INSTANT ANSWERS, 5W'S & H TOOL

Muhammad Faisal[1], Usman Waheed[2] and Muhammad Nabeel Arif[3]
Computer & Software Engineering Department; Bahria University Karachi Campus, Pakistan,
Corresponding author's email: [1]{ mfaisalpk@outlook.com }, [2]{uwaheed@bimcs.edu.pk}, [3] {mnabilarif@outlook.com}

ABSTRACT: *Instant answers search engine provides solution to the problems of visiting multiple sources, compilation effort, time consumption and incompleteness of information while using Crawler-based search engines, encyclopedias, video portals, image portal etc. for the purpose of making assignments, reports and in other research works. Instant answers search engine works with both structured and unstructured data and uses Five W's & H to provide completeness. It involves multiple levels of classifications, selection of the correct data sources, strategies for the extraction of data with the use of multi layered Pipeline Architecture. Its ultimate goal is to provide fast, accurate, complete and compiled information for the students, researchers and knowledge seekers.*

Keywords: *Instant Answer (IA); Knowledge-Base (KB); Natural language processing (NLP); Named Entity Recognition (NER); 5 W's & H Model (What, Who, Where, Why, When, How);*

1. INTRODUCTION

In order to get the complete information of a desired person, place, organization, and concept or object you need to look for it on multiple sources which requires a lot of time and effort. At first you do so much effort in finding the right data and then again you put your effort in the compilation of the data that you have collected but it still misses the completeness. Such type of search engines are known as question answer based search engines.

Question answer based search started in early nineties when text based data started growing exponent. First recognized text based question answer search engine START [1, 2] is developed in 1997. Then series of TREC Question Answering [3] systems emerged with natural language queries, which resulted formatted passages of desired answers. These searches were based extracting answers from many sources.

The second era of question answering was based on knowledge base (KB), and QALD (Question answer over linked data). These researches are based on recognized knowledge-bases DBpedia [4,5], which is based on a wide range of properties from Wikipedia info boxes and Freebase [6] which provides "base" domain information in verity of areas. These searches are based generating SPARQL [7] query on given question's entities and relations from knowledge bases. And the resultant parts of query are organized to make a statements and paragraphs of answer. Recognized KB projects are AutoSPARQL [8,9,10] system, DEANNA [11], and Xser[12]. Beside research some commercial projects are evolved. Most famous is Wolfram alpha [13, 14], which provides general knowledge question answers, based on their own knowledgebase along with existing knowledgebase. Second recognized is Facebook graph search [15], which is based on relationship between user and their friends in context of persons, places, pictures, and different tags. Google search, especially knowledge graph [16, 17] is based on freebase KB.

Latest research is based on question answers on combined data means multiple knowledge bases or semantic web data sources. Most popular is Watson system [18, 19] which uses both text, multiple knowledge bases, encyclopedias, and different directories to answer the user question. It uses multilayer architecture to answer the complex questions in presentable format. It uses natural language processing with machine learning using unstructured data.

2. CHALLENGES

Five W's & H is a model to completely understand about any incident, event, topic, location, person, organization and etc. is also known as reporter's question. This model makes it sure that not a single bit of information gets misses. By automating this saves a lot of time and effort but this process incurs three challenges in the way of its automation.

(i) *Identifying different parts of question, or input keyword*
(ii) *Identifying relevant sources*
(iii) *Same words with different meanings*

First challenge is catered by Natural language processing, which uses statistical models to understand and parse the naturally written language. Named-entity recognition (NER) is also know by the names of entity extraction, entity identification and entity chunking is subtask in information extraction from given text into pre-defined categories like person, organization and etc. [20]. NER involves issues in itself to exactly identifying the right entities; like many statistical models it uses only useing local structure that limits its many tasks and constraint them in tractable model inference. A general method for solving this method is to relax the requirement of exact inference, substituting approximate inference algorithms instead, thereby permitting tractable inference in models with non-local structure. One of such algorithm is Gibbs sampling [20], a simple Monte Carlo algorithm [21] that is appropriate for inference in any factored probabilistic model, including sequence models and probabilistic context free grammars.

