FINDING STRUCTURED AND UNSTRUCTURED FEATURES TO IMPROVE THE SEARCH RESULT OF COMPLEX QUESTION

Dewi Wisnu Wardani

Department of Information Technology, Universitas Sebelas Maret, Jalan Ir Sutami No 36-A, Surakarta, 57126, Indonesia

E-mail: dww.okok@gmail.com

Abstract

The current researches on question answer usually achieve the answer only from unstructured text resources such as collection of news or pages. According to our observation from Yahoo!Answer, users sometimes ask in complex natural language questions which contain structured and unstructured features. Generally, answering the complex questions needs to consider not only unstructured but also structured resource. In this work, researcher propose a new idea to improve accuracy of the answers of complex questions by recognizing the structured and unstructured features of questions and them in the web. Our framework consists of three parts: Question Analysis, Resource Discovery, and Analysis of The Relevant Answer. In Question Analysis researcher used a few assumptions and tried to find structured and unstructured features of the questions. In the resource discovery researcher integrated structured data (relational database) and unstructured data (web page) to take the advantage of two kinds of data to improve and to get the correct answers. We can find the best top fragments from context of the relevant web pages in the Relevant Answer part and then researcher made a score matching between the result from structured data and unstructured data, then finally researcher used QA template to reformulate the questions.

Keywords: *structured feature, complex question, question answering*

Abstrak

Penelitian yang ada pada saat ini mengenai *Question Answer* (QA) biasanya mendapatkan jawaban dari sumber teks yang tidak terstruktur seperti kumpulan berita atau halaman. Sesuai dengan observasi peneliti dari pengguna Yahoo!Answer, biasanya mereka bertanya dalam *natural language* yang sangat kompleks di mana mengandung bentuk yang terstruktur dan tidak terstruktur. Secara umum, menjawab pertanyaan yang kompleks membutuhkan pertimbangan yang tidak hanya sumber tidak terstruktur tetapi juga sumber yang terstruktur. Pada penelitian ini, peneliti mengajukan suatu ide baru untuk meningkatkan keakuratan dari jawaban pertanyaan yang kompleks dengan mengenali bentuk terstruktur dan tidak terstruktur dan mengintegrasikan keduanya di web. Framework yang digunakan terdiri dari tiga bagian: *Question Analysis*, *Resource Discovery*, dan *Analysis of The Relevant Answer*. Pada *Question Analysis* peneliti menggunakan beberapa asumsi dan mencoba mencari bentuk data yang terstruktur dan tidak terstruktur. Dalam penemuan sumber daya, peneliti mengintegrasikan data terstruktur (*relational database*) dan data tidak terstruktur (halaman web) untuk mengambil keuntungan dari dua jenis data untuk meningkatkan dan untuk mencapai jawaban yang benar. Peneliti dapat menemukan fragmen atas terbaik dari konteks halaman web pada bagian *Relevant Answer* dan kemudian peneliti membuat pencocoka skor antara hasil dari data terstruktur dan data tidak terstruktur. Terakhir peneliti menggunakan *template* QA untuk merumuskan pertanyaan.

Kata Kunci: *structured feature, complex question, question answering*

1. Introduction

Analyzing the focus of question is not a new issue on question analysis research. 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 key features of questions are the prominent works of those researches for reach user information's need. This topic becomes more interesting to face the long and complex questions. In some of the researches, complex questions often refer to long answer questions. On complex question's research, an answer of 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 search over text documents (e.g HTML documents) have been proposed, they all produce a list of individual pages as results [2].

Automatic Question Answering System usually give a document or a passage that contain the answer as the result. For the example of the question is, "Who is president of USA" then we usually find the results as given by figure 1. We can see that the result usually returns a bag of words. The asker's intention is actually quite clear that they need the name of current president of USA. The results from search engines used to be a bag of words that contain a relevant answers.

Sometimes, it is difficult to achieve the answer of one complex question since the answer can not be retrieved from only one web page or one resource. In fact, it is very common that the answer of one complex question is possibly separated in several web pages. Recently, the research of Question Answering got a challenge of complex question [3][4][5][6]. The detail of our observation will be described on next section.

In this work, the complex question is a natural language question that contains structured and unstructured features. Thus, researcher propose an idea to integrate structured and unstructured data on the web to answer those questions. It is effective to improve the search result of the question. The resources are need to consider not only unstructured data but also structured data. One example is, "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, researcher can find "the capital city" as the structured feature of question and "that the largest country in Arabian Peninsula" as an unstructured feature of question. By these features researcher can effectively retrieve the relevant resource data to answer from both structured data and unstructured data.

For comparison, figure 2 shows the result from search engine Bing usually a relevant passage that contains the needed answer. The factual answer is Riyadh.

