.

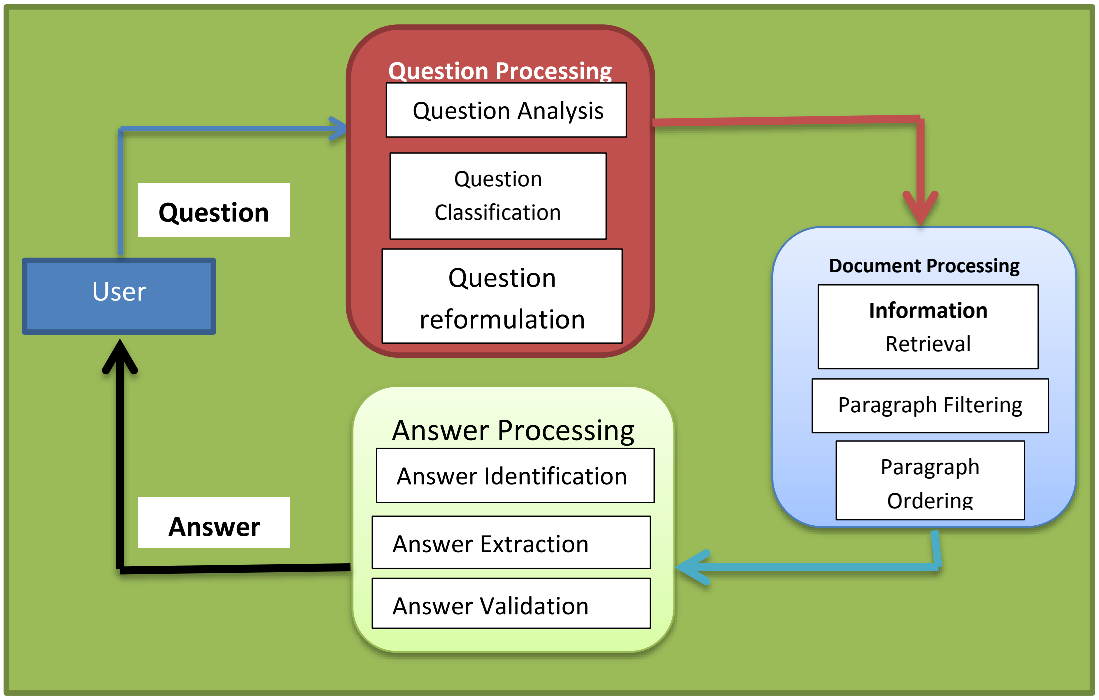


Figure 1.1: Architecture of Question Answering System

1.5 Need of Collaborative Question Answering Systems

To find the answer of the question on the web, numerous type of search engines are available there. Most of the search engines provide us the link of web pages where the user can get the links with answers rather than a specific and particular answer. For instance, if a user wants to search for birthplace of ex USA president George W. Bush, or he wants to search the total number of employees in Microsoft Company, in such cases the exact and right answer would be more useful rather than the links of other web pages, which might contain the required information. For this purpose, an expert system is required, which provides the answers of this type of question. For some type of questions, the user needs a direct answer rather than the links, which takes us to other webpages where the answer is available in somewhere in the body of that webpage itself. Even sometimes, the users are required to compute the answer from the available information, but this is not what the user required. Instead of this, the user needs an interactive system of question answering. Therefore, the direct answers of the questions of the user are better than finding the answer in the web page links provided by the search engine. In general, it means that the information required by the user is not well taken by the system of question answering. The processing part of the question might be failed to comprehend the question appropriately or in some cases, the information required for creating the answer is not collected certainly. In such cases, the system must be capable to rearrange the question and display a dialog box. For instance, the system could not find the answer of given question, rewrite the question in different phrase or write another question. Even the direct answer is more appropriate when the users have smart devices such as tablet or cell phone, because it is difficult to find out what the user needs on short screen. Therefore such smart system is required to understand the question, feelings behind the question and rendering to this as asked by the users. It will find out the answer in knowledge base system and provide us the direct answer of the question. Collaborative Question Answering (CQA) Systems play important roles in this situation. In CQA, the users share their knowledge and information. The user who knows the answer of the given question can write its answer. In this way, a huge collection of questions and related answers is being collected and can be used for future needs. The procedure of CQA is elaborated in the following figure 1.2.

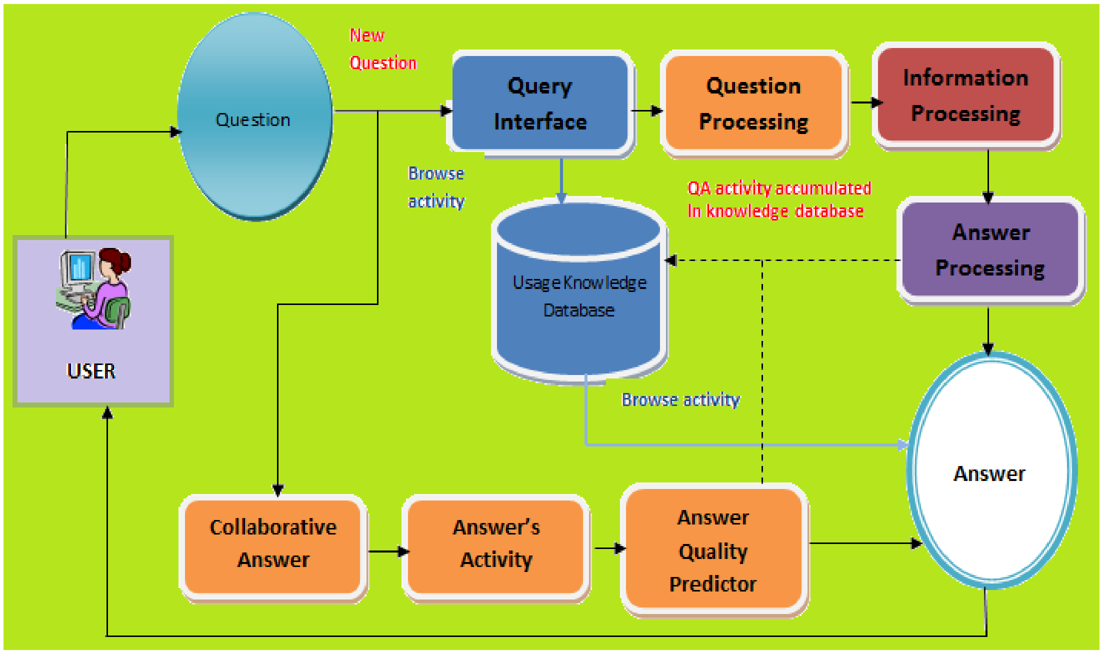


Figure 1.2: Mechanism of Collaborative Question Answering System

1.6 Research Objectives

This thesis focuses on Collaborative Question Answering Systems (QAS). As discussed earlier, there are many Collaborative Question Answering Systems available on the Web containing millions of questions and answers. However, they are not able to return specific answers of the questions posed by the users. Therefore, we state the research objectives of this thesis in terms of following hypotheses:

H1: Accurate prediction of type of entity (Person, Place, Time etc.) expected to be present in the answer can increase the accuracy of answer retrieval process.

H2: Expanding the user query by adding semantically equivalent keywords can improve the question matching process of Collaborative Question Answering Systems.

1.7 Organization of the Thesis:

This research work carried out in this dissertation has been divided into five chapters as described below.

Chapter-1 Introduction

This chapter provides information about information retrieval and their applications, history of the developments taken place in the field of Question Answering Systems in general and collaborative Question Answering Systems in particular. This chapter also discusses the structure and need of a Question Answering System. The research objectives have also been formulated in this chapter.

Chapter-2 Related Work

In this chapter, a detailed literature review of Collaborative Question Answering Systems has been provided. This review focuses on various aspects of Collaborative Question Answering Systems, particularly in e-learning process through collaborative Question Answering System.

Chapter-3 Design and Development of Collaborative Question Answering System

This chapter presents the work related to hypothesis one (H1). It describes the work done in field of question classification and prediction of answer entities. This chapter describes methods, tool and linguistic resources required to conduct experiments. The results have been presented at the end of the chapter.

Chapter-4 Findings/Result

This chapter presents the work related to hypothesis two (H2). It describes the work done in the field of query expansion. This chapter describes methods, tools and linguistic resources required to conduct experiments. The results have been presented at the end of the chapter.

Chapter-5 Conclusion & Future work

This chapter provides the conclusion of this research and presents possible research directions in the field of Collaborative Question Answering Systems.

CHAPTER TWO

LITERATURE REVIEW

Collaborative Question Answering (CQA) Systems are allowing users to ask questions. Answers of these questions are provided by other users in the community. There has been lot of development in the field of Collaborative Question Answering. This chapter presents the work done in the field of Collaborative Question Answering.

2.1 Social Networks and Collaborative Question Answering Systems

When people communicate with each other over the Internet with some common interest, this is called online social networking. A social network provides a well-organized structure where the people are connected directly or indirectly. There are two essential components in a social network, people or users considered as actors and their links called ties. According to the kinds of relations there are various networks even the similar group of actors have controlled observations (Knoke & Yang, 2008). Since the idea of internet is being applied in the real world, the popularity of social networks has been increasing in all over the world. There are various types of social networks available on the web to provide an opportunity to the people to communicate with each other according to their own interest. Some common examples of social networks are Face book, Twitter, internet forums and blogs. Through these online social media, people belonging to different groups communicate with each other according to their interest. Through these mediums, people share their knowledge with each other. (Cheung, 2011).

In Collaborative Question Answering (CQA) systems, the users communicate with each other through chatting and discussion. The main objective of this kind of systems is to share the information and exchange their ideas and views with each other. The examples of CQA Systems include Yahoo! Answers, Wikipedia and many others. It is claimed that Yahoo! Answers is a shared brain of coming generation, it served as database of information to its users. This is a repository of rich information. This website is full of knowledge that is managed and run by its community. The users interconnect with each other and share their knowledge by asking and putting the answers of the questions. This is completely collaborative online portal as compared to other QA portals, which are not really collaborative. Such as ask.com, that is not collaborative in its nature. These types of portal collect their knowledge from web databases. These portals act as simple search engine. They find their related information from the web and then return an appropriate answer.

The features and properties of online social networks can be examined through the analysis and exhibited of graph structure of social networks. Detailed sympathetic of graph structure of available social networks is very important to assess the social networks as well as to recognize their influence on internet (Mislove et al, 2007).

2.1.1 Graph Structure

The theory of graph structure greatly helped to discover the mistakes in the existing procedures of online social networks and assisted to recommend methods for these social networks. Such kind of research of web technology directed to develop algorithms to find the sources of rights in the internet system. With the application of graph structure theory, it is possible to find out the power rules, scales free networks and small networks in the social networks that are available on the Internet as well as it can display the distribution of users in Yahoo Answers. This theory is helpful to display strongly connected component (SCC) and weakly connected component (WCC) as well. This concept provides an understanding of connections of all users in social networks, how the users acquire support from others (Mislove et al, 2007).

Most of the online social networks follow power law, such as the Web, Flicker, Live-Journal and YouTube. In power law network the possibility of a node with degree k is relative/proportional to $K^{-\gamma}$, for greater value of K and γ is more than 1, where γ is considered as parameter and called power law coefficient (Mislove et al, 2007).

The networks, which have short diameter and large clusters, are called small world networks. In their graphs, the nodes do not appear as neighbor of each other but most of the nodes reach from other nodes through slight number of steps or phases (Adamic et al, 2003). In web structure there are two important parts known as strongly connected component and weakly connected component. A theory of web structure demonstrates that shape of the web is similar to a bow tie and it contains one big component that is connected strongly. That strong component can be touched by other clusters of different nodes. Whereas a weakly connected component is like a digraph where each node is accessible from other collections of nodes. On the other hand it is not essentially each node in the weakly connected component can extent to other collection of nodes (Tarjan, 1972).

2.2 Related Work

In the field of information retrieval and Human Computer Interaction (HCI) community, the Collaborative Question Answering Systems are becoming very attractive and a lot of efforts in research have been done. An important factor of success this Collaborative Question Answering (CQA) System is to offer an efficient and useful service in the limited time that inquirers need to collect good results or answers for questions.

In general, there are three different approaches to accomplish this objective. The first one is to reuse the huge volume of available contents of CQA systems to satisfy the user requirements, based on real retrieval of appropriate questions and answers to the required information. In second approach that is more attractive to increase the quality of the answer for CQA requester to forward the questions towards the specialists of that topic of the question that has been an enthusiastic field of research. Such as, Jurczyk et al (2007) developed graph structure for collaborative Question Answering Systems and used an algorithm of web link examination to find influential users in contemporary groups. The question viewed by Wenyin (2009) as a request to access the profiles of users as documents, applied various language models to collect proficient ranking, and declared the skilled finding problem as a problem of information retrieval. Kelly (2009) examined the procedure of automatically differences concerning authoritative and non-authoritative operators by exhibiting their right marks as a combination of gamma dispersals for respective subject. Away from CQA framework, an extensive work has been performed related to proficient result in accessible mediums. Different methods of question routing related to the user interests have been introduced to minimize the response time of the questions in CQA system. Such as, Guo et al (2008) created a probabilistic procreative model to find concealed subjects for requests and users and integrated information of both subject and term levels to endorse new problems to prospective answerers. Chang & Pal

(2013) solved the problem related to the questions route in real time Collaborative Question Answering System by means of considering the user interest and social associations.

In third approach to minimize, the response time for askers in CQA is merely to appreciate the additional answerers that related to good understanding of the answerer attitude. A lot of work is being done to understand user attitude in CQA. Such as, Adamic et al (2003) examined the content characteristics and arrangements of user interaction through various categories of Yahoo Answers. Gyongyi et al (2007) observed different features of user response in Yahoo! Answers, like level of users' activities, their interests and status. Guo et al (2008) examined the configurations of users in online social networks including Question Answering Systems related to knowledge and information sharing. Liu et al (2008) discovered consequences of the answerer's web exploring perspective on the usefulness of CQA systems.

