Intelligent Information Access: A Survey

Nupur Choudhury
Department of CSE & IT
Assam Don Bosco University,
Guwahati, India

Rupesh Mandal
Department of CSE & IT
Assam Don Bosco University,
Guwahati, India

Vikas Sharma
Department of Space
North Eastern Space Applications
Centre, Meghalaya, India

ABSTRACT
This paper deals with a generic study about basic computational techniques which are required to develop systems that deal with Intelligent Information Access. This paper primarily deals with the better understanding of the underlying working of the concept of Web 3.0 or the Semantic web. Moreover this study related to the survey also describes how data can be maintained to reduce the conflicts related to information and enhancement of user experience. In addition to it, this paper also deals with various technologies which makes use of human like intelligence that provides efficient and effective access to huge, distributed, heterogeneous and multilingual information resources. It also describes various technological standards that are based on resource development framework, SPARQL, Web Ontology language and different techniques that are related to Intelligent Information Access.

Keywords
Information Access, Semantic Web, Web 3.0.

1. INTRODUCTION
In the present world complete and accurate information is crucial in all aspects. The resources that are available are flooded with various information that impacts the quality of the output of a process and its related decision that has been processed. Generally more amount of time and effort is invested on a daily basis which is needed in locating the "critical" piece of information which are needed for their current demands. Intelligent Information Access (IIA) refers to various technologies which makes use of human –like intelligence in providing efficient and effective access to large, distributed, heterogeneous and multilingual information resources. In other words any technology related to Information access which might involve applying human knowledge in order to retrieve, understand, synthesize or extract information are considered as Intelligent Information Access. The primary goal of IIA is to provide a personalized access to information by using various information retrieval techniques. This process produces an intelligent representation of the information content.

2. BACKGROUND STUDY

2.1 Requirement of intelligent information access
Since a long time, in the middle period of the twentieth century, computers were utilized primarily for storing, managing and retrieving the data. However since then at about decades later the increasing data storage and the related information to it created the need of utilizing this exponential growth of data that has been stored by the databases. This data if not extracted and utilized would be useless for people if they are not obtained in desired manner. The primary issue is not storing or the data search, but what is the objective of the data search and what meaning it holds. Data storage should not only be the primary aim but information retrieval should be the objective i.e. both the data and its associated meanings. Which means that Information systems rather than database systems are needed. However technology to manage data was available but technology to manage information was limited. Similar was the case with the experts which were primarily focused with data retrieval and acquisition but not their underlying meanings. Hence in order to manage the underlying semantics the concept of knowledge representation was introduced in the AI community which made use of various formalisms like frames, semantic networks, fuzzy systems, knowledge bases etc. in order to describe the meaning of the data. Presently knowledge bases are named as ontologies which are used to describe the data sources, and the knowledge management systems to deal with them has been assisting to provide semantics management to information systems since then. In the present world where the access to the required information is very crucial, ontology permits the software designers to relieve the users from having to deal with large quantity of data. In other words it allows the viewing of huge information systems from the point of view it is required.

2.2 Semantic Web
Semantic Web was first coined by Sir Tim Berners-Lee who was also the originator of the World Wide Web in 1989. Semantic Web is also known as Web 3.0, Web of data, Linked or connected data web etc. The term semantic represents meaning or understanding.

The primary difference between technologies that are related to data like relational databases and the semantic web technologies is that the Semantic web is basically concerned with the underlying meaning and not the data structure. The important advantage of Web 1.0 was that is abstracted and hid away the physical storage and networking layers which are involved in information exchange in between two machines. This process gave a look as if the documents are directly connected to one another. Similarly, the application layers involved in the exchange in information is abstracted away by the Semantic web. The Semantic Web is used to connect facts, hence instead of connecting to a specific document or application, it can be referred to a specific piece of information that might be present in a document or an application. Hence whenever that information is updated, the advantage of the update can be taken automatically. Semantic web can generate the next major evolution in information connection. It enables the data to be connected from one source to any other source and in a way which can be understood by the machines in order to perform increasingly sophisticated task thereby replacing human labor by computation.

