A Systematic Hybrid Approach in building an Effective Answer Retrieval System

1Kiran B. Malagi2Jaya S. Pujar,3 Karuna C. Gull, 4Akshata B. Angadi

1,2,3,4 K.L.E. Institute of Technology, Department of Computer Science and Engineering, Hubli, India

malagikiran@gmail.com,2 jayapooja@rediffmail.com 3 karuna7674@gmail.com 4 akshata_angadi@yahoo.co.in

ABSTRACT
Google is one of the top ranked search engine that give results for a query, asked by user in a form of list of URL’s / websites. It uses Page Rank Algorithm. Finding answers from those websites is time consuming as n pages and n websites list are displayed based on the factors it considers. The question is “Can we trust the content displayed by any website just by its number of hits/occurrences alone??” No. Considering this as our initial question we thought of throwing an idea on this and proposed a framework that help user to get the answer quickly and trust it without having a second thought. The paper present an algorithm of believing the website based on ‘n’ factors. Survey in a paper details the work done by scholars. A novel approach in a paper makes people to think in different way and implement a system that gives better experience to users. What a user needs? A quick answer and a trustful result. Concentrating on these two keywords, paper framed a model that helps in relevant access of the content. It’s a novel technique of retrieval system that interacts with user to rephrase a question in better way if system is unable to get it. System needs to be trained before the query is posed. The approach combines the concept of framing the question and finding trustworthiness, this hybrid approach builds an effective and healthy move towards the development of search engines.

Keywords: IR (Information Retrieval), QA (Question Answering), Trustworthiness(Tw), URL(Uniform resource Locator),WWW(world wide web)

1. INTRODUCTION
With the explosion of WWW, the people across the world expect that for every question, they can find an answer through the internet. No doubt large amount of data is stored in the databases, web pages and the documents in many servers. The search engines search for the answer in the resources mentioned. But the question is: Is a user satisfied with the answer returned by the search engine when posed a question? No, as he will be getting the links to the answer but not the exact answer. Again one has to traverse through the links to find answers of their need. It’s time consuming. Lots of work is being carried out for factoid questions where only one word answer is expected. But same procedure can not applied for descriptive answer.

To overcome this we are proposing a model where we try to extract the exact answer from the list of links provided by the search engines based on ‘n’ factors. We have introduced an algorithm for trust worthiness and confidence of the websites by doing a survey. One cannot say that a web site/page is trustworthy just by knowing the number of hits or visits made. Along with that we may consider the age, popularity factor etc.

Many IR engines will be trained with these data and accordingly the information is fetched. The resulting information is stored back in the local database for future reference in question-answer pair format along with a rank assigned to it so that the next time when the same type of question is asked we can get the answer from the database itself just like working of cache. To extract an answer, basically we should understand the question being asked. A user asks question in natural language which may be ambiguous/unclear. By applying the NLP techniques we formulate question which fits to a particular predetermined category. The system is trained to answer to a question based on the knowledge given. That is, the user ask a question , if the question is not clear to the system the system lists question menu to the user and help to form correct sentence , this helps user to learn to question and the system to search properly.

2. LITERATURE SURVEY
Research in Question Answering (QA) Technique helps user to get a relevant result for their questions. But the data in the web is often incomplete; sometimes contain duplicates with different vocabularies. Datasets may also contain conflicting information for same question.

Dragomir R. Radev et al., (2001) [1], proposes a probabilistic algorithm called QASM (Question Answering using Statistical Models) that learns the best query paraphrase of a natural language question. System validate approach for both local and web search engines using questions from the TREC evaluation. QASM has been implemented in a natural language question answering system with two different backends - one based on a local corpus and another - using an external search engine which retrieves answers from the World-Wide Web.

Fabio Rinaldi et al.,(2003), [2] Exploiting Paraphrases in a Question Answering System, the system implements a simple and efficient logic representation of questions and answers that maps paraphrases to the same underlying semantic representation.