Second challenge is regarding Answering to natural language question involves two main challenges: recognizing the questions meaning and grounding to the relevant data source or precisely data set to answer the questions. The existing semantic parsers try to model the two aspects in a nice uniform model [22], but usually have difficulties when: Simultaneously learning the meaning representations and the mappings against KB items will lead to a huge search space, thus it is often inefficient to train such a parser in open domain.

Research finds that recognizing the meaning representation of user's intention in a question is naturally KB (Knowledge Base)-independent, while the mapping phase is indeed KB-related. We thus propose a pipeline paradigm involving two

steps; first recognizing the KB-independent meaning representations inherent in the questions, and then converting the meaning representations into KB-related structured queries. Most of existing sources are KB, but they have some limitations such as:

(i) Search engines these days are considered as the most popular source for finding information on the internet. Example Google [23] and Bing [24]. Major limitation of using search engine is that they only provide you the web links that misses vital portions of your information which make it complete.

(ii) Encyclopedias plays very important part in fulfilling the need of information needed by the knowledge seekers. Example Wikipedia [25] and webopedia [26]. They only cover three W's. What, When and where. In most cases, it usually misses rest of the other W's.

(iii) Knowledge Portals which are fully dedicated for the information gathering with their relevant videos and images. Example InstaGrok [27]. They only covers "What" part of the information and their videos and images are most of the time seems irrelevant.

Firstly, using the basic simple Lambda Dependency-Based Compositional Semantics as the KB-independent meaning representation language, except that each predicate is a natural language phrase [22]. Secondly, probabilistic model to determine the probabilities of mapping between natural language phrases and KB items as well as aggregation functions.

Third challenge is Finding 5W's and H for any concept, term or other entity is even more difficult because here we don't have any structured data sources to get the data of related queries and interpretation of these becomes ambiguous some time e.g. OOPs can be an expression and on the other hand it can be interpreted as Object oriented programming. Ambiguity problem is although solved by Wikipedia disambiguate architecture. We are using two strategies mainly to cater these issues: first, to use crawler to a specific source for the extraction of information with the help of keywords with their synonyms which categorized under W's and H. second, use the links which are provided by search engines and again extract information with keywords which are already categorized.

3. INSTANT ANSWERS

Research designed an Instant Answers tool is to answer about person, place, organization and concept on separate containers of single web page. Framework uses Pipeline architecture to answer user query in separate 5 W's & H containers in detailed paragraph form. Instant Answers tool also provide images and videos related to user question or query. Before the implementation of this strategy, it is very important to classify the inputs that user can be able to given. So, the classification of the inputs and scope of algorithm is in this way that it can include person, place, organization, and concept or object. After this classification further more classification is required for each W and H so that data can be extracted from the web for a topic/concept is represented in Table 1 to Table 4. Whereas Table 1 is for Topic 5W's & H keywords that are worth searching regarding each W's & H similarly Table 2 is for person, Table 3 is for place and Table 4 is for Organization.

WHAT	WHEN / WHO	WHO	WHERE	WHY / WHY NOT	HOW TO
DEFINITION	HISTORY	INVENTOR	APPLICABILITY	SIGNIFICANCE	FRAMEWORK
DESCRIPTION	BACKGROUND	USERS	USES	ADVANTAGES	METHODOLOGY
FEATURES	LITERATURE		APPLICATIONS	DISADVANTAGES	TECHNIQUES
ABOUT	PREVIOUS WORK		DOMAINS	PROBLEMS	STEPS
TYPES /CATEGORIES	TIMELINE		REFERENCES HELP	FEATURES	PROCEDURES AND PROCESS
	EVOLUTION		DOCUMENTATION	BENCH MARK / COMPARISON	IMPLEMENTATION
			VIDEO TUTORIALS		TOOLS
			TECHNOLOGY		VIDEOS

Table.1 KEYWORDS OF 5WH (TOPIC)

WHAT	WHEN	WHO	WHERE	WHY / WHY NOT	HOW TO
ABOUT	ACHIEVEMENTS	RELIGION	BORN	LEFT	DISCOVERED
ROLES		BIRTH NAME	LIVED	DEFAMED	DREAMS GOT REALITY
		PROFESSION	TRAVELED	IGNORED	PHONE NUMBER
		RELATIONSHIP	DIED	BECOME FAMOUS	
		FAMILY			

