Personalized Information Seeking Powered by Mobile Ontology for Sightless Residents of under Developing States

Ahmad C. Bukhari, Mehtab Afzal, and Yong-Gi Kim

Abstract—**Nowadays internet is considered as one of the prevalent sources of information. Due to the unimpeded augmentation of the internet, billions of bytes of data are becoming the part of the internet on a daily basis. You can hardly find a topic about which you cannot get information from internet. But at the same time, it's extremely difficult to find the accurate personalized information from a large pool of data. Several conventional information fetching tools are currently working on the World Wide Web (WWW) but they are not powerful enough to search out user centric information. On the other hand, the sightless residents of underdeveloped countries are deprived from getting the required information from the internet due to unavailability of internet and special technical aid which is required for them to get along with any computer system. In the past no serious measures were taken to facilitate the special citizens and the research work which we found were mostly for academic purpose. So, there has been a strong impetus in the area of information engineering to develop and improve techniques to extract precise information from huge volumes of extraneous data. Ontology is one of the emerging solutions which can intelligently express and fetch the domain knowledge. The proposed architecture introduces an intelligent mechanism for the extraction of highly precise information for the visually impaired people through voice SMS. The architectural simplicity also does the proposed solution prominent from the available solutions.**

Index Terms—**Mobile ontology, information extraction, ontology based extraction for blind users.**

I. Introduction

With the advancement in web technologies, the heterogeneity in current internet is increasing rapidly. In every hour thousands of web pages are being added to the internet and millions of transactions take place using E-commerce applications. The relevant information extracted from this large pool of data is considered to be very difficult by using conventional technologies. Researchers introduced several techniques in the field of information engineering to find the precise information from the net. The current web is no more than the repository of millions of web pages and most of the existing information retrieval mechanisms are using keyword base conventional search engines. The conventional search engines are not able to find exact information according to user's requirements. The search engine can be divided into two categories: crawler based search engine and human powered directories. The crawler based search engine such as Google and Yahoo; automatic index the web pages which are being uploaded on the internet. Search engines collect and index the web pages after applying the optimization criteria which varies from search engine to search engine. When an internet user wants some information from the internet, he/she enters the searching string in the search engine's query field. The backend mechanism of the search engine converts the string into formalized query and forwarded the request to searching algorithm which further displays the crawled results. The human powered directories consist of several storage directories which are organized in hierarchical fashion. The website's authors usually have to add their website details manually in repositories. The website's detail includes topic terms, linked keywords and associated metadata. The keywords based search engines follow the match and show rule. By using the keywords based search engine, we usually get high recall of data but the precision level of fetched information is very low. This is the age of dynamic script which can update the website contents automatically. This poses a problem for static keyword storage repositories as it's hard to keep updated the repositories with ongoing changes in websites. Tim Burners' Lee presented his idea of intelligent web which is currently known as Semantic web. The semantic web layered architecture produces machine readable and human understandability. The basic theme behind the semantic web is to make the web intelligent enough so it can understand and conceive the meaning of resided data [1].The visually impaired people are part of civil society, these special people were ignored in the past as there were no serious actions taken to facilitate the blind users regarding relevant information extraction. The ontology based highly precise information extraction architecture for blind users provide a mechanism to extract precise information using the vocal command system for blinds.

II. Related Past Research

In the near past, several techniques were developed to extract relevant information from large scattered internet. Natural language processing (NLP), Part of speech tagging, Name entity relationships, and ontology based information extraction are some of the techniques. In this section, we elaborate the working functionality of these techniques.

A. Web Content Mining Techniques for Information Extraction

As the internet presents several useful information

Manuscript received August 20, 2012; revised October 10, 2012.

Ahmad C. Bukhari and Yong-Gi Kim are with the Artificial Intelligence Lab, Department of Computer Science, Gyeongsang National University Jinju, Kyungnam 660-701, Korea.

