Semantic Web based Information Extraction

Nimeshkumar Arvindbhai Patel[1] Sanjay M. Shah[2]
[1]P. G. Scholar [2]Professor
[1,2]Department of Computer Engineering
[1,2]Government Engineering College, Modasa, India

Abstract— Information extraction system is playing an important role now days. Currently search engines search on keyword based, which turn to give lots of enormous data available to the user and user cannot get the required information. Such limitation may be solved by new web architecture, which is known as semantic web architecture. The semantic web is an extension of W3C (World Wide Web Consortium) which promote common data formats & exchange protocols on the web which overcome the limitation of Knowledge Base searching also known as searching based on conceptual or keyword based. Natural Language processing is a mostly useful technique which is used in QA (Question Answer) system where a user asks a query and many steps are performed on such query and after that proper result is available. The conceptual search engine identifies the meaning of each tokenization query term and interprets it with the relation among document concept related to a particular domain which gives a proper answer instead of giving a list of answers. In this paper, we used semantic search based on ontology, Quepy API, SPARQL Query, Wordnet.

Key words: Resource Description Framework; SPARQL Protocol and RDF(Resource Description Framework) Query Language; Quepy API; Ontology; Knowledge Base; Word Net

I. NOMENCLATURE

KB	Knowledge Base
SW	Semantic Web
QA	Question Answer
SSE	Semantic Search Engine
RDF	Resource Description Framework
GUI	Graphical User Interface
WWW	World Wide Web

II. INTRODUCTION

Now a day all information is available on the internet. All domains data are available and show relevant information. Such information is huge in amount and dispreads over the internet. So, the user cannot get easily required information or can't get whole data or even get that data in the unstructured or semi-structured form. So, one of the system required which cover all the above problems.

Everyone requires quick information even also it required in a structured format, when searching it on any common search engine it wouldn't get it because information and meaning of information are not preserved, for example a query like "Mobile phone with red cover". It produces relevant as well as irrelevant result like different mobile phone models with specifications, red lotus, flower and cover. Here system lost the significant of the term "with" so such types of irrelevant results produced. In order to reduce ambiguity of result intelligent search concept evaluate that is SW (Semantic Web). SW came in 1996 by Tim Berners Lee [1]. Develop SSE (Semantic Search Engine) for implementation of ontology, SW, KB (Knowledge Base) into one system.

A. Challenges of SW

- It required extensive knowledge by users for well-defined SW technologies and components.
- Searching cost and scalability are higher [2].

B. Important to develop SSE

- Accuracy
- Time retrieval
- Efficient ranking of web pages [3].

C. Objective of SSE are as followings

- Create ontology database for storage of ontology specific knowledge.
- Generation of SPARQL of user's Natural Language query.
- Semantic search on Natural Language query.
- To get answer from ontology (domain specific) [3].

This paper gives the guidance for the user to get semantic information from sources like web, documents, etc…

This paper is organized as follow: section I introduces about semantic web and its importance, section II is literature survey and classification of semantic similarity methods and section III is Proposed SSE, section IV concludes paper.

III. LITERATURE SURVEY

Jishma & Sunith [4] described ontology based Abstractive Summarization. They present the meaning of summarization i.e. summarization is the process of extracting important information from the source text & to present that information to the user in the form of summary and if this done by the computer automatically then it is called Automatic Text Summarization [5]. Summarization of document can be done in two way, one is Extractive which extracts the meaningful portion of text and recreates them in the format which given as input so it is useful where more depth analysis required on work this method using we can generate new sentence which improves the focus of summary and reduce redundancy [5]. Second it Abstractive Summarization. It has two technique one is structured based summarization that is a tree, template, ontology, rule-based method which depend on structured and other is semantic based summarization that is Multimodal semantic, graph-based method.

Ontology + instance of class = Knowledge Base [6] for example class is vehicle then the specific vehicle is an instance of the class. Reasons of developing ontology i.e. reuse data, time complexity, and shared basic common related information of structure to the people. Ontology provides capability of reuse of domain knowledge, makes change easier, provide knowledge separating among domain & operational, Facility of reusing & extracting them.

Now a day semantic relatedness is less studied [7]. It shows the improvement in semantic relatedness assessment means the knowledge which Taxonomically & Non-Taxonomically similar by aggregating the ontology-based semantic estimation with the distributional resemblance of textual terms. Wordsim 353 benchmark [8] which contained 353 English word pairs which check semantic relatedness also used Wordnet [8] to described more than 100000 general concepts which are in semantic relation.

