ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A Survey on Information Extraction in Web Searches Using Web Services

Maind Neelam R., Sunita Nandgave

Department of Computer Engineering, G.H.Raisoni College of Engineering and Management, wagholi, India

ABSTRACT: Now- a- day efficient searching is having the primary concern in every transaction. Most of the search engine will work only on server side i.e. if we want to search a particular keyword, then the web crawler will search only at the server side & returns the result .So for every time we have to search in the server thereby increasing the processing time. Many existing crawler will search the data from server but doesn't returns any source at client side. Existing search engine takes more time to answer the query because the total no. of transaction is more. For example if an user wants to search a song of particular singer the current web crawler systems will give the proper result but if the user wants the singer of some specific song , then the Web service cannot be called, even though the underlying database might have the desired piece of information. All the existing extraction systems are based on textual query therefore we cannot search the required results from any visual inputs. The limitation of any normal web application is that, they cannot keep track of dynamic data because of stateless protocol therefore our proposed system will prove to efficient because it supports web services and many stateful protocols .So our system will search the textual query [4],[5] along with keeping track of visual queries[9]. Due to web services when we are searching any data its references are also stored in sub servers so whenever we search the same query at second time it will return quickly through sub servers rather than contacting to main server.

KEYWORDS: web crawler, textual query, visual query, web services etc.

I. INTRODUCTION TO EXISTING SYSTEM

Existing system considers conjunctive query plan over the views that is corresponding to the entered query is NP-hard in the size of the query. This postulation is impractical in our setting with Web services, where sources may overlie or go together with each other but are usually incomplete. When sources are incomplete, one aims to find maximal controlled rewritings of the original query, in order to provide the maximum number of results to create accessible functions to compute results, which often consume the total budget before any answer is returned. In existing system, for any visual content retrieval we have to extract both the audio & video features. The extraction could be done by the method such as

1. Temporal Features:

The temporal domain [9] is the native domain for audio signals. All temporal features Have in common that they are extracted directly from the raw audio signal, without any preceding transformation. Consequently, the computational complexity of temporal features tends to be low. To extract the temporal features we partition the group of temporal features into three groups, depending on what the feature describes

1. zero crossings feature
2. Amplitude Base features
3. power base features

In zero crossing features, Zero crossings are the basic property of an audio signal that is often employed in audio classification. Zero crossings [9] allow for a rough estimation of dominant frequency and the spectral centroid it's having three different phases

- Zero Crossing Rate
- Linear Prediction Zero crossing Rate
- Zero crossing peak amplitude

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Where as in amplitude base features, many features are directly computed from the amplitude i.e. the pressure variation of a signal. Amplitude-based features [5],[9] are easy and fast to compute but limited in their articulateness. They represent the temporal envelope of the audio signal. It has two phases

1: MPEG-7 audio waveform (AW)
2: Amplitude descriptor (AD).

And in power based features, the energy of a signal is calculated as the square of the amplitude represented by the waveform. The power of a sound is the energy transmitted per unit time. Consequently, power is the mean-square of a signal. Many times the root of power (root-mean-square) is used for feature extraction. The generalized block diagram for feature extraction is given in figure 1.

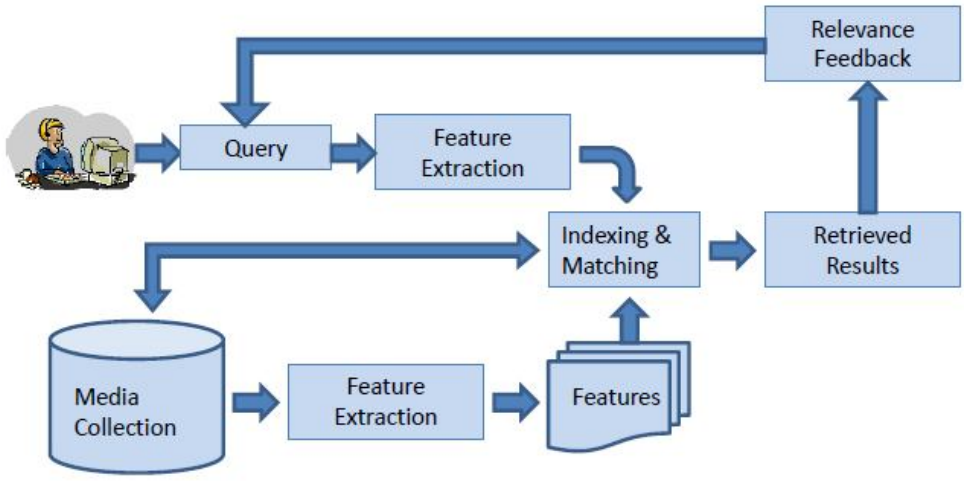

Figure1:- Block Schematic of Feature Extraction in Audio Visual Based Searching

In case of existing textual search engine, a web crawler [2], [3] is maintained to search the document. The queried keyword is searched against database documents by measuring the semantic similarities parameters as

