國立成功大學
資訊工程學系

利用結構化與非結構化特徵改善複雜問答

Finding Structured and Unstructured Features to Improve the Search Result of Complex Question

研究生：韋絲若
指導教授：盧文祥

Abstract

Finding Structured and Unstructured Features to Improve the Search Result of Complex Question

Dewi Wisnu Wardani --- Wen Hsiang Lu

Recently, search engine got challenge deal with such a natural language questions. Sometimes, these questions are complex questions. A complex question is a question that consists several clauses, several intentions or need long answer.

In this work we proposed that finding structured features and unstructured features of questions and using structured data and unstructured data could improve the search result of complex questions. According to those, we will use two approaches, IR approach and structured retrieval, QA template.

Our framework consists of three parts. Question analysis, Resource Discovery and Analysis The Relevant Answer. In Question Analysis we used a few assumptions, and tried to find structured and unstructured features of the questions. Structured feature refers to Structured data and unstructured feature refers to unstructured data. In the resource discovery we integrated structured data (relational database) and unstructured data (webpage) to take the advantaged of two kinds of data to improve and reach the relevant answer. We will find the best top fragments from context of the webpage In the Relevant Answer part, we made a score matching between the result from structured data and unstructured data, then finally used QA template to reformulate the question.

In the experiment result, it shows that using structured feature and unstructured feature and using both structured and unstructured data, using approach IR and QA template could improve the search result of complex questions.

Table of Content

List of Figure

List of Table

Chapter 1. Introduction

1.1 Motivation

Analyze the focus question is not a new on question analysis's research area. A big part of the purposes of those researches are to achieve the information of question type or user intention clearly and definitely. Understanding the features of questions are the prominent works of those researches for guiding to reach the user information's need. This topic become more interesting to face the long and complex questions. Some of the researches, complex questions, on the other hand, often refer to long-answer questions. On complex question's research, a definition that it means an answer to a complex question is often a long passages, a set of sentences, a paragraph, or even an article [1]. Although many prior studies of keyword over text documents (e.g HTML documents) have been proposed, they all produce a list of individual pages as result [2].

Sometimes, it is difficult to achieve the answer of one complex question, since the answer can not only be retrieved from one webpage or one resource. In fact, it is very common that the answer of one complex question, is possibly separated in several webpages. Recently, the research of Question Answering got a challenge of complex question [3-6]. The detail of our observation will be described on chapter 2.

In this work, the complex question is a natural language question that contains structured and unstructured features and using the integration of structured and unstructured data on the web to answer those questions. It is more to improve the search result of the question. The resources are needed not only consider the unstructured data but also structured data.

One example:

> ***"What is the capital city of the country that the largest country in Arabian peninsula"***

The focus of this question is to know clearly capital name of the country that the country is largest in Arabian peninsula. From this question, we can find ***"the capital city"*** as the structured information of question and ***"that the largest country in Arabian peninsula"*** as an unstructured information of question. By these features we can start and retrieve the resource data to answer from both structured data and unstructured data.

For comparation, the result of search usually a relevant passage that contains the needed answer. The factual answer is Riyadh.


Saudi Arabia - Wikipedia, the free encyclopedia
Capital (and **largest city**) Riyadh 24°39′N 46°46′E &#20;/&#20; 24.65°N 46.767°E &#20;/&#20; 24.65; 46.767 Official languages Arabic Demonym Saudi, Saudi **Arabian** ... The Kingdom of Saudi Arabia, KSA (Arabic: المملكة العربية السعودية , al-Mamlaka al-'Arabiyya as-Su'ūdiyya), is an Arab **country** and the **largest country** of the **Arabian Peninsula**.
en.wikipedia.org/wiki/Saudi_Arabia · cached page

Figure 1.1 The example result (rank no.5) from Bing[1] Beta version

In another example, in topic "movie", we can find the database of movie on the web as structured data and also webpages that contain information of movie are huge amount exist on the web. Actually, many domains data that considered as structured data on the web. Thus, these are all of our motivations in this work and the major concentration is about how to find the structured and unstructured features of the question and use two kinds of data as the resource to improve the answer of the question.

Most of information is stored in semi-structured or unstructured documents. Making this information available in a usable form is the goal of text analysis and text mining system [7]. The importance of unstructured data as an information source is increasing steadily, both in the scientific and business communities. Recent Gartner research reports suggest that 80% of data is unstructured [8]. The other side of this condition that huge amount of unstructured documents go untapped whereas actually those documents contain useful hidden information.

Structured data on the web is prevalent but ignored often by existing information search [9]. Moreover, the quality of structured data on the web used to pretty high-quality content such as flight schedules, library catalogs, sensor readings, patent filings, genetic research data, product information, etc. Recently, the world wide web is witnessing an increasing in the amount of structured content-vast heterogeneous collections of structured data. Such as product information, google base, table on the webpage, or the deep web [10].

[1] www.bing.com

According to the above brief explanations, that an huge amount of informations on the web are structured and unstructured or semi-structured. According to the characteristics of two kinds of data, it will be pretty good to take the advantages of them. The searcher will not care about from which kind of the resource of the information can be found, they only want to get the better answers of their questions.

1.2 The Considered Problem

The existing search engines cannot integrate information from multiple irrelated pages to answer queries meaningfully [2]. On the other case, they usually only consider from one kind of resource, unstructured data such webpages or structured data such as freebase (Powerset[2] uses it).

1.2.1 Question Analysis

In this first step, we need to know the structured feature and unstructured feature that exist on the question. Simplification, for this initial work we only consider one kind of complex question that might contain structured and unstructured feature. As be known that natural language question has many forms of syntac and expression. Hence, we put some assumptions in this step according to our observation of the question from Yahoo! Answer (in English). Besides to find those features, we also want to find the question focus and question subfocus of the question. From the same example:

"What is the capital city of the country that is the largest country in arabian peninsula"

Question topic = ***"country"***

Question focus = ***"the capital city"***

Question subfocus = ***"that that is the largest country in arabian peninsula"***

Structured feature = ***"the capital city"***

Unstructured feature = ***"country that is located on a long boot shaped peninsula"***

We can see that the structured features is the question focus. This condition is one of situation that is issued in dealing with question analysis. Our question data are mostly

[2] www.powerset.com

about entity question, because we more want to see the answer tends to structured data. The reason is if the answer tend to structured data whereas the question is a complex question that usually tend to unstructured data in the common search engine, means our idea goes well. We will explain it more on the chapter 3.

1.2.2 Resource Discovery and Reach The Relevant Answer

Using two kinds of data, for the structured data, the form of this data is simple relational data. Single table with attribute name and attribute value. For Unstructured data we crawl webpages from several websites included using Wikipedia, even though not all will be used.

For this initial work we call it light integration according to the integration of the resource be used to improve the result question and do not to be the factual answer of question. One of the basic problem of integration is matching problem. In our work this matching mostly about the matching terms of both two recources. We will propose the simple model to reach the score matching between the unstructured data and structured data by considered the given question. Finally, by this score then can be reached the advanced information from both two kinds of resources, hence we can improve the result answer of the question as well.

1.3 Organization

The rest of this chapter, we will describe more detail of our work in several chapters. Chapter 2 is our related and previous work. It will be our first foundation idea on our work. Chapter 3 will talk about the detail problem definitions and our method to deal those problems. According to our observation, this work is quite new idea, so actually we put several constrains to make our work easier as the beginning of work. We propose the modified previous model in our work. Chapter 4 is our experiment description, we will try to do the experiment in several conditions to make sure and improve our confidently on our idea. The last section, chapter 5 we will briefly explain our summary and sure, the future work of this initial work.

Chapter 2. Related Works

2.1 Question Analysis on Question Answering

Since the question is the primary source of information to direct the search for the answer, a careful and high-quality analysis of the question is of utmost importance in the area of domain-restricted QA. [11] explain 3 mains question-answering approaches based on Natural Language Processing, Information Retrieval, and question templates. [12] proposed another approaches according to the resource on the web. Lin proposed federated approach and distributed approach. Federated approach is techniques for handling semistructured data are applied to access Web sources as if they were databases, allowing large classes of common questions to be answered uniformly. In distributed approach, large-scale text-processing techniques are used to extract answers directly from unstructured Web documents.

2.1.1 NLP Approach

NLP techniques are used in applications that make queries to databases, extract information from text, retrieve relevant documents from a collection, translate from one language to another, generate text responses, or recognize spoken words converting them into text. [13] explain QA based on NLP is the systems that allow a user to ask a question in everyday language and receive an answer quickly and succinctly, with sufficient context to validate the answer. [14] distinguish questions by answer type: factual answers, opinion answer or summary answer. Some kinds of questions are harder than others. For example, “why” and “how” questions tend to be more difficult, because they

require understanding causality or instrumental relations, and these are typically expressed as clauses or separate sentences summary [13].

2.1.2 IR Approach

IR systems are traditionally seen as document retrieval systems, i.e. systems that return documents that are relevant to the user's information need, but that do not supply direct answers. [15] The Text REtrieval Conferences (TREC) aim at comparing IR systems implemented by academic and commercial research groups also the previous version of them. The best performing system within the two latest TREC, Power Answer [16] had reached 83% accuracy in TREC 02 and 70% in TREC 03. A further step towards the QA paradigm is the development of document retrieval systems into passage retrieval systems [17-22].

2.1.3 Template-based QA

Template-based QA extends the pattern matching approach of NL interfaces to databases. It does not process text. Like IR enhanced with shallow NLP, it presents relevant information without any guarantee that the answer is correct. This approach mostly useful for structured data, as mentioned on [12]. [23] proposes a generic model of template-based QA that shows the relations between a knowledge domain, its conceptual model, structured databases, question templates, user questions, and describes about 24 constituents of template-based QA. [24] used a kind template and used ontology on question analysis, and work on structured information on the text.

2.2 Complex Question

Recently, some Question Answering researches have interested challenging of Question Answering research such complex questions. It is a complex sentence that have intention as question sentence. We can observe first about complex question. The definition from Wikipedia of complex sentence on English grammar definition, a complex sentence is sentence with an independent clause and at least one dependent clause (subordinating clause). The dependent clause is introduced by either a subordinate

conjunction such as although or because, or a relative pronoun such as who or which. For the examples as follows:

When he handed in his homework, he forgot to give the teacher the last page.

The teacher returned the homework *after she noticed the error.*

The clause with underline is main clause, and the italic one is subordinate clause. This rule also can be found on the question sentence.

In the Question Answering researches some have proposed another definitions approaches of Complex Questions. In the question answering researches a complex question is more about the complex answer. NTCIR in their workshop paper [25] described that their complex question related to definitions, biographies, relationships, and events. The task is "complex," and assume the answers are summarization-oriented, which means they contain various types of information that requires careful filtering. [26] explores the role of information retrieval in answering "relationship" questions, a new class complex informatiosn needs formally introduced in TREC 2005. [5] described that a complex questions are a kind of context questions, need a list of answer and a set of complex sentences. As follows their examples of complex questions:

"What is semolina?", ***"What is Wimbledon?"*** tend to a definition questions

"How do you measure earthquakes?" tend to the answer is a summary of passage

"Which museum in Florence was damaged by a major bomb explosion in 1993? On what day did this happen? Which galleries were involved?", it is similar like a definition of complex sentence from English grammar.