Dell ZhangA et al., (2003), [3] Web-based Question Answering System, describes a Web-based question answering system LAyMP, which is publicly accessible. A particular characteristic of this system is that it only takes advantage of the snippets in the search results returned by a search engine like Google.
Silviu Cucerzan et al., (2005), [4] Factoid Question Answering over Unstructured and Structured Web Content system was developed from scratch for this year’s TREC evaluation and was initially tested on the TREC 2004 factoid QA evaluation. For unstructured content, authors used a web-based system with novel features such as web snippet, pattern matching and generic answer type matching using web counts. And also experimented with a new, complementary question answering approach that uses information from the millions of tables and lists that abound on the web.

Luiz Augusto Sangoi Pizzato et al., (2005).[5] Extracting Exact Answers using a Meta Question Answering System, this work focuses a question answering tool that uses multiple Web search engines and Web question answering systems to retrieve snippets of text that may contain an exact answer for a natural language question. A meta-QA system that combines the results of different Web search/QA systems in order to provide exact answers for natural language questions.

Radu Soricut et al.,(2006), [6]Automatic Question Answering Using the Web: Beyond the Factoid questions, uses a Main approach where QA assumes no restrictions on the type of questions that are handled, and no assumption that the answers to be provided are factoids. This paper attempts to explore the portion related to answering FAQ-like questions, without restricting the domain or type of the questions to be handled, or restricting the type of answers to be provided.

Håkan Sundblad (2007), [7] Question Classification in Question Answering Systems, focuses on five different machine learning algorithms. The algorithms are k nearest neighbors, naïve bayes, decision tree learning, sparse network of winnows, and support vector machines. These algorithms have been applied to two different corpora, one of which have been used extensively in their work and has been constructed for a specific agenda.

Nguyen Le Nguyen (2010), [8]Applying Semantic Analysis to find Similar Questions in Community Question Answering Systems is done by a semantic role labeling system by leveraging on grammatical relations extracted from a syntactic parser and combining it with a machine learning method to annotate the semantic information in the questions. Then author utilizes the similarity scores by using semantic matching to choose the similar questions.

3. METHODOLOGY

The proposed system is in Fig.1. Procedure of proposed system (in steps):

Step 1: Query Input [Search]
Step 2: Searches results from Local Database if present and goto Step 6 else continue
Step 3: Link to Google Search Engine
Step 4: Extract data from Websites given by Google
Step 5: Calculate trustworthiness of websites based on factors specified
Step 6: Retrieve the Answers of top ranked sites w.r.t trustworthiness Algorithm
Step 7: Display top ranked answer to the user and place the answer in the Local Database.

The main components in system are:
1) Framing the Question: When a user pose a question it will be in the natural language but the way of questioning will be completely different from the way they want to present/ask. So framing the question plays an important role in our system. The Local Database is used to store the pattern of questions i.e. for a query ‘n’ number of different patterns. The drop up menu is available to make a match to a query given by user. Drop up menu is the lists of question flavors already present in local database and is displayed as per user question placed in search text box. User is provided facility to pick the pattern and search the required information. The importance is given for framing of question because the retrieval of answer is dependent on the user query. So the system is trained to work according the question formed. Hence when a query is posed the question is selected and is processed further.

2) Categorization: Answers are retrieved based on the category it falls. The user has to choose the category before the question is popped. The different categories considered here are: Business/Management (C1), Technology (C2), Medical (C3). These are the main categories which can be further sub divided into sub categories. If user knows the sub category of the query he/she can select that cluster else choose the main cluster (C1/C2/C3). As per the selected cluster the retrieval process carry on.