In another example, in topic "movie", researcher can find the database of movie on the web as structured data. web pages that contain information of movie are also huge amount exist on the web. Actually, many domain data are stored 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 integrate two kinds of data as the effective resource to improve the answer of the question.

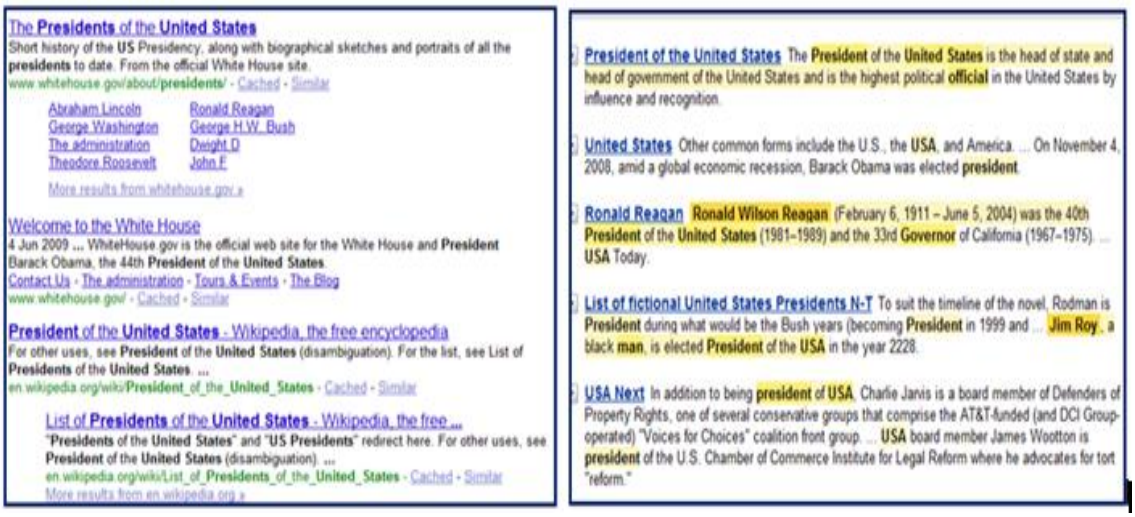

Figure 1. The example result Google and Powerset.


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 2. The example result (rank no.5) from Bing Beta version.

Structured data on the web is prevalent but ignored often by existing information search [7]. Moreover, structured data on the web usually have 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 heterogeneous collections of structured data. Such as product information, Google base, tables on the web pages, or the deep web [8].

According to the complementary characteristics of two kinds of data, it will be very useful to take the advantages of them. The user will not care about from which kind of the resource the relevant information can be found, they only want to get the better answers of their questions.

Since a 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. [9] explains 3 mains question-answering approaches based on Natural Language Processing, Information Retrieval, and question templates. [10] proposed another approaches according to the resource on the web. Lin [11] proposed federated approach and distributed approach. Federated approach is techniques for handling semistructured data 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.

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. [12] explains 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.[13] distinguishes questions by answer type: factual answers, opinion answers or summary answers. 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 [12].

IR systems are traditionally seen as document retrieval systems, i.e. systems that return documents that are relevant to user's information need, but that do not supply direct answers. The Text Retrieval Conferences (TREC) aim at comparing IR systems implemented by academic and commercial research groups. The best performing system within the two latest TREC, Power Answer[14] 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 [15][16][17][18] [19][20][21].

Template-based QA extends the pattern matching approach of NLP 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 is mostly useful for structured data, as mentioned on [10]. [22] propose 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.[23] used a kind template and used ontology on question analysis, and work on structured information on the text.

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

Question Analysis: In the beginning of researcher's idea, researcher only consider the question whose prefix has a question word (What, Who, Where, When, Which, Why, How) for each of topic domain, including Book, Country, and Movie.

In this first step, researcher need to know the structured feature and unstructured feature that exist on the questions. For the sake of simplification, in this initial work researcher only consider one kind of complex question that might contain structured and unstructured feature. As had been known, a natural language question has many forms of syntax and expression. Hence, researcher put some assumptions in this step

according to our observation of the questions from Yahoo!Answer (in English). Besides finding those features, researcher also want to find the focus and subfocus of the question. From the same example, "What is the capital city of the country that is the largest country in Arabian Peninsula?". Where Question Topic is "country", Question Focus is "the capital city", Question Subfocus is "that is the largest country in Arabian Peninsula", Structured feature is "the capital city", and Unstructured feature is "country that is located on a long boot shaped peninsula".

We can see that the structured features are the question focus. This condition is one of situation that is issued in dealing with question analysis. Our question data are mostly about entity question. We want more to see the answer tends to structured data.