2.3 Common Approaches in Collaborative Question Answering Systems

Several Collaborative Question Answering (CQA) Systems are introduced to assess different tasks of CQA. All these systems covered broad range of diverse methods and structural designs, for example ontology of question, databases of peripheral knowledge, exploration for mining answers of specific types, creation of answers, rationalization of answers, inference procedures, response loops, rational analysis and machine learning, so as incredible to encapsulate all differences in a particular architecture. However, the overall approach of CQA system is to propose the type of the job itself. There are a number of features, which are very common in most of CQA systems and help to present an overall architecture of a classical Collaborative Question Answering System. In general, approach of CQA, the system must offer some options for evaluating the question as well as to understand the question and its type. To search the right answers of the question the systems must be able to search the relevant documents speedily and proficiently. At the end, the systems are required to find the ranges of all these answers and select the best one to make available to the user.

Question answering system of social network based on the relationships among the users. In such social networks need to find out the specialists who are competent and eager to answer the questions depend on their social interactions with questioner (Du et al, 2013). In a collaborative learning environment, a system of question answering used a quality predictor. There are many answers related to one question, one of them must be selected. Each answer has its own weight and its subject to choose it. Quality of the answer must be determined by using answer quality predictor. This model might be beneficial for predictor excellence information as an acclaimed system to accomplished collaborative learning (Arai & Handayani, 2013). Arai & Handayani (2013) proposed that in Collaborative learning environment an answer quality predictor and domain knowledge is used to select the best answer. A model of PageRank based was proposed to solve the issue of estimating question difficulty in Collaborative Question Answering systems. This approach considerably performed very well (Liu, 2013). A conceptual thesaurus was built based on semantic relationships collected from knowledge base of Wikipedia. An integrated framework was developed to influence such semantic relations to increase the similarity measures for question retrieval process in the model space (Zhou, 2013). Toba et al (2014) stated that quality of the answer delivered by CQA could be analyzed through its textual illustration structures. Hierarchy of classifiers can be used to find a higher quality answer in the archives of CQA systems. San Pedro & Karatzoglou (2014) proposed Rank SLDA that joins supervised rank subject modeling. This approach can be applied to recommend the question where the subject text and community response are together modeled for user ranking in terms of their importance for a new question.

2.3.1 Analysis of Question

The process of analyzing the natural language question presented input to the system is considered as the initial step to search the answer of the question. The significant part of analysis module is to know the object of the question. To classify the goal of the question, it is evaluated in different ways. In the beginning a morph syntactic analysis of different words of the question is performed. In this process, a tag of part of speech is assigned to each word of the question to declare whether the word is verb or singular noun or plural noun and so on. After giving the tag of part of speech to the words, it becomes very easy to check the type of information the question is requesting. To extract the real answer of the given question it is very important to find the semantic kind of the question. The term question classification means to put the questions into numerous semantic groups can enforce some checks on the reasonable answers as well as it will propose various processing approaches. The collection of probable classes is described already and the choices from some elementary sets only reliant on observing at the key word of the question, for example when looks for date or time, where for some place, who for some person (Al-Chalabi et al, 2015). For example a system recognizes the question "Who was the first man landed on the surface of the moon?" it assume the name of the person as answer, in this case the exploration space of reasonable answers will be compressed significantly? In practical, all types of Question Answering Systems (QA) include a module of question classification. The whole performance and functioning of the QA system really depend on the correctness of classification of the question. However, this is not sufficient; meanwhile different words of English question do not convey much more semantic capturing information for example which, what etc. do not convey good semantic information. Some questions are related to some entities like which organization …? or what kind of building …?, likewise easy to conclude. For other types of questions, which are syntactically complicated like how many degrees in Mathematics, were awarded at Oxford University last year? Situation turns out to be more problematic. Therefore, most of the systems possibly do more by examination of the question. Which concludes extra restrictions on the entity answer by finding significant terms in the question which will be applied in matching applicant response bearing sentences, or recognizing associations, semantic or syntactic that have to keep between an applicant answer object and other objects or proceedings given in the problem.

So different systems have been assembled in the order of various types of the questions according to the kinds of answers required and try to keep the question into suitable group in hierarchy. Most of the systems practice a rough grained group definition. Let us consider different classes of the questions in the following table, usually practiced in an online Question Answering System. However, these collections of classes are related to a specific system but it shows an effective similarity with many classes of the questions of several existing online Question Answering Systems. Moldovan et al. (2003) developed hierarchy of question around 25 types himself from the examination of the TREC-8 training dataset.

Table 2.1: Collection of question types

Number	Definition	organization
Rate	Length	Money
Person	Place	Date
Description	Abbreviation	Known for
Nominal	Duration	Purpose
Reason	Other	

In the above table the term "Description" is practiced for such questions looking for its description of some individual, for example "Who was Newton?", whereas another question like "Who invented the glass?" it must be labeled as type of the person. Another term "Nominal" defines questions where nominal expression as answers then it cannot be allocated to some other particular classes for example a person or an organization. If the questions do not fitting to none of these types go in "other".

To classify the questions can be applied in different ways. The simple and very effective method is to implement to the question to find out its category is known as "pattern matching". The technique of classification is very complex to the sequence in which the patterns are implemented. For example the additional definite patterns birth date and date of death are implemented first, formerly the common pattern date (Bronner & Monz, 2012). There are more classy methods as substitute to pattern matching for question category (Suzuki et al., 2003). There is another method known as heuristic rules based algorithm and it needs to write heuristic rules by hand for the classification of the question, even though it is marvelous amount of tiresome work.

2.3.2 Methods to Recognize the Quality of the Answer of the Question

To analyze the quality of the answer of question there are different approaches. Generally, these approaches are divided into three types built on:

a) Information Retrieval and Natural language processing techniques
b) Link analysis that comprises of hits and PageRank kind of check

c) Arithmetical analysis

Content oriented approach (Information Retrieval and Natural language processing techniques)

In this method to find the answer of the question that is selected from the web or from available database, it depends on the implication of the question. In content based approach there are two most important methods, known as Natural Language Processing (NLP) and Information retrieval (IR). TREC is known as one of the best conferences and also held QA path since 1999. In the beginning of the Question answering way in TREC, the tentative competition was organized to resolve the simple type of question, for example description type. In advanced version of competition attempts to provide the answers of some challenging types of questions, for example more descriptive and expressive type. In TREC, all proposed techniques are combination of Natural language processing and of Information retrieval techniques. Information retrieval uses the keywords to track down the related document. But such questions which request for some particular information cannot be answered through information retrieval. For instance, a question requests, "Who is the Canadian Prime Minister?" In this case there are two answers, first answer: Canadian Prime Minister David visited American….., second answer: Canadian Prime Minister is the head of the government of Canada. Both of these answers have the key words – Canadian Prime Minister. Therefore, information retrieval approaches cannot distinguish two answers and select the best answer. The questions usually begin with wh-words like when, what, where, why, which, who and how. In Information retrieval processing, these words are considered as stop words and in the pre-processing these words are removed. That is why there is no right answer to the question. In such cases natural language processing is required to respond the natural language question. This kind of computation is expensive to make the processing of thousands number of documents reply to the question out of borders. In best Information retrieval systems specifically practiced at finding only few documents out of massive corpora which have the greatest contribution of request expressions. Therefore, in information retrieval process the most common technique is to fine downward the search to a comparatively less number of documents and then process the outstanding documents using NLP approaches to mine and grade the answer (Prager, 2006).

2.3.2.1 Information Retrieval

The process of Information Retrieval (IR) is very compactly associated with Collaborative Question Answering (CQA) Systems. The main idea of Information retrieval is to search such documents, which are related to the asked question. The objective of CQA system is to make available the accurate answer to the question. In terms of providing information that is more precise for the process of Question Answering System is more challenging as compare to Information retrieval. However, the quality and performance of CQA system is based on the efficiency of Information retrieval process as well. Providing the extremely related documents from CQA System that makes use of information retrieval approaches would outcome in a more precision of responses from the system.

2.3.2.2 Natural Language Processing (NLP)

The important objective of applying information retrieval approaches in a Question Answering System is to seek for and explore related information that must be comprised in the answer delivered. Such information that is provided by the finding methods may be different from one result, to some results or much more. However, how can a CQA System recognize the results it has gained and then conclude the best one that fulfills the user's requirements. For this purpose, the Natural Language Processing is developed to perform such functions. It was developed to choose and determine the quality of the outcomes provided by the search operation (Shaalan, K., 2014).

The NLP is an advanced computational tool along with a technique of inspecting and assessing claims related to human understandable language. It is related to the common study of reasoning operations by computational progression where the importance is usually given to the task of knowledge representation. There are two different types of techniques in NLP for CQA system:

- Shallow technique
- Deep technique

The first technique is usually uses a keyword technique to find the stimulating segments and sentences from the recovered documents and then collects the similarities between the applicant text and the required type of answer. This approach is more appropriate for factoid styles of question, which request about simple realities or relations and can be responded simply with only a few words usually as noun phrase (Prager, 2006). Whereas in deep approach that is more complex, concerns numerous processes including, analysis of question types, assessment of answer type, request extension, analysis of answer and needs a number of NLP approaches (Corston et al. 2005). There are numerous types NLP techniques, for example, Part of Speech (POS), Name Entity Recognition (NER), Apposition, Relation, Co-reference, and Word Sense Disambiguation (WSD).

2.3.2.3 Link Analysis

This kind of technique analyzes the links among the actors of social network to show their association. Such associations can be applied to develop rank esteem in social network study. In case of collaborative social network, such as Yahoo! Answers, the link analysis can be applied to rate the answers corresponding to the status of the users. In terms of web, exploring the link analysis has been effectively used. An excellent example of link analysis application is in search engine called "PageRank", because of this "Google" come to be very famed (Liu, 2007).

2.3.2.4 Statistical Technique

To determine the worth of answers the statistical assessment approach can be used on non-textual characteristics like number of occurrences of the questions, the length or size of the answer and its rate of acceptance. These methods are as difficult as the application of complex machine learning approaches on non-textual and appropriate characteristics in order to study the pattern of question answer to evaluate the worth of the answer. There are numerous types of probability-centered classification and clustering approaches to solve partial or whole question for instance Bayesian, Maximum Entropy and Markov Chain (Florian et al, 2003).

Zhang (2007) examined the "Java Forum", it is a big online group of users searching for help, and suggested a Z-score calculate according to the examination that the users asking the question lack awareness whereas the people providing answers of asked questions have more knowledge. Z-score is used to joined with users asking and answering patterns.

Guo (2008) identified that less contribution rate of users in collaborative Question Answering System is a critical issue. He recommended all probable answer suppliers for the question, as an alternative of selecting the top answer from the collection of answers proposed for a question.

In this modern age of science, the advance communal, methodological and financial growth have made it promising a new standard of software improvement and problem solving through lightly structured cooperation of individuals on the world wide web. A huge amount of data, knowledge is available on the internet in the form of electronic copy but there is no real information access tool present to offer the people with appropriate information access. Internet technologies in different application areas have greatly transformed the life styles and communications. With fast development in E-learning, combined learning is essential for sharing knowledge. Users can ask their questions from other users when they like to collaborate, asking from others for knowledge, evaluating other's information and knowledge. One question will has many answers that must be picked. There is a technique in collaborative learning to predict the quality of the answer. In the process of combined learning, the "knowledge-base" will be improved/enhanced for Question Answering Systems in future. In this case, the users will be able to get the answers from others as well as offered by the system (Arai & Handayani, 2013).

CHAPTER THREE

QUESTION CLASSIFICATION IN COLLABORATIVE QUESTION ANSWERING SYSTEMS

3.1Introduction

To find the answer of a question in natural language, Question Answering Systems analyze the given question to create some description of required information. This task involves a number of sub-tasks such as question classification, identification and extraction of important words in the question, prediction of type of entities in anticipated answer, and query expansion. In this chapter, we concentrate primarily on question classification. A detail description of question classification and its importance in the process of Question Answering System is provided in this chapter. Different methods or models used by various types of Question Answering Systems are discussed in brief. In this chapter, we discuss different models of question classification, but mainly focus on semantic approach for question classification using WordNet and Wikipedia. The processes of query expansion shall be explained and analyzed in next chapter.

3.2 Question Classification

The process of Question classification involves defining its class. The Question Answering Systems usually define a definite set of question classes though there is always some possibility of adding new classes. The objective of this process is to match a question with one of the prearranged classes. In context of CQA systems, the process of question classification allows systems to conclude a source of information that is required to be explored for the asked question. Ontologies can be used to represent different types and sub-types of questions. Such ontology is typically a hierarchical organize, it is a taxonomy where its higher level ideas seem to be relatively distinctive based on the purpose of a question. Researchers have proposed different question categories. The matter of organizing requests into a classification was first deliberated and printed in late 1970. Lehnert (1977) presented a set of thirteen types of all questions that can probably be asked by a user fall into. All these categories are presented in Table 3.1, with some questions as examples related to them (Lapshin, 2012).

Table 3.1: Question categories (Lehnert, 1977)

Question category	Definition
Causal Antecedent	Why did Vasiliy come to Moscow?
Goal Orientation	Why did Ivan take this book?
Enablement	What does Oleg have to do in order to leave?
Verification	Has Ivan left?
Disjunctive	Were Osya or Kisa here?
Instrumental/Procedural	How did Vasiliy get to Moscow?
Concept Completion	What did Ivan eat?
Expectation	Why didn’t Ivan come to Moscow?
Judgmental	What does Vasiliy have to do so that Masha will not leave?
Quantification	How many people gather at this stadium?
Feature Specification	What is the color of Ivan’s eyes?
Request	Will you please pass me the salt?
Causal Consequent	What happened when Martin left?