Note: If the main cluster is selected the searching process is carried in voluminous data as shown in Fig.2. If the sub cluster is selected the data searching will be restricted within the sub clusters range. The Question generated (includes question and category) from the previous phase will be combined with category selected/given for further processing i.e. Q+C is an input to the next phase

3) Search Engine: A question, consisting of question and category is given as an input to the search engine i.e. QAS (Question Answering System). This proposed search engine searches the answer in the Local Database, if answer is found in database it checks for its updates. Updates are checked w.r.t date and time of the entry of answers in database with current date.
Figure 1: Proposed Architecture
If it is greater than the threshold value set by user then it executes trustworthiness algorithm [9] and updates its data for the query entered. That updated answer will be provided to the user. If not, the engine gets linked up with the Google behind the scene and gets ‘n’ result sites, which are processed in the next phase to get the top most result of worthy websites.

Google Search Engine: When the query gets linked to Google, the sites given by it cannot be trusted as the answers will be in ‘n’ numbers and cannot believe all the sites’ answers or read all sites (time consuming). Finding the trustworthiness of the sites come into picture.

Trustworthy site: The Google uses a Page Ranking Algorithm to get the results. But these days there are n websites giving answers, which to believe and not!!! Moreover its time consuming to search/read answers in all sites and user itself selecting/thinking about what’s right. So here we propose/frame a model that searches automatically and displays the answers for the query. Reading from all sites is reduced. By considering the following definitions we begin to find the trustworthy sites from multiple answering sites [10].

The Algorithm is as follows:

```
Variables:
Storages / Database: (url, Tr, T, flag)
Integer I;
Float Or;
Float array T[n];
String Temp;
String Array Tw[n], url[n], Tr[n];
//Tw[i]=Trustworthiness of url[i];
//url[i]=Uniform Resource Locator;
//Tr[i]=Trustiness of url[i];
//T[i]=Time of url[i];
For each url[i] from n number of URL’s
Begin
  if url[i] is in a database/storages and is flag set to 0 & its
  T[i] < Threshold(set by user)
  begin
    Tw[i]=Tr[i] picked from db
    Sort the urls based on Tr[i]
    End
  else
    if url[i] is in db and is flag is set to 1
    begin
      Calculate value for Tw[i] by retrieving Origin
      information and assign it to Tr
      Tw[i]=Tr[i]+Or
      End
    else
      Calculate value for Tw[i] by retrieving Origin
      information and assign it to Tr
      Tw[i]=Or
      End
  Tw[i]=Factconf(Tw[i])               //Call function Factconf by passing
  Tw[i] which returns new value for Tw[i]
  Tr[i]=Tw[i]                         //Assign new Tw[i] into Tr[i]
  Store new calculated values of Tw[i] in db related to url[i]
Call Outlink(Out)
End
```

```
//Calculation of Fact confidence

Function Factconf(Tw[i])
Variables:
Float Fc1, Fc2, Fc3, Sum, S1,S2;
String Temp;
File Out;
begin
  Fc1 = auth(url[i])  //Retrieval of Auth information of url[i]
  Fc2 = age(url[i])  //Retrieval of Age information of url[i]
  Fc3 = pop(url[i])  //Retrieval of Pop information of url[i]
  Out=outlink(url[i]) //Retrieval of outlinks of url[i]
  Repeat until website url[i] becomes stable
  Begin
    Temp=Tw[i]
    Sum=\sum_{i=1}^{n}(Fc[i])
    If (sum<1)
      S1= Find logarithm of sum and Tw[i] which
      results in negative value
      Else
        S2 = Find logarithm of sum and Tw[i] which
        results in positive value
        Tw[i] = Take summation of exponent of each S1 and S2
        Diff = Temp – Tw[i]
      End
    End
    Return Tw[i]
End
```

```
//Finding the belongingness of the outlinks in the database

Function Outlink(Out)
Begin
Variables:
Integer n;
String ourl;
//ourl=url picked from Out;
//n= Number of factors considered;
For each ourl in Out of url[i]
begin
  If(ourl in database S)
    Break;
  else
    begin
      Tr[i]=Tw[i]/n;
      Flag=1
      Store information of ourl in db S
    End
End
```

Definition 1: (Confidence of facts) The confidence of a fact f is the probability of f being correct, according to the best of our knowledge.

Definition 2:(Trustworthiness of web sites) The trustworthiness of a web site w is the expected confidence of the facts provided by w.