Mehtab Afzal is with the School of Information Science and Technology Southwest Jiaotong University, Chengdu, Sichuan, China.

repositories e.g.: product catalogue, current news, weather forecasting and event listing etc. So in a start, web content mining techniques were introduced to fetch relevant information from large, multidimensional web. Web content mining can extract, integrate the useful data, information and knowledge from a web page. There are two widely used methods in web content mining one is wrapper induction information extraction and second is automatic information extraction. In wrapper induction technique [2], researchers use machine learning techniques to make rules for information extraction. The Wrapper induction technique is specifically designed to extract data from any type of websites. The wrapper developer generates the logical rules for information extraction which can extract products, name and specific contents information. We have to manually tag the contents of the web pages at first then the machine learning techniques analyze the contents and define the extraction rules. But to write wrappers and keep them functional is very tough job as the resided information on the internet is so dynamic and changes continuously. With the change in website contents we have to update the wrapper rules against new data structure. At the same time the efficiency of wrapper induction technique is not so high and it faces the data error problem usually [2]. The automatic information extraction [3] from a large website is another technique to extract information from large website. In this technique a sample page is given for wrapper learning instead of writing a new wrapper. The issues of disjunction are very hard to handle in automatic information extraction and it is very difficult to generate attribute names for extended data [3].

B. *Information Extraction for Blind User*

A lot of information is available on the internet but visually impaired people cannot take benefit from these resources as visioned people can. Several softwares have been developed so far for blinds to search efficiently for relevant information for quick and timely decision making. But several of these were desktop oriented and most of them use a screen reading technique like BrookesTalk [4]. BrookesTalk is one of the tools which can generate a summary of the page for the blind user vocally. It uses different sounds to distinguish between the sections of any webpage, so the blind user can understand the architecture of the webpage and with the help of this tool. By using this tool user can easily navigate between the sections of web pages and can extract desired information. Its ability to use different sounds for different parts of the web page may irritate the users. Its design is based on the standard tool for blind pwWebSpeak (ä). Another approach is introduced in [5] which claim to facilitate the blind users in website searching and in content navigation. The areas mostly focused while developing a solution for blind users are: providing guidance for blind users; empowering blind users; and reducing cognitive load. The result of this research appeared in the form of prototype named as NavAccess. It provides a pleasant, effective and efficient environment for blind users. The server searches all the desired pages of the website, mark, and link and inform the user through vocal notification. The user agent accessibility guideline (UAAG) 1.0 and web content accessibility guideline (WCAG) are the standards, which are given by W3C and the NavAccess structure follows these.

C. *Voice Browsing Techniques*

The voice user interface (VUI) [6] is rapidly making its place in current web engineering technologies. Many E-commerce companies are seemed to be in the race of interactive voice portal development to facilitate its customers and capture the market. Usually in the development of voice portal developer’s use the Voice Xml. Voice XML is developed by the W3C to expedite the use of voice on the internet. Voice XML is very powerful tool and recently it proves that it can be used in the development of complex cooperate portals. The web services powered by voice XML are considered to be speaker independent these can activate the hyperlinks can send and receive requests on the user's behalf. Two techniques are very famous in voice browsing: Phone browsing and transcoding. In transcoding technique all HTML based pages can be converted into sound clips using text to speech mechanism (TTS) defined in [6].

III. Semantic Web and Ontology for Information Extraction

The invention of internet has changed our ways to communicate with each other and sharing of information. Easy access to internet played a major role in its infinite growth. Everyone can design a web page and can upload it on the internet without any prior approval. As a result, we can find information almost about any kind of topic from the internet. Trillion bytes of data transmit every day on the internet and hundreds of thousands of companies have shifted their businesses on internet [1]. When the internet was invented, its vision was to create a virtual world in which human and computer work together while sharing information, which means computers also having same contents understanding as human beings have. Later on its continuously unchecked growth derailed its structure from its vision. Due to its gigantic structure and tendency towards human understandability created several problems regarding relevant and correct information retrieval. Currently one has to work a lot to find correct and relevant piece of information in shorter span of time. Because current web is designed for human consumption and it does not provide any help for machine process ability. The computer does not have ability to understand the actual meaning of sentence for example Computer cannot distinguish between i.e. “Man eats chicken” and “chicken eats man”. Several statistical and natural language processing techniques have been developed to create machine readable. With natural language processing techniques, we can extract nouns, verbs and classify them but to get inference we have to use other techniques. Web mining techniques also failed to produce inference because data on the internet is in bulk and scattered.