"***Name of 15 religious cult?***" tent to a list of answer

From a few above observations of a complex questions, except the definition of complex sentence (included question sentence) in English grammar, Question Answering researcher put their own definitions of complex questions. Mostly of their definitions refers to that the questions need a complex, long or complex process to answer.

To improve the result of answer as the common Question Answering researches, as already explained on the above paragraph, moreover for complex question also used to use an external source knowledge to improve their system result. [27] issued question

decomposition can be approached in one of two ways: either by approximating the domain-specific knowledge for a particular set of domains, or by identifying the decomposition strategies employed by human users.
In our work, we will try to face another kind of complex questions, it is more about factoid question. We will more explain it on Chapter 3.

2.3 Structured Information to Improve Question Answering

Some researcher believe that finding the structured information or semi-structured information can improve the result of Question Answering particularly for factoid question. Mostly the IR and Question Answering work on unstructured data. Some researches proposed techniques to find structured information over the unstructured data. [28] proposed segmentation, classification, association and normalization to find structured information over text. The others work, [29] proposed the the novel structured query to reach structured information over unstructured data, they continue working on [30], more work on the table (structured information on webpage), according to high information inside the table on the webpages. [31] also still work on web table data, it is more how to find hidden structured information that very usefull to answer the question particularly on Factoid Question. [32] proposed using both resource structured (mostly a table) and unstructured information (mostly text on webpage) to improve the result of factoid information. They used TREC 2005 as main question ask which use complementary models of answering questions over both structured and unstructured content on the Web. Their system attempts to answer factoid questions by guessing relevant rows and fields in matching web tables and integrating the results. They also proposed rule on the question analysis then implemented to their system. Another similar work, [33] they proposed Question Answering system to deal the semi structured data. Their aim is to answer factual questions by exploiting the structure inherent in documents found on the World Wide Web (WWW). In this work they did a kind of segmentation, documents are indexed into smaller units and associated with metadata. This aim the segmentation is to find the best one segment's score of the document to reach the answer. One the other, [34] implemented approach for domain-restricted question answering from structured knowledge sources, based on robust semantic analysis in a hybrid NLP system

architecture. They use TREC question and huge structured knowledge. All the above previous research are the evidences that structured information pretty useful to improve the result of QA particularly for Factoid Question.

2.4 Structured Retrieval

Some researches showed that structured approach retrieval on Question Answering can improve accuracy of the result. [35] made reformulation of query, executes separate queries for each structure and merges the lists of results. They believe that that structured retrieval is capable of retrieving more relevant sentences at higher ranks, compared with bag-of-words. [36] propose poses a novel approach wherein the application specifies its information needs using only a SQL query on the structured data, and this query is automatically "translated" into a set of keywords that can be used to retrieve relevant unstructured data. Another similar work, unstructured data often contains a substantial amount of implicit structure, much of which can be captured using information extraction (IE) algorithms [29].

A few above observations showed that structured approaches retrieval pretty useful to improve the result answer both on Information Retrieval and Question Answering.

2.5 Integration Information

[10] and severals previous researches have done novelty prominent idea of the integration resources [9, 10, 37-41]. The main reason of their work is try to find the advanted on each of them. Richer their resources mean better answer. Particularly [10] said that asker do not care the resource, they only want find the better answer. Another work [7, 32] talk about using both structured and unstructured to improve the answer. [2] first work on the keyword search on integration data: structured, semi structured and unstructured data with graph approach. [42] proposed a kind of integration entities that exist on table-like format on the webpages. It is integration information on the unstructured data.

Using the structured data moreover unstructured data in Information Retrieval or Question Answering researches are not new research issue. Since the the size of hingh quality structured data on web is increasing and not yet be optimum explored but using

the combination of them seem a pretty new area on Question Answering. Some previous proposed a prominent work. [31]. [43] proposed the integration of web document and myriad structured information about real-word object embedded in static web and online web database. It said that hybrid approach, using both structured and unstructured feature gave the best result on object information retrieval.

Chapter 3. Idea and Method

3.1 Observation of Question and Assumptions

3.1.1 Complex Question

According to the definition of complex question some of them are a kind of complex question, as have described on the above pages. Complex Question from TREC, cluster ciQA, start from TREC- 2007, also contain the factoid questions. They used template approach for complex factoid question.

Complex Question from NTCIR, cluster ACLIA, start from NTCIR – 8 (it started on June 2009). Its more about Japanese and Chinese Question, cover the area of complex factoid question, and multilingual complex question. Their focus on:

List/event questions , example: ***List major events in Saddam Hussein's life***

Relations questions, example: ***What is the relationship between Saddam Hussein and Jacques Chirac***

Biography questions , example: ***Who is Kim Jong Il*?** and

Definition questions , example: ***What is ASEAN?***

Our questions are a kind of complex question, a natural language question that contains structured and unstructured features. The resource of questions in our work is a real question from Yahoo!Answer (in English language).

We input one topic keyword like book, movie, country to the site and then we collected some of a real question we would like to study in this work.

As follows are the examples:

what are all the names of the books in the golden compass trilogy?

what are some of the books i can find about the causes of global warming?

what is the name of the book about japanese high school kids fighting each other to survive?

what is the capital of the country that has a border with hungary and a coastline on the black sea?

which countries of southeast asia are affected by monsoons?

What movie has people drinking in a house then the girl goes in the bedroom when a cartoon comes out tv?

3.1.2 Question from Yahoo!Answer

Some of the questions from Yahoo!Answer are usually long sentence questions and have different pattern rule of the question. Our observation on long and complex question from Yahoo!Answer, as in Table 3.1 is a brief summary of the information from 100 questions on 3 topics, book, country and movie

Type	Description	Book	Country	Movie
1	Prefix question using question tag (5W1H)	92 %	90 %	91 %
2	Prefix question using modal (can, should, etc)	1 %	0 %	2 %
3	Prefix question using particle (is, are, do, does, etc)	2 %	4 %	3 %
4	Others, long sentence	5 %	6 %	4 %

Table 3.1. Statistics of different types of prefix of question in 3 topics

Note: 5W1H (What, Who, When, Where, Why, How)

The example of those questions:

What was the book where the clue was hidden in the back of a famous statue in Europe? (example no. 1)

Who is the lead actor in the movie Fireeproof? (example no.1)

Can anyone recommend a really good love poems book? (example no.2)

Is the book Masquerade the last book in the blue blood series? (example no.3)

Zimbabwe now has the world's highest inflation rate, so which country had the highest rate before? (example no.4)

From above observation, we give a constrain in this work that we will only consider on the question that using question words, ,"What", "Who", "When", "Where", "Why", "How"

3.2 Idea

3.2.1 A Bag of Words of Question Answering Result

Automatic Question Answering usually give a document or a passages that contain the answer as the result. Or the answer is a bag of words. For the example

"Who is president of USA"

Usually we found the result as follows:

President of the United States The **President** of the **United States** is the head of state and head of government of the United States and is the highest political **official** in the United States by influence and recognition.

United States Other common forms include the U.S., the **USA**, and America. ... On November 4, 2008, amid a global economic recession, Barack Obama was elected **president**.

Ronald Reagan **Ronald Wilson Reagan** (February 6, 1911 – June 5, 2004) was the 40th **President** of the **United States** (1981–1989) and the 33rd **Governor** of California (1967–1975). ... **USA** Today.

List of fictional United States Presidents N-T To suit the timeline of the novel, Rodman is **President** during what would be the Bush years (becoming **President** in 1999 and ... **Jim Roy**, a black **man**, is elected **President** of the **USA** in the year 2228.

USA Next In addition to being **president** of **USA**, Charlie Jarvis is a board member of Defenders of Property Rights, one of several conservative groups that comprise the AT&T-funded (and DCI Group-operated) "Voices for Choices" coalition front group. ... **USA** board member James Wootton is **president** of the U.S. Chamber of Commerce Institute for Legal Reform where he advocates for tort "reform."

Figure 3.1. The result answer from Powerset[3]

The **Presidents** of the **United States**
Short history of the **US** Presidency, along with biographical sketches and portraits of all the **presidents** to date. From the official White House site.
www.whitehouse.gov/about/**presidents**/ - Cached - Similar

Abraham Lincoln — Ronald Reagan
George Washington — George H.W. Bush
The administration — Dwight D
Theodore Roosevelt — John F

More results from whitehouse.gov »

Welcome to the White House
4 Jun 2009 ... WhiteHouse.gov is the official web site for the White House and **President** Barack Obama, the 44th **President** of the **United States**.
Contact Us - The administration - Tours & Events - The Blog
www.whitehouse.gov/ - Cached - Similar

President of the **United States** - Wikipedia, the free encyclopedia
For other uses, see **President** of the **United States** (disambiguation). For the list, see List of **Presidents** of the **United States**. ...
en.wikipedia.org/wiki/**President**_of_the_**United**_**States** - Cached - Similar

List of **Presidents** of the **United States** - Wikipedia, the free ...
"**Presidents** of the **United States**" and "**US Presidents**" redirect here. For other uses, see **President** of the **United States** (disambiguation). ...
en.wikipedia.org/wiki/List_of_**Presidents**_of_the_**United**_**States** - Cached - Similar
More results from en.wikipedia.org »

Figure 3.2. The example result from Google[4]

We can see that the result usually a bag of words. The asker's intention actually quite clear, they need the name of current president of USA. The result used to a bag of words that contain a relevant answer.

[3] www.powerset.com
[4] www.google.com

The others examples for answer results more complex question,

"What is the capital of the country that has a south-eastern border with burma?"

are illustrated in Figure 3.3

Figure 3.3. The example result for longer question from Powerset

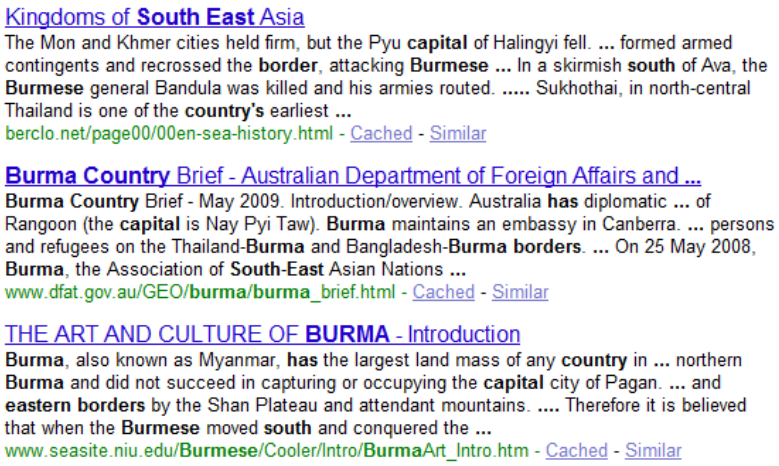

Figure 3.4. The example result for longer question from Powerset and Google

From the rank number 2 of Powerset's result answer snippet, one of close relevant answer:

"Burma The main player in the country's drug market is the United Wa State Army, ethnic fighters who control areas along the country's eastern border with Thailand, part of the infamous Golden Triangle. ..."