Resource Discovery and Reach the Relevant Answer: Figure 3 show a framework that use in this work. We take advantage for two kinds of data. For the structured data, the form of this data is simple relational data, e. g single table with attribute name and attribute value. For unstructured data researcher crawl web pages from several websites included Wikipedia. For this initial work, researcher tried to integrate the answer result from two different types of data resource. One of the basic problems of integration is relevant answer matching problem. In our work this answer matching is mostly about the matching terms of both two resources. We will propose a simple linear combination model to reach the score matching between the unstructured data and structured data for a given complex question. Finally, based on the simple answer matching model, it can be reached from both two kinds of resources. Hence researcher can improve the result answer of the question.

We focus on two main works, the first step is finding the structured and unstructured features on the question. The second step 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 part, 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 $\{S_u\}$.

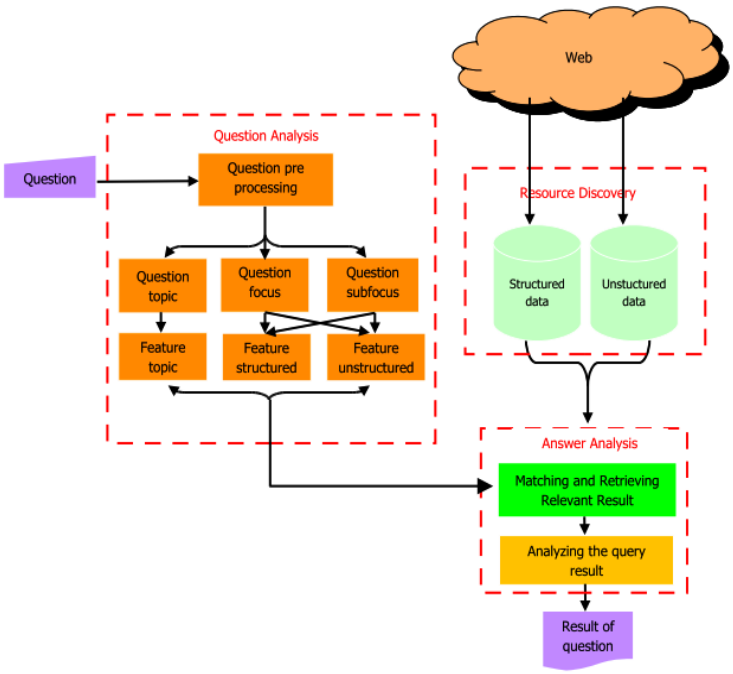

Figure 3. Framework of finding structured and unstructured features to improve result of complex questions.

2. Methodology

Question Analysis: In the beginning of our idea, researcher only consider the question whose prefix has a question word (What, Who, Where, When, Which, Why, and How). We observed 100 questions of three topics, Book, Country, and Movie. We consider on the question that has phrase "of a" or "of the" or has main clause and subordinate clause. We proposed the Algorithm Finding Structured-Unstructured Feature, consists first step of finding the Question topic (*Qt*), Question focus (*Qf*) and Question sub focus (*Qs*) and the second step finding the Feature topic (*Fs*), Feature structured (*Fs*) and Feature unstructured (*Fu*) from the question.

To measure whether the Qf is Fs or Fu researcher use this equation:

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

Where, Fs is Feature_structured, Qf is Question_focus, Ds is Data_structured, An is Attribute_name, and Av is Attribute_value.

Next, to measure whether the *Qf* can become the Focus of Attributes (*FAn*) researcher use this equation.

$$FAn^* = \arg\max_{FAn} \sum_{FAn \in An} P(FAn \mid Qf) \tag{2}$$

Where *An* is Attribute_name. Figure 4 is an algorithm of finding structured-unstructured features.

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 [24]. In this prominent work researcher use on the Data_structured *(Ds)* side, the relational database single table, and as usually the Data_unstructured *(Du)* side, the web pages [25].

```
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
```

Figure 4. The algorithm of question analysis.

[8] and several previous researches have proposed idea of the integration resources [1][8] [22] [26][27][28][29][30] . The main reason of their work is try to find the advantage on each of resources. Richer their resources mean better answer. Particularly [8] said that asker do not care the resource, they only want find the better answer. Another works [31][32] about using both structured and unstructured data to improve the answer.[2] first work on the keyword search on integration data: structured, semi structured and unstructured data with graph approach. Proposed a kind of integration entities that exist on table-like format on the web pages. It is the integration of information on the unstructured data.

Using the structured data and unstructured data in Information Retrieval or Question Answering researches are not new research issue. Since the size of high quality structured data on web is increasing and not yet be optimum explored, using the combination of them seems a new research issue on Question Answering. One previous proposed a prominent work, find structured content over text [33]. [34] 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.

The question 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_name (*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 or fragment on the Du.

To find the relevant page Du_j by the cosine similarity measure which defines in Equation (3), and use the *Fu* to find the annotated snippet.