Table.2 KEYWORDS OF 5WH (PERSON)

WHAT	WHEN	WHO	WHERE	WHY / WHY NOT	HOW TO
DESCRIPTION	TIMELINE	DISCOVERED	ESTABLISHED	FAMOUS	TO REACH
CATEGORY		CONQUERED	DO INVENTED		TO CALL
LANDMARKS		NEIGHBOURS			
POPULARITY					

Table.3 KEYWORDS OF 5WH (PLACE)

WHAT	WHEN	WHO	WHERE	WHY / WHY NOT	HOW TO
PURPOSE	FOUNDED	FOUNDED	LOCATED	VISION	SUCCESS STORY
VISION	ESTABLISHED	CREATED		CLOSED	FAILURE STORY
MISSION		TEAM		BANKCROUPTED	
REVENUE		STAKEHOLDERS			

Table.4 KEYWORDS OF 5WH (ORGANIZATION)

4. ARCHITECTURE AND FRAMEWORK

Instant Answers search engine is built on the pipeline architecture based on six layers which includes Object Identification Layer: responsible of identifying the user input, its auto complete and correction, Data Sources Layer: responsible for identifying multiple sources for identified input, Data Tags Association Layer: responsible for assigning attributes, and tags to each 5W's & H keywords, Data Fetching Layer: responsible for fetching data defined in previous two layers, Syntactic Layer: responsible for phrasing and paragraphing based on data acquired from previous layer also include data formatting which is responsible for placing statements in predefined containers of 5W's & H on page, and Aesthetics Layer: responsible for formatting text, images, videos, and info boxes. Framework is illustrated in Figure 1.

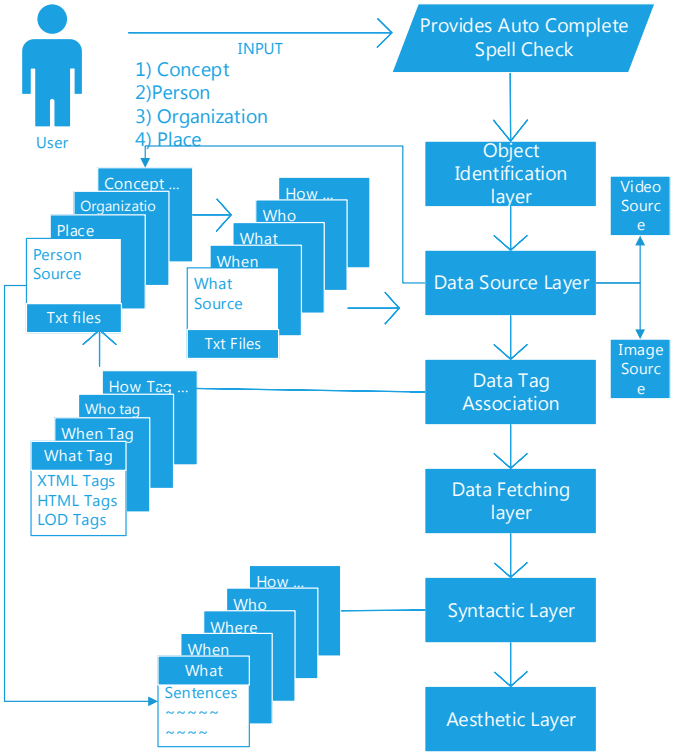


Figure 1: Framework of Instant Answers

User can input concept, person, organization, place or any phrase that asks question regarding any area mentioned. If user input is incomplete or requires spellcheck, two facilities will activate.

Autocomplete Facility: Autocomplete provides suggestions to the user for completing their input. It is consisting of application program interfaces of Google [23] and Wikipedia [25] which provides suggestion either directly from their database or by fetching it from the internet and by doing it provide convenience in the user input.