This snippet is not the clear answer of the question, even though contain the relevant result answer. The answer that needed is the name of capital city of Thailand. Thailand is the country on the eastern border of Burma, and the capital city is Bangkok, but no information about Bangkok on the result. To reach "Bangkok" as the answer need to reach another resource data or additional approach.

This is our motivation, how to improve result of a pretty complex questions from Yahoo!Answer. Sometimes, to answer one question, moreover a kind of complex question, we can not only find the answer from one webpage, it used to the answer actually exist on several webpages. Another approach, to answer one question, can not only retrieve answer from one kind of resource, but sometimes also need to consider from different kind of resource. Our approach is finding structured and unstructured features of question and using both structured and unstructured data as resources and reach the better answer.

3.2.2 Finding Answer Using Structured and Unstructured Data

Previous researches usually use only unstructured data to answer the question or use external resource knowledge such as WordNet. Using the external knowledge is very useful to improve the answer result, but it has ome disadvantages, that for using external resource of knowledge need developing time till the source ready to be used. It is not direct resource such webpages or another unstructured resources. Recently the growing of structured data on the web is very rapid and the number of online database also increases significantly. The research on deep web leads on this area. As described before, this kind of data usually contain high quality information, but used to be ignored by common search engines. Particularly, a few of search engine have a prominent work on structured data but still less of them using both structured data and unstructured data. We hypotized that by using structured data and structured retrieval will improve the result of complex question. We will use two approaches simultaneously to improve the answer result by using two kind of resources, unstructured and structured data.

Actually as be known that historically, retrieval information in these two kinds of data have involved separately.

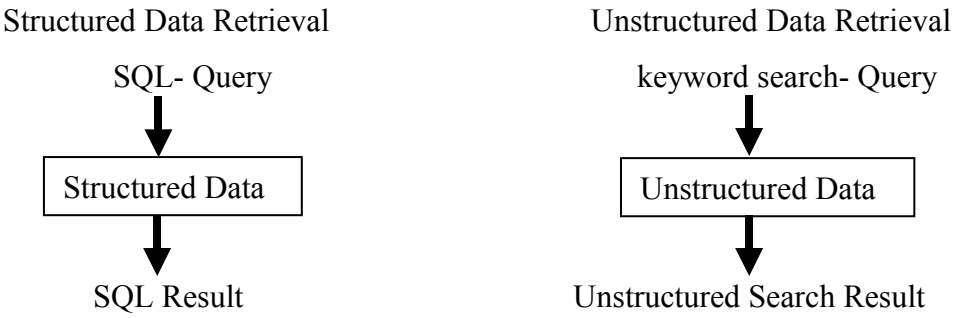

Figure 3.5. Information Retrieval of Structured Data and Unstructured Data

In our work we will try to combine two approaches to take the advantages of two kinds of data. The Question Answering based on template have been introduced on several previous works and proved were quite useful in structured information. In unstructured data, IR approach in Question Answering is well known. We will try to use it in the initial work on a few topic question as prominent research.

3.2.3 Framework

Our framework consists of three main works:

1. Question Analysis. In this part we introduced structured features and unstructured features. The others are Question focus and question subfocus
2. Resource Discovery. The resource on both structured data and unstructured data are explored for answering of complex questions
3. Answer Analysis. We introduce the simple ranking of our matching of answer candidates to choose the best answer

This is the framework of our work

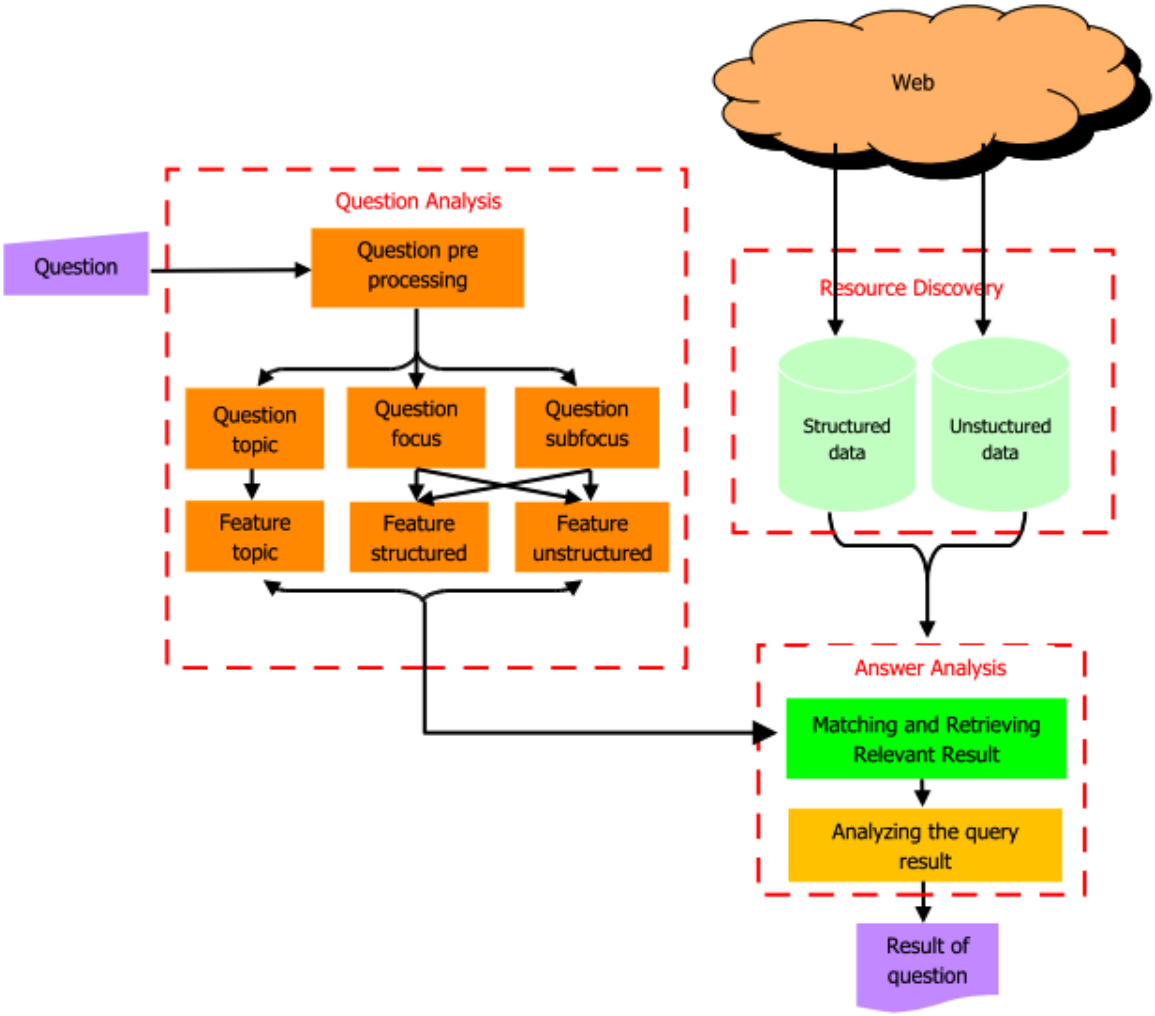

Figure 3.6. Framework of finding structured and unstructured features to improve result of complex questions

3.2.4 A Problem Definitions

We focus on two main works, the first is finding the structured and unstructured features on the question. The second, is retrieving the relevant information over structured data and unstructured data to achieve the exact answer. Some notations and definitions that would be used in this work are listed below.

For the Question Analysis, let ***Q*** is **Question**, ***Qt*** is **Question_topic**, ***Qf*** is **Question_focus** and ***Qs*** is **Question subfocus**. Then, ***Ft*** is **Feature_topic**, ***Fs*** is **Feature_structured** and ***Fu*** is **Feature_unstructured**. Next for Resource Discovery, consider two kinds of data. On the **Data_structured *(Ds)*** side, is used the relational

database. It has a set of **record** ***{R_i}***. Record ***i*** contain a set of **Attribute_value** ***{Av_{ij}}*** a set of **Attribute_name** ***{An_k}***. The **Focus of Attribute_name** ***(FAn)*** and the **Focus of Attribute_value of record** ***i (FAv$_i$)***. On the side of **Data_unstructured** ***(Du),*** is used the text documents. It has a set of **terms** ***{t_m},*** a set of **Attribute_unstructured** ***{Au_n}*** and a set of **snippet** ***{S_u}.***

3.3 Question Analysis

3.3.1 A Survey of Question

In the beginning of our idea, we only consider the question whose prefix has a question word ("What", "Who", "Where", "When", "Which", "Why", "How"). After observing 100 questions for each of topic domains, included Book, Country and Movie, the statistics of question type is shown in Table 3.2

No	Prefix Question Word Type	Book	Country	Movie
1.	What	69 %	36 %	93 %
2	Who	0 %	4 %	2 %
3	Where	1 %	8 %	1 %
4	When	1%	3 %	0 %
5	Why	0 %	29 %	2 %
6	How	27 %	0 %	2 %
	*Which	1 %	20 %	0 %

Table 3.2. The summary of question type of the questions from Yahoo!Answer

The question type that usually appear in 3 topics. We can see that the most appearing question word is "What". In country topic, "Which" and "Why" also have pretty high percentage, but "Which" usually refers to "What".