3.2.1 Significance of Question Classification

In Question Answering Systems, the question classification plays a key role. Even though different kinds of Question Answering Systems have dissimilar architecture, mainly adhere to a structure where question classification performs a critical role. It has been observed that the functioning of question classification has substantial effect on the whole function of QA system (Voorhees, 2001). Generally, there are two important purpose of question classification (Zhou et al, 2012):

- Finding the answer entity: The question classification process supports in forecasting the kind of entities required to be existing in the applicant by classifying the question into several query classes. Prior knowledge of the class of the question helps not only minimize the search space require to get the answer, it also helps to find a correct answer in existing collection of candidate answers. This supportive information becomes helpful in ranking the user answer. To search the answer of asked question the initial or basic information of entity type like place, person, event etc. estimated to be existed in the answer is quite vital. For example consider the question, “who is the secretary general of UNO in 2015?” There are two types of classes, human (person) and year, which should be looked for while searching the answer of this question. Question Answering System will give more rank to such documents having the information about person and year.
- Creating answer pattern: The second important purpose of Question Classification is to create semantic patterns for the user answers. Such patterns are very useful in matching process and finding the user answers. The class of a question can be used to select the searching approach as soon as the question is restructured to a query completed Information Retrieval (IR) engine.

3.3 Question Classification Methods

In question classification, there are two different methods / approaches:

▪ **Rule based methods** attempt to match the asked question through some manually created rules. Such methods face some problems, need to describe a number of rules. Moreover, this type of approach might perform very well on some specific dataset, but they cannot show the same performance on different or new dataset and subsequently it is hard to measure them.

- **Learning based methods,** on other side, accomplish the classification by separating some characteristics from the questions; develop a classifier and expecting the class marker by using the prepared classifier. A number of effective learning based classification methods have been projected.

There is another approach called hybrid approach that combines both of these rule based and learning based approaches. A research study of Silva et al. (2011) is known as successful works on this method of question classification where, first compare the question with some already defined rules and after that apply the matched rules by means of features in the method of learning based classifier. Subsequently learning based and hybrid approaches are very successful in the process of question classification. Most of the modern works are built on these methods (Ray et al., 2010).

There are different retrieval models available for question retrieving process. In this research, some of these models are being studied. Description of these models is given below (Cao et al, 2010):

- **Vector Space Model**

In the process of question retrieval, Vector Space Model has been applied very extensively. Let us consider a general deviation of this model. Assumed a query q and question d, the grade score $S_{q,d}$ of the question can be calculated as given below:

$$\mathrm{Sq, d} = \frac{\sum t\epsilon\ q\cap d\ w\ q,t\ wd,t}{WqWd} \qquad (1), \qquad \text{Where}$$

$$w_{q,t} = \ln(1+\frac{\mathrm{N}}{ft}),\quad w_{d,t} = 1+\ln(t\,f\,t,\mathrm{d})$$

$$\mathrm{Wq} = \sqrt{\sum_{\mathrm{t}} \mathrm{w}^2\ \mathrm{q,t}}$$

$$Wd = \sqrt{\sum_{t} w^2\ d, t}$$

Where N is total number of questions in the whole dataset, ft shows the questions containing the term t and t f t,d is considered as frequency of term t in d. Where term Wq is ignored, as it is persistent for a specified query and it does not disturb the rankings of old questions. It is noted that wq,t catches the inverse document frequency of t in the group and wd,t catches the term frequency of t in d.

There are some advantages and disadvantages of Vector Space Model as given below.

Vector Space Model supports in ranked reclamation, the terms are weighted by position and it offers limited matches. Similarly this model has some disadvantages like assumes terms are self-determining and weighting is spontaneous but not very strict.

- **Okapi BM25 Model**

Vector Space Model is appropriate and favorable for small questions whereas OkapiBM25 Model solves this problem. This model is used to retrieve the questions. For a query q and question d the score of rank $S_{k,d}$ is computed as following:

$$Sq, d = \sum_{t \in q \cap d} w\, q, t\, w\, d, t \qquad (2)\,,$$

Where

$$w_{q,t} = \ln\left(\frac{N - ft + 0.5}{ft + 0.5}\right) \frac{(k3+1)tf\ t,q}{k3 + t\ f\ t,q}$$

$$w_{d,t} = \frac{(k1+1)tf\ t,d}{Kd + t\ f\ t,d}$$

$$K_d = k1((1-b) + b\frac{Wd}{WA})$$

Where N is the number of total questions in the group, f_t represents the number of questions in term t, $tf_{t,d}$ is the frequency of t in d, k1, b, and k3 are parameters need to set to 1.2,0.75, and ∞, correspondingly.

- **Language Model**

This language model is applied in past work to retrieve the question. The central concept of this language model is to assess language model for every question and then find the rank of the questions with the help of probability of the query agreeing to the projected model for the questions. For a given query q and question d the score of ranking $S_{q,d}$ is calculated as:

$$\boldsymbol{S}\ \boldsymbol{q},\boldsymbol{d} = \prod\nolimits_{t\in \boldsymbol{q}}((1-\lambda)Pml(t|\mathrm{d}) + \ \lambda Pml(t\ |\mathrm{Coll})) \qquad (\mathbf{3}), \text{ Where}$$

$$Pml(t|d) = \frac{tf\ t,d}{\sum_{t'\in d} tft',d}$$

$$Pml(t|Coll) = \ \frac{tft,Coll}{\sum_{t'\in Coll} tft',Coll}$$

Where $Pml(t|\mathrm{d})$ is calculated as maximum likelihood approximation of the word t in d, $Pml(t|\mathrm{Coll})$ is calculated as the maximum likelihood approximation of the word t in collection Coll, λ is taken as leveling parameter.

- **Translation Model**

In the past it was reported that Translation Model produces persistently extra ordinary. This model exploits word translation possibilities. For a given query q and question d the score S $_{\mathrm{q,d}}$ of rank is calculated as following:

$$\mathrm{S\ q,d} = \prod_{\mathrm{t}\in\mathrm{q}} \left((1-\lambda) \sum_{\omega\ \in \mathrm{d}} \mathrm{T(t|w)Pml(w|}d \right) + \lambda\mathrm{Pml(t|Coll))} \quad (4)$$

Pml (*w*|d) and *Pml* (*t*|Coll) can be calculated same as in equation 3 for Language Model and $T(t|w\ )$ represents the possibility that word $w$is interpretation of word t.

- **A semantic approach for question classification using WordNet and Wikipedia**

In the process of Question Answering System (QA), question classification performs a critical role in defining the expectations of the people. It is observed that some existing Question Answering System do not perform very well because of poor designing of question classification. Therefore, a perfect question classification is very important for an ideal Question Answering System. Ray, et al (2010) proposed an approach on question classification that manipulates the strong semantic characteristics of Word Net and huge information storage of Wikipedia to define enlightening terms clearly. They used a huge number of questions, more than five thousand and then verified it over collections of five TREC questions set. While performing these experiments / tests a critical enhancement in the reliability of question classification was being observed. The accuracy of question classification recommends the usefulness of this approach in the discipline of open domain question classification.

3.4 Question types Organization

When we assign semantic type to a question, it is called question classification (Zhang & Lee, 2003). Zhang and Lee (2003) suggested a two layered taxonomy of question as shown in the following table 3.2. There are six coarse grained types and fifty well grained sub types. However, the coarse grained types can do for problem analysis of question type but a well grained type description is extra helpful in tracing and confirming reasonable answers.

Table 3.2: The coarse and fine types of question (Zhang & Lee, 2003)

Coarse	Fine
ABBR	acronym, expansion
DESC	explanation, description, method, reason
ENTY	animal, form, color, currency, creation, medical/disease, event, food, instrument, language, other, letter, product, plant, sport, religion, substance, symbol, technique, term, word, vehicle
HUM	group, description, title, individual
LOC	country, city, other, mountain, state
NUM	count, date, code, distance, money, other, order, period, percent, temperature, period, speed, weight, size

The type of questions could be evaluated by categorizing the question into well-defined question types. Later on, the output of categorizing of the question can minimize amount of time applied on choosing the finest answer.

3.5 Research Methodology for Proposed Question Classification

In this section, the proposed question classification method is explained. In this research, we applied a semantic approach for question classification. WordNet and Wikipedia have been used as semantic resource in the process of question classification in a Collaborative Question Answering (CQA) System. We have also described the sources of data for conducting the experiments, the tools required and results.

3.5.1 Tools and Ontological resources

3.5.1.1 Wikipedia

Wikipedia is considered as a huge information database available on the Web. It exceeds other types of information and knowledge sources in its analysis of conceptions, productive semantic information and advanced content. In the field of Information Retrieval (IR), Wikipedia has been used to solve various problems like classification of documents and text clustering. So it became very useful in the field of IR (Wang & Domeniconl, 2008).

In Wikipedia each topic is described by a particular article, the title of the topic is a concise and well-fashioned expression that looks like a phrase in a traditional vocabulary (Milne et al., 2006). Every article is connected to minimum one type and there are hyperlinks among articles to acquire their semantic associations. All such semantic relationships contain these types of relations equivalence, hierarchical types of relation and association types of relations. Wikipedia is developed for the purpose of common human uses so it is considered as an open data resource. That is why it is unavoidable contains more noise and semantic knowledge in it and not appropriate for direct applicable in the process of question retrieval in collaborative Question Answering System. To develop it clean and easily accessible as thesaurus, it is required to preprocess the data of Wikipedia to access its impressions and then explicitly develop association between Wikipedia built on the fundamental understanding of Wikipedia. In Wikipedia, an article explains a separate topic and its label can be applied to exemplify the concept for example, "United Kingdom". Though some articles are pointless, they are just applied for management and organization of Wikipedia, for example "1999s", "List of magazines", etc.

The structural relation of Wikipedia is very productive, for instance the synonym, polysemy, hypernym and associative relationship. These semantic relations are completely stated in terms of hyperlinks between the articles of Wikipedia (Milne et al., 2006).

- **Synonym**

In Wikipedia, there is only single article for a given topic by applying forward hyperlinks to group correspondent ideas to the favorite one. These forward links deal with capitalization and spelling differences, acronyms, synonyms and common terms. In Wikipedia, the synonym mostly derives from these forward links. For instance "IBM" is an item with a huge number of forward links, synonyms (I.B.M, Big blue, IBM Corporation) (Cai et al., 2011). Along with this, the articles of Wikipedia usually express other ideas as well, which previously have equivalent articles in Wikipedia. However the anchor text on every hyperlink might be dissimilar with the label of related article. Consequently anchor texts can be applied as additional source of synonym (Hu et al., 2008).

- **Polysemy**

The impression of polysemous is provided by disambiguation pages in Wikipedia. All conceivable meanings related to the equivalent idea are listed on disambiguation page and every meaning is deliberated in the form of article. For instance a disambiguation page of word "IBM" records three related concepts, containing "Inclusion build myositis", "Injection blow molding", and "International Business Machine" (Cai et al., 2011).

- **Hypernym**

In Wikipedia, the concepts and categories belong to as a minimum one group and groups are organized in the form of ordered structure. The consequential hierarchy formed a directed graph where many classification schemes exist at the same time. To obtain the actual classified associations from Wikipedia groups, an approach to develop general hierarchical relation from group links is used. In this way hypernym for every Wikipedia perception can be achieved (Milne et al., 2006).

- **Associative relation**

Each article of Wikipedia holds a number of hyperlinks and these hyperlinks describe the connection among them. The links between the terms are related distantly. Let us consider an example to compare the two links, one from the object of "IBM" to the object "Apple

Inc.", and the second is from an article "IBM" to "Software engineer". It is very vibrant that first two articles are closely related with each other than the second pair. Thus it is very critical matter to measure the association of hyperlinks inside the articles in Wikipedia (Cai et al., 2011).

For research, data about Wikipedia can obtained from its link[3] easily; data is accessible in the form of dump databases, which released occasionally.

3.5.1.2 WordNet

For the English language, there are various dictionary tools are available, which work as vocabulary database. WordNet is one of them. These tools group the English words in the form of sets of synonyms known as synsets. It offers brief complete definitions and keeps the numerous semantic relationships between these sets of synonyms. WordNet is a philological reference system available on the web. Its design stimulated by contemporary psycholinguistic concepts of human being lexical memory. Traditional in-order processes for organizing lexical data set together words, which are spelled in the same way and disseminate words with related or associated meanings randomly across the list (Miller, 1990). WordNet is an excellent and simply available ontological source to discover series of a word in order. Therefore, we use WordNet to discover the greatest rank descriptor for question expressions and contain them as supplementary exploration terms (Prager, 2001).

WordNet is considered as strongly used accessible semantic resources in the process of Question Answering System exclusively in the field of query expansion and question classification. To classify the questions, the head words and respective hypernym from Word Net were applied and correctness of 89.2% observed (Huang et al, 2008). The tests with WordNet identify that application of semantic knowledge for question classification highly increases the efficiency of Question Answering (QA) Systems (Ray et al, 2010).

[3] www.download.wikipedia.org

3.5.2 Data Collection

In order to conduct the experiments for the proposed research methods in this chapter and next chapter, we search the Internet for set of standard question on the pattern of TREC or CLEF questions. However, there were no such question sets for CQA systems. The questions in TREC and CLEF are usually smaller in size (number of words in questions) and their answers are in document collections prepared for this purpose. However, users in typical CQA environment ask long questions. These questions may sometimes consist of 3-4 sentences. Thus, TREC or CLEF questions are not suitable for experiments with CQA systems. This compels us to prepare our own question dataset. We collected one thousand questions from different online CQA systems. The questions are related to different topics of daily life so that heterogeneity of questions could be maintained. Some example topics are games (Football, cricket, Olympics), computer security, programming languages etc. We are giving a brief description of CQA systems from where questions in our dataset were collected. :

- **Allexperts[4]:** Allexperts was developed in the beginning of 1998. In the history of Internet, it was the first large scale service of question and answer provided on the net. There are a huge numbers of experts including expert lawyers, engineers, doctors and scientists who provide the answers of questions asked by users. Most of the answers are provided within a day. All answers are freely available on this website.
- **Answerbag[5]:** Answerbag is known colloquially as "AB". It was developed in 2003 by Joel Downs and acquired by Info-search Media in the beginning of 2006 and later on in October 2006 it joined Demand Media. It is a combined online databank of frequently asked questions where the questions are asked and answers by different users. More than one answers of the question are provided there. In December 2006 it became second biggest social Q&A website after Yahoo Answers.
- **Ask MetaFilter[6]: Ask MetaFilter** began in December 2003. It is a question, answer (Q&A) site, and at present comprises of million questions through twenty different types.