Spell check Facility: Spell check helps the system in spell checking user's input in both the cases when the input is given right or wrong. It also has ability to return the right version of wrong spelled words. Spell check consists of external spell check libraries like aspNETSPELL [28] along with search engine spell correction application programming interface.

Object identification layer: first layer takes the input from spell check, which returns corrected version of the input and it tries to understand it semantically using natural language algorithms (NLP) and finally identify the entity as person, place, organization or concept/topic. It consists of Stanford NLP NER [26] libraries. Figure 2 shows if user input Imran Khan identifies as a person, it association with Pakistan and its current status as Chairman.

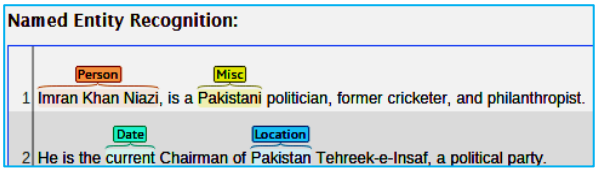


Figure 2. Response of Stanford NLP NER

Whereas Alchemy NER [29] application programming interface (API) which gives output in the form of standard JSON (JavaScript Object Notation). Figure. 3 shows that Imran Khan identifies as Person.

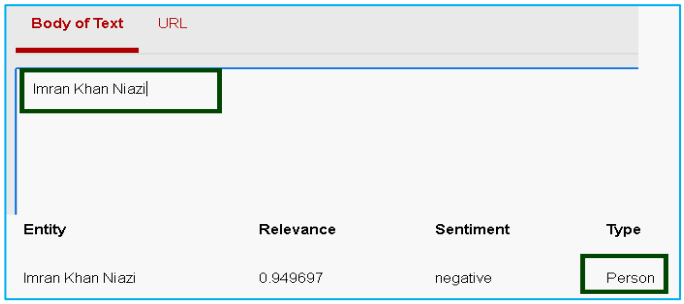


Figure 3. Response of Alchemy NER

And If LOD (Linked Opened Data Sources) [30] used F which is queried online using SPARQL (SPARQL Protocol and RDF Query Language) queries. Figure 4. Shows the "Imran Khan" as person just below his name.

Figure 4. Response of SPARQL

Object identification layer's main purpose is to recognize the given input either as place, person, organization or topic. So, that next layer can fetch right data from their respective sources.

- **Data Sources identification layer**: This layer totally dedicated for the identification of predefined data sources or their identification at the runtime; it takes input from the object identification layer and according to the identified entity it makes their decision of choosing particular data source. Selection of the data sources layer is designed in very sophisticated manner to make the whole process accurate and fast.

In order to identify data sources, it uses multiple options to make it happen. It uses files which already defining the few of the generalized sources illustrated in Figure 6. It makes the use of search engine API (Application Programming Interface) like Bing to get sources at runtime from unstructured data sources. As Instant answers search engine is multisource search engine, it configures separate configuration files for each area (person, place, organization and concept) and each file contains list of web pages and directories similar to shown in figure 5.

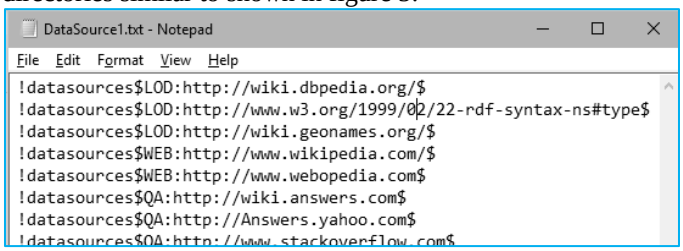


Figure 5. Generalized Data Sources

It also has video aggregator which again uses configuration file for its further operations as illustrated in Figure 6.

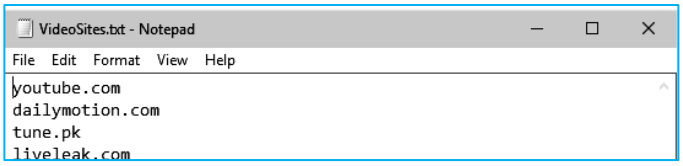


Figure 6. Data Sources for Video sites

It also uses LOD (linked open data sources) [30] to identify the right source for particular entity which is hub of knowledge bases (structured data sources). Figure 7 shows linked open data sources hub of knowledge bases.