The real questions from Yahoo!Answer usually have many kind of patterns, and they are complex questions according to above definitions. The next, a survey surface of syntactic pattern of questions, randomizing on 3 topics and 100 questions. We used Stanford Tagger[5] to obtain POS Tags of the questions

[5] http://nlp.stanford.edu/

No	Pattern	Example questions	%
1.	_WP/_WR/_WD+ _VB+ [A*] + [“of the”/ ”of a”] + _N+ [B*]	- what_WP is_VBZ the_DT name_NN of_IN a_DT book_NN that_WDT came_VBD out_RP toward_IN end_NN of_IN 2008_CD ._.it_PRP was_VBD a_DT mystery_NN with_IN clues_NNS in_IN the_DT front_NN ? - what_WP is_VBZ the_DT name_NN of_IN a_DT book_NN in_IN which_WDT there_EX is_VBZ a_DT safari_JJ guide_NN in_IN nairobi_NN and_CC a_DT character_NN named_VBN delilah_NN ? - what_WP are_VBP the_DT titles_NNS of_IN the_DT books_NNS where_WRB the_DT carradigne_NN king_NN looks_VBZ in_IN wyoming_VBG for_IN an_DT heir_NN to_TO the_DT throne_NN ? - what_WP is_VBZ the_DT capital_NN of_IN the_DT country_NN that_IN borders_NNS the_DT ivory_NN coast_NN and_CC through_IN whichthe_JJ prime_JJ meridian_NN passes_NNS ? - what_WP are_VBP the_DT names_NNS of_IN the_DT countries_NNS that_WDT have_VBP free_JJ education_NN ? - what_WP is_VBZ the_DT name_NN of_IN the_DT movie_NN that_IN a_DT guy_NN lost_VBD his_PRP$ job_NN so_IN he_PRP start_VB selling_VBG his_PRP$ sperm_NN and_CC having_VBG sex_NN with_IN girls_NNS ?	
2	_WP/_WR/_WD+ _VB+ _N+[A*]	- What_WP was_VBD the_DT book_NN where_WRB the_DT clue_NN was_VBD	

		hidden_VBN in_IN the_DT back_NN of_IN a_DT famous_JJ statue_NN in_IN Europe_NNP ? - What_WP is_VBZ a_DT good_JJ girls_NNS book_VBP club_NN read_NN relating_VBG to_TO London_NNP or_CC England_NNP ? - What_WP was_VBD the_DT book_NN where_WRB the_DT clue_NN was_VBD hidden_VBN in_IN the_DT back_NN of_IN a_DT famous_JJ statue_NN in_IN Europe_NNP ? - What_WP is_VBZ the_DT richest_JJS country_NN in_IN the_DT world_NN ? - What_WP is_VBZ a_DT country_NN where_WRB you_PRP can_MD get_VB plastic_JJ surgery_NN cheaply_RB and_CC safely_RB ? - what_WP were_VBD all_DT of_IN the_DT countries_NNS the_DT nazi_NN 's_POS invaded_VBN ? - what_WP is_VBZ that_DT movie_NN where_WRB a_DT guy_NN goes_VBZ back_RB in_IN time_NN somehow_RB and_CC near_IN the_DT end_NN of_IN the_DT movie_NN ?	
3	_WP/_WR/_WD+ _N+ [A*]	- What_WP book_NN about_IN Wildlife_NNP will_MD inspire_VB me_PRP and_CC provide_VB me_PRP with_IN an_DT insight_NN into_IN the_DT subject_NN ? - What_WP book_NN that_WDT has_VBZ the_DT hero's_NNS journey_NN format_NN would_MD you_PRP recommend_VB ? - What_WP book_NN should_MD I_PRP read_VB on_IN the_DT 20th_JJ century_NN ? - Which_WDT Harry_NNP Potter_NNP book_NN	

		do_VBP we_PRP find_VB out_RP about_IN the_DT enchantment_NN on_IN the_DT girls_NNS dormitory_NN stairs_NNS ? - What_WP countries_NNS accept_VBP military_JJ officers_NNS from_IN another_DT country_NN ? - What_WP country_NN will_MD pay_VB for_IN your_PRP$ health_NN care_NN costs_NNS if_IN you_PRP get_VBP sick_JJ when_WRB traveling_VBG ? - What_WP European_JJ country_NN has_VBZ temperature_NN above_IN 20_NNP C_NNP during_IN daytime_JJ in_IN November_NNP ?	
4	_WP/_WR/_WD+ _MD+ _PR+ [A*]	- How_WRB can_MD you_PRP sell_VB your_PRP$ own_JJ book_NN online_NN ? - How_WRB can_MD I_PRP make_VB a_DT stapleless_JJ book_NN at_IN home_NN ? - How_WRB can_MD a_DT country_NN be_VB referred_VBN to_TO as_IN an_DT entity_NN in_IN itself_PRP ? - What_WP would_MD you_PRP do_VB for_IN your_PRP$ country_NN if_IN you_PRP become_VBP the_DT Prime_NNP Minister_NNP ?	
5	_WP/_WR/_WD+ VB+ [A*]	- How_WRB does_VBZ a_DT person_NN get_VB their_PRP$ book_NN published_VBN ? - How_WRB does_VBZ one_CD become_VB a_DT childrens_JJ book_NN illustrator_NN ? - Why_WRB are_VBP supporting_VBG characters_NNS in_IN a_DT book_NN important_JJ ? - How_WRB can_MD a_DT country_NN pull_NN out_IN of_IN a_DT War_NNP without_IN jeopardizing_VBG the_DT lives_NNS of_IN	

		the_DT remaining_VBG soldiers_NNS ?	

Table 3.3. The summary of patterns of complex questions from Yahoo!Answer

Table 3.3 show five types of surface and POS pattern, as well as some examples. We can find structured feature and unstructured features existing on the complex questions which consist of at least 2 clauses and mostly appear on the type 1,2 and 3 but if we look more detail then we find that the type 1 is the best example complex questions. For type 2 and 3 actually pretty similar, so we generate them as the same type. Finally, in this prominent wor we will only consider of 2 types.

We waived type 4 and 5 because almost all need the definitions or a summarization passages. This kind of question have been used by previous complex question's researches on NTCIR and TREC.

This is our second assumption that we will only consider on those patterns. We consider on the question that has main clause and has subordinate clause.

Additional, we also ignored the questions are not complex questions or have already done as the questions resources in the previous researches or it is not a complex questions. For the examples:

"What is the genre of the movie Twilight"

"Who is the author the book Harry Potter"

"What is capital city of country "

"Where is country Taiwan"

"What do you think the best James Bond movie was"

"What is the genre of the book "skinny"?"

"What is the genre of the book "the five people you meet in heaven"?"

"What is the capital of the country jordan?", etc

The complex question that has at least 2 usually contain dependent clause, the explaination as follows:

No	Subordinate	Example questions

	annotate term	
1.	[A*]+[“has”/ “had”]	- what v.c. andrews book is it, has cathy going back to foxworth hall? - what is the country has the longest border with honduras? - What country had the most people enslaved and sold in the New World during the 1600s?
2.	[A*]+[“have”]	- what are the names of the countries have rights to have nuclear weapons?
3.	[A*]+[“that”]	- what is the country that has a border with hungary and a coastline on the black sea? - what is the movie that 2 soldiers on an island fight representing there countries than war?
4.	[A*]+[“where”]	- what book where everyone is made surgically blind at birth?
5.	[A*]+[“when”]	- what is the name of the movie when jet li was a bodyguard ?
6.	[A*]+[“which”]	- what is the name of the book in which brothers battle over revealing documents about scripture? - what is the name of the movie in which a very heavy person goes into a restaurant and eat everything in sight?
7.	[A*]+[“whose”]	- what is the capital of the country whose neighbors are colombia brazil and guana?

Table 3.4. The example of existing subordinate indicate complex questions

3.3.2 Algorithms and Method of Question Analysis

We will introduce our approaches of Question Analysis (first part of our work). In general, our approach used the simple approach, consider the surface syntactic pattern of the questions.

We proposed the Algorithm Finding Structured-Unstructured Feature, consists first step of finding the Question topic (Qt), Question focus (Qf) and Question subfocus(Qs) and the second step finding the Feature topic(Fs), Feature structured (Fs) and Feature unstructured (Fu) from the question.

ALGORITHM OF FINDING STRUCTURED-UNSTRUCTURED FEATURES

Input : Question (Q)
Output: Question_topic (Qt), Question_focus (Qf), Question_subfocus (Qs)
Feature_topic (Ft), Feature_structured (Fs), Feature_unstructured (Fu)
Step :
Begin
Use POS Tagger to get POS tag for each question
if (rule of tag sentence question,
Type 1: WP_tag+[A*]+["of a"|"of the"]+NP_tag+[B*]) **then**
//NP_tag is the nearest NP after ["of a"|"of the"]
NP_tag is Question_topic (Qt)
[A*] is Question_focus (Qf)
[B*] is Question_subfocus (Qs)
end if
if (rule of tag sentence question,
Type 2: Wp_tag+[A*]+NP_tag+[B*]) **then**
//NP_tag is the nearest NP before [B]*
//[B] phrase that contain the annotated term of subordinate clause*
NP_tag is Question_topic (Qt)
[A*] is Question_focus (Qf)
[B*] is Question_subfocus (Qs)
end if
Question_topic is Feature_topic (Ft)
if (Match (Qf,Ds)) **then**
Feature_structured (Fs) is Question _focus(Qf) and
Feature_unstructured(Fu) is Question_subfocus (Qs)
else
Feature_structured (Fs) is Question _subfocus(Qs) and
Feature_unstructured(Fu) is Question_focus (Qf)
end
end

And this equation is the measure whether the Qf is Fs or Fu

$$
\begin{aligned}
Fs &= Match(Qf, Ds) \\
&= \arg\max_{Fs} \sum_{Fs \in Ds} P(Ds \mid Qf) \\
&= \arg\max_{Fs} \sum_{Fs \in \{An_i, Av_i\}} P(An_i, Av_i \mid Qf) \\
&= \arg\max_{Fs} \sum_{Fs \in \{An_i, Av_i\}} P(An_i \mid Qf) P(Av_i \mid Qf)
\end{aligned}
\tag{1}
$$

Note:

Fs : Feature_structured

Qf : Question_focus

Ds : Data_structured

An : Attribute_name

Av : Attribute_value

Step 1, finding the Qt is very important, because this term will be a key of matching data between structured and unstructured data. Question topic is the domain problem in general of the questions. Question focus (Qf) is the most information that be needed by the asker. Question subfocus (Qs), actually question subfocus is a part of question as additional information to answer the main focus of question. According our constrain, we propose the algorithm of Finding Question topic(Qt), Question focus(Qf) dan Question subfocus(Qs).

Step2, find the Feature topic(Fs), Feature structured (Fs) and Feature unstructured (Fu) from the question. Feature topic (Ft) is tent to the Question topic. Feature structured (Fs) is the feature tent to the structured information/ data on the question and Feature unstructured is feature that tent to the unstructured information/data.

We propose algorithm for finding Feature topic (Ft), Feature structured (Fs) and Feature unstructured (Fu)

That algorithm can be explained more clearly in the following example:

Step 1

No	Question	Qt	Qf	Qs
1	what is the capital of the country that is located on a long boot-shaped state or country near by africa?	country	capital	that is located on a long boot-shaped state or country near by Africa
2	what is that movie called? the one where people went down a cave and were attacked by bat like creatures?	Movie (as a topic, tent to entity)	Movie (as a focus, tent to attribute)	where people went down a cave and were attacked by bat like creatures
3	what is the author/title of the book where a guy goes back in time to give lee & the confederate army ak-47s?	book	Author/title	where a guy goes back in time to give lee & the confederate army ak-47s
4	What font is used for the cover tittle of the movie "John Woo Presents Blood Brothers"?	movie	font is used for the cover tittle	"John Woo Presents Blood Brothers"

Table 3.5. The example result of algorithm finding Question_topic, Question_focus and Question_subfocus

After this step 2 we will have result of Question Analysis as follow for the example

No	Question	Step 1	Step 2
1	what is the capital of the country that is located on a long boot-shaped state or country near by africa?	Qt : country Qf : capital Qs : that is located on a long boot-shaped state or country near by africa	Ft : country Fs : capital Fu : that is located on a long boot-shaped state or country near by africa
2	What summary of the movie "John Woo Presents Blood Brothers"?	Qt : movie Qf : summary Qs : "John Woo Presents	Ft : movie Fs : "John Woo Presents Blood Brothers"

		Blood Brothers"	Fu : summary

Table 3.6. The result example for finding topic, structured and unstructured feature

From Table 3.6, we can see that Qf could be Fs or Fu, and Qs could be Fs and Fu also. According to this, we put our thirth assumption that we will only work on Fs is Qf. The reason is we can see on the question no.2. This condition is common question that can be answerd only consider all as unstructured feature, because the answer is "need the summary". This answer used to already exist on unstructured data.