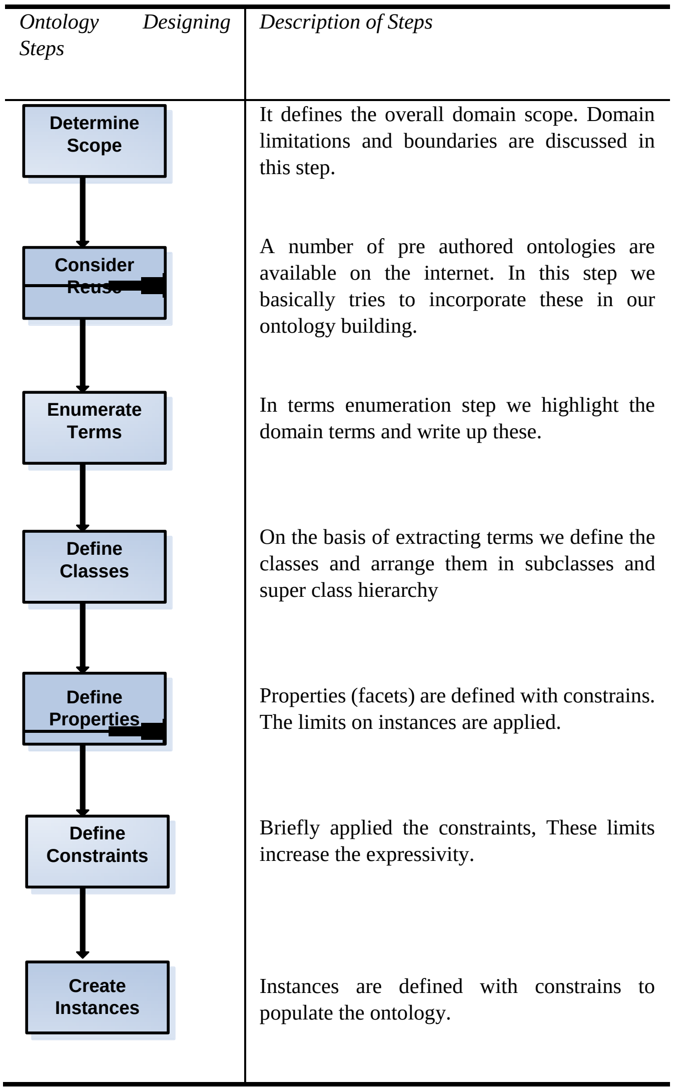
In 2001 Berners-Lee [1], [7] presented a vision of a semantic web which can transform the current interlinked directories into arranged knowledge presentation. This can be done to add the semantic annotation in a web page which creates the ability for machine to understand the underlying meaning among concepts and relationships. The layered architecture of semantic web divides the whole process into seven functional steps from Unicode /URI to trust. The layered architecture in [1] clearly defines all steps. To add semantic annotation, different techniques were developed. In start XML code had been used to add metadata in web pages which increased the content's readability. After that, the resource description framework (RDF) was being used for several years. The RDF arranges the contents and concepts into triples. Subject predicate object logic is working behind the RDF. RDF is currently used in many industrial and professional tools. Ontology [8] is the branch of metaphysics which focuses on the study of existence. Ontology can also be defined as an explicit specification of a shared conceptualization of any domain. In ontology, we basically concentrate on concepts (classes) of the domain, their relationships and their properties. Ontology arranges the classes in a hierarchical structure, in the form of subclasses and super-classes hierarchy and also defines which property has constraints on its values. Ontology is authored to share a common understanding of the domain among people and software. Ontology helps the developers to reuse the classes of a domain instead of rewriting them. Ontology is written in most widely used language OWL. OWL is developed by W3C and is specially designed for ontology authoring. The ontology designing process can be divided into seven steps for simplicity [8]. Fig. 1 depicts the steps.