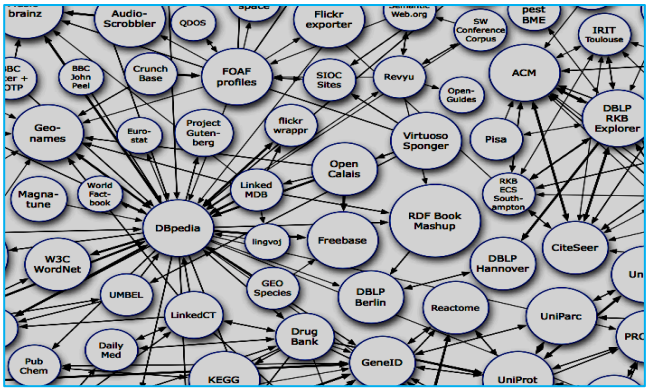


Figure 7. Open LODs network

Data tags association layer: It plays vital role in the creation of 5W's (What, When, Where, Who, Why) and 1H (How). It uses multiple techniques for preserving the tags which come handy for the next layer. It uses files with distinct format for the structured data sources like Dbpedia, Yago, Freebase and WebQuestions [31]. Figure 8 illustrates that the tags for retrieval of person's Who from dBpedia.

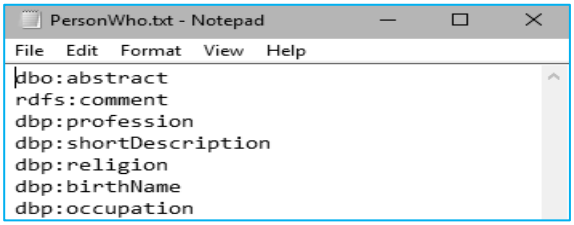


Figure 8. Tags for Person who of 5W's & H

Similarly, JSON tags are used for retrieving data from the know sources and these sources provides semi-structured data as representing in Figure 9.

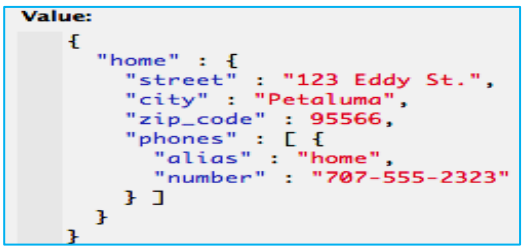


Figure 9. Tags of Place where of 5WH

Web pages, PPT files, pdf files, simple text files and even text inside images provides unstructured data. Figure 10 Shows unstructured data on MSDN [32] webpage

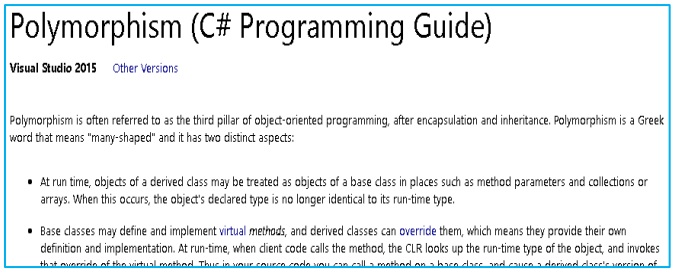


Figure 10. Other format of extracting text from webpages

Data fetching layer: Exactly after the selection of the data source data fetching layer starts doing its work and started fetching data from the various source defined in data source layer along with tags defined in data tags association layer. It not only solicits data in text format but also in the form of videos, images and maps. It uses SPARQL [33] queries to fetches data from the Linked Open Data Sources.

In context of LOD's, Instant Answers uses separate six configuration files for each W and H. which contains attributes of each 5W's & H as showing in Figure 13 for Example for 'Who' of the Person attributes can be (religion, name, birth name, profession etc.) based on Table 2 Who column (RELIGION, BIRTH NAME, PROFESSION, RELATIONSHIP, FAMILY) Search six separate files are maintained for each person, place, organization and concept.

It using XML [34] and JSON parsing [35, 36] from different source pages where APIs are used which shows JSON client (Instant Answers) and client in Figure 11.