Equestion (2) is used to measure whether the Qf can become the Focus of Attributes. It will simply work after the thirth assumptions

$$FAn = Match(Qf, An_i) = \arg\max_{FAn} \sum_{FAn \in An} P(An_i \mid Qf) \quad (2)$$

The result of Question Analysis, will be very important to reach the better result answer, we will explain later on the Finding the Relevant Answer (in section 3.5).

3.4 Resource Discovery

Most of information on the Web is stored in semi-structured or unstructured documents. Making this information available in a usable form is the goal of text analysis and text mining system [44]. In this prominent work we use on the Data_structured *(Ds)* side, the relational database single table, and as usually the Data_unstructured *(Du)* side, the webpages [4].

The example:

"What is the capital of the country that is located on a long-boot shaped peninsula?"

Question_focus (Qf) is the same as Feature_structured (Fs), and ***"capital"*** is Focus_Attribute_name (FAn) which is one of Attribute_nane (An) on Data_structured (Ds)

Question_subfocus is identified as Feature_unstructured (Fu), ***"that is located on a long-boot shaped peninsula"***, is annotated as terms on Data_unstructured (Du)

From the annotated term on Du, some useful attributes names and their corresponding values can be extracted from term around the annotated terms, and find the best snippet/ fragment on the Du.

The illustration as follow in Figure 3.7

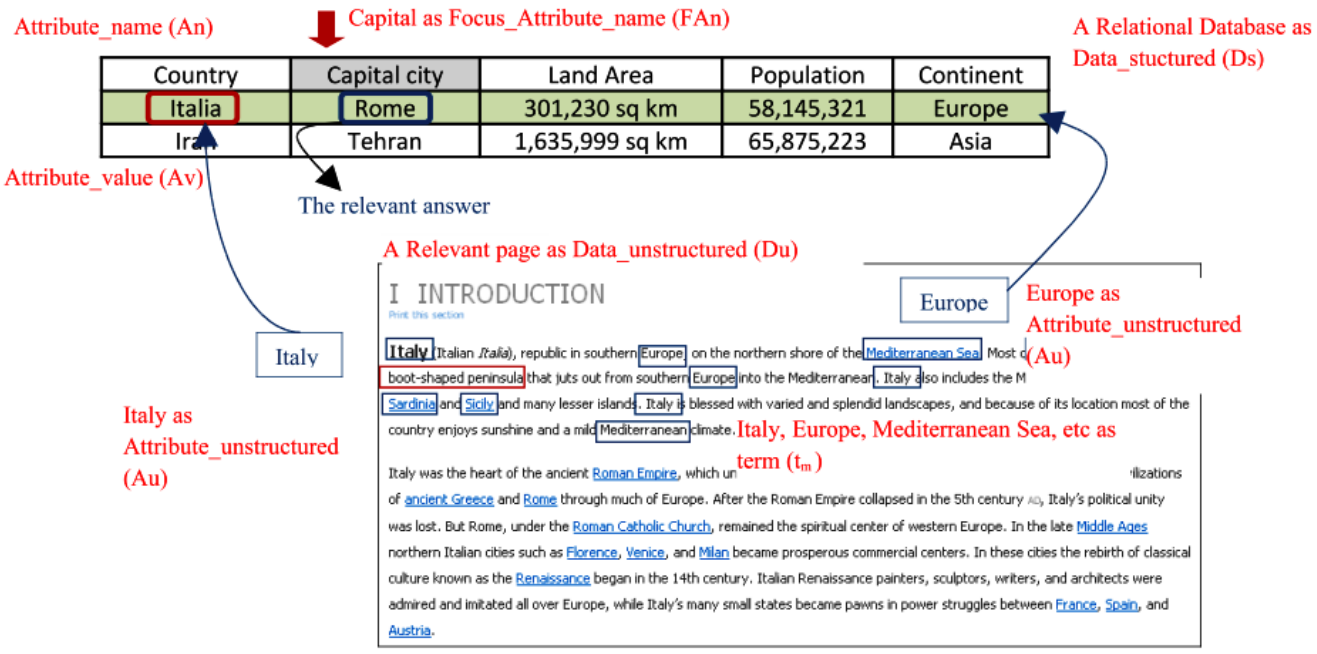

Figure 3.7. Example of Resource Discovery

We use the Feature topic (Ft) and Feature_unstructured (Fu) the general question *Q*, to find the relevant page *Du* by the cosine similarity measure which define in Equestion (3), and use the *Fu* to find the annotated snippet.

$$S(Du_j, q) = \frac{\sum_{i=1} wi(Du)_j \cdot wi(q)}{\sqrt{\sum_{i=1} wi(Du)_j^2 \cdot \sum_{i=1} wi(q)^2}}$$

(3)

S : Score cosine similiraty between Du and q

Du : Data_unstructured

q : Feature_topic and Feature_unstructured

where the weight (w) is based on weighting scheme

$$wi = tf \bullet idf = tf \bullet \log\left(\frac{N}{n}\right) \quad (4)$$

Be inspired from previous work [44], we want to find the relevant snippet of *Du*, where N is the number of total attributes value in *Ds*, and n(t) is the number of total attribute value (Av) that contain t on Du.

$$S_{snippet}(Av, S) = \sum_{t \in T(Av,S)} tf(t, S) \cdot w(t) \quad (5)$$

$$w(t) = \begin{cases} \log(N / n(t)) \; if \; n(t) > 0 \\ 0, \; otherwise \end{cases} \quad (6)$$

Here, consider the score of snippet/ fragment have found of a relevant documents. As follow the example of fragmentations of context of the webpage

From the question: *"What countries are not a member of coalliton of willing"*

(it is a complex questions, need a list of answer result to answer)

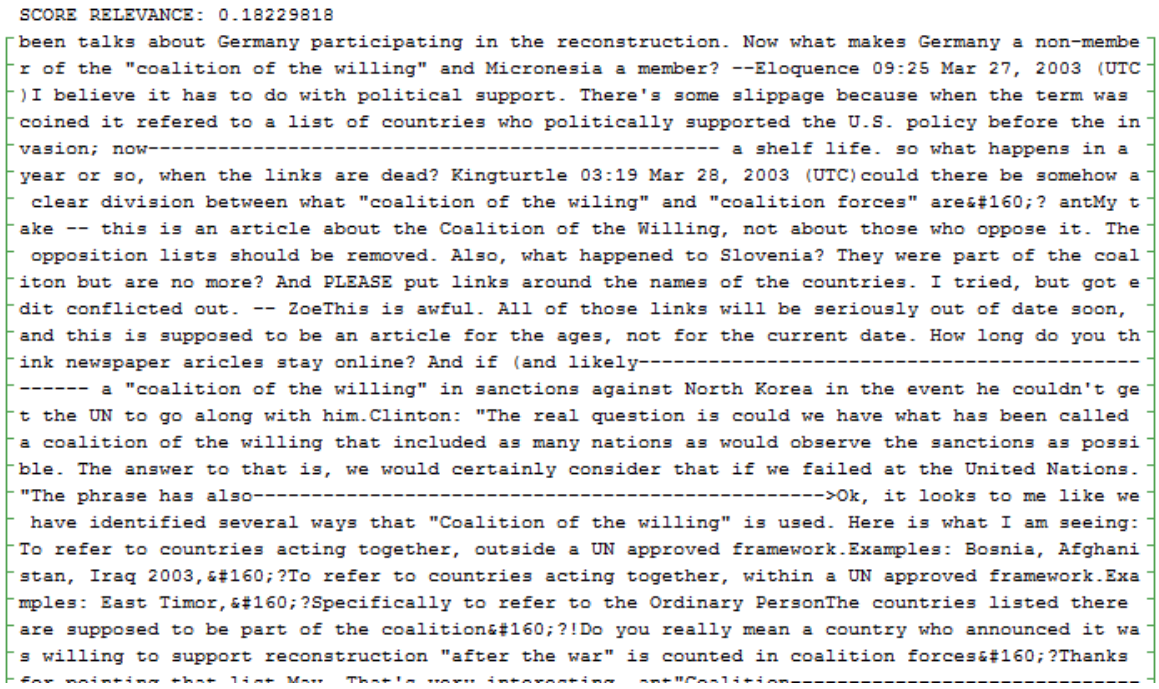

Figure 3.8. Example of fragmentations, the "---" is the boundary of the fragments

3.5 Finding The Relevant Answer

Analyze all terms on the relevant snippets, choose the terms (*t*) that contain *Av* as Attributes_unstructured *(Au),* Equestion. (7). Around n-gram term *"long boot shaped peninsula"* we would get another term such as *"Italy*", *"Sicilia", "Roman Empire", "Renaissance", "Sardinia", "Mediterranean" etc.*

$$Au = \sum_{t_i \in Au} P(t_i \mid Av) \qquad (7)$$

Consider all terms on the snippet that could be the candidates of Attribute unstructured (Au) and calculate the score matching of Unstructured data and Structured data, the get the score of record. Proposed score matching inspired from full string matching based Jaccard coefficient and q-gram matching.

$$Score1 = J(R, Au) = \frac{|R \cap Au|}{|R \cup Au|}$$

(8)

Q-grams are typically used in approximate string matching by "sliding" a window of length q over the characters of a string to create a number of 'q' length grams for matching a match is then rated as number of q-gram matches within the second string over possible q-grams.

Using q-grams in a DBMS for Approximate String Processing

$$\text{Score2} = S(R, Au) = S_q(\tau_{k,l}) = \max_{i,j}\left(S_q(\tau_{i+q-1,j+q-1}) + S_q(\tau^*_{i,j})\right) \quad (9)$$

where

$$\tau^q_{i,j} = (x_{i+1}\ldots x_{i+n}, y_{j+1}\ldots y_{j+n}) \quad (10)$$

We used Q-gram, to find the similarity between Du and Ds and consider the position oof letter so we will find similarity even not really exact. Those all about the matching score. The matching score is very important to match the unstructured data and structured data. It is all use IR approach Then, the score simply as linear combination as follow

$$\text{Score\_Match} = \alpha.\text{score1} + (1-\alpha)\ \text{score2} \quad (11)$$

Where α.is weighting parameter (0.1 – 0.9)

After we get the matching score and got the relevant record, then we will find the final answer by implementing QA template over all data we have.from Question Analysis and Resource Discovery. We have described in the beginning of this thesis that Template-based QA approach extends the pattern matching approach of NL interfaces to databases. It does not process text. Like IR enhanced with shallow NLP, it presents

relevant information without any guarantee that the answer is correct. This approach mostly useful for structured data

In the Question Analysis we got already Question_topic (Qt), Question_focus (Qf), Question_subfocus (Qs), Feature_topic (Ft), Feature_structured (Fs), Feature_unstructured (Fu) and finally we have Focus_Attribute_name (FAn).