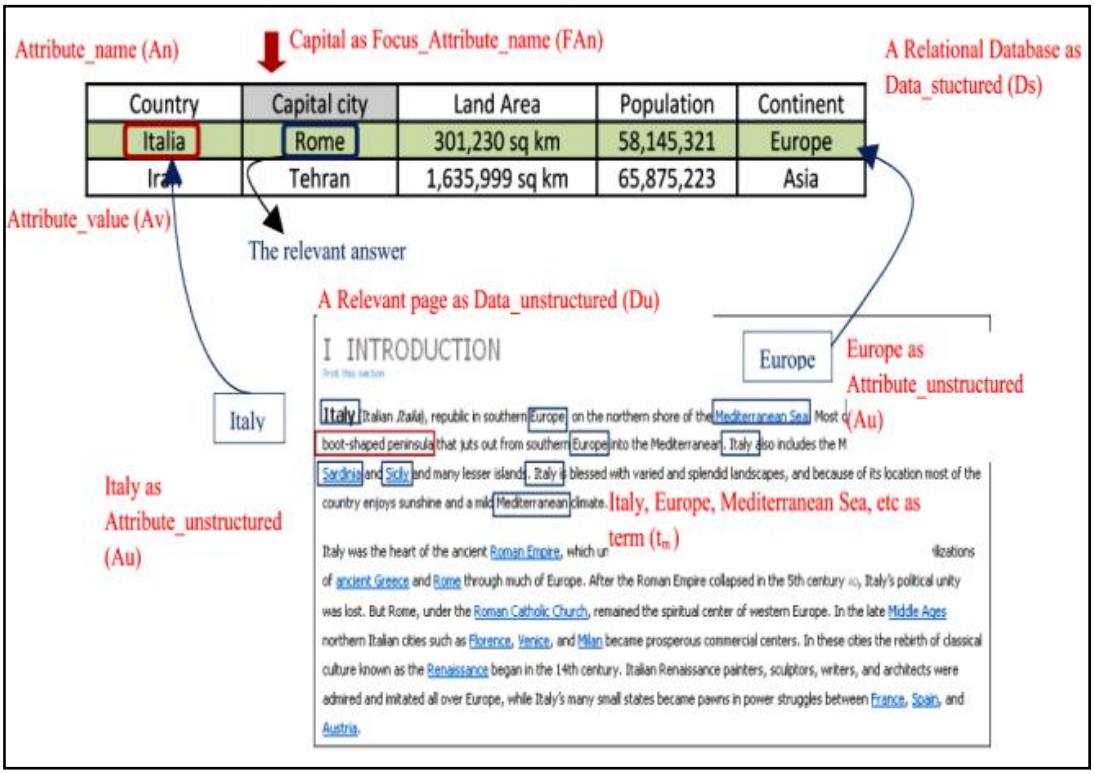

Figure 5. Example of resource discovery.

$$S(Du_j, q) = \frac{\sum_{i=1} wi(Du_j) \cdot wi(q)}{\sqrt{\sum_{i=1} wi(Du_j)^2 . \sum_{i=1} wi(q)^2}} \quad (3)$$

Where *S* is Score of cosine similarity between Du_j and *q*, *Du* is Data_unstructured, and q is Feature_topic and Feature_unstructured. Where the weight (*w*) is based on *TFIDF* weighting scheme.

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

Be inspired from previous work [15], researcher want to find the relevant snippet of Du_j, 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_j.

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

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

Where, *Av* is Attributes_value of *Ds* and *S* is the choosen snippet of *Du*. Here, consider the score of snippet or fragment have found of a relevant documents.

Finding The Relevant Answer: To analyze all terms on the relevant snippets *Du* and then choose the terms t_i that contains a set *Av* as Attributes_unstructured *(Au)*. For the question, "What is the capital of the country that is located on a long-boot shaped peninsula?" around n-gram term "long boot shaped peninsula" we would get another term such as "Italy", "Sicilia", "Roman Empire", "Renaissance", "Sardinia", "Mediterranean" etc.

$$Au^* = \arg\max_{Au} P(Au \mid Av) \quad (7)$$
$$= \arg\max_{Au} \sum_{t_i \in Au} P(t_i \mid Av)$$