IV. The Proposed Information Seeking Architecture

Our proposed Ontology based highly precise information extraction architecture is a novel methodology which facilitates the blind users in the extraction of the highly precise information from heterogeneous web resources using vocal command system for timely decision making. The proposed architecture is based on already developed technologies, so a little effort is needed to implement it publicly. According to World Health Organization (WHO) [9], some key facts of blinds are:

- The 314 million people of world population are the victim of blindness and 45 million of these are completely blind.
- The blind people are mostly old citizen in every part of the world. The developing countries have 87% of blindness ratio.
- The ratio of age related blindness increases over the few years as compared to blindness due to infection.
- Correction of refractive errors could give normal vision to more than 12 million children (ages 5 to 15).

Currently researchers are working to bring the blind in social circle so they can spend their life as normal human beings can. The proposed architecture gives a ray of hope to blind with its novel architecture. The next section explores the stages of the proposed architecture.



Ontology Designing Steps	*Description of Steps*
Determine Scope	It defines the overall domain scope. Domain limitations and boundaries are discussed in this step.
Consider Reuse	A number of pre authored ontologies are available on the internet. In this step we basically tries to incorporate these in our ontology building.
Enumerate Terms	In terms enumeration step we highlight the domain terms and write up these.
Define Classes	On the basis of extracting terms we define the classes and arrange them in subclasses and super class hierarchy
Define Properties	Properties (facets) are defined with constrains. The limits on instances are applied.
Define Constraints	Briefly applied the constraints, These limits increase the expressivity.
Create Instances	Instances are defined with constrains to populate the ontology.

Fig. 1. Ontology designing steps [8].

V. The Blind User interaction Stage

According to the architecture it is supposed that the blind user has a mobile phone Voice SMS facility. The user records his voice in interactive voice user interface (VUI) of mobile. The user's voice will automatically be converted into text and sent to the central server for further processing. This scenario needs a real time continuous speech recognition system (RTCSRS) which should be effective, fast and must have simple algorithm so that mobile device can easily use this with low battery consumption. For this purpose we selected the CMU Pocket SPHINX-2 [10]. The CMU Pocket SPHINX-2 is an open source large vocabulary continuous speech recognition system. There are many Texts to Speech (TTS) and Speeches to Text (STT) systems are available on internet under paid license. The Pocket CMU SPHINX-2 is the first known open source continuous speech recognition system till date. The CMU-SPHINX-2 [10] is lightweight and specially designed for mobile and low processing handheld devices. The CMU-SPHINCX-2 converts the recorded voice into text and sends to the central server for further processing.