In Resource Discovery we got the relevant document, the relevant snippet/fragment Attribute_unstructured (Au), and finally we have Score_Match. Mostly in this step the approach is IR approach.

To reach the final answer we use QA template approach that have modified by IR approach as structured retrieval. QA template approach be used to build the reformulation of question and make structured retrieval

For the example of the question :

"What is <u>the capital city</u> of <u>the country</u> that the largest country in Arabian peninsula"

This QA template is simple like this one:

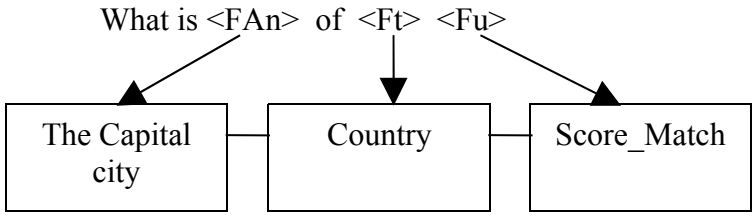

Figure 3.9 Question template in this work

FAn is Focus_Attribute_name, could be change by any Attribute_name on the Data_structured. FAn always a Feature_structured

Ft is Feature_topic that could be changed by any domain database from Data_structured

Fu is Feature_unstructured that could be change by Score_Match, the result of matching step from Data_structured and Data_unstructured.

Reformulation of the questing become like:

SELECT <capital_city> FROM <country> WHERE <that the largest country in Arabian Peninsula.

<capital_city> is FAn, Focus_Attribute_name on Data_structured

<country> is Ft, Feature topic

<that the largest country in Arabian peninsula> is Fu, Feature unstructured the Fu will be changed by Score_Match

According to structured retrievel that be used by QA template, we proposed the general QA template in this work. The reformulation of question as follow:

SELECT <FAn> FROM <Ft> WHERE < (Score_Match > T) >

Note: T as a threshold of score.

Or

SELECT <FAn> FROM <Ft> WHERE < (max(Score_Match)) >

Chapter 4. Experiment

In this chapter, we conducted several experiments to show how our simple approach could improve the result of complex question by finding the structured and unstructured features and using light combination structured data and unstructured data. The experiment is devided into two sections, in the Question Analysis and the result answer.

We will show a result of our Question Analysis and some experiments on finding the relevant answer with some different variable.

4.1 Experimental Setup

4.1.1 Dataset

In our work, for Question Analysis we used real questions from Yahoo!Answer. We collected the questions in 3 topics including book, country and movie questions. We collected those complex questions in April 2009, the question only in English.

	Training	Testing
Book	65	33
Country	65	35
Movie	65	34

Table 4.1 Dataset of question

As in the very beginning of our explaination, we used two kind data. As follows our data in 3 topics. Structured Data is single table relational database and unstructured data is a webpage from websites.

The attribute on the table of structured data:

Book → [id, isbn, title_name, author, year_publication, publisher, url_image]

Country→ [id, country_name, capital_city, government_form_country, area, population, religion, language, currency, trading_partner, primary_product, major_industries, export, mass_communication]

Movie → [id, name_title, year_release, director, genre]

Both the structured data and unstructured data were crawled on March 2009

No	Topic	Structured Data	Unstructured Data
1	Book	10,378 rows From Amazon[6]	~ 800 KB From Infoplease[7] ~ 238 GB From Wikipedia[8]
2	Country	196 rows From About[9]	
3	Movie	10,978 rows From IMDB[10]	

Table 4.2 Description of Dataset

4.1.2 Experiment Metrics

In Question Analysis we use evaluation metrics Recall (R), Precision (P) and F-Measure (F-Measure).

$$R = \frac{Tp}{Tp \cap Fn} \tag{10}$$

$$P = \frac{Tp}{Tp \cap Fp} \tag{11}$$

$$F - Measure = 2\frac{P \cdot R}{P + R} \tag{12}$$

where Tp is true positive result, Fn is False negative result and Fp is False positive result.

[6] www.amazon.com
[7] www.infoplease.com
[8] www.en.wikipedia.org
[9] www.about.com
[10] www.imdb.com

In the Resource Discovery and reach the relevant answer, besides use the Precision, Recall and F-Measure, we will use MRR in different fragment size, different threshold of match_score and different α.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{13}$$

4.2 Experimental Result

4.2.1 Question Analysis

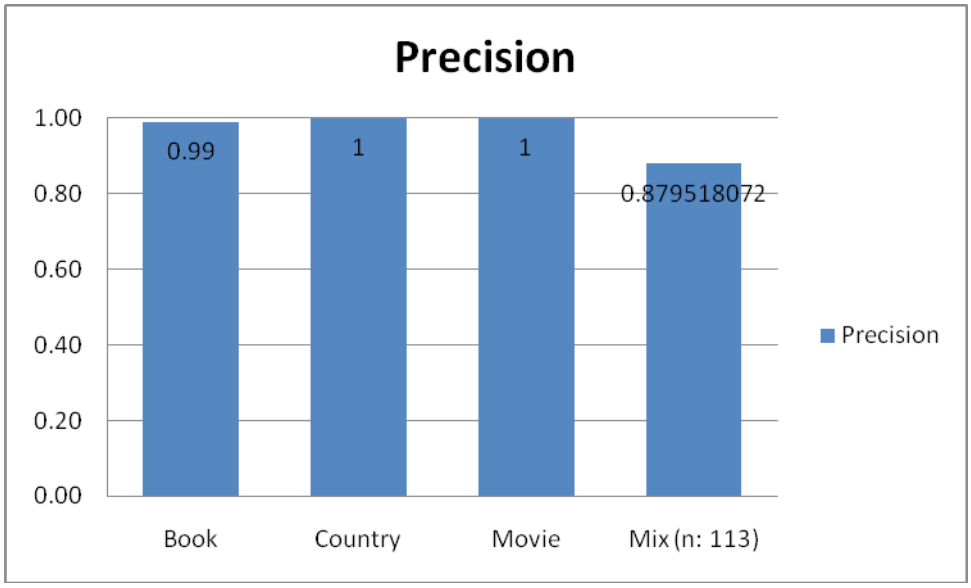

Figure 4.1 Precision of finding Qt, Qf, Qs and finding Ft, Fs and Fu

Figure 4.1, due to our assumptions, we got pretty good precision of Question Analysis's result. The same conditions of pretty good result on Recall and F-Measure. The result of Precision, Recall and F-Measure in single topic were pretty good, because we had a few

assumptions in choosen question as we have explained in the previous pages, we were not deal to all kind of question's type and all situations of a complex questions.

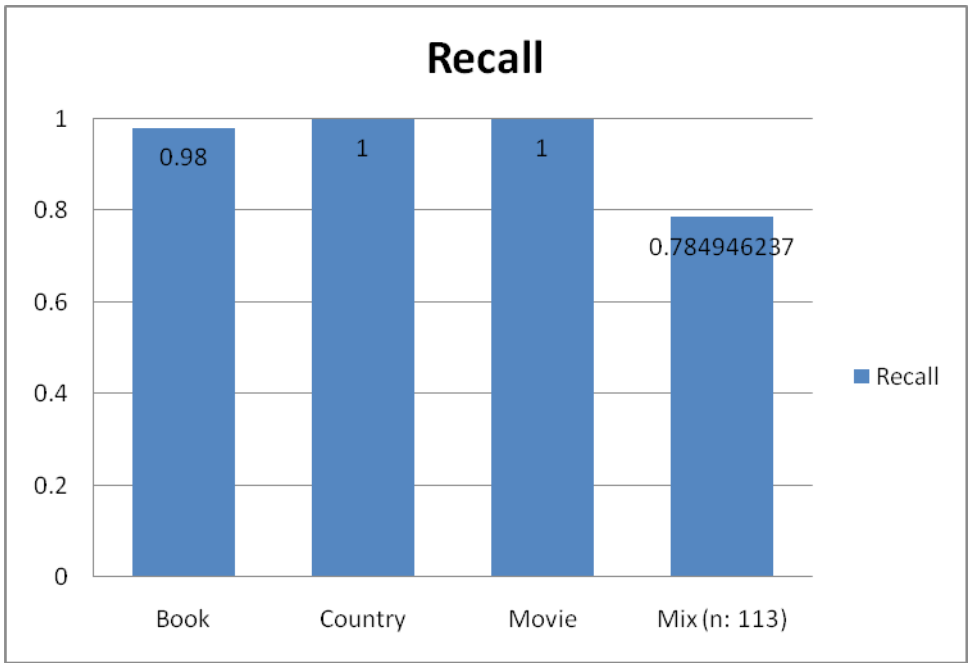

Figure 4.2 Recall of finding Qt, Qf, Qs and finding Ft, Fs and Fu

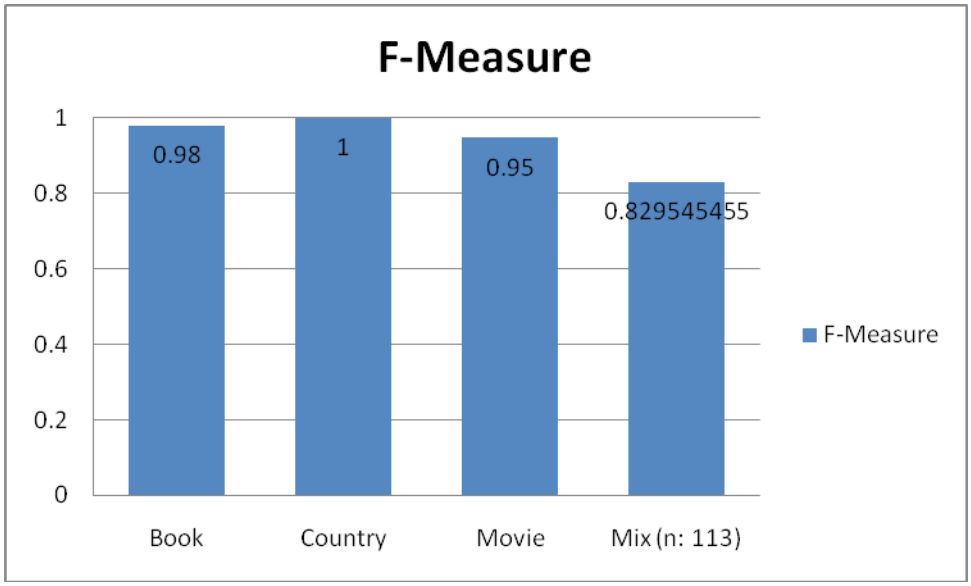

Figure 4.3 F-Measure of of finding Qt, Qf, Qs and finding Ft, Fs and Fu

From the above figures, in the mix topics of questions the result a lower than single topic, it was because several questions gave error in finding Feature_structured (Fs). Several questions contain more than one we try in the combination questions. We choosed the questions randomly and only consider the questions that use prefix 5W1H.