[4] http://www.allexperts.com/central/service.htm
[5] http://en.wikipedia.org/wiki/Answerbag
[6]http://ask.metafilter.com/about.mefi

In this website the members provide mutual supports to each other where the users post questions and finds answers provided by other users of the community. Questions can also be unidentified, by means of the distinct form for that job.

- **Quora**[7] : Quora is one of the rich website of question and answer where the questions are produced, replied, modified and systematized by its communal of users. This company was established in June 2009, and this website was prepared accessible to the community on June 21, 2010. Quora accumulates questions and answers to subjects. The users can cooperate by checking questions and advising corrects to other users answers.
- **Answers**[8]**:** Answers is a web based information exchange website, which contains Wiki Answers, Reference Answers, Video Answers. It provides questions and answers in five international languages. In 1996, this domain name "Answers.com" was bought by businesspersons Bill Gross and Henrik Jones at idea lab. This domain name was developed by Net Shepard and successively purchased by Guru Net. This website is the most important product of Answers Corporation (formerly known as Guru Net), question and Answers Company with office in the City of New York. It was established in 1999 by Bob Rosenschein. The website supports different languages of the world such as English, Italian, French, German, Spanish, and Tagalog.
- **Answers Corporation**[9]: Answers Corporation main task is to authorize customers, products, and establishments by linking them with the facts their requirement to make well-informed judgments. Answers Platform influences the substantial range of the top twenty websites. This company in coordination with other companies like ForeSee, Web-collage, and Reseller-Ratings is trying to provide an opportunity to all business organizations to get advantages of cloud based solutions to be involved with their customers at every communication point. This technique offers an opportunity to the customers to take investment decision and enhances their experience as well.

We collected 1000 questions from the above CQA systems. We used 700 of these questions for studying and writing the question patterns. The remaining 300 questions were used for

[7] http://en.wikipedia.org/wiki/Quora
[8] http://en.wikipedia.org/wiki/Answers.com
[9] https://www.crunchbase.com/organization/answers-corporation

testing these rules. This set of 300 questions is called test dataset Table 3.3 presents domains from which questions were collected and number of questions in each domain.

Table 3.3: Types of questions

Category	Size
Computer science	150
Programming	150
Olympics	300
Football/Soccer	150
Cricket	150
General	100

3.5.3 Proposed Question Classification Method

We carefully reviewed the group of collected questions and recognized different patterns in them. Rule based classification technique was applied to develop different types of patterns related to the questions. Seven patterns for (wh) questions and one for other types of questions were design from syntactic point of view. A detailed explanation of these patterns is given below.

- Questions of functional words: This category includes non-wh questions excluding how. Such types of questions usually begin with insignificant verb expressions. For example, Name past and present LPGA commissioners.
- When questions: This category includes all questions start with keyword "When" and are time-based. The common pattern of this category is "When (do | does | did | AUX) VP NP X", where AUX, VP and NP signified auxiliary verbs, verb phrases and noun phrases correspondingly will contain the similar meaning in the whole research. The character "|" is considered as Boolean "OR" operator and "X" mean any grouping of words. For example, when was Microsoft established?
- Where questions: These category of questions start with keyword "Where" and are associated to some places. It represents natural entity like terrestrial boundaries, man-

made places (buildings) and mountains. The pattern of this type is "Where (do | does | did | AUX) VP NP X". For example, where are zebras most likely found?

- Which questions: In general, this type of question starts with "Which" keyword. Its pattern is "Which NP X". The answer type of this kind of question is determined by entity type of NP. For example, Which British team has Manchester United played?
- Who/Whom/Whose questions: The pattern of such questions fall in this group is, "(Who | Whom | Whose) (do | does | did | Aux) (VP) (NP) X?" Such type of questions usually related to a person or about an organization. For example, Who were leading players for Manchester United in the 1990's?
- Why questions: These types of questions ask for particular reasons or descriptions. Its general pattern is, " Why (do | does | did | Aux) VP (VP) (NP) X". For example, why can't ostriches fly?
- How questions: There are two types of patterns in this group. "How (do/ does / did / Aux) VP NP X?", "How (big | long | fast | much| many| afr| ,,,) X?" In the first pattern, the probable answer is explanation of some procedure whereas the second type questions return some numeric type answers. For example, how much did Mercury spend on advertising in 1993?
- What question: There are several kinds of patterns in this category. Very common expression for this type of questions is written as, "What (NP) (do/does/did/ AUX) (functional words) (NP) (VP) X?" These types of questions usually ask about virtual thing. For example, what are the titles of songs written by John Prine?

All these patterns of questions are very useful for further classification of the questions and summarized in table 3.4

Table 3.4: Classification of Questions

Question Classification	Type/Pattern	Examples
When	Date	When was Microsoft established
Where	Location	Where are zebras most likely found?
Which	Entity type	Which British team has Manchester United played?
Who / Whom/Whose	Person/organization	Who were leading players for Manchester United in the 1990's?
Why	Reason	Why can't ostriches fly?
How	Explanation	How hot does the inside of an active volcano get? How much did Mercury spend on advertising in 1993?
What	Definition Number Title	What is the name of the managing director of Apricot Computer? What was the monetary value of the Nobel Peace Prize in 1989? What are the titles of songs written by John Prine?

Proposed Method

We adopted the Question Classification method as described in (Ray et al, 2010). A brief description of the above method is given here:

Step (1) Make a Question Taxonomy containing Coarse-grained type and fine grained subtypes.

Step (2) Define Question patterns (what, when, where, which, who, why, how).

Step (3) Define expected answer entity for each question pattern as described in step 2.

Step (4) For "what" type of question use Wikipedia and WordNet to derive expected entity type.

Step (5) Prepare question collection.

Step (6) Conduct experiments.

Step (7) Present the result.

3.6 Experimental Results

The efficiency and performance of question classifier is assessed by computing the accuracy of some specific classifier in test data set. It can be obtained by using the following equation:

$$\text{Accuracy} \ = \frac{\text{no.of accurate classified samples}}{\text{Total tested samples}} \qquad (5)$$

Accuracy is the whole perfection of the model and it is computed as the sum of accurate classification divided by total number of classifications. We tested the proposed question classification method over 300 questions in our test question dataset. The expected answer types were predicted for each of these questions using Wikipedia and WordNet. We correctly predicted 273 questions. The accuracy of the method was 0.91 as shown below.

$$\text{Accuracy} = \frac{273}{300} = 0.91$$

To the best of our knowledge in CQA systems, we could not find related result using rule based question classification technique. However, most of the work is being done using various machine learning approaches.

CHAPTER FOUR

QUERY EXPANSION

4.1 Introduction to Query Expansion

The process of adding semantically equivalent words appropriate to a query to make the query to be very clear and distinctive, is called query expansion. Different types of techniques are available to improve the query interpretation by re-expressing the queries. query expansion is possibly one of the most efficient approaches. In the process of information retrieval (IR), query expansion is represented to as algorithms, techniques or approaches that re-create the actual query by including new terms in the query so that to accomplish an improved retrieval successfulness (He & Ounis, 2009).

The process of IR emphasis on searching documents where the contents to be identical with consumer query from huge collection of documents. As well designed and accurate query is challenging for most of the users, so it is important to use query expansion to extract appropriate information. Query expansion approaches are extensively used for improving the competence of written information retrieval system. All these approaches assist to minimize vocabulary mismatch problems by expanding the actual query with extra-related terms and reassessing such terms in the extended query (Rivas et al, 2014).

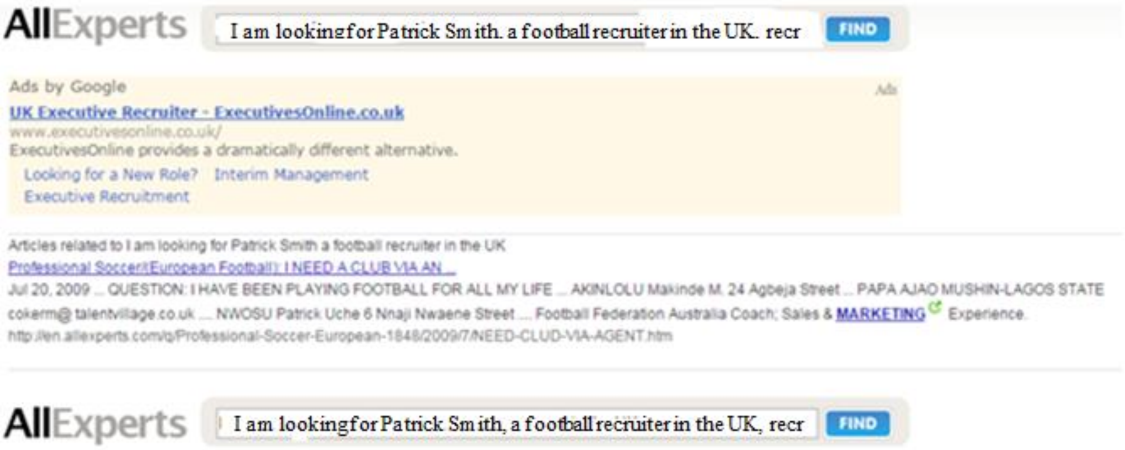
4.2 Need for Query Expansion

The process of adding supplementary terms or expressions to the primary question (query) to enhance the retrieval performance is known as query expansion. Query expansion plays a very important role in Information Retrieval process. Query expansion boosts up the recall function of information retrieval process of CQA Systems. For example, Information Retrieval emphases on retrieving documents whose words match with user question from a huge collection of documents. Formulating well-prepared queries is hard for most of the users. Therefore, it is important to use query expansion to find appropriate information.

Let us consider the following example, which elaborates the importance of query expansion. This question was taken from AllExperts.com:

Question Statement

"I am looking for Patrick Smith, a football recruiter in the UK, recruiting for clubs in the US. I have not been able to find him listed in any organization; is it possible to be an independent recruiter? He has some odd stories and I am suspicious he is a scammer. His assistant is James Sapseed."



I am lookingforPatrick Smith. a footballrecruiterin the UK. recr

Ads by Google

Ads

UK Executive Recruiter - ExecutivesOnline.co.uk

www.executivesonline.co.uk/

ExecutivesOnline provides a dramatically different alternative.

Looking for a New Role? Interim Management

Executive Recruitment

Articles related to I am looking for Patrick Smith a football recruiter in the UK

Professional Soccer(European Football): I NEED A CLUB VIA AN ...

Jul 20, 2009 ... QUESTION: I HAVE BEEN PLAYING FOOTBALL FOR ALL MY LIFE ... AKINLOLU Makinde M. 24 Agbeja Street ... PAPA AJAO MUSHIN-LAGOS STATE cokerm@ talentvillage.co.uk .... NWOSU Patrick Uche 6 Nnaji Nwaene Street .... Football Federation Australia Coach; Sales & MARKETING Experience.

http://en.allexperts.com/q/Professional-Soccer-European-1848/2009/7/NEED-CLUD-VIA-AGENT.htm

I am lookingforPatrick Smith, a footballrecruiterin the UK, recr

Figure 4.1: Example of unmatched question

This question was collected from existing CQA system (Allexperts.com). However, when we fed the same question in the input system of this CQA system, the system did not return any result.

Query expansion techniques are extensively applied for refining the competence of the written information retrieval methods. These techniques support to solve vocabulary mismatch matters by expanding the primary query with added applicable expressions and reweighting the words in the extended query (Rivas et al, 2014).

4.3 Query Expansion Approaches

There are three basic ways to expand the query (Carpineto et al, 2012). These methods are given below:

- **Manual:** This type of query expansion needs the involvement of the users. This is related with Boolean online search. This type of query expansion is accomplished by choosing the terms of query to expand it manually, and explaining the topic of the query through thesaurus such as GNU Collaborative International Dictionary of English.

- **Interactive:** In this type of query expansion, both the system and user are responsible to specify and choose the required terms by query expansion. This task can be carried out in two steps. At first the system need to select, retrieve and then rank the terms of expansion. In second step, the user must decide that which supportive terms are essential for the query from the listed ranked terms.
- **Automatic**: In this type of query expansion, the retrieval system is liable for enhancing the basic or following queries.

4.4 Experimental Setup for Query Expansion

4.4.1 Data Collection

For testing the query expansion approach, we have used the same dataset as described in Chapter 3.

4.4.2 Tools Used

This section describes various types of tools and their applications used in the process of expanding query for CQA systems. Different types of tools are used in the process of matching the questions, For example WordNet can be used to collect equivalent words for query expansion to increase the chances to match the questions with existing questions in the corpus of CQA system. We are using DocFetcher and GNU Collaborative International Dictionary of English (CIDE) to conduct the experiments.