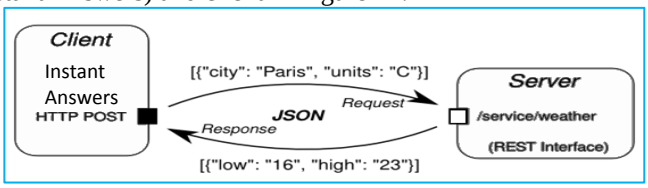


Figure 11. Jason Parsing

It also uses scraping [37] and html parsing [38] kind of thing where usually we have no other options available. Figure 12 shows HTML tags of Wikipedia page of Imran Khan.

Figure 12. Scraping from HTML tags

Figure 13 shows images which are accumulated by Instant Answers from the web sources.

Figure 13 Image extraction from their sources

Figure 14 shows one of the video results of Instant Answers Video Aggregator with the help of configuration files earlier mentioned in Figure 7.

Figure 14 Video extractions from their sources

Syntactic layer: It helps in creating sentences from the tags and retrieval of data from the appropriate sources. Sentence creation requires special format of the sentences which predefined or either they are being manufactured at the run time. For that purpose, Instant Answers created spate files of each 5WH of each area (person, place, organization and concept). That means there are six separate file covering each area ant total of twenty-four configuration file. Creation of sentences can be done in the same way but when it comes fully unstructured data it remains a big challenge.

Figure 15 illustrates predefined sentences for person's Who in special format, where $ is used for tag parsing.

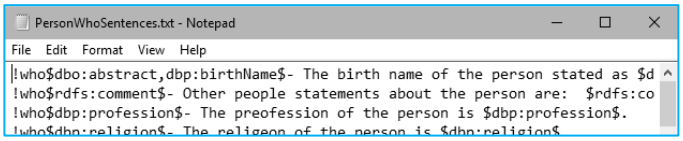


Figure 15. statement maker with tags parser

The Figure 16 exemplify the results for the persons who Sentences generated by Instant Answers.

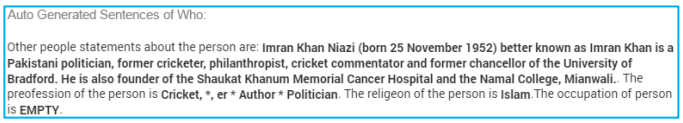


Figure 16. formatted statements after tag parsing

Data formatting: Data formatting works entirely on the distribution of the data into their right place using five W's and one H model. Existing format contains five different containers, each representing 5W's and 1H. This algorithm is designed and implemented using design patterns so that it can become adaptable and be able to add new components in the future. Its architecture involves async /parallel call to make the data extraction process faster. Finally Figure 18 illustrates the Who, Where and When of person and so on.

Aesthetic layer: It provides a very vital role in giving a good finish to this whole big process. It actually formats the text in which format you required like in sentences from values and in others as well. This layers also handles the formatting of the videos, images, maps as well using their erudite talent. Display formats are given below:

Text – info box, auto generated sentences, short and long paragraphs

Image – single image, multiple images in rows and columns

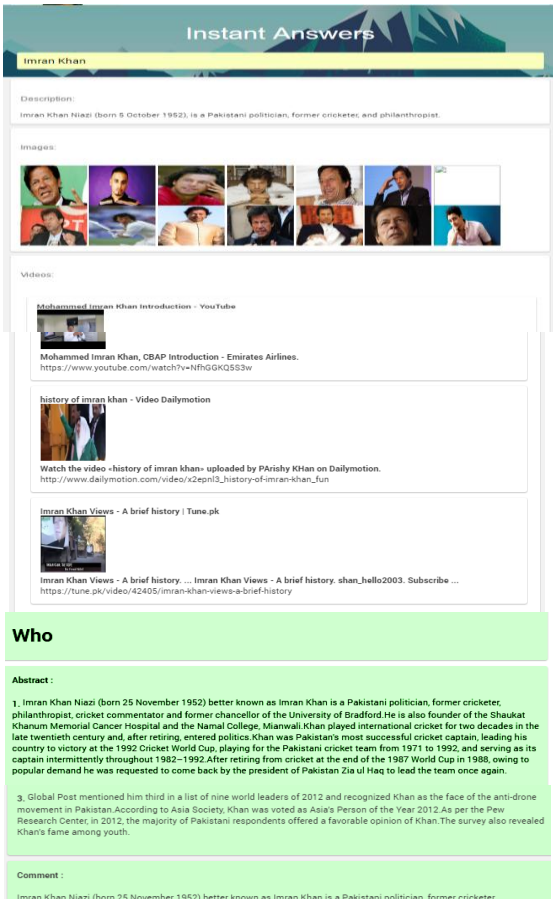
Video – thumbnail with source link, description, title