A. about Semantic Similarities

Wikipedia is the world's largest collaboratively edited source of encyclopedic information, which provides important semantic information for us. So we can get external information about words from Wikipedia to examine semantic similarity between words. Firstly, we must decide which part in Wikipedia for a word is useful for us. For example, if we search word "car" in Wikipedia, we can get much information about "car", such as car's history, its production and its safety, and so on. But we can't use all of them for not all snippets are useful for us to analyze semantic similarity. Usually, Wikipedia return some top result for the word for which we search information in Wikipedia. These snippets use simply vocabulary to explain the word, or give simply definition or some description about the word [2]. We select these for investigative target to measure semantic resemblance between the words.

B. Preprocessing the snippets from Wikipedia

We can't use the snippets [4],[5] downloaded from Wikipedia directly because it may contain lot of semantic-unrelated words and word in different form will bring in Negative impact in our calculating. Therefore, we must deal with the snippets by following steps:

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

1) Removing stop words. Words like “a”, “of” and so on called stop words which are worthless for semantic analysis. So, before making any calculation we delete those stop words first.

2) Because we will do some statistical work on the snippets from Wikipedia, words in different form will bring in disadvantage influence. We can use algorithm like Stemmer algorithm3 gives us critical help to deal the text. We use the algorithm to deal with every word in the snippets from Web Documents.

C. About TF-IDF

The TF–IDF (term frequency–inverse document frequency) is a weight [5] often used in information retrieval and text mining. This weight is an arithmetical measure used to evaluate how important a word is to a document in a compilation or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the TF–IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. If we consider a document containing 100 words where in the word cow appears 3 times. The term frequency (TF) for cow is then 0.03 (3/100).

The TF-IDF weighting scheme is often used in the vector space model together with cosine similarity to determine the similarity between two documents. We will use TF-IDF and cosine similarity to analyze the text which is from Web Documents after preprocessing.
D. Calculate the Semantic Similarity by cosine similarity

In this section a method which integrates TF-IDF and cosine similarity is proposed in details to measure semantic similarity between words.

II. LITERATURE SURVEY

1. Web Services

The protocol used in web services to maintain the synchronous communication among server & sub servers is Simple Object Access Protocol (SOAP), WSDL or UDDI. SOAP is a stateful protocol used to keep track of server & sub server’s content. It means when we would like to add any data on server side it can be replicated on the sub server automatically with the help of web services. In existing systems mostly web services are used to keep track of backup of information so than in case of node failure we can recover the data from other node. But in our system we use web services to maintain the track of data & storing of server information onto sub server.

Service-oriented architecture (SOA) is an evolution of distributed computing based on the request/reply design paradigm for synchronous and asynchronous applications. An application's business logic or individual functions are modularized and presented as services for consumer/client applications. The service interface is self-governing of the execution. Application developers can build applications by composing one or more services without knowing the services' underlying implementations. For example, a service can be implemented either in .Net or J2EE, and the application consuming the service can be on a different platform or language.

Service-oriented architectures have the following key characteristics:
1. SOA services have self-describing interfaces in platform-independent XML credentials. Web Services Description Language (WSDL) is the standard used to describe the services.
2. SOA services commune with messages formally defined via XML Schema (also called XSD). Communication among consumers and providers or services typically happens in diverse environments, with little or no knowledge about the provider. Messages between services can be viewed as key business documents processed in an enterprise.
3. SOA services are maintained in the enterprise by a registry that acts as a directory listing. Applications can look up the services in the registry and invoke the service. Universal Description, Definition, and Integration (UDDI) are the standard used for service registry.
Following figure 2. Shows the typical service oriented architecture.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