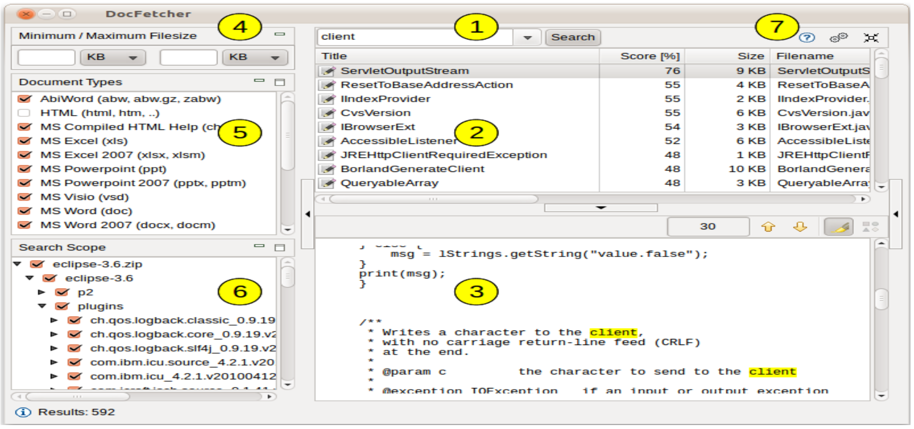
- **GNU Collaborative International Dictionary of English[10] (CIDE)**

The CIDE was developed from Webster's Dictionary and it has been enhanced using some of description from WordNet. It was prepared and checked by different people from all over the world. An electronic version of CIDE dictionary is also available as a free on internet to develop a modern inclusive encyclopedia dictionary. It was developed by combined efforts of all persons willing to assist build a great and freely accessible knowledge base. There are a number of derivatives types of this kind of dictionary on internet available with good user interface.

[10] http://en.wikipedia.org/wiki/Collaborative_International_Dictionary_of_English

- **DocFetcher**

DocFetcher is an important application used to search the documents. It is an open source desktop application used to examine the contents of different files on our computer. It can be considered as Google search engine for local existing files. DocFetcher supports various kinds of archive formats, like zip, 7z, rar and tar.* and it has powerful query syntax. Main screen of this application is shown in the following figure 3.0. Queries are written in a field at (1). After searching the files, it displays the results in pane (2). In pane (3) it displays text only preview of file being selected in the output pane. The matched texts are emphasized in yellow color. The results can be filtered by choosing minimum\ maximum size of file at (4), type of file at (5), and place or location at (6). Buttons at (7) are used to open the manual, open preferences and to minimize the program at status bar. The following screenshots are taken from the web link of DocFetcher on source-forge[11].



client

☑ AbiWord (abw, abw.gz, zabw)
☐ HTML (html, htm, ..)
☑ MS Compiled HTML Help (ch
☑ MS Excel (xls)
☑ MS Excel 2007 (xlsx, xlsm)
☑ MS Powerpoint (ppt)
☑ MS Powerpoint 2007 (pptx, pptm)
☑ MS Visio (vsd)
☑ MS Word (doc)
☑ MS Word 2007 (docx, docm)

```
▼ ☑ eclipse-3.6.zip
  ▼ ☑ eclipse-3.6
    ▶ ☑ p2
    ▼ ☑ plugins
      ▶ ☑ ch.qos.logback.classic_0.9.19
      ▶ ☑ ch.qos.logback.core_0.9.19.v2
      ▶ ☑ ch.qos.logback.slf4j_0.9.19.v2
      ▶ ☑ com.ibm.icu.source_4.2.1.v20
      ▶ ☑ com.ibm.icu_4.2.1.v20100412
```

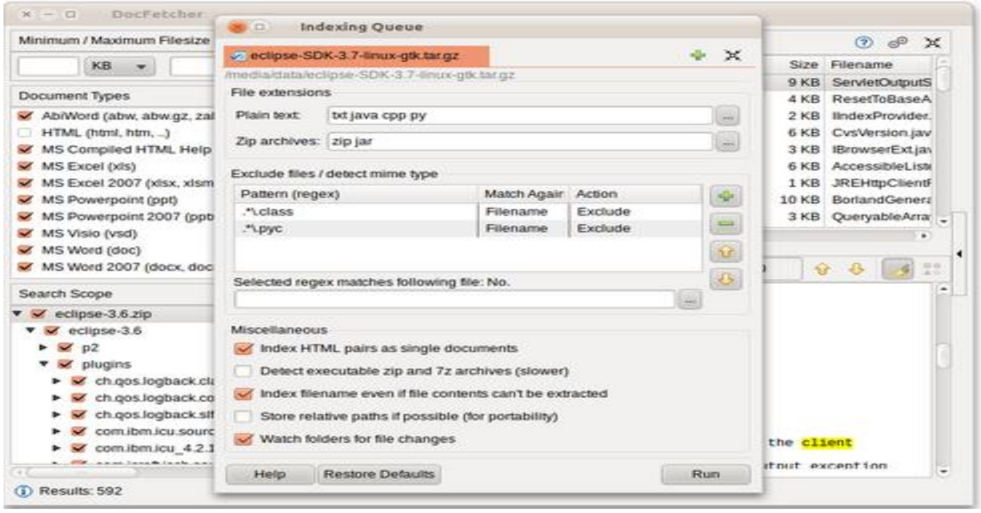
Title	Score [%]	Size	Filename
ServletOutputStream	76	9 KB	ServletOutputS
ResetToBaseAddressAction	55	4 KB	ResetToBaseA
IIndexProvider	55	2 KB	IIndexProvider.
CvsVersion	55	6 KB	CvsVersion.jav
IBrowserExt	54	3 KB	IBrowserExt.jav
AccessibleListener	52	6 KB	AccessibleListe
JREHttpClientRequiredException	48	1 KB	JREHttpClientF
BorlandGenerateClient	48	10 KB	BorlandGenera
QueryableArray	48	3 KB	QueryableArra

```
        msg = lStrings.getString("value.false");
    }
    print(msg);
    }


    /**
     * Writes a character to the client,
     * with no carriage return-line feed (CRLF)
     * at the end.
     *
     * @param c          the character to send to the client
     *
     * @exception IOException   if an input or output exception
```

Figure 4.2: Different components of DocFetcher

[11] http://sourceforge.net/projects/docfetcher/#screenshots



eclipse-SDK-3.7-linux-gtk.tar.gz
/media/data/eclipse-SDK-3.7-linux-gtk.tar.gz

File extensions

Plain text: txt java cpp py

Zip archives: zip jar

Exclude files / detect mime type

Pattern (regex)	Match Agair	Action
.*\.class	Filename	Exclude
.*\.pyc	Filename	Exclude

Selected regex matches following file: No.

Miscellaneous

☑ Index HTML pairs as single documents
□ Detect executable zip and 7z archives (slower)
☑ Index filename even if file contents can't be extracted
□ Store relative paths if possible (for portability)
☑ Watch folders for file changes

Minimum / Maximum Filesize: KB

Document Types

☑ AbiWord (abw, abw.gz, zal
□ HTML (html, htm, ...)
☑ MS Compiled HTML Help
☑ MS Excel (xls)
☑ MS Excel 2007 (xlsx, xlsm
☑ MS Powerpoint (ppt)
☑ MS Powerpoint 2007 (ppt
☑ MS Visio (vsd)
☑ MS Word (doc)
☑ MS Word 2007 (docx, doc

Search Scope

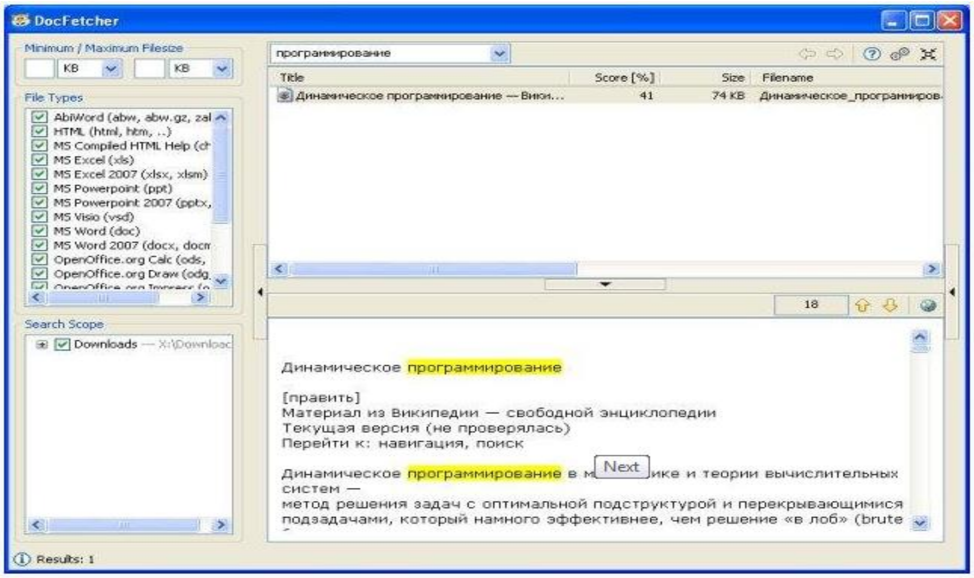
- ☑ eclipse-3.6.zip
  - ☑ eclipse-3.6
    - ☑ p2
    - ☑ plugins
      - ☑ ch.qos.logback.cl
      - ☑ ch.qos.logback.co
      - ☑ ch.qos.logback.sl
      - ☑ com.ibm.icu.sourc
      - ☑ com.ibm.icu_4.2.1

Size	Filename
9 KB	ServletOutputS
4 KB	ResetToBaseA
2 KB	IIndexProvider.
6 KB	CvsVersion.jav
3 KB	IBrowserExt.jav
6 KB	AccessibleList
1 KB	JREHttpClientF
10 KB	BorlandGenera
3 KB	QueryableArra

the client
tput exception

Results: 592

Figure 4.3: Indexing queues



Minimum / Maximum Filesize: KB KB

File Types

☑ AbiWord (abw, abw.gz, zal
☑ HTML (html, htm, ..)
☑ MS Compiled HTML Help (ch
☑ MS Excel (xls)
☑ MS Excel 2007 (xlsx, xlsm)
☑ MS Powerpoint (ppt)
☑ MS Powerpoint 2007 (pptx,
☑ MS Visio (vsd)
☑ MS Word (doc)
☑ MS Word 2007 (docx, docm
☑ OpenOffice.org Calc (ods,
☑ OpenOffice.org Draw (odg,

Search Scope

☑ Downloads — X:\Downloac

программирование

Title	Score [%]	Size	Filename
Динамическое программирование — Вики...	41	74 KB	Динамическое_программиров

18

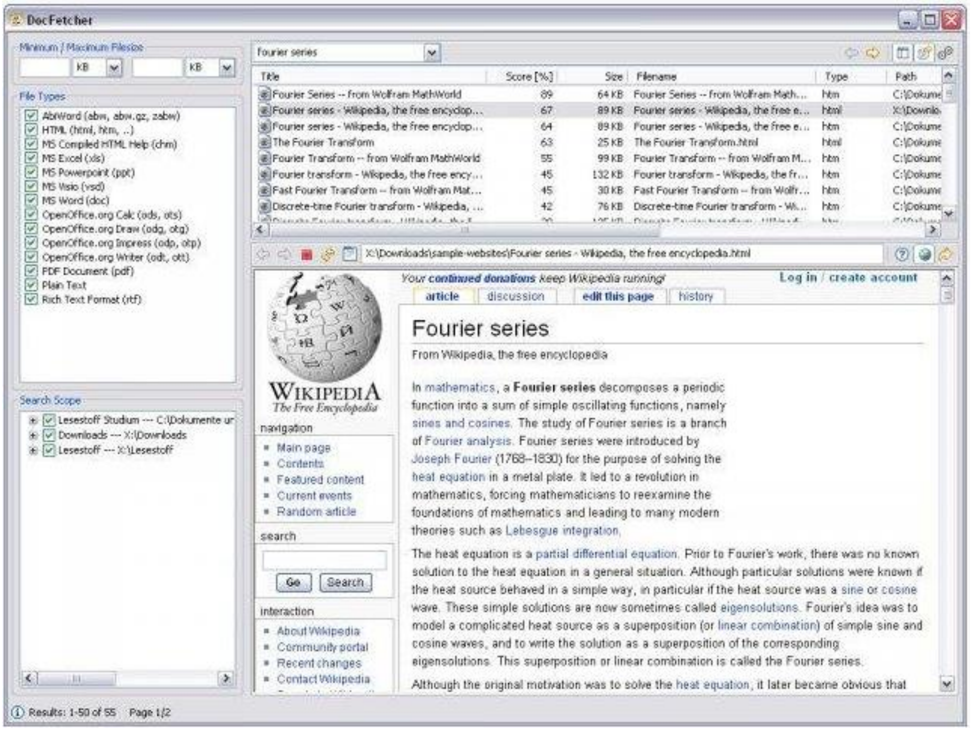
```
Динамическое программирование

[править]
Материал из Википедии — свободной энциклопедии
Текущая версия (не проверялась)
Перейти к: навигация, поиск

Динамическое программирование в м     ике и теории вычислительных
систем —
метод решения задач с оптимальной подструктурой и перекрывающимися
подзадачами, который намного эффективнее, чем решение «в лоб» (brute
```

Results: 1

Figure 4.4: Result of matched question/ answer



☑ AbiWord (abw, abw.gz, zabw)
☑ HTML (html, htm, ..)
☑ MS Compiled HTML Help (chm)
☑ MS Excel (xls)
☑ MS Powerpoint (ppt)
☑ MS Visio (vsd)
☑ MS Word (doc)
☑ OpenOffice.org Calc (ods, ots)
☑ OpenOffice.org Draw (odg, otg)
☑ OpenOffice.org Impress (odp, otp)
☑ OpenOffice.org Writer (odt, ott)
☑ PDF Document (pdf)
☑ Plain Text
☑ Rich Text Format (rtf)