Info Boxes – it takes image if available and variable and values to form info box

It provides final touch to this whole process by creating W's of card which are coded in CSS (Cascade Style Sheets). They are able to generate dynamically as data of particular search increases and decreases.

5. INSTANT ANSWERS MAIN INTERFACE

Final version of Instant Answers is shown in figure 17, that represents the information in the form of 5W's & H.



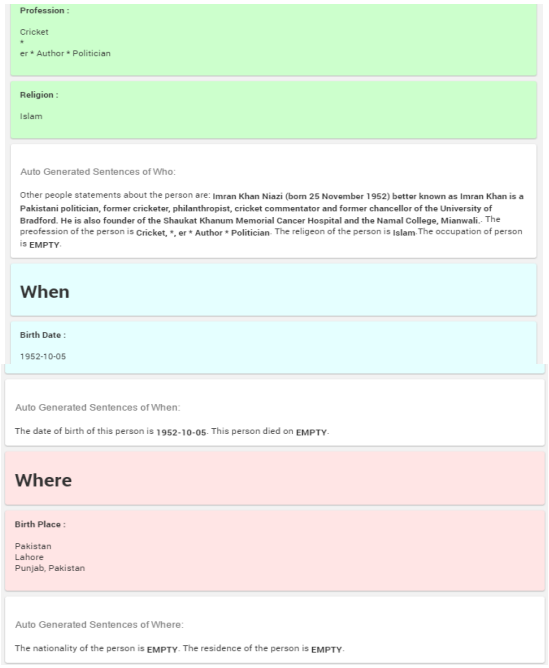


Figure 17 Final view of Instant answers

6. CONCLUSION

Research is producing search engine (Instant Answers) with solution to the problems of multiple sources, time consumption, compilation effort and incompleteness of knowledge seekers by helping them in their research works. Research also introducing first ever architecture to automate 5W's & H model. Instant Answers uses multilayer architecture with configuration files while making it more flexible and extendable. Configuration files in data source layer will enable tool to add new sources of websites, LODs, and directories. Whereas tags configuration file will make each W&H more detailed and evolutionary in nature. However, syntax parser configuration file will enhance it to make more concise and knowledgeable paragraphs. On the other hand, aesthetic layer will further equip tool to modern layouts and presentations and if there is need to add new area such as products, we just have to add separate configuration file in three layers without modifying code. Instant Answers is more powerful than the existing question answer tools which provides answer of single question, although Instant Answers is capable to answer series of questions and also covers complete 5W's & H model for single input.

REFERENCES

[1] http://start.csail.mit.edu/index.php, *"START, the world's first Web-based question answering system"* 1993-2013

[2] Borchardt, Gary C. *"Understanding causal descriptions of physical systems."* AAAI. 1992.

[3] Moldovan, Dan I., Christine Clark, and Moldovan Bowden. *"Lymba's PowerAnswer 4 in TREC 2007."* TREC. Vol. 1. No. 5.3. 2007.

[4] http://dbpedia.org, Feburary 2014, *Introduced by [Lehmann, Isele, Jakob, Jentzsch, Kontokostas, Mendes, Hellmann, Morsey, Kleef, Auer, and Bizer,]*.

[5] http://wiki.dbpedia.org/about , January 2016.

[6] http://www.freebase.com , January 2016, *Introduced by [Bollacker, Evans, Paritosh, Sturge, and Taylor*, 2008].