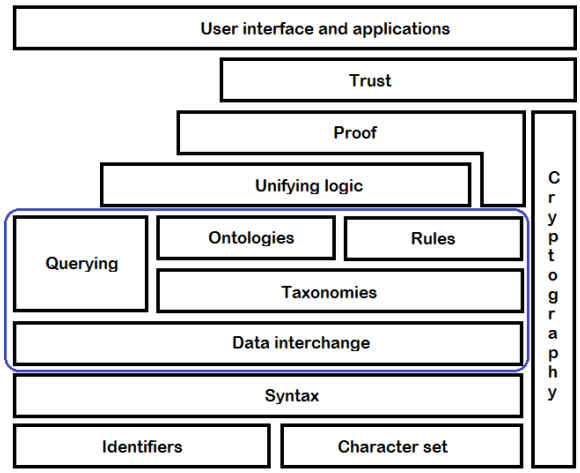

Fig 1: Architecture of the semantic web [3]

The figure 1 above represents the semantic web architecture which primarily consists of querying, ontologies, rules, taxonomies and the data interchange. It is an inbuilt part of the existing web.

Technically, there are three technical standards of the Semantic Web:

- RDF (Resource Description Framework): RDF is used as the data modelling language for Semantic Web. It is used for storing and representing Semantic web information.
- SPARQL: This is used as the query language for the Semantic Web. It is designed specifically to generate query across different systems.
- OWL (Web Ontology Language): This is used to represent the schema language or the knowledge of the semantic web. This language helps to define the concepts in a composed way so OWL enables you to define concepts in a composed way so that they can be utilized and re-utilized as much and as often as possible. Composibility means that each and every concept is carefully defined so that it can be selected and arranged in different combinations with other ideas as required for various other applications and activities.

3. TECHNIQUES FOR DEVELOPMENT OF INTELLIGENT INFORMATION ACCESS SYSTEM

3.1 Text Summarization

In today's world a huge number of searches is done by business analysts, leaders various students and academicians on a regular basis to progress in their own respective domains and a major portion of their research and search is wasted by simply trying to figure out which document is relevant and which is not. The readability of a document is judged quickly and can be assessed in a short span of time if the important sentences can be extracted and comprehensive summaries can be generated. Text summarization is the process of shortening a text document or an article so that a summary can be created that would contain all the major points of the initial document of origin. The primary concept of text summarization is to generate a data subset which would possess the "information" or the "meaning of the entire set. This type of processes are widely used in the present world industries. It is used to find out the information content of the documents by identifying the most informative sentences that can be represented in a concise and precise manner to enhance the search process. Text summarization can be categorized primarily into 2 different approaches: extraction and abstraction.

a. **Extraction**: These methods are used to form the summary in the original text document by creating a subset of the informative words, phrases or the sentences.
b. **Abstraction**: Abstractive methods are used to create an internal semantic web representation after which Natural Language generation techniques are used which would enable to create a summary that would resemble what a human might think or what he/she might express. These type of summarization includes numerous verbal innovations

Summarization of the textual work is done by a mathematical process which is initiated by calculation of the word frequencies for the entire document. After that the identification of the most common words, n, that are stored and sorted is done. (The n variable might change or vary for various situations). Based on the occurrence frequencies each and every sentence is then stored based on the higher frequency. The last step involves identification of the top X sentences which are sorted that has some resemblance to the original text considering their position.

3.2 Information Extraction (IE):

There are various machine readable documents that contains unstructured and semi structured formats. Information Extraction is the process of automatically generating structures information from those type of documents. World Wide Web contains enormous information storages which are available for use. Generally these storages are focused to be utilized and investigated by various human users with the help of a web browser. However using machine intelligence in these data sources gives rise to various interesting applications, for eg. Using web data in various software agents. However this has certain limitations. Whenever a machine is given the rights to generate or extract information based content from the web pages (which are primarily in HTML format), in order to restructure them into some format that can be understood by machines becomes a cumbersome process. Structured data consists of entities and relationships that are organized in a predictable manner.

For example, considering the relationship between certain companies and their locations, if the goal required is to find out the locations where that particular company does business or conversely if the reverse is required which means given a particular location the identification of the company which does the business is required then straightforward answers are derived if the data is available in some structures or tabular format. For the datasets which are unstructured in nature, the basic and primary technique used for entity detection is chunking. This enables the segmentation and labelling of multi-token sequences. This process is started off by generating the noun phrase chunking or NP chunking where searching for chunks correlating to the individual noun phrases is done. NP-chunks are basically smaller phrases or pieces as compared to the original noun phrases. These are then evaluated by converting the Inside-Outside Beginning (IOB) format to Natural Language Toolkit (NLTK) format for further processing.