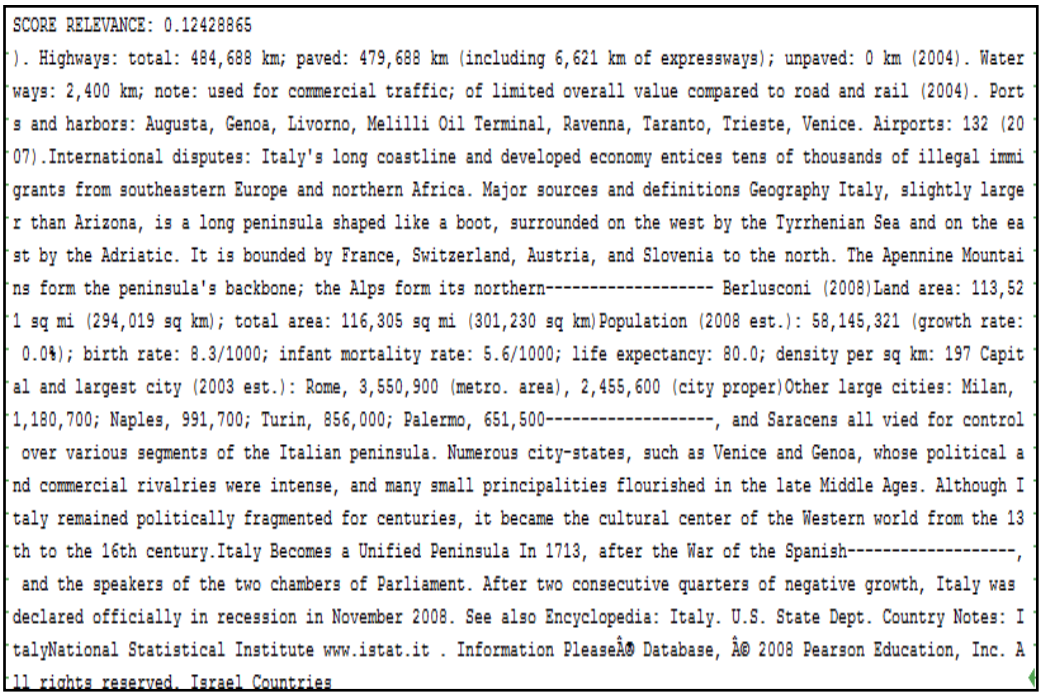
SCORE RELEVANCE: 0.12428865
). Highways: total: 484,688 km; paved: 479,688 km (including 6,621 km of expressways); unpaved: 0 km (2004). Water ways: 2,400 km; note: used for commercial traffic; of limited overall value compared to road and rail (2004). Port s and harbors: Augusta, Genoa, Livorno, Melilli Oil Terminal, Ravenna, Taranto, Trieste, Venice. Airports: 132 (20 07).International disputes: Italy's long coastline and developed economy entices tens of thousands of illegal immi grants from southeastern Europe and northern Africa. Major sources and definitions Geography Italy, slightly large r than Arizona, is a long peninsula shaped like a boot, surrounded on the west by the Tyrrhenian Sea and on the ea st by the Adriatic. It is bounded by France, Switzerland, Austria, and Slovenia to the north. The Apennine Mountai ns form the peninsula's backbone; the Alps form its northern---------- Berlusconi (2008)Land area: 113,52 1 sq mi (294,019 sq km); total area: 116,305 sq mi (301,230 sq km)Population (2008 est.): 58,145,321 (growth rate: 0.0%); birth rate: 8.3/1000; infant mortality rate: 5.6/1000; life expectancy: 80.0; density per sq km: 197 Capit al and largest city (2003 est.): Rome, 3,550,900 (metro. area), 2,455,600 (city proper)Other large cities: Milan, 1,180,700; Naples, 991,700; Turin, 856,000; Palermo, 651,500----------, and Saracens all vied for control over various segments of the Italian peninsula. Numerous city-states, such as Venice and Genoa, whose political a nd commercial rivalries were intense, and many small principalities flourished in the late Middle Ages. Although I taly remained politically fragmented for centuries, it became the cultural center of the Western world from the 13 th to the 16th century.Italy Becomes a Unified Peninsula In 1713, after the War of the Spanish----------, and the speakers of the two chambers of Parliament. After two consecutive quarters of negative growth, Italy was declared officially in recession in November 2008. See also Encyclopedia: Italy. U.S. State Dept. Country Notes: I talyNational Statistical Institute www.istat.it . Information PleaseÂ® Database, Â© 2008 Pearson Education, Inc. A ll rights reserved. Israel Countries

Figure 6. Example of fragmentation of unstructured data, the dash as boundary between fragment.

Consider all terms on the snippet that could be the candidates of Attribute unstructured (*Au*) and calculate the score of answer matching of Unstructured data and Structured data in order to get the answer matching score of record *R*. We proposed score matching inspired from full string matching based Jaccard coefficient and *n*-gram matching. First, researcher use Jaccard coefficient to calculate the answer matching score between a record *R* in *Ds*.

$$Score1 = J(R, Au) = \frac{|R \cap Au|}{|R \cup Au|} \quad (8)$$

Second, n-grams are typically used in approximate string matching by "sliding" a window of length n over the characters of a string to create a number of 'n' length grams for matching a match is then rated as number of n-gram matches within the second string over possible n-grams. Inspired from [35], researcher use equation (9) to calculate the answer matching score between *R* and *Au*. *R* contains a set of *Av* and *Au* is sequence of text, they are be a pair of *n*-grams in *X* and *Y*. Let $R: x_1 \ldots x_k$ and $Au: y_1 \ldots y_l$