[7] Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. "Semantics and Complexity of SPARQL." *International semantic web conference. Springer Berlin Heidelberg*, 2006.

[8] Unger, Christina, and Philipp Cimiano. "Pythia: Compositional meaning construction for ontology-based question answering on the semantic web." *International Conference on Application of Natural Language to Information Systems. Springer Berlin Heidelberg*, 2011.

[9] Unger, Christina, et al. "Template-based question answering over RDF data." *Proceedings of the 21st international conference on World Wide Web. ACM*, 2012.

[10] Lopez, Vanessa, et al. "Evaluating question answering over linked data." *Web Semantics: Science, Services and Agents on the World Wide Web* 21 (2013): 3-13.

[11] Yahya, Mohamed, et al. "Natural language questions for the web of data." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics*, 2012.

[12] Xu, Kun, et al. "Answering natural language questions via phrasal semantic parsing." *Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg*, 2014. 333-344.

[13] Wolfram, Stephen. Cellular automata and complexity: collected papers. Vol. 1. Reading: Addison-Wesley, 1994.

[14] http://www.wolframalpha.com/faqs.html , Frequently Asked Questions,

[15] Ugander, Johan, et al. "*The anatomy of the facebook social graph." arXiv preprint arXiv*:1111.4503 (2011).

[16] Wical, Kelly. "Concept knowledge base search and retrieval system." *U.S. Patent No. 6,038,560*. 14 Mar. 2000.

[17] Eder, Jeffrey Scott. "Knowledge graph based search system." *U.S. Patent Application No. 13/404,109.*

[18] Ferrucci, David, et al. "Building Watson: An overview of the DeepQA project." *AI magazine 31.3 (*2010): 59-79.

[19] Ferrucci, David A. "Introduction to "this is watson*"." IBM Journal of Research and Development* 56.3.4 (2012): 1-1.

[20] Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, 2005.

[21] Myllymaki, Jussi. "Effective web data extraction with standard XML technologies." *Computer Networks* 39.5 (2002): 635-644.

[22] Xu, Kun, et al. "What Is the Longest River in the USA? Semantic Parsing for Aggregation Questions." *AAAI*. 2015.

[23] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." *Computer networks 56.18 (2012): 3825-3833.*

[24] Graepel, Thore, et al. "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine." *Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.*

[25] Hahn, Rasmus, et al. "Faceted wikipedia search." *International Conference on Business Information Systems. Springer Berlin Heidelberg, 2010.*

[26] by Webopedia, Cloud Computing. "*Online Resource."*

[27] Bishop, Christopher M. "Pattern recognition." Machine Learning 128 (2006).

[28] Team, Wrox Author, et al. Professional ASP. net web services. *Wrox Press Ltd., 2001*.

[29] Johnston, Melissa P. "instaGrok: a (re) search engine for learning." *Knowledge Quest 43.4 (2015): 78.*

[30] Manning, Christopher D., et al. "The Stanford CoreNLP Natural Language Processing Toolkit." *ACL (System Demonstrations). 2014*.

[31] Jain, Prateek, et al. "Linked Data Is Merely More Data." *AAAI Spring Symposium: linked data meets artificial intelligence. Vol. 11. 2010.*

[32] Kirtland, Mary. "A platform for web services." Microsoft Developer Network (2001).

[33] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data-the story so far." *Semantic Services, Interoperability and Web Applications: Emerging Concepts* (2009): 205-227.

[34] Seaborne, Andy, et al. "SPARQL/Update: A language for updating RDF graphs." *W3c member submission 15* (2008).

[35] Bray, Tim, and Jean Paoli. "C. Sperberg-McQueen," Extensible Markup Language (XML)." *World Wide Web Consortium Recommendation REC-xml-19980210 (1998).*

[36] Parr, Terence. The definitive ANTLR 4 ,*Pragmatic Bookshelf, 2013*.

[37] Aho, Alfred V., Ravi Sethi, and Jeffrey D. Ullman. Compilers, Principles, *Techniques. Addison wesley, 1986.*

[38] Carmel, David, et al. "Juru at TREC 10-Experiments with Index Pruning*." TREC. 2001.*