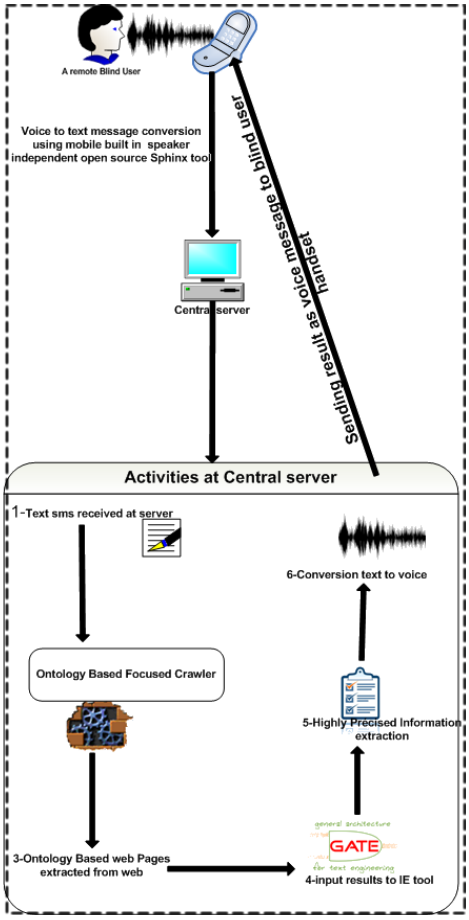

A remote Blind User
Voice to text message conversion using mobile built in speaker independent open source Sphinx tool
Sending result as voice message to blind user handset
Central server
Activities at Central server
1-Text sms received at server
Ontology Based Focused Crawler
3-Ontology Based web Pages extracted from web
GATE
general architecture for text engineering
4-input results to IE tool
5-Highly Precised Information extraction
6-Conversion text to voice

Fig. 2. Mobile ontology based information seeking architecture.

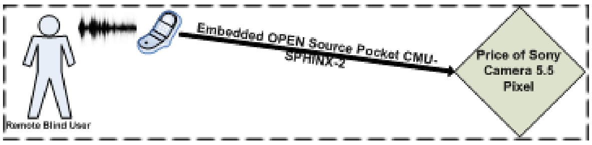
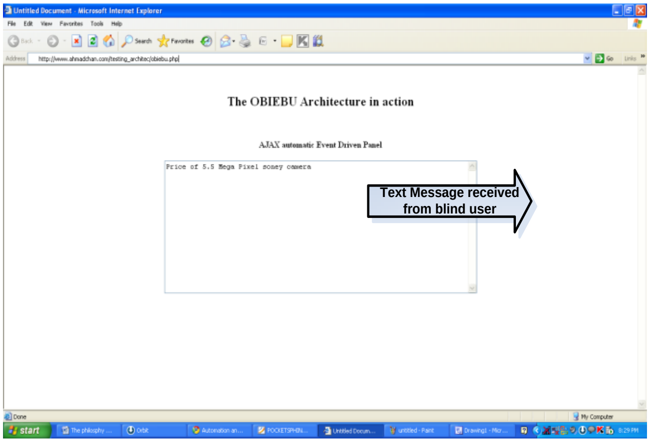

Embedded OPEN Source Pocket CMU-SPHINX-2
Price of Sony Camera 5.5 Pixel
Remote Blind User

Fig. 3. The speech to text conversion using pocket CMU SPHINX-2.

VI. The Central Server Processing Activities

A series of activities are being executed at the central processing server level. The blind user's text is received at the central server. The blind user can send his message either by using GPRS or simple text messaging utility. The received text message forwards to specially designed AJAX (Asynchronous JavaScript and XML) based dynamic web panel [11]. AJAX is not a technology but it is a combination of several technologies. Before developing AJAX web forms were dependent on manually triggered events. The user has to send a request for further processing by pressing the button. In AJAX, server and client remain in connection with the help of XMLHttpRequest method, which is specially designed for seamless communication purpose. The underlying intelligent mechanism converts the user SMS into formalized query and forward the query to the ontology based web crawler. Relevant information extraction from web has always been a difficult process in information engineering domain. A web crawler is a program that browses the web automatically [11]. The difference between a simple web crawler and focused web crawler is that simple web crawler searches the information from the web and indexed the results regardless the results are relevant or not while the focused crawler only indexes those web pages which have some relevancy with the search string. Many web crawler developed in the past had different criteria of searching based on page relevance, link relevance etc.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go Links
Done
My Computer
start