- Assumption1: This true fact appears to be the same or similar on different web sites.
- Assumption2: The false facts about a certain object are much less consistent than the true facts.
- Assumption3: A web site that provides mostly true facts for many objects will likely provide true facts for other objects.

4. EXPECTED OUTPUT

The output of the QAS is the answer given by the user by following the steps mentioned. The Fig.3.a) shows the page to enter the question and select the category. The given query is microprocessor which is grammatically
incorrect word and is incomplete question. User needs to select the category (Fig. 3b) and type question and submit it. A query like “microprocessor” with category like “Technical” is entered, the result of which is either picked from local database of system designed or from trustworthy site and displayed on the screen. The list of questions can be seen from which the user can select the clear question as per his requirement and when selected the answer of the query is displayed to user as well as last modified date and the source from where it has been picked shown in Fig. 4.

Figure 3. a) Selection of Category  b) Selection of Questions from the menu given by system

Figure 4. Final Answer given by system
5. CONCLUSION

Today there is a need of good crawling technology to gather the documents of web. The indexing system should process thousands of gigabytes of data efficiently. Queries need to be handled quickly at rate of hundreds to thousands per second and should be trustful. These tasks are becoming increasingly difficult as the Web grows.

Searching data in huge web should be reduced this can be achieved when system data is divided into clusters or is categorized and system should interact with user in phrasing the right question. Considering this, model is designed and the implementation of the system will be shown in further study. Answers are displayed as considering the confidence and trust score of the sites. The trustworthy websites can be known better when ‘n’ queries are executed on the system i.e. we can state the website is trustworthy if for ‘n’ query, particular site is having top score always eg: if wiki is getting the top rate for all queries asked by user that the site has a high score (factors need to be considered). The proposed system reduces the time and area of search. The local database acts as a cache.

REFERENCES


[3] Silviu Cucerzan, Eugene Agichtei, “Factoid Question Answering over Unstructured and
Structured Web Content system”, Microsoft Research, 2005.


AUTHOR PROFILES

Kiran B. Malagi has received the B.E degree in Computer Science and Engineering from Visvesvaraya Technological University, India in the year 2005 and the M.Tech degree in Computer science and Engineering from the Visvesvaraya Technological University, India in the year 2010. He has been working in the area of data mining and social networking since 2010. He worked as Lecturer and head about 6 years .He is currently working as an Assistant Professor in K.L.E.I.T, Hubli, India since 2010.

Jaya S. Pajar has received the B.Tech. degree in Computer Science and Engineering from Acharya Nagarjuna University in the year of 2012 and pursuing M.Tech degree in Computer science and Engineering at Jawaharlal Nehru Technological University Hyderabad. He has been working in the area of Political Science, Data Mining and Networking field.

Akshata B. Angadi has received the B.E. degree in Computer Science & Engineering from TCE affiliated to Visvesvaraya Technological University, India in the year 2011. She has published 4 National and 2 International papers in Conference Proceedings. She has also attended many of the workshops. She has been working in the area of data mining and social networking since 2011. She has published 6 papers in International Journals. She has also attended many workshops. She is working as Lecturer in Computer Science Department in K.L.E.I.T, India.

Karuna C. Gull has received the B.E. degree in Electronics and Communication from Karnataka University, India in the year 1996 and the M.Tech degree in Computer science and Engineering from the Visvesvaraya Technological University, India in the year 2008. She has been working in the area of data mining and social networking since 2009. She has published 1 International paper on Data Mining (2009), 2 International papers on Image Processing (2011). She has also published 6 National and 4 International papers in Conference Proceedings. She has also attended many of the workshops and conferences held in different places on High Impact Teaching Skills, Embedded System Using Microcontroller, Information Storage and Management (ISM), Data Mining, and many more. She worked as a Lecturer and Senior Lecturer for about 10 years. She is currently working as an Assistant Professor in Computer Science Department in K.L.E.I.T, Hubli India since 2011.