3.3 Automatic Classification:

The mechanical way of identifying the subject category of an object is termed as Automatic classification. The object which undergoes the process of classification can be anything which can be processed and identified by a computer which might include documents, images, text, research articles or projects. The pioneer areas of automatic classification are book processing, papers of journals, documents, scientific articles etc. Automatic Classification can be broadly categorized into 2 major categories:

a. **Rule based automatic classification:** This type of classification makes use of human based expert knowledge in order to generate the knowledge base and creation of the rules.
b. **Machine learning techniques:** These techniques helps to enable the system in order to learn from various examples or the patterns that were used for training which aids to extract and obtain the various knowledge contained in the item into the index entries that can be searched autonomously. Hence these techniques have a wider range of usability because of their speed, precision and scalability.

3.4 Rule Based Method:

This method enables the system to resemble a person during classification of a document by leveraging the natural Language understanding capability of a system and generating the linguistic rules. This also means automatic categorization of relevant information considering the semantic aspect. Unlike the previous methods which uses statistics or mathematics, this method has an added advantage as it uses the open box approach and constantly improves the performance. Moreover in complex situations or scenarios as well, it produces higher quality performance.

3.5 Artificial Neural Network:

This type of networks is used for text classification and have two different layers for execution: hidden layer and an output layer, where inputs are given. Neurons are connected based on the weights, this allows the expression of the strength of connections within definite elements. ANN is widely used for text classification since it can work with huge data sets of various features and can have accurate classification even if noise is present. Moreover ANN is capable of performing parallel computations where all the neurons in the layers is capable of processing data independently.

3.6 k-Nearest Neighbours Method

It belongs to the category of nonparametric regression algorithms. Its primary advantage is that it does not require any initial training for the classifiers and is purely based on the relations within the text documents. But this algorithm has limitations in the form that it has extensive running time of execution. The resemblance or the similarity of various tested documents and the k training documents is tested using the KNN method. Depending on the features the stored knowledge helps in the classification of the documents. KNN is considered as the fastest machine learning algorithms. The classification result for a resultant document is the class in which the largest number of neighbors would belong to a minimum positive number, generally k, which is also termed as majority voting. If k=1, the object is generally assigned to the class of which the nearest neighbor belongs to. The primary advantage of the K-NN based classification method is that it is simple and has easy to comprehend and use. Moreover whenever classification of documents of multiple categories is considered KNN gives good results. This classification scenario can also be considered as classification of more than 2 categories. Hence whenever the training set is growing as well, i.e. in computationally demanding situations, this methods considers the distance between features and gives accurate results.

3.7 Decision Rule

This algorithm represents the category profiles by creating a set of rules. These rules are created as follows:

IF condition

THEN conclusion

Where features that are based on certain category are represented in the form of conditions and the conclusion is represented by predicting the category. AND or OR logical operators are used to create the rule dataset by connecting the atomic rules with the logical operators. The rules created from in the knowledge base do not require to be fired altogether during the classification process. While handling huge datasets, the redundant rules are removed and the system efficiency is adjusted by using heuristics. During the phase of feature extraction, the primary advantage of this method is its ability to generate a local dictionary for each different category. These type of local dictionaries discriminates the meanings of homonyms (particular words) for various other categories. For example in English language homonyms can be a word stalk that can be comprehended as a herb, part of a plant, threat as a plot or any danger, skate as gliding on ice or a fish. Since the similar type of keywords can be in different rules for various classes, this algorithm is limited as it is not capable of assigning a document to a single category. Moreover human intervention and expertise is required to create or actualize the rule set in the knowledge base which helps in updating of the decision rules and learning. Moreover accurate results cannot be expected when working with large datasets.

3.8 Cross-Language Information Retrieval (CLIR)

Cross Language Information Retrieval system is capable of providing information to the users which is in a dissimilar language than the language from the given queries. For example, users who speaks English is capable of mining information based on Chinese language using English queries with the help of systems like English-Chinese cross language information retrieval (ECCLIR). The primary working strategy for information retrieval in modern world is to match the queries with the documents. The primary-strategy for retrieval of information in the present world scenario is to match various documents along with their queries. The complexity increases on either side if the documents that needs to be classified and investigated and the queries are in dissimilar languages, which is taken care of by a transformation closely resembling the situation of a CLIR since the matches are not direct in nature.