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

Where and contains at least one complete *n*-gram.

$$S_n(\tau_{k,l}) = 0 \; if \; (k = n \wedge l < n) \vee (k < n \wedge l = n) \quad (10)$$

And if both strings exactly one n-gram, the initial definition is strictly binary: 1 if the n-gram are identical and 0 otherwise.

$$S_n(\tau_{n,n}) = S_n(\tau^n_{0,0}) = \begin{cases} 1 \; if \; \forall_{1 \le u \le n} \; x_u = y_u \\ 0, \; otherwise \end{cases} \quad (11)$$

Researcher used n-gram, to find the similarity between *Du* and *Ds* and consider the position of letter so researcher will find similarity even not really exact. Those all about the answer matching score. The answer matching score is very important to match the unstructured data and structured data. It is all use IR approach then the score is a linear combination as follows:

$$Answer\_Match\_Score = \alpha \bullet Score1 + (1 - \alpha) \bullet Score2 \quad (12)$$

Where α, is weighting parameter (0.1 to 0.9).

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

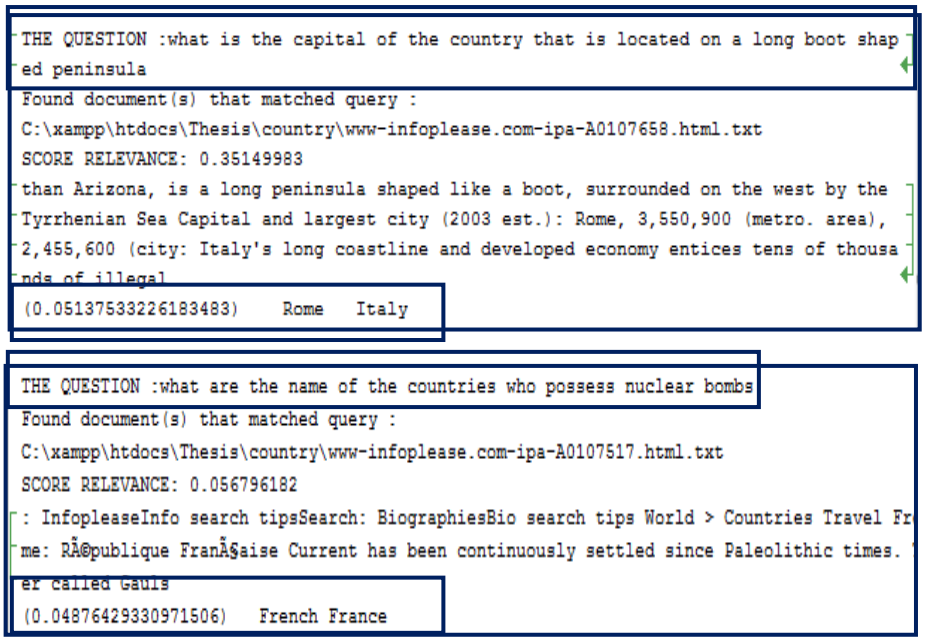

Figure 7. Example of final result of this system.

For the example of the question, "What is the capital city of the country that the largest country in Arabian peninsula", the QA template is like figure 8.

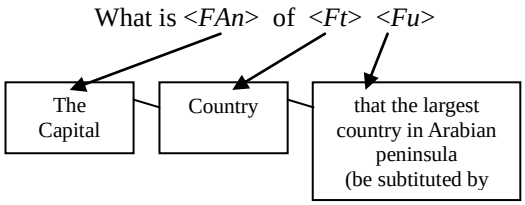

Figure 8. Question template approach in this work.

We can see from figure 7, from two question this system can give the accurate suggestion answer and both of them are true.

Dataset: In our work, for Question Analysis, researcher used real questions from Yahoo!Answer and the chosen question from TREC 2005-complex question track. The question only in English.

TABLE I
DATASET OF QUESTIONS

Topics	Training	Testing
Book	65	40
Country	65	40
Movie	65	40

As in the very beginning of our explanation researcher used two kind of data. As follows our data in 3 topics. Structured Data is single table relational database and unstructured data is a web page from websites.

The attributes on the table of structured data are 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], and Movie → [id, name_title, year_release, director, genre]. Table I and II show the the dataset question and the description of dataset.

TABLE II
DESCRIPTION OF DATASET

No	Topic	Structured data	Unstructured data
1	Book	10,378 rows From Amazon	~ 800 KB From Infoplease
2	Country	196 rows From About	~ 238 GB From Wikipedia
3	Movie	10,978 rows From IMDB	

3. Result and Analysis

In Question Analysis researcher use evaluation metrics *Recall (R)*, *Precision (P)* and *F-Measure (F-Measure)*. In the Resource Discovery and reach the relevant answer, besides use the *Precision*, *Recall* and *F-Measure*, researcher will use *MRR* in different fragment size, different threshold of match_score and different α.