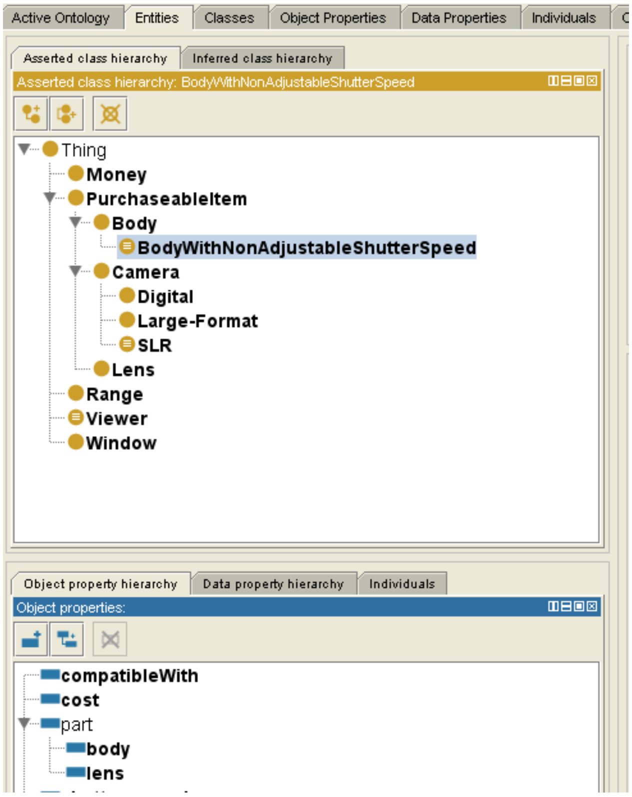
The OBIEBU Architecture in action

AJAX automatic Event Driven Panel

Price of 5.5 Mega Pixel sony camera

Text Message received from blind user

Fig. 4. The AJAX based user query panel.



Asserted class hierarchy: BodyWithNonAdjustableShutterSpeed

- Thing
  - Money
  - PurchaseableItem
    - Body
      - BodyWithNonAdjustableShutterSpeed
    - Camera
      - Digital
      - Large-Format
      - SLR
    - Lens
  - Range
  - Viewer
  - Window

- compatibleWith
- cost
- part
  - body
  - lens

Fig. 5. The camera ontology.

The proposed architecture uses the ontology based crawler to index the relevant webpage from the scattered web. The ontology based web crawler first fetches the Web Pages from the web with its fetcher component and then parses and stores the web pages on cache. Ontology allies against the fetched web pages and the most relevant web pages are indexed. We used Prot ég é 4.0.2 to write up the camera ontology. The camera ontology defines the classes and subclasses and properties of camera domain. In Fig. 3 we suppose that a user sends a query from voice user interface to AJAX based query panel after processing from Pocket CMU Pocket Sphinx-2 mechanism which converts the speech into text. The AJAX based panel automatically routes the query to the ontology based crawler which is loaded with our modified camera ontology. The ontology based crawler then starts its crawling and the finds the relevant information which is further indeed. Fig. 5 expresses the Protégé based ontology.

GATE [12] is the generalized architecture for text engineering. In the information extraction process, we take unstructured and scattered data as input and generate the fixed length output. The generated out can be used for analysis purpose. The information extraction process is not a simple task as most of the input is in human language format. The information extraction tools provide the facility of tokenization, part-of-speech tagging, and named entity recognition. GATE is an architecture, development environment, and framework for building systems that process human language queries. A nearly new information extraction (ANNIE) is an information extraction tool which is a domain and even application independent. These two utilities used in tagging of information on the basis of defined criteria. The name entity utility of ANNIE is most widely used and most reliable in information engineering circle. It extracts the information about person, place, price, items etc. We populate the GATE corpus on the behalf of result which is extracted by ontology based crawler. The GATE rules are based on JAPE (Java Annotation Patterns Engine) which defines all the rules and regulations which help the extraction process. The information is extracted at this stage and passed to the last stage to reconvert the text message into a voice. At last stage the architecture sends the voice SMS back to the user’s handset.