☑ Lesestoff Studium --- C:\Dokumente ur
☑ Downloads --- X:\Downloads
☑ Lesestoff --- X:\Lesestoff

fourier series

Title	Score [%]	Size	Filename	Type	Path
Fourier Series -- from Wolfram MathWorld	89	64 KB	Fourier Series -- from Wolfram Math...	htm	C:\Dokume
Fourier series - Wikipedia, the free encyclop...	67	89 KB	Fourier series - Wikipedia, the free e...	html	X:\Downlo
Fourier series - Wikipedia, the free encyclop...	64	89 KB	Fourier series - Wikipedia, the free e...	htm	C:\Dokume
The Fourier Transform	63	25 KB	The Fourier Transform.html	html	C:\Dokume
Fourier Transform -- from Wolfram MathWorld	55	99 KB	Fourier Transform -- from Wolfram M...	htm	C:\Dokume
Fourier transform - Wikipedia, the free ency...	45	132 KB	Fourier transform - Wikipedia, the fr...	htm	C:\Dokume
Fast Fourier Transform -- from Wolfram Mat...	45	30 KB	Fast Fourier Transform -- from Wolfr...	htm	C:\Dokume
Discrete-time Fourier transform - Wikipedia, ...	42	76 KB	Discrete-time Fourier transform - Wi...	htm	C:\Dokume

X:\Downloads\sample-websites\Fourier series - Wikipedia, the free encyclopedia.html

Fourier series

From Wikipedia, the free encyclopedia

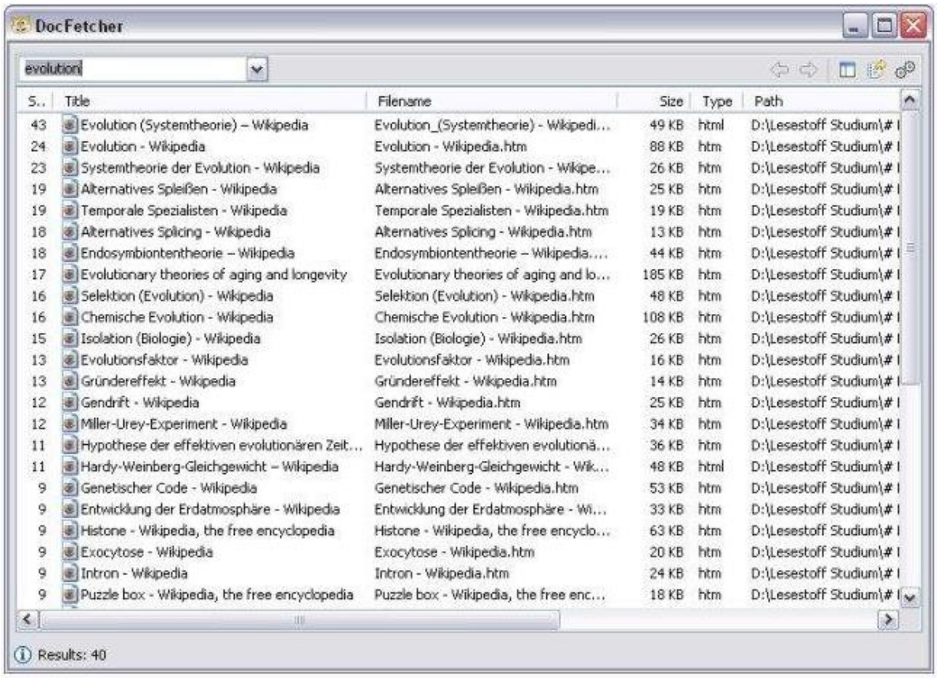
In mathematics, a **Fourier series** decomposes a periodic function into a sum of simple oscillating functions, namely sines and cosines. The study of Fourier series is a branch of Fourier analysis. Fourier series were introduced by Joseph Fourier (1768–1830) for the purpose of solving the heat equation in a metal plate. It led to a revolution in mathematics, forcing mathematicians to reexamine the foundations of mathematics and leading to many modern theories such as Lebesgue integration.

The heat equation is a partial differential equation. Prior to Fourier's work, there was no known solution to the heat equation in a general situation. Although particular solutions were known if the heat source behaved in a simple way, in particular if the heat source was a sine or cosine wave. These simple solutions are now sometimes called eigensolutions. Fourier's idea was to model a complicated heat source as a superposition (or linear combination) of simple sine and cosine waves, and to write the solution as a superposition of the corresponding eigensolutions. This superposition or linear combination is called the Fourier series.

Although the original motivation was to solve the heat equation, it later became obvious that

Results: 1-50 of 55 Page 1/2

Figure 4.5: Matched questions

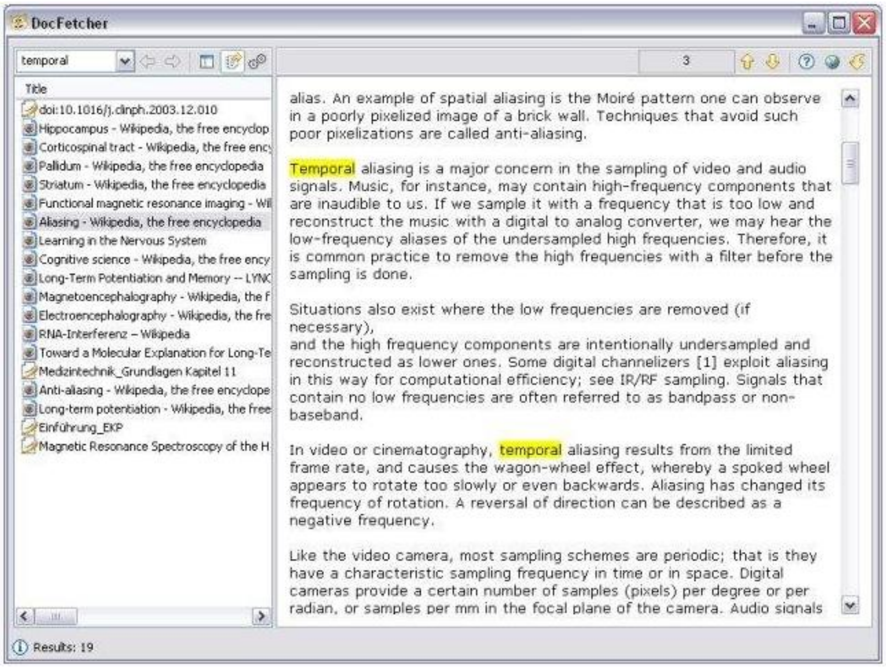


evolution

S..	Title	Filename	Size	Type	Path
43	Evolution (Systemtheorie) – Wikipedia	Evolution_(Systemtheorie) - Wikipedi...	49 KB	html	D:\Lesestoff Studium\# I
24	Evolution - Wikipedia	Evolution - Wikipedia.htm	88 KB	htm	D:\Lesestoff Studium\# I
23	Systemtheorie der Evolution - Wikipedia	Systemtheorie der Evolution - Wikipe...	26 KB	htm	D:\Lesestoff Studium\# I
19	Alternatives Spleißen - Wikipedia	Alternatives Spleißen - Wikipedia.htm	25 KB	htm	D:\Lesestoff Studium\# I
19	Temporale Spezialisten - Wikipedia	Temporale Spezialisten - Wikipedia.htm	19 KB	htm	D:\Lesestoff Studium\# I
18	Alternatives Splicing - Wikipedia	Alternatives Splicing - Wikipedia.htm	13 KB	htm	D:\Lesestoff Studium\# I
18	Endosymbiontentheorie – Wikipedia	Endosymbiontentheorie – Wikipedia....	44 KB	htm	D:\Lesestoff Studium\# I
17	Evolutionary theories of aging and longevity	Evolutionary theories of aging and lo...	185 KB	htm	D:\Lesestoff Studium\# I
16	Selektion (Evolution) - Wikipedia	Selektion (Evolution) - Wikipedia.htm	48 KB	htm	D:\Lesestoff Studium\# I
16	Chemische Evolution - Wikipedia	Chemische Evolution - Wikipedia.htm	108 KB	htm	D:\Lesestoff Studium\# I
15	Isolation (Biologie) - Wikipedia	Isolation (Biologie) - Wikipedia.htm	26 KB	htm	D:\Lesestoff Studium\# I
13	Evolutionsfaktor - Wikipedia	Evolutionsfaktor - Wikipedia.htm	16 KB	htm	D:\Lesestoff Studium\# I
13	Gründereffekt - Wikipedia	Gründereffekt - Wikipedia.htm	14 KB	htm	D:\Lesestoff Studium\# I
12	Gendrift - Wikipedia	Gendrift - Wikipedia.htm	25 KB	htm	D:\Lesestoff Studium\# I
12	Miller-Urey-Experiment - Wikipedia	Miller-Urey-Experiment - Wikipedia.htm	34 KB	htm	D:\Lesestoff Studium\# I
11	Hypothese der effektiven evolutionären Zeit...	Hypothese der effektiven evolutionä...	36 KB	htm	D:\Lesestoff Studium\# I
11	Hardy-Weinberg-Gleichgewicht – Wikipedia	Hardy-Weinberg-Gleichgewicht - Wik...	48 KB	html	D:\Lesestoff Studium\# I
9	Genetischer Code - Wikipedia	Genetischer Code - Wikipedia.htm	53 KB	htm	D:\Lesestoff Studium\# I
9	Entwicklung der Erdatmosphäre - Wikipedia	Entwicklung der Erdatmosphäre - Wi...	33 KB	htm	D:\Lesestoff Studium\# I
9	Histone - Wikipedia, the free encyclopedia	Histone - Wikipedia, the free encyclo...	63 KB	htm	D:\Lesestoff Studium\# I
9	Exocytose - Wikipedia	Exocytose - Wikipedia.htm	20 KB	htm	D:\Lesestoff Studium\# I
9	Intron - Wikipedia	Intron - Wikipedia.htm	24 KB	htm	D:\Lesestoff Studium\# I
9	Puzzle box - Wikipedia, the free encyclopedia	Puzzle box - Wikipedia, the free enc...	18 KB	htm	D:\Lesestoff Studium\# I

Results: 40

Figure 4.6: Showing total indexed files



Title

doi:10.1016/j.clinph.2003.12.010
Hippocampus - Wikipedia, the free encyclop
Corticospinal tract - Wikipedia, the free ency
Pallidum - Wikipedia, the free encyclopedia
Striatum - Wikipedia, the free encyclopedia
Functional magnetic resonance imaging - Wi
Aliasing - Wikipedia, the free encyclopedia
Learning in the Nervous System
Cognitive science - Wikipedia, the free ency
Long-Term Potentiation and Memory -- LYNC
Magnetoencephalography - Wikipedia, the f
Electroencephalography - Wikipedia, the fre
RNA-Interferenz – Wikipedia
Toward a Molecular Explanation for Long-Te
Medizintechnik_Grundlagen Kapitel 11
Anti-aliasing - Wikipedia, the free encyclope
Long-term potentiation - Wikipedia, the free
Einführung_EKP
Magnetic Resonance Spectroscopy of the H

alias. An example of spatial aliasing is the Moiré pattern one can observe in a poorly pixelized image of a brick wall. Techniques that avoid such poor pixelizations are called anti-aliasing.

Temporal aliasing is a major concern in the sampling of video and audio signals. Music, for instance, may contain high-frequency components that are inaudible to us. If we sample it with a frequency that is too low and reconstruct the music with a digital to analog converter, we may hear the low-frequency aliases of the undersampled high frequencies. Therefore, it is common practice to remove the high frequencies with a filter before the sampling is done.

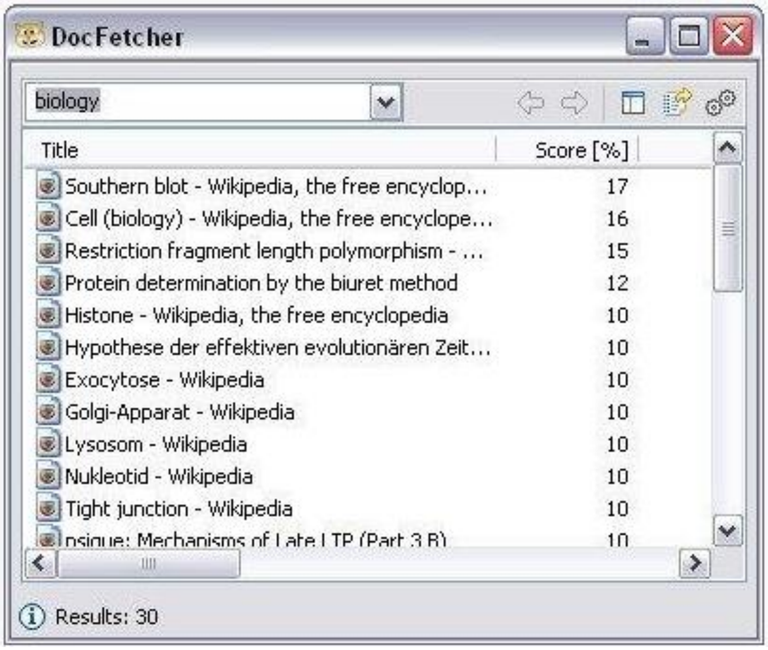
Situations also exist where the low frequencies are removed (if necessary),
and the high frequency components are intentionally undersampled and reconstructed as lower ones. Some digital channelizers [1] exploit aliasing in this way for computational efficiency; see IR/RF sampling. Signals that contain no low frequencies are often referred to as bandpass or non-baseband.

In video or cinematography, temporal aliasing results from the limited frame rate, and causes the wagon-wheel effect, whereby a spoked wheel appears to rotate too slowly or even backwards. Aliasing has changed its frequency of rotation. A reversal of direction can be described as a negative frequency.

Like the video camera, most sampling schemes are periodic; that is they have a characteristic sampling frequency in time or in space. Digital cameras provide a certain number of samples (pixels) per degree or per radian, or samples per mm in the focal plane of the camera. Audio signals

Results: 19

Figure 4.7: Result of matched words in the text



Title	Score [%]
Southern blot - Wikipedia, the free encyclop...	17
Cell (biology) - Wikipedia, the free encyclope...	16
Restriction fragment length polymorphism - ...	15
Protein determination by the biuret method	12
Histone - Wikipedia, the free encyclopedia	10
Hypothese der effektiven evolutionären Zeit...	10
Exocytose - Wikipedia	10
Golgi-Apparat - Wikipedia	10
Lysosom - Wikipedia	10
Nukleotid - Wikipedia	10
Tight junction - Wikipedia	10
psique: Mechanisms of Late LTP (Part 3 B)	10

Results: 30

Figure 4.8: Rank/Score of given question in different files

Searching the required file in huge collection of documents might slow down this process. DocFetcher needs to create indexes for those folders we like to search in. Such indexes let DocFetcher to fast search for the huge collection of files by using some keywords. This process of creating indexes may take some time, it depends on the size and number of files is being indexed. It is finished only one time for each folder and then we can search those indexed folders a number of times as we need.