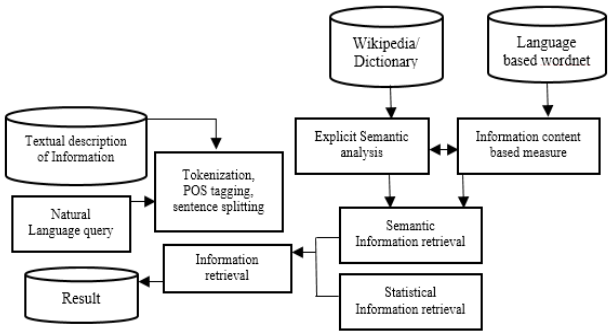

Fig 2: Cross-Language Information Retrieval (CLIR)

The figure represents how knowledge bases along with the human generated rules are used for intelligent Information access in CLIR systems where they make use of language based word nets exclusively to translate the queries to derive the required results. The transformations in CLIR can be categorized into 3 primary classes: query translation, translation of documents and techniques based on different linguistics. In case of query translation the CLIR system is responsible of translating the queries into the languages that the document is written in. Whereas the reverse happens in document translation. Here the document language is translated into the language in which the query is asked for. However in the third technique the language and the query both are translated into a different representation. However out of the three approaches CLIR computational systems mostly make use of translation of the query because it is simple to use and is effective in execution. It makes use of different knowledge bases like dictionaries that follow different linguistics, Machine translation systems (MT), and parallel texts or sources which are generated by combining more than one source. These are then used to translate the queries into the collected documents after which monolingual information retrieval is conducted. This makes execution of information retrieval efficient and effective.

4. CONCLUSION

It is believed that IIA technology is the future in improving user services and minimizing conflicts in information. This paper helps to better understand the underlying working of the concept of Web 3.0 or Semantic web. The main purpose of this survey is to understand how data can be maintained to reduce conflicts in information and better user experience. It also deals with technologies that make use of human-like intelligence to provide effective and efficient access to large, distributed, heterogeneous and multilingual information resources. It also describes about the technical standards based on resource description framework, SPARQL, Web Ontology Language and various techniques used for Intelligent Information Access.

5. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Armano G, de Gemmis M, Semeraro G, editors. Intelligent information access. Springer; 2010 Jul 8.

[3] Berry MW, Dumais ST, Letsche TA. Computational methods for intelligent information access. InSupercomputing, 1995. Proceedings of the IEEE/ACM SC95 Conference 1995 (pp. 20-20). IEEE.

[4] Onlineresource:https://www.obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html , Last Access: December 2017.

[5] Allan J, Aslam J, Belkin N, Buckley C, Callan J, Croft B, Dumais S, Fuhr N, Harman D, Harper DJ, Hiemstra D. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. InACM SIGIR Forum 2003 Apr 1 (Vol. 37, No. 1, pp. 31-47). ACM.

[6] Aparicio F, De Buenaga M, Rubio M, Hernando A. An intelligent information access system assisting a case based learning methodology evaluated in higher education with medical students. Computers & Education. 2012 May 31;58(4):1282-95.

[7] Malone TW, Grant KR, Turbak FA, Brobst SA, Cohen MD. Intelligent information-sharing systems. Communications of the ACM. 1987 May 1;30(5):390-402.

[8] Sycara K, Pannu A, Willamson M, Zeng D, Decker K. Distributed intelligent agents. IEEE expert. 1996 Dec;11(6):36-46.

[9] Choo CW. Information management for the intelligent organization: the art of scanning the environment. Information Today, Inc.; 2002.

[10] Flake GW, Lawrence S, Giles CL, Coetzee FM. Self-organization and identification of web communities. Computer. 2002 Mar;35(3):66-70.

[11] Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the world wide web. InTools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on 1997 Nov 3 (pp. 558-567). IEEE.

[12] Lawrence S. Context in web search. IEEE Data Eng. Bull.. 2000 Sep;23 (3):25-32.