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 of structured data and unstructured data. The experiment is devided into two sections, in the Question Analysis and the result answer.

In table III, researcher obtained high precision of Question Analysis's result. The same conditions on *Recall* and *F-Measure*. The result of *Precision*, *Recall* and *F-Measure* in single topic were high, because researcher had a few assumptions in chosen questions as researcher have explained in the previous pages, researcher do not deal to all kinds of question's type and all situations of a complex questions. In the mix topics of questions the result is lower than single topic because several questions gave errors in finding Feature_structured (*Fs*). Several questions contain more than one *Fs* in the combination questions. We chose the questions randomly and only consider the questions words, 5W1H.

TABLE III
PRECISSION, RECALL AND F-MEASURE OF OF FINDING QT, QF, QS AND FINDING FT, FS AND FU

	Book	Country	Movie	Mix
Precission	0.88	0.89	0.87	0.87
Recall	0.87	0.88	0.80	0.85
F-Measure	0.87	0.88	0.83	0.86

TABLE IV
MRR OF THREE TOPICS

α	Book	Country	Movie
0.1	0.665531	0.562879	0.629573
0.2	0.544161	0.558594	0.659150
0.3	0.550361	0.559016	0.683141
0.4	0.549761	0.553989	0.695557
0.5	0.547021	0.549287	0.702760
0.6	0.546361	0.531259	0.701841
0.7	0.546008	0.527401	0.693414
0.8	0.521202	0.521100	0.681369
0.9	0.454650	0.518998	0.665531

We did the experiments on the small unstructured data. According to this condition researcher 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).

For the above results, on topic "Movie" and "Book", the *MRR* values as show in table IV, not really high but very promising for this initial work that used shallow approach on Question Analysis and Relevant Answer.

We also have compared our approach to the other systems, QuALiM and Powerset. We compared to them because of the resource data of unstructured data were alike, from Wikipedia. Since the result of them is a snippet of result that contains the answer, researcher manually calculate the *MRR* of their result. We examine whether the answer exist on the snippet. The answer is correct if researcher could find the correct answer on the snippet.

TABLE V
COMPARISON MRR OF QUALIM, POWERSET AND THE PROPOSED APPROACH

	QuALiM	Powerset	Proposed approach
MRR	0.1730769	0.4539103	0.5847888

Table V shows that this approach could improve the search result. One note that our approach not only give a snippet result but also an exact suggestion's answer as already explained on the previous pages.

4. Conclusions

We have proposed the preliminary work of finding structured and unstructured features on complex questions. The complex question in this work is a natural language question that contains structured features and unstructured features. Structured feature refers to Structured data and Unstructured feature refers to unstructured data. Structured data grows rapidly on the web but usually be ignored by existing search engine. In this work show that combination structured and unstructured data. Besides use two kinds of data, researcher also use two approaches, IR approach tend to unstructured data and QA-Template approach tend to Structured data. Actually, historically those two approach worked separately. The other idea of this work, researcher tried to use structured approach on unstructured approach.

This work gives a pretty good result on the Question Analysis in all evaluation metrics, Precision, Recall and F-Measure. In the finding the relevant answer, the result was not really high but still promising, the average > 0.5. Also the comparison with two other systems, QuALiM and Powerset, our approach outperforms both systems. We compared it because they use the similar unstructured data, Wikipedia (english version).

According to our knowledge, the idea on this work is novel, because the previous relevant researches used to worked on unstructured data or structured data. We believe it will very useful. Since this work is our preliminary work, researcher still have many things to do. Our future work will emphasis on Question Analysis and matching measure parts. Improving Question Analysis to handle many kinds of complex questions, even long questions.

Improving the scoring measure, as far as our observation, the main work of integrated structured and unstructured features is matching problem. This part still have a long journey on the integration data. In the unstructured data, work on bigger unstructured data and not really related with structured data and in the structured data side, work on more complex structured data, multi table, and multi scheme.

Acknowledgement

Thanks to Yahoo, Amazon, Infoplease, About, and IMDB that let the author crawl their data also Wikipedia that let the author using latest data of Wikipedia.

Reference

[1] J. Lin, "The Web as a resource for question answering: Perspectives and challenges" *In Proceeding of LREC*, pp. 555–562, 2002.

[2] J. Lin, "The role of information retrieval in answering complex questions" *In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 523–530, 2006.

[3] L. Bovens & W. Rabinowicz, "Democratic answers to complex questions–an epistemic perspective" *Synthese*, vol. 150, pp. 131-153, 2006.