Positive result of Question Analysis is obtained in terms of Qt, Qf , Qs and Ft, Fs and Fu on each of questions, some examples of positive result are listed in Table 4.2:

No	Question	Qt = Ft	Qf	Qs	Fs	Fu
1	What book mentions karl marx going to church with his daughter just to listen to the music	Book ~entity	book ~ name/ title	karl marx going to church with his daughter just to listen to the music	book	karl marx going to church with his daughter just to listen to the music
2	what is the name of the movie about a kid who get a million dollars	Movie	Name	who get a million dollars	name	who get a million dollars
3	what are the characteristics of the country Turkmenistan	country	characteristics	Turkmenistan	Turkmenistan	characteristics

Table 4.2 The true positive example

Some example of negative result are listed in Table 4.3:

No	Question	Qt = Ft	Qf	Qs	Fs	Fu
1	what did you think of the movie disturbia i dont know if thats how you spell it but yea	movie	You think	disturbia i dont know if thats how you spell it but yea	-	-
2	What are the best ways that are the best ways to learn a new language without actually being in that country?	-	The best way	the best ways to learn a new language without actually being in that country	-	-

Table 4.3 The false negative example

Table 4.3, some features can not be obtained. The question have more difficulties to identified their features in Feature_structured and Feature_unstructured, because actually those question only have Feature_unstructured. We can not obtained the Feature_structured.

4.2.2 Resource Discovery and The Relevant Answer

Our first experiment is in country topic. The reasons was country has smaller data structured and this topic has a lot of factoid complex questions. We did the experiments on the small data unstructured. We could not find the relevant unstructured data for some questions. According to this condition we firstly only consider the first top rank document and did the experiment on different fragment_size (fragment size: 50, 75 and 100) and different number of fragment (n: 3, 5, 7 and 10).

We did not define the threshold score_match because some of questions still give a correct answer in the first rank even though the score pretty low. In the first experiment we assumed that small fragment of unstructured data that contain worth of the candidate

of Attributes_unstructured (Au) should give a high score matching than the longer fragment. The result also will show how the scoring of fragments of the webpage is pretty good approach than scoring and consider tho a whole context of webpage. Hence, firstly we only consider in the small size fragment.

In the topic "Country", we can see actually the result pretty similar on different fragment size and different α value. After α value is 0.5 the MRR a little bit lower because the bag of Au wider than previous α, so it cause a little bit decreasing of MRR result

MRR "Country"

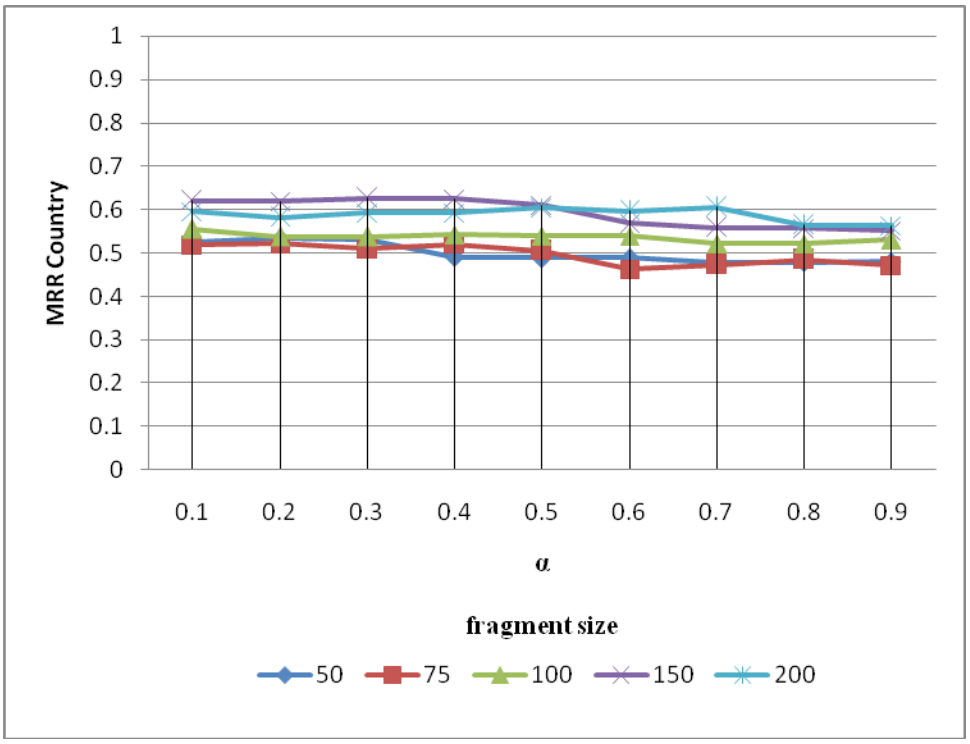

Figure 4.4. MRR for topic "Country"

Figure 4. 4 shows that on the short fragment size (50) the result answer average similar and good in α = 0.1 – 0.3. For wider fragment number 7 and 10 the value decreased along with the wider candidate Attribute_unstructured (Au).

Here is the simple example of our result:

```
THE QUESTION :what are the name of the countries who possess nuclear bombs
Found document(s) that matched query :
G:\wikipedia-en-html.tar\en\articles\n\u\c\Nuclear_proliferation.html
SCORE RELEVANCE: 0.20225996
;Other states known to possess nuclear weapons     &#16
dopted them">nuclear weapons states&quot; from the NPT    &#
(0.04416666785255075)   United States of America USA
```

Figure 4.5 The example of Score_Match and answer result in topic "Country"

We can see beside the passage that contain fragment from Data_unstructured discovery, we also calculate with Data_structured and give some suggestion answer. France is the correct answer of the question and as follow the example result from Bing beta version

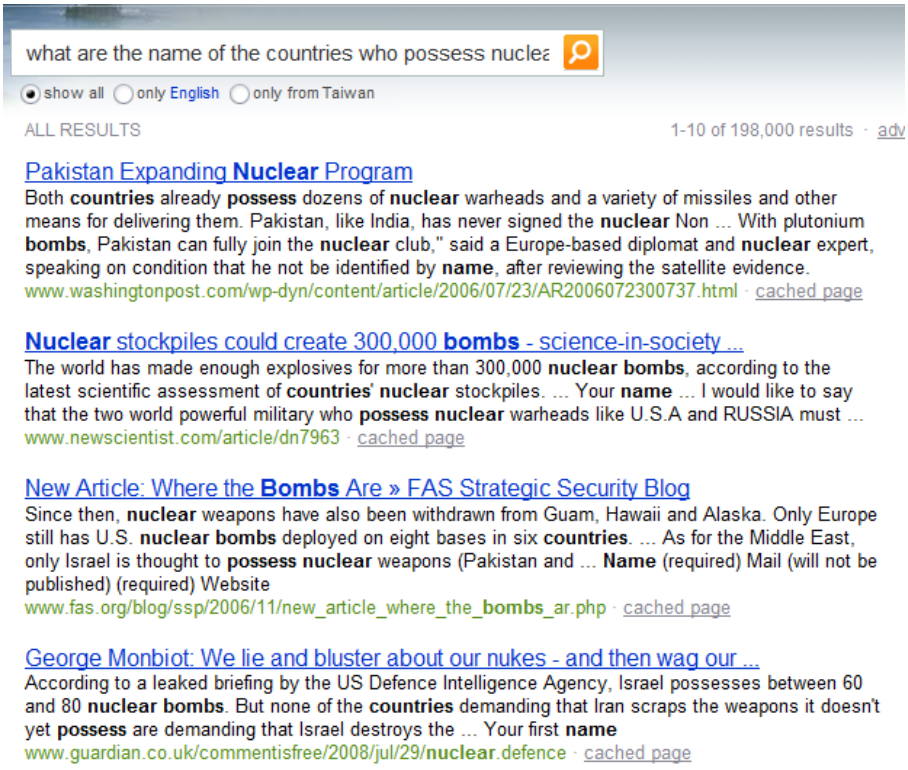

Figure 4.6 The example result from Bing beta version

Figure 4.7 The example result from Google

Next, the experiment for another topics.

MRR "Movie"

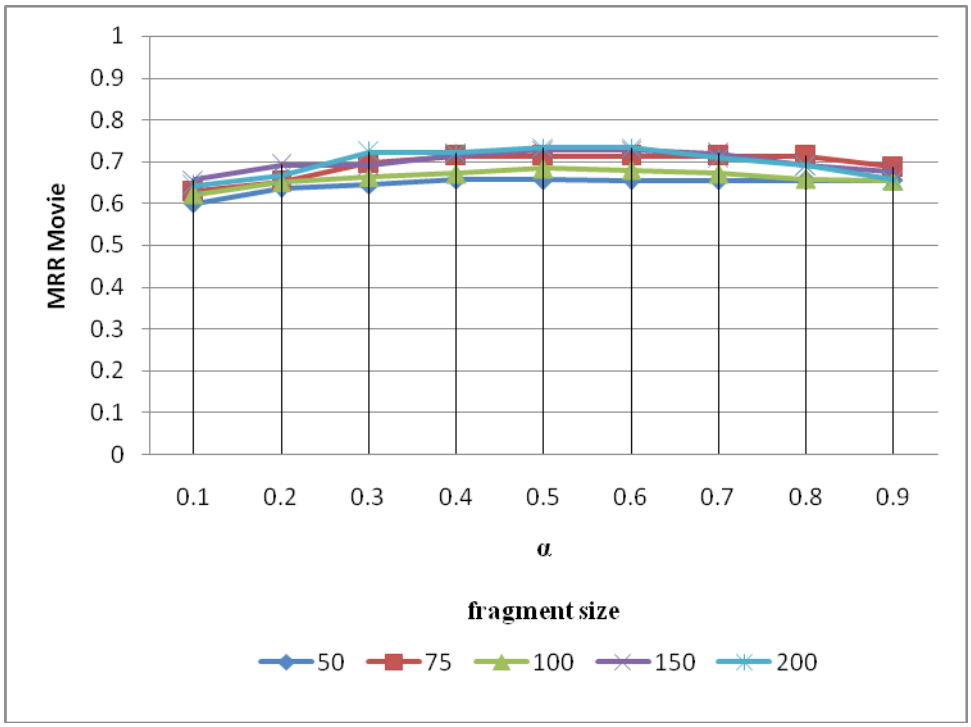

Figure 4.8. MRR "Movie"