Name Entity Recognition (NER) is applied to recognize names, associations and places. When QA system identifies the type of expected answer, in this case NER may help to find some certain words to correlate it well with right form of answer (Shaalan & M, 2014). Usually, NER is very suitable for when, where and who kinds of questions. Such as, consider a question, which requests, "Who is the President of US?" A user answer is that "Barak Obama is chosen as 44th President of US"? In this question the name of the US president is required. NER is capable to select Barak Obama as a name from the applicant answer and consequently the result be well suited the question (Floridan et al, 2003).

4.5 Research Methodology for Proposed Query Expansion

As we discussed already, there are different approaches of query expansion. In this research we follow a manual approach of query expansion for Collaborative Question Answering (CQA) System. We are describing the steps of the proposed query expansion method. The Proposed research for query expansion consists of:

1.Finding semantically equivalent words for the keywords from a Thesaurus.
2.Reformulate the equivalent queries.
3.Prepare the experimental dataset.
4.Test the proposed method with experimental dataset and produce the result.

We have already described the experimental dataset in chapter 3. Next subsection describes the methods to find equivalent words and reformulate the query.

4.5.1 Proposed Query Expansion Method

The proposed algorithm of query expansion is described as below:

Input: Question entered by the user (Q)

Output: Semantically enriched query (query expansion) (Q_E)

Step 1: To extract the keywords from the user query (Q) W_1, W_2, …., W_n .

Step 2: For k= 1 to n

From ontological resource to extract n semantically equalant terms for keywords

(W_1 to W_n). For keyword W_k semantically equivalent words are W_{k1}, W_{k2}, W_{k3}…W_{kn}.

Step 3: Develop a query by using Boolean operators, "AND", "OR". Such as

(W_{11} OR W_{12} OR… OR W_{1n}) AND (W_{21} OR W_{22} OR… OR W_{2n}) AND …. AND (W_{m1} OR W_{m2} OR… OR W_{mn})

Step 4: Perform experiments and evaluate the results.

Step 5: End

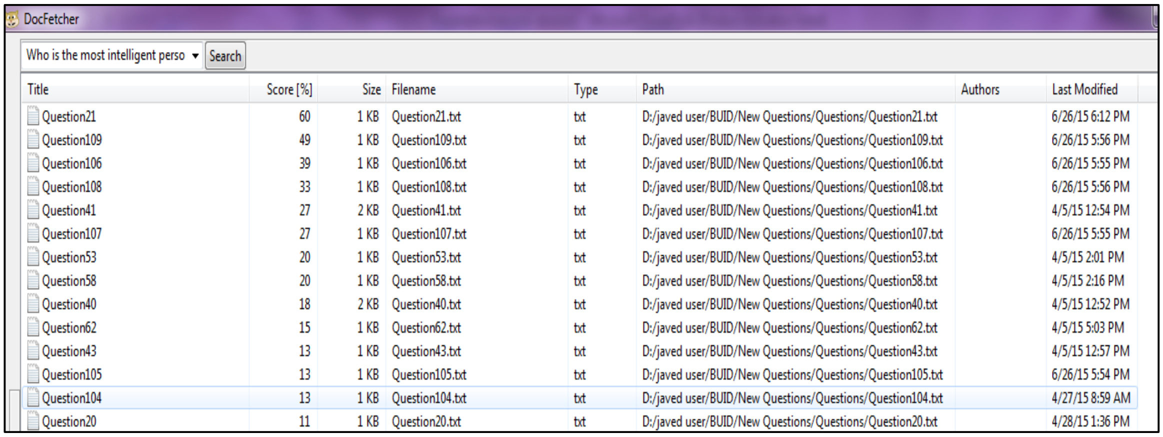
From the user query (Q) the important/keywords are extracted. We used dictionary (CIDE) to find the semantically equivalent words/terms for each keyword. A new search query (semantically equivalent) is built by using Boolean operators, "AND", "OR".

4.5.2 Experimental Results

DocFetcher tool is used to create index and retrieve the answers of the questions.. The answers for each question were retrieved and stored into separate files. DocFetcher is used to index these files and later was used to retrieve documents for questions (in both cases, before and after query expansion)

Consider the following question that is a part of our test dataset. It is observed that how Query expansion technique affects the process of question matching.

“Who is the most intelligent person in the world's history?”



Title	Score [%]	Size	Filename	Type	Path	Authors	Last Modified
Question21	60	1 KB	Question21.txt	txt	D:/javed user/BUID/New Questions/Questions/Question21.txt		6/26/15 6:12 PM
Question109	49	1 KB	Question109.txt	txt	D:/javed user/BUID/New Questions/Questions/Question109.txt		6/26/15 5:56 PM
Question106	39	1 KB	Question106.txt	txt	D:/javed user/BUID/New Questions/Questions/Question106.txt		6/26/15 5:55 PM
Question108	33	1 KB	Question108.txt	txt	D:/javed user/BUID/New Questions/Questions/Question108.txt		6/26/15 5:56 PM
Question41	27	2 KB	Question41.txt	txt	D:/javed user/BUID/New Questions/Questions/Question41.txt		4/5/15 12:54 PM
Question107	27	1 KB	Question107.txt	txt	D:/javed user/BUID/New Questions/Questions/Question107.txt		6/26/15 5:55 PM
Question53	20	1 KB	Question53.txt	txt	D:/javed user/BUID/New Questions/Questions/Question53.txt		4/5/15 2:01 PM
Question58	20	1 KB	Question58.txt	txt	D:/javed user/BUID/New Questions/Questions/Question58.txt		4/5/15 2:16 PM
Question40	18	2 KB	Question40.txt	txt	D:/javed user/BUID/New Questions/Questions/Question40.txt		4/5/15 12:52 PM
Question62	15	1 KB	Question62.txt	txt	D:/javed user/BUID/New Questions/Questions/Question62.txt		4/5/15 5:03 PM
Question43	13	1 KB	Question43.txt	txt	D:/javed user/BUID/New Questions/Questions/Question43.txt		4/5/15 12:57 PM
Question105	13	1 KB	Question105.txt	txt	D:/javed user/BUID/New Questions/Questions/Question105.txt		6/26/15 5:54 PM
Question104	13	1 KB	Question104.txt	txt	D:/javed user/BUID/New Questions/Questions/Question104.txt		4/27/15 8:59 AM
Question20	11	1 KB	Question20.txt	txt	D:/javed user/BUID/New Questions/Questions/Question20.txt		4/28/15 1:36 PM

Figure 4.9 Retrieved question with low score.

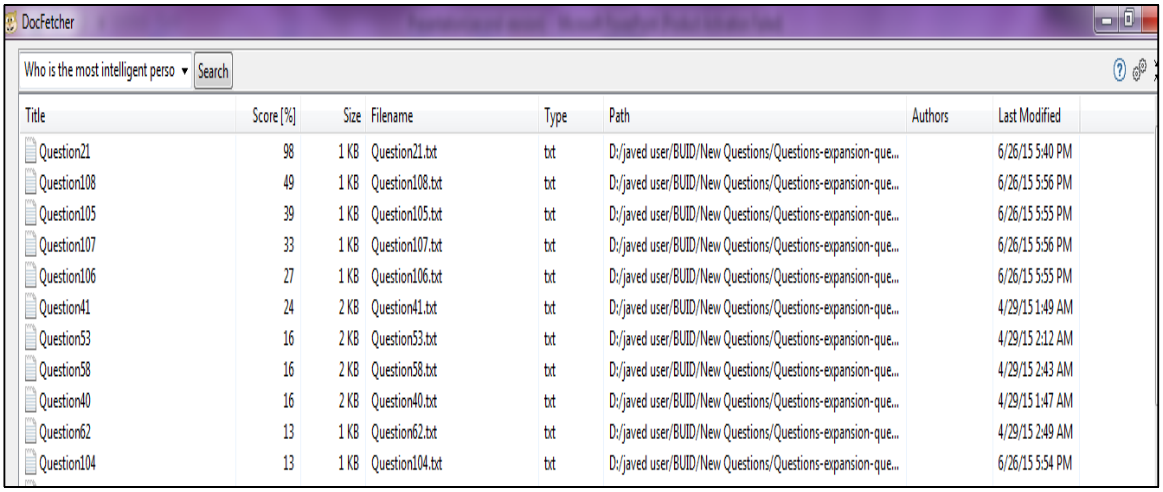
It searched and retrieved only few questions with low score as shown in figure 4.9.

After query expansion:

Who is the most intelligent person in the world's history?

who AND world AND person AND history AND (most OR greatest)AND

(intelligent OR Sensible OR understanding)



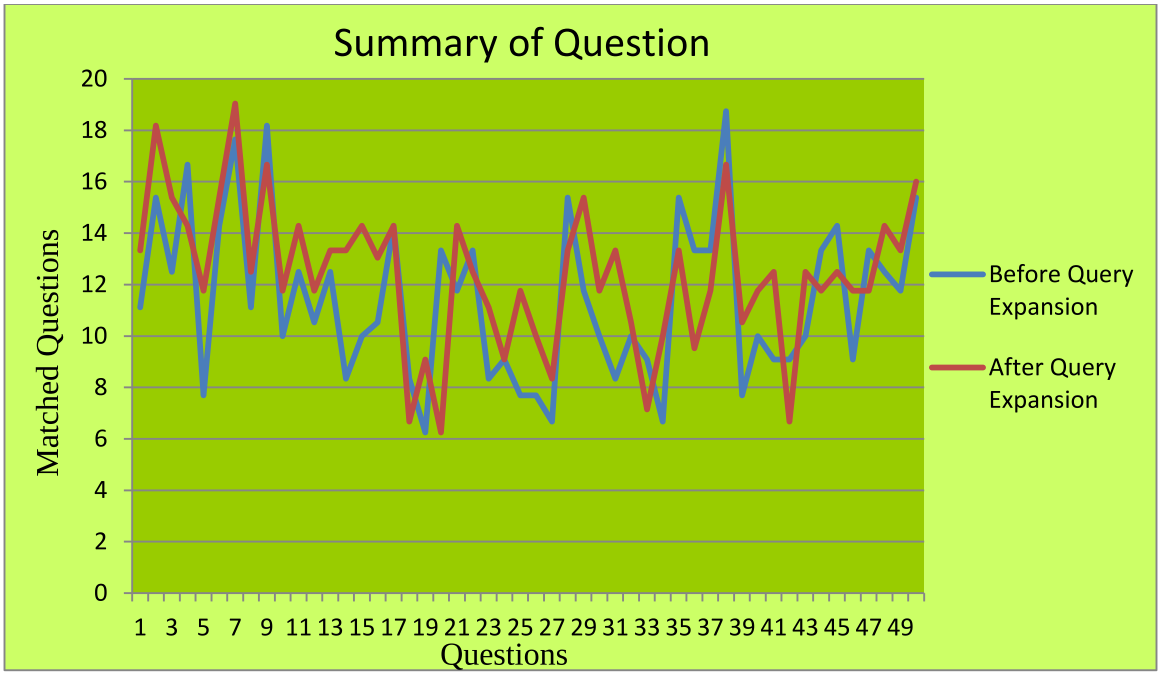
Title	Score [%]	Size	Filename	Type	Path	Authors	Last Modified
Question21	98	1 KB	Question21.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:40 PM
Question108	49	1 KB	Question108.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:56 PM
Question105	39	1 KB	Question105.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:55 PM
Question107	33	1 KB	Question107.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:56 PM
Question106	27	1 KB	Question106.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:55 PM
Question41	24	2 KB	Question41.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		4/29/15 1:49 AM
Question53	16	2 KB	Question53.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		4/29/15 2:12 AM
Question58	16	2 KB	Question58.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		4/29/15 2:43 AM
Question40	16	2 KB	Question40.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		4/29/15 1:47 AM
Question62	13	1 KB	Question62.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		4/29/15 2:49 AM
Question104	13	1 KB	Question104.txt	txt	D:/javed user/BUID/New Questions/Questions-expansion-que...		6/26/15 5:54 PM

Figure 4.10 Retrieved question with high score.

More questions are retrieved with high score as shown in figure 4.10.

In the first phase of testing the proposed query expansion method on our test dataset as mentioned earlier, the questions in their original form were entered in DocFetcher. The DocFetcher returns the document containing answers of the questions with certain score assigned for each answer. In the second phase, we reformulated the query as described in previous section. The modified query was entered into DocFetcher again to retrieve relevant answers. The DocFetcher returns the documents containing possible answers with score assigned to each answer. The experiment was conducted on test dataset. All these results were tabulated and analyzed as shown in figure 4.11.

Form our test data set we fed the question in existing CQA system before query expansion and results were collected and tabulated. System returned the number of matched question from where the correct number of questions was noted for each question. After that query expansion technique was applied on all questions of test dataset. Each question was entered in respective CQA system and results were collected. After analysis of experiments in both cases before and after query expansion the result was compared, as given below. In most of the questions query expansion helped to retrieve the questions successfully.



. **Figure 4.11: Summary of Matched Questions**

Results of query expansion as shown in figure 4.11 reveal that query expansion impact positively in the process of question matching. In this experiment, query expansion increase the score of answers for most of the questions except some of them, which are 4, 9, 18, 20,28,33,35,36,38,42 and 47 as shown in figure 4.11. For some questions query expansion did not affect their matching result and it was same as before query expansion. But overall it increased the matching score of number of questions as it is shown in the figure 4.11. Therefore, we can conclude that query expansion is very helpful for matching process of questions in Collaborative Question Answering System.