[4] L. Hirschman & R. Gaizauskas, "Natural language question answering: The view from here," *Natural Language Engineering*, vol. 7, pp. 275-300, 2002.

[5] G. Kondrak, "N-gram similarity and distance, " *Lecture notes in computer science*, vol. 3772, p. 115, 2005.

[6] V. Tablan, D. Damljanovic, & K. Bontcheva, "A natural language query interface to structured information" Lecture notes in computer science, University of Sheffield

Department of Computer Science, p. 361, 2008.

[7] K.C.C. Chang & B. He, Z. Zhang, "Toward large scale integration: Building a metaquerier over databases on the web" *In Proceeding of CIDR*, pp. 44-53, 2005.

[8] D. Moldovan, et al., "LCC tools for question answering" *TREC*, pp. 144-154, 2003.

[9] A. Andrenucci & E. Sneiders, "Automated question answering: review of the main approaches". *In Proceeding of ICITA*, pp. 514-519, 2005.

[10] J. Madhavan, et al., "Web-scale data integration: You can only afford to pay as you go" *In Proceeding of CIDR*, 2007.

[11] X. Liu & W.B. Croft, "Passage retrieval based on language models" *In the Proceedings of CIKM '02 conference*, pp. 375-382, 2002.

[12] A. Levy, "The Information Manifold approach to data integration," *IEEE Intelligent Systems*, vol. 13, pp. 12-16, 1998.

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

[14] G. Salton, J. Allan, & C. Buckley, "Approaches to passage retrieval in full text information systems" *In Proceeding of* SIGIR, pp. 49-58, 1993.

[15] A. Halevy, A. Rajaraman, & J. Ordille, "Data integration: the teenage years" *In Proceeding of VLDB Endowment*, pp. 9-16, 2006.

[16] A. Doan & A.Y. Halevy, "Semantic-integration research in the database community," *AI magazine*, vol. 26, pp. 83-94, 2005.

[17] V. Ganti, A.C. Conig, & R. Vernica, "Entity Categorization Over Large Document Collections" *In Proceeding of KDD*, pp. 274–282, 2008.

[18] E. Mittendorf & P. Schäuble, "Document and passage retrieval based on hidden Markov models" *In Proceeding of SIGIR*, pp. 318–327, 1994.

[19] Z. Nie, et al, "Web object retrieval" *In Proceeding of WWW*, pp. 81-90, 2007.

[20] E. Sneiders, Automated question answering using question templates that cover the conceptual model of the database, Lecture notes in computer science, pp. 235-240, 2002.

[21] C. Yao, et al, "Towards a global schema for web entities" *In Proceeding of WWW*, 2008.

[22] S. Tellex, et al, "Quantitative evaluation of passage retrieval algorithms for question answering" *In Proceeding of SIGIR*, 2003.

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

[24] A. Halevy, A. Rajaraman, & J. Ordille, "*Data integration: the teenage years*" *In Proceeding of VLDB Endowment*, pp. 9-16, 2006.

[25] J. Allan, et al. "Challenges in information retrieval and language modeling" *In Proceeding of SIGIR*, pp. 31–47, 2003.

[26] C.L.A. Clarke & E.L. Terra, "Passage retrieval vs. document retrieval for factoid question answering" *In Proceeding of SIGIR*, p. 2, 2003.

[27] S. Cucerzan & E. Agichtein, "Factoid Question Answering over Unstructured and Structured Web Content," *In The Text Retrieval Conference (TREC)*, pp.1-6, 2005.

[28] L. Gravano, et al., "Using q-grams in a DBMS for approximate string processing," *IEEE Data Engineering Bulletin*, vol. 24, pp. 28-34, 2001.

[29] S. Harabagiu, et al., "Answering complex, list and context questions with LCC's Question-Answering Server" *In Proceedings of The 10th Text Retrieval Conference (TREC-2001)*, pp. 355-361, 2001.

[30] V. Tablan, D. Damljanovic, & K. Bontcheva, "A natural language query interface to structured information" Lecture notes in computer science, University of Sheffield Department of Computer Science, p. 361, 2008.

[31] D. Bitton, et al, "One platform for mining structured and unstructured data: dream or reality?" *In Proceedings of the 32nd international conference on Very large data bases*, pp. 1261-1262, 2006.

[32] H. Cui, et al, "Question answering passage retrieval using dependency relations" *In Proceeding of SIGIR*, pp. 400-407, 2005.

[33] E. Agichtein, C. Burges, & E. Brill, U.S. Patent No. 7873624, October 2006.

[34] E. Saquete, et al, "Splitting complex temporal questions for question answering systems" *In Proceeding of ACL*, 2004.

[35] S. Harabagiu, F. Lacatusu, & A. Hickl, "Answering complex questions with random walk models" *In Proceeding of SIGIR*, 2006.