The example result as follow:

```
THE QUESTION :what is the name of the movie in which a father takes an entire hospital hostage to get a heart for his son
Found document(s) that matched query :
G:\wikipedia-en-html.tar\en\articles\j\o\h\John_Q_9c5e.html
SCORE RELEVANCE: 0.282738
">hostage</a> situation. He gathers the hostages and sets his demands: his son's name on the recipient list as soon as poss
otiator, Lt. Frank Grimes (<a href="../../../../articles/r/o. Smith) runs around the bases and heads to second, he grabs h
. After a series of tests at the hospital, John is informed that Michael has an enlarged heart and will need to release so
 them is about to give birth &amp;amp; needs to go to the maternity ward but he doesn't give the police this reason) in return
ing on the list an hour into the building via an air shaft. Meanwhile, John speaks with his wife and then his son, telling
on will be all right, unaware that a hidden camera in the hospital has been hacked by a news, and that whatever course he t
is imprisonment or his death. John nevertheless demands that his son be brought to the emergency room, apparently for him t
 The police agree</a> so his heart can be used to save his son. He persuades Dr. Turner (<a href="../../../../articles/j/a/
ml" title="James Woods">James Woods</a>) to perform the operation that his gun was unloaded the entire time he held them ho
ingle bullet into the gun and pulls the trigger, only to learn that the safety was on. As he holds the gun to his own head'
idnapping</a> and although it is unknown what his sentence for the crime will be, his attorney says it won't be more than 3
me="Reception" id="Reception"></a></p="../../../../articles/d/e/n/Denzel_Washington_8230.html" title="Denzel Washington">De
s John Quincy Archibald, a father and husband whose son is diagnosed with an enlarged <a href route to the hospital (this t
persed throughout the movie).</p><p>John Quincy Archibald (<a href="../../../../articles/d/e/n/Denzel_Washington_8230.html'
gton
(0.018729264364272608)    John Q
```

Figure 4.9. The example of Score_Match and answer result in topic "Movie"

MRR "Book"

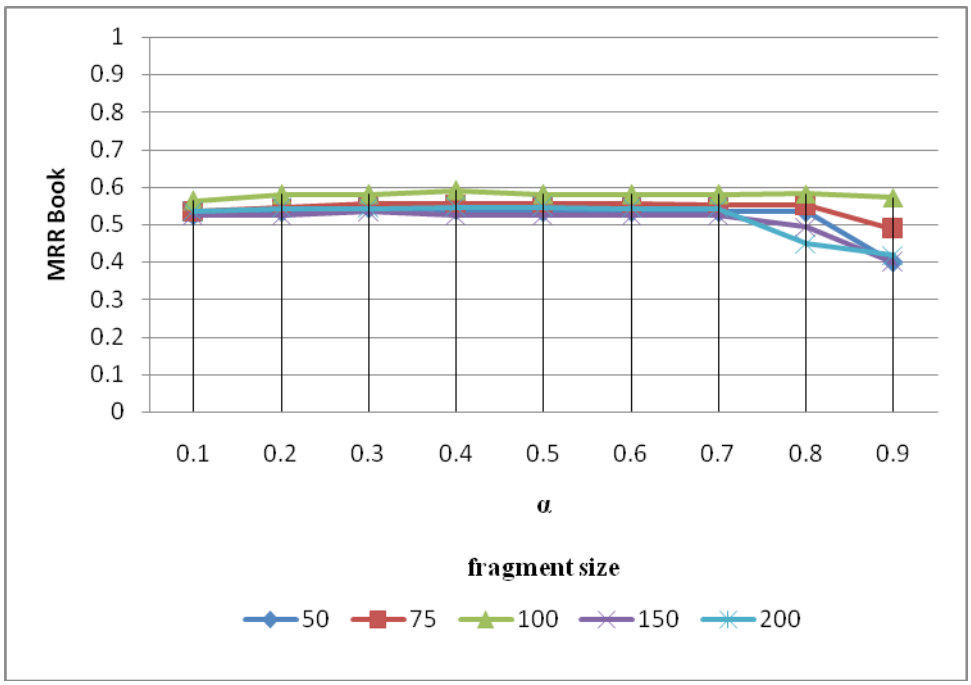

Figure 4.10. MRR "Book"

The example result as follow:

```
THE QUESTION :what book mentions karl marx going to church with his daughter just to listen to the music
Found document(s) that matched query :
G:\wikipedia-en-html.tar\en\articles\g\r\o\Groucho_Marx_8301.html
SCORE RELEVANCE: 0.094280906
by child-like adoring fans. Marx and his brothers</a>. Throughout the rest of his life, Marx would in his
s and he on television.</p><p>One quip from Marx concerned his and did not match the rest of his face, so
G:\wikipedia-en-html.tar\en\articles\k\a\r\Talk~Karl_Marx_Archive_3_79a0.html
SCORE RELEVANCE: 0.059234887
="../../../../misc/favicon.ico"/> <title>Talk:Karl Marx="firstHeading">Talk:Karl Marx/Archive 3</h1> <div
</a></span="Talk:Karl Marx">current talk page</a>.</td></tr Marx. Thus, Marx's mentions of class warfare
(0.08595845406608922)   Karl Marx: A Life
```

Figure 4.11. The example of Score_Match and answer result in topic "Book"

Those above the others result, on topic "Movie" and "Book". The MRR value not really high but pretty good for this initial work that used shallow approach on Question Analysis and Relevent Answer.

Chapter 5. Conclusions and Future Work

5.1 Conclusion

1. We have proposed Finding Structured and Unstructured features to improve a complex questions, and the preliminary work give a pretty good result, can improve the search result answer of a complex question
2. Combining Structured retrieval approach into common documents retrieval can improve the result of questions

5.2 Future Work

1. Improve Question Analysis so can handle many kinds of a complex questions
2. Improve the scoring measure
3. In the unstructured data, work on bigger unstructured data and not really related with structured data.
4. In the structured data, work on more complex structured data, muti table and and multi scheme.

References

1. Lin, C.J. and R.R. Liu. *An analysis of multi-focus questions*. 2008.

2. Li, G., et al. *EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data*. 2008: ACM New York, NY, USA.

3. Bovens, L. and W. Rabinowicz, *Democratic answers to complex questions–an epistemic perspective.* Synthese, 2006. **150**(1): p. 131-153.

4. Harabagiu, S., F. Lacatusu, and A. Hickl. *Answering complex questions with random walk models*. 2006: ACM New York, NY, USA.

5. Harabagiu, S., et al., *Answering complex, list and context questions with LCC's Question-Answering Server.* NIST SPECIAL PUBLICATION SP, 2002: p. 355-361.

6. Saquete, E., et al. *Splitting complex temporal questions for question answering systems*. 2004.

7. Bitton, D., et al. *One platform for mining structured and unstructured data: dream or reality?* 2006: VLDB Endowment.

8. Bhattacharya, I., S. Godbole, and A. Joshi. *Structured Entity Identification and Document Categorization Two Task with One Joint Model*. in *KDD*. 2008.

9. Chang, K.C.C., B. He, and Z. Zhang. *Toward large scale integration: Building a metaquerier over databases on the web*. 2005.

10. Madhavan, J., et al. *Web-scale data integration: You can only afford to pay as you go*. 2007.

11. Andrenucci, A. and E. Sneiders. *Automated question answering: review of the main approaches*. 2005.

12. Lin, J. *The Web as a resource for question answering: Perspectives and challenges*. 2002.

13. Hirschman, L. and R. Gaizauskas, *Natural language question answering: The view from here.* Natural Language Engineering, 2002. **7**(04): p. 275-300.

14. Burger, J., et al., *Issues, tasks and program structures to roadmap research in question & answering (Q&A).* Document Understanding Conferences Roadmapping Documents, 2001.

15. Voorhees, E., et al., *TREC: Experiment and evaluation in information retrieval*. 2005: MIT Press.

16. Moldovan, D., et al., *LCC tools for question answering.* In TREC 2002 Proceedings, 2003.

17. Clarke, C.L.A. and E.L. Terra. *Passage retrieval vs. document retrieval for factoid question answering*. 2003: ACM New York, NY, USA.

18. Cui, H., et al., *Question answering passage retrieval using dependency relations.*

19. Liu, X. and W.B. Croft. *Passage retrieval based on language models*. 2002: ACM New York, NY, USA.

20. Mittendorf, E. and P. Schäuble. *Document and passage retrieval based on hidden Markov models*. 1994: Springer-Verlag New York, Inc. New York, NY, USA.

21. Salton, G., J. Allan, and C. Buckley. *Approaches to passage retrieval in full text information systems*. 1993: ACM New York, NY, USA.

22. Tellex, S., et al. *Quantitative evaluation of passage retrieval algorithms for question answering*. 2003: ACM New York, NY, USA.

23. Sneiders, E., *Automated question answering using question templates that cover the conceptual model of the database.* Lecture notes in computer science, 2002: p. 235-240.

24. Tablan, V., D. Damljanovic, and K. Bontcheva, *A natural language query interface to structured information.* Lecture notes in computer science, 2008. **5021**: p. 361.

25. Lee, Y.H., et al., *Complex Question Answering with ASQA at NTCIR 7 ACLIA.* Entropy. **1**: p. 10.

26. Lin, J. *The role of information retrieval in answering complex questions*. 2006.

27. Hickl, A., et al. *Experiments with Interactive Question-Answering in Complex Scenarios*. 2004.

28. McCallum, A., *Information extraction: distilling structured data from unstructured text.* 2005.

29. Cafarella, M., et al. *Structured Querying of Web Text*. 2007: CIDR.

30. Cafarella, M.J., et al., *Webtables: Exploring the power of tables on the web.* Proceedings of the VLDB Endowment archive, 2008. **1**(1): p. 538-549.

31. Agichtein, E., C. Burges, and E. Brill. *Question Answering over Implicitly Structured Web Content*. in *Web Intelligence*. 2007.

32. Cucerzan, S. and E. Agichtein, *Factoid Question Answering over Unstructured and Structured Web Content.* Microsoft http://research. microsoft. com/users/silviu/Papers/trec05. pdf, 2005.

33. Pinto, D., et al. *Quasm: A system for question answering using semi-structured data*. 2002: ACM New York, NY, USA.

34. Frank, A., et al., *Question answering from structured knowledge sources.* Journal of Applied Logic, 2007. **5**(1): p. 20-48.

35. Bilotti, M.W., et al. *Structured retrieval for question answering*. 2007: ACM New York, NY, USA.

36. Roy, P., et al. *Towards automatic association of relevant unstructured content with structured query results*. 2005: ACM New York, NY, USA.

37. Cody, W.F., et al., *The integration of business intelligence and knowledge management.* Management, 2002. **41**(4).

38. Doan, A. and A.Y. Halevy, *Semantic-integration research in the database community.* AI magazine, 2005. **26**(1): p. 83-94.

39. Halevy, A., A. Rajaraman, and J. Ordille. *Data integration: the teenage years*. 2006: VLDB Endowment.

40. Levy, A., *The Information Manifold approach to data integration.* IEEE Intelligent Systems, 1998. **13**(5): p. 12-16.

41. Williams, D. and A. Poulovassilis. *Combining data integration with natural language technology for the semantic web*. 2003.

42. Yao, C., et al., *Towards a global schema for web entities.* 2008.

43. Nie, Z., et al. *Web object retrieval*. 2007: ACM New York, NY, USA.

44. Ganti, V., A.C. Conig, and R. Vernica. *Entity Categorization Over Large Document Collections,*. in *KDD*. 2008.