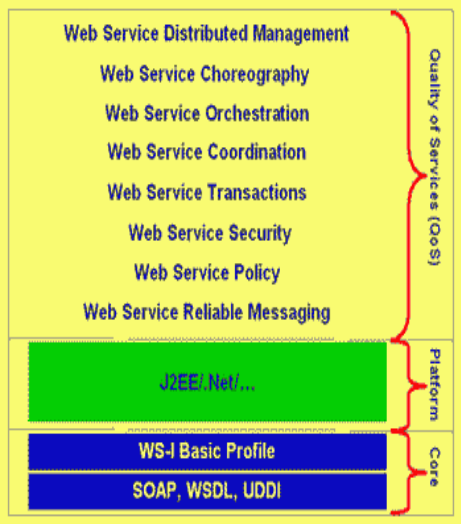

Figure 2. A typical SOA infrastructure

2. HTTP VS SOAP Protocol

Http stands for hypertext transfer protocol. I.e. HTTP protocol is responsible to transfer the request & response for any web based material such as documents, audio files, video files etc. Http can only be used for transferring the request or response but it cannot hold any data information. i.e. Http will transfer user's request to any website like google, gmail etc. but user can do their modification on their own account only but cannot reflect the changes done at server side on client side because HTTP is a stateless session protocol so it cannot hold the state of any network data whereas SOAP(Simple Object Access Protocol) is a stateful session protocol therefore we can hold the network data for updating the client side also. I.e. when a user adds any data at server side its automatic replications can be done at client sides. We can configure the SOAP or any stateful protocol using web services. All of these rules for messages are defined in a file called WSDL (Web Services Description Language). Shown in figure 3.

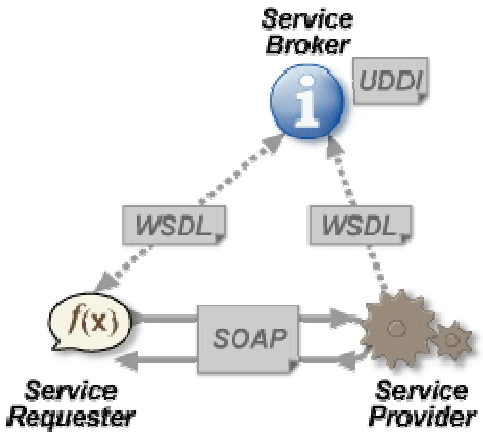

Figure 3. Web Services Communication

Web services architecture: the service provider sends a WSDL file to UDDI. The service requester contacts UDDI to find out who is the provider for the data it needs, and then it contacts the service provider using the SOAP protocol.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

The service provider validates the service request and sends structured data in an XML file, using the SOAP protocol. This XML file would be validated again by the service requester using an XSD file.

III. PROPOSED SYSTEM

In our system we are going to implement a system which can handle any type of user query along with reduction of transaction time thereby shifting the transaction load to local database instead of central server. In means that if a user enters any text queries [4], [5] it will be pre-processed to remove the stop words and after pre processing [2], [3] along with textual query search we also search for snippets. After searching on the basis of TF-IDF [3] we will find out the weight of searched keyword & ranked accordingly and if the query is a audio visual query we will extract the features on the basis of temporal features such as zero crossing, amplitude base and power base as explained above. The web services [7] are used in textual query as well as in visual query for communication between local data & global data i.e. if user enter any query, for first time it will searched in central server but during searching the data sub server will keep track of same data & make copy of same into sub server so, for second time the searched data will be quickly retrieved through local sub server. Along with this in case of audio visual query sub server & servers are maintaining index track [9] of whole data so when user can go through forward or reverse engineering also.

IV. CONCLUSION

In this paper, we had gone through the various existing information extraction techniques such as textual extraction technique, audio-visual data extraction technique but for every technique we have to gone through lots of transactions thereby results in delayed output but in our proposed system we had keep track of features of both the techniques along with reducing the total no. of transactions and keeping track of every index of textual or audio or visual data for faster output

REFERENCES

1. Nicoleta Preda, Fabian Suchanek, Wenjun Yuan, Gerhard Weikum, “SUSIE: Search Using Services and Information Extraction”, IEEE transactions on knowledge and data engineering year 2013
2. F. M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A Core of Semantic Knowledge,” in WWW, 2007.
3. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “DBpedia: A nucleus for a Web of Open Data,” Semantic Web, 2008.
4. C. Li and E. Y. Chang, “Query planning with limited source capabilities,” in ICDE, 2000.
5. C. Li, “Computing complete answers to queries in the presence of limited access patterns,” VLDB J., 2003.
6. S. Kambhampati, E. Lambrecht, U. Nambiar, Z. Nie, and G. Senthil, “Optimizing recursive information gathering plans in EMERAC,” J. Intell. Inf. Syst., 2004.
7. S. Ran, "A model for web services discovery withQoS," ACM SIGecom Exchanges, vol. 4, Spring 2003
8. S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In Proc. of the 8th Int. World Wide Web Conference (WWW8), May 1999
9. Dr. H. B. Kekre, Tanuja K. Sarode, “New Clustering Algorithm for Vector Quantization using Rotation of Error Vector ”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010