We also used Mean Reciprocal Ratio (MRR) metric to show the impacts of query expansion on retrieval process. . MRR for a given question Q can be described as

$$\text{MRR (Q)} = \sum_{i} 1/i$$

Where *i* is considered as rank of the correct matched question. For example, if the correct matched for a question is found in documents ranked 2, 4, 5 and 8, then MRR will be

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{5}+\frac{1}{8}=1.08$$

We evaluated the results of the question matching with and without query expansion using MRR as shown in figure 4.12.

It is observed that ranked MRR values before query expansion varied from 0.61 to 3.92 approximately, after applying query expansion it improves the result as it can be seen in graph, MRR values varied from 0.77 to 4.34 approximately. MRR Average before Query Expansion was 2.43, which went up to 2.69 after query expansion. Therefore, we can conclude that query expansion helps to improve the results.

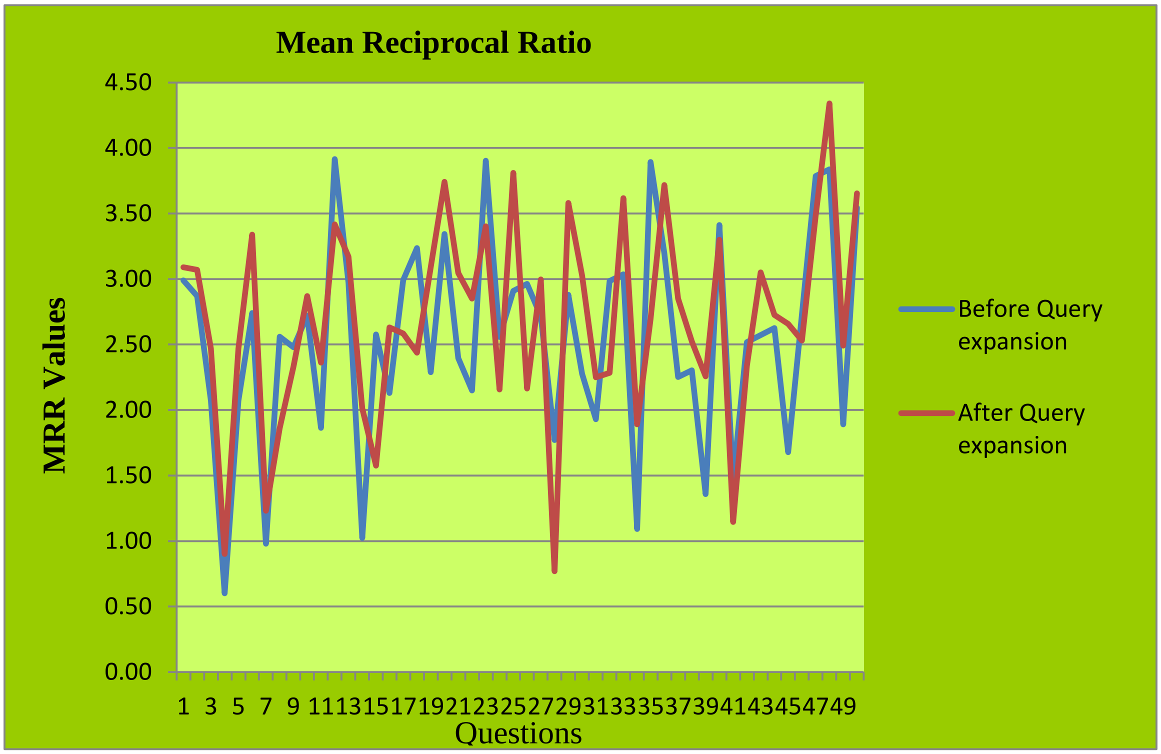


Figure 4.12: MRR of Matched Questions

Chapter Five

Conclusion & Future Work

5.1 Conclusion

Nowadays there are a number of different types of social networks on the internet where each of them has its own structure and characteristics. By analyzing the structure and characteristics of Yahoo! Answers, we tried to explain the generic structure of existing Collaborative Question Answering Systems. Yahoo! Answers is known as one of the famous social network having a huge size of database of question answer. The foremost intention behind the work in this research was to examine a structural approach to improve the matching process of user's question in conventional Collaborative Question Answering Systems. Specifically, it was observed that when a user entered a question in online collaborative question answering system where the same question existed but the system could not find that question. In some cases the large scale Collaborative Question Answering System produced a huge number of related documents instead of that specific question. For this purpose, classification of question plays a very important role. Correct classification improves the question retrieval process in Information Retrieval process. A number of different retrieval models for question retrieval were studied. A semantic approach of question classification using WordNet and Wikipedia was used in this research.. The proposed approach gave encouraging results for for question classification in open domain Collaborative Question Answering System.

As deliberate and precise query is challenging for the common users, so it is essential to use query expansion technique to extract appropriate information. Query expansion approaches are extensively used for improving the competence of written information retrieval system. In this research, we applied the query expansion technique using GNU Collaborative International Dictionary of English and DocFetcher. The results of the query expansion method show positive results.

5.2 Future Work

In this research, we kept our focus on the questions analysis, to improve the performance of Collaborative Question Answering (CQA) System by using semantic resources. We proposed some rules for question classification and query expansion processes to enhance the question retrieval procedure in question answering system. Such work of question classification can be extended to generate new patterns by using machine learning. In this research, we used a set of fixed classes that is not enough for open domain question answering. Therefore, some new approaches are required that could introduce new classes when needed. In query expansion method, we faced some limitations in GNU Collaborative International Dictionary of English (CIDE) while collecting similar semantic equivalent keywords. Improved version of these resources would be used. In future, other two phases of question answering system, information retrieval and answer analysis can be carried out in this research.

References

Adamic, L., Buyukkokten, O., & Adar, E. (2003). A social network caught in the web. *First monday*, *8*(6).

Al-Chalabi, Hani, Ray Santosh K., Shaalan Khaled. (2015). Question Classification for Arabic Question Answering Systems. Accepted for publication in Proceedings of International Conference on Information and Communication Technology Research , IEEE explore, Dubai.

Andrenucci, A., & Sneiders, E. (2005, July). Automated Question Answering: Review of the Main Approaches. In *ICITA (1)* (pp. 514-519).

Arai, K., & Handayani, A. N. (2012). Question Answering System for an Effective Collaborative Learning. *IJACSA Journal*, *3*(1).

Arai, K., & Handayani, A. N. (2013). Question Answering for Collaborative Learning with Answer Quality Predictor. *International Journal of Modern Education and Computer Science (IJMECS)*, *5*(5), 12.

Arai, K., & Handayani, A. N. (2013). Collaborative Question Answering System Using Domain Knowledge and Answer Quality Predictor. *International Journal of Modern Education and Computer Science (IJMECS)*, *5*(11), 21.

Bronner, A., & Monz, C. (2012, April). User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 356-366). Association for Computational Linguistics.

Cai, L., Zhou, G., Liu, K., & Zhao, J. (2011, October). Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge. InProceedings of the 20th ACM international conference on Information and knowledge management (pp. 1321-1330). ACM.

Cao, X., Cong, G., Cui, B., & Jensen, C. S. (2010, April). A generalized framework of exploring category information for question retrieval in community question answer archives. In Proceedings of the 19th international conference on World wide web (pp. 201-210). ACM.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, *44*(1), 1.

Cowie, J., & Lehnert, W. (1996). Information extraction. Communications of the ACM, 39(1), 80-91.

Chang, S., & Pal, A. (2013, August). Routing questions for collaborative answering in community question answering. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*(pp. 494-501). IEEE.

Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010, March). A first step towards understanding popularity in YouTube. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010* (pp. 1-6). IEEE.

Cheung, C. M., Chiu, P. Y., & Lee, M. K. (2011). Online social networks: Why do students use facebook?. *Computers in Human Behavior*, *27*(4), 1337-1343.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493-2537.

Corston, S. H., Dolan, W. B., Vanderwende, L. H., & Braden-Harder, L. (2005).*U.S. Patent No. 6,901,399*. Washington, DC: U.S. Patent and Trademark Office.

Du, Q., Wang, Q., Cheng, J., Cai, Y., Wang, T., & Min, H. (2013, September). Explore Social Question and Answer System Based on Relationships in Social Network. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2013 Fourth International Conference on* (pp. 490-495). IEEE.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003, May). Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 168-171). Association for Computational Linguistics.

Guo, J., Xu, S., Bao, S., & Yu, Y. (2008, October). Tapping on the potential of q&a community by recommending answer providers. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 921-930). ACM.

Gyongyi, Z., Koutrika, G., Pedersen, J., & Garcia-Molina, H. (2007). Questioning yahoo! answers.

He, B., & Ounis, I. (2009). Studying query expansion effectiveness. In *Advances in Information Retrieval* (pp. 611-619). Springer Berlin Heidelberg.

Hu, J., Fang, L., Cao, Y., Zeng, H. J., Li, H., Yang, Q., & Chen, Z. (2008, July). Enhancing text clustering by leveraging Wikipedia semantics. InProceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 179-186). ACM.

Jurczyk, P., & Agichtein, E. (2007, July). Hits on question answer portals: exploration of link analysis for author ranking. In *Proceedings of the 30th annual international*

ACM SIGIR conference on Research and development in information retrieval (pp. 845-846). ACM.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, *3*(1—2), 1-224.

Knoke, D., & Yang, S. (Eds.). (2008). *Social network analysis* (Vol. 154). Sage.

Lapshin, V. A. (2012). Question-answering systems: Development and prospects. Automatic Documentation and Mathematical Linguistics, 46(3), 138-145.

Lehnert, W. G. (1977, August). A conceptual theory of question answering. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1* (pp. 158-164). Morgan Kaufmann Publishers Inc..

Li, B., Jin, T., Lyu, M. R., King, I., & Mak, B. (2012, April). Analyzing and predicting question quality in community question answering services. In*Proceedings of the 21st international conference companion on World Wide Web* (pp. 775-782). ACM.

Liu, J., Wang, Q., Lin, C. Y., & Hon, H. W. (2013). Question Difficulty Estimation in Community Question Answering Services. In *EMNLP* (pp. 85-90).

Liu. Q., & Agichtein E. (2011). Modeling Answerer Behavior in Collaborative Question Answering Systems. Emory University, Atlanta, USA.

Li, X., & Roth, D. (2002, August). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.

Liu, Y., Bian, J., & Agichtein, E. (2008, July). Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 483-490). ACM.

Milne, D., Medelyan, O., & Witten, I. H. (2006, December). Mining domain-specific thesauri from wikipedia: A case study. In Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence (pp. 442-448). IEEE Computer Society.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, *3*(4), 235-244.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*(pp. 29-42). ACM.

Moldovan, D., Paşca, M., Harabagiu, S., & Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, *21*(2), 133-154.

Prager, J. M. (2006). Open-Domain Question-Answering. *Foundations and Trends in Information Retrieval*, *1*(2), 91-231.

Prager, J., Chu-Carroll, J., & Czuba, K. (2001). Use of WordNet Hypernyms for Answering What-Is Questions---DRAFT.

Qu, B., Cong, G., Li, C., Sun, A., & Chen, H. (2012). An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, *63*(5), 889-903.

Ray, S. K., Singh, S., & Joshi, B. P. (2008, December). Question Answering Systems Performance Evaluation-To Construct an Effective Conceptual Query Based on Ontologies and WordNet. In *SWAP*.

Ray, S. K., Singh, S., & Joshi, B. P. (2010). A semantic approach for question classification using WordNet and Wikipedia. Pattern Recognition Letters,31(13), 1935-1943.

Rivas, A. R., Iglesias, E. L., & Borrajo, L. (2014). Study of query expansion techniques and their application in the biomedical information retrieval. *The Scientific World Journal*, *2014*.

Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K., & Lin, C. Y. (2011, February). Using graded-relevance metrics for evaluating community QA answer selection. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 187-196). ACM.

San Pedro, J., & Karatzoglou, A. (2014, October). Question recommendation for collaborative question answering systems with rankslda. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 193-200). ACM.

Shaalan, K. (2014). A survey of Arabic named entity recognition and classification. *Computational Linguistics*, *40*(2), 469-510.

Shaalan, K., & Oudah, M. (2014). A hybrid approach to Arabic named entity recognition. *Journal of Information Science*, *40*(1), 67-87.

Suryanto, M. A., Lim, E. P., Sun, A., & Chiang, R. H. (2009, February). Quality-aware collaborative question answering: methods and evaluation. In*Proceedings of the second ACM international conference on web search and data mining* (pp. 142-151). ACM.

Suzuki, J., Taira, H., Sasaki, Y., & Maeda, E. (2003, July). Question classification using HDAG kernel. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12* (pp. 61-68). Association for Computational Linguistics.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM journal on computing*, *1*(2), 146-160.

Toba, H., Ming, Z. Y., Adriani, M., & Chua, T. S. (2014). Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences*, *261*, 101-115.

Voorhees, E. M. (2001, October). Question answering in TREC. In Proceedings of the tenth international conference on Information and knowledge management (pp. 535-537). ACM.

Wang, P., & Domeniconi, C. (2008, August). Building semantic kernels for text classification using wikipedia. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 713-721). ACM.

Wenyin, L., Hao, T., Chen, W., & Feng, M. (2009). A web-based platform for user-interactive question-answering. *World Wide Web*, *12*(2), 107-124.

Zhang, J., Ackerman, M. S., & Adamic, L. (2007, May). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). ACM.

Zhang, D., & Lee, W. S. (2003, July). Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 26-32). ACM.

Zhou, G., Liu, Y., Liu, F., Zeng, D., & Zhao, J. (2013, August). Improving question retrieval in community question answering using world knowledge. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 2239-2245). AAAI Press.

Zhou, T. C., Lyu, M. R., & King, I. (2012, April). A classification-based approach to question routing in community question answering. In Proceedings of the 21st international conference companion on World Wide Web (pp. 783-790). ACM.