VII. Conclusion and Future work

In this article we proposed mobile ontology based information seeking architecture for visually impaired masses of under developing countries. The architecture takes the input through voice SMS and perform the information extraction steps automatically. The architecture is novel and simplest which is based on existing technologies. The automatic execution of this novel architecture can bring change in vision-less people’s life. The adaptation of this architecture publicly can bring millions of people in social circle who are partially or completely blind. A few weak things still exist in the architecture which needs to be improved. As a future work we will integrate the automatic vocal ontology designer with our architecture.

References

[1] A. J. Gerber, A. B. Brand, and A. J. Merwe, “Towards a semantic web layered architecture,” *Proceedings of the 25th Conference on IASTED International Multi-Conference: Software Engineering*, pp. 353-362.

[2] K. Nicholas, “Wrapper induction: Efficiency and expressiveness,” *Journal of Artificial Intelligence*, vol. 118, pp. 15-68.

[3] C. Valter, M. Giansalvatore, and P. M. Roadrunner, “Towards automatic data extraction from large websites,” *Proceedings of the 27th VLDB Conference Roma*, pp. 109-118.

[4] H. Hans and E. Vanessa, “Web navigation for blind users,” *University of Amsterdam*, pp. 105-175

[5] Z. Mary and P. Chris, “A web navigation tool for the blind,” *ACM Press*, pp. 204-206.

[6] K. B Michael, C. G Stephen, and C. S Brian, “Web page analysis for voice browsing. Interaction design: Beyond human-computer interaction,” *John Wiley and Sons, Ltd*. Sharp Rogers and Preece 2007.

[7] N. Sahar, N. Mahdi, and M. B Mehrdad, “The semantic web: A new approach for future world wide web,” *World Academy of Science, Engineering and Technology*, vol. 58, 2009.

[8] F. N. Natalya and L. M. Deborah, “Ontology development 101: A guide to creating your first ontology,” *Stanford University, Stanford, CA*, vol. 94, no. 305.

[9] How Many Blind People Are There In The World - Ask Community. [Online]. Available: http://www.numberof.net/number-of-blind-people-in-the-world/.

[10] D. Huggins, M. Kumar, A. Chan, and A. I Rudnicky, “Pocket-sphinx: A free, real-time continuous speech recognition system for hand-held devices,” in *Proceedings of ICASSP*, pp. 41-46, 2006.

[11] E. Gatial, Z. Balogh, M. Laclavik, M. Ciglan, and L. H, “Focused web crawling mechanism based on page relevance,” in *Proceedings of ITAT 2005 Information Technologies -Applications and Theory*, 2005, pp. 41-46.

[12] Cunningham, “GATE, a generalized architecture for text engineering,” *Computers and the Humanities*, vol. 36, pp. 223-254, 2002

Ahmad C. Bukhari received his B.S. Degree in Mathematics from University of Punjab, Lahore Pakistan in 2005 and M.Sc. Degree in information Technology in 2008 from PUCIT, Lahore Pakistan. Besides he holds a M.S. Degree in computer science which he has completed from Gyeongsang National University Korea. His area of research includes intelligent system, Semantic Web, Soft computing. Collision Avoidance System for Autonomous Underwater Vehicle.

Mr. Mehtab Afzal received his Master’s degree in Computer Science from COMSATS Institute of Information Technology Abbottabad, Pakistan in 2009. He worked as a Research Associate for two years and as a Lecturer for one year. Currently, he is a PhD Scholar at Southwest Jiaotong University, Chengdu, China. His research interests include Information Retrieval, Semantic Web, Concept based Video Retrieval and Web Video Annotation and Classification.

Yong Gi Kim received the B.S. Degree in electrical engineering from Seoul National University, Seoul, Korea in 1978, the M.S. Degree in computer science from University of Montana, U.S.A. in 1987, and the Ph.D .Degrees in computer and information sciences from Florida State University, U.S.A. in 1991. He was a Visiting Scholar in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, U.S.A., from 2008 to 2009. He is currently a Professor in the Department of Computer Science, Gyeongsang National University, Korea. His current research interests include soft computing, intelligent systems and autonomous underwater vehicles.