The semantic web is a solution of Multilingual Natural Language Queries means used for multiple languages like if query fired in the Arabic language then also get the semantic result of that query. Each language has its own way of processing and according to that framework need to develop and after on that SPARQL Query generated. SPARQL Query has 3 steps 1. Query Processing 2. Query Mapping 3. SPARQL Query generation.

The database contained information but it is problematic if it contained indirect knowledge of the domain this information in the form. Information about value in the field. So first every common field gathered and distinct values are found and analyzed then find a key candidate after that foreign key is identified. Here they also used a C4.5 algorithm for calculating distinct value [9] .

Query extraction step is a sub-task of like query processing which extract the user's query and finding similar words. In Query processing, remove stop word or noise word, Parse the query and then create small token. It used Stanford parser then it searches for semantically related terms. It used Wordnet 3.0 and after that import existing ontology for domain specific [10].

Semantic web represents the extension of the www will provide web exchange data between two different users the same work also done by many methods and one of the Extensible Markup Language (XML). XML has built up as a non-specific procedure to store compose and recover information from the web [11]. XML provides self-defined tags and data model is a tree. No semantic data are added [12]. It uses DTD, XSLT, and XSD as a Schema definition. Whereas resource description framework (RDF) is a typical dialect that empowers the office to store asset's data that are accessible on the World Wide Web utilizing their own particular space vocabularies [11]. RDF provides fixed and pre – defined tags and model are graph. It uses RDFS and OWL RDF schema. Here semantic data is added [12]. Table 1 shows classification of Semantic Similarity Methods.

Method	Goal	Description
Thesaurus Based Similarity	Helps to identify word similarity.	Lexical resourced such as dictionary which specify semantic relationship for example a web source like Wordnet [8]. Drawback – it unable to show extent of similarity.
Empirical method	Helps to identify word similarity.	It solved above drawback by using page count and text snippet method [13].
Empirical using corpus of web documents	Helps to identify term similarity and find meaning of same terms presented in different forms	Following factors are considered 1. Documents length 2. Number of common terms 3. Whether terms are common or unusual. 4. Terms occurrences [2].
Logic form transformation (LFT)	Helps to identify degree of semantic equivalence between two sentences	Pairing and knowledge deriving from different source here pairing of each word of one sentence to most similar one in the second sentence after word to word taken up. The same process repeats from second sentence to first one. Drawback – web based search related to structured data i.e. ONGC and Oil and Natural Gas Corporation is considered as a different [14].
Candidate generation selection & noise cleaning	Helps to identify entity based semantic search	Entity synonyms various classes Normalization – difference due to stops words and punctuation symbol. Spelling- a difference due to typing or spelling errors. Subset- includes main parts of the string. Superset-additional information with the subset.

		Abbreviation – Abbreviation such as ONGC, TCS, etc. [15].
Document Vector Model (DVM)	Helps to find out document based semantic similarity	1. Define a form for similar text content. 2. Define semantic similarity metric based upon form. 3. Convert the document into requisition form and after for finding semantic similarity score [2].

Table 1: Classification of Semantic Similarity Method

IV. PROPOSED SSE

In our system, User Interface is created which deal with end user who enters a query in the form of Natural Language. It is pass to the parser for query processing. It removes stop word, unwanted word and keeps the meaningful word of the query. It also analyzes the query stream to identify question type, answers type, subject, verb, noun, phrases, etc… Tokens are extracted from the question, meaning analyzed and then passed to next stage. Protégé tool used for ontology creation and edition purpose, Quepy API is used for mapping SPARQL Query to ontology [11].

A. Expansion

In this phase, Wordnet or Local domain specific dictionary is used. It covered lots of similar words related to token then after SPARQL generated. Wordnet gives synonyms words for example start and begins so it matches a question with appropriate rules. A wordnet is a data dictionary which gives 100000 of the specific meaning of the word in the English language. It is used to match token with Wordnet dictionary token and get all word from Wordnet and search result in Knowledge Base [16].

B. Knowledge Base

Knowledge Base = Ontology + Instance of class [6]
It is domain specific here we can add one or more ontologies and it operate on KB and get the specific results [17].

C. Query API

Quepy API is a python based framework which transfer natural language question to database query language. It has facility to customized different types of questions in natural language and database. Natural language to SPARQL transformation is done by special form of regular expression. Part Of Speech (POS) tagset used by Quepy in regular expression. For language independent Quepy uses an abstract semantics which mapped to query language. This allows mapping of your question to different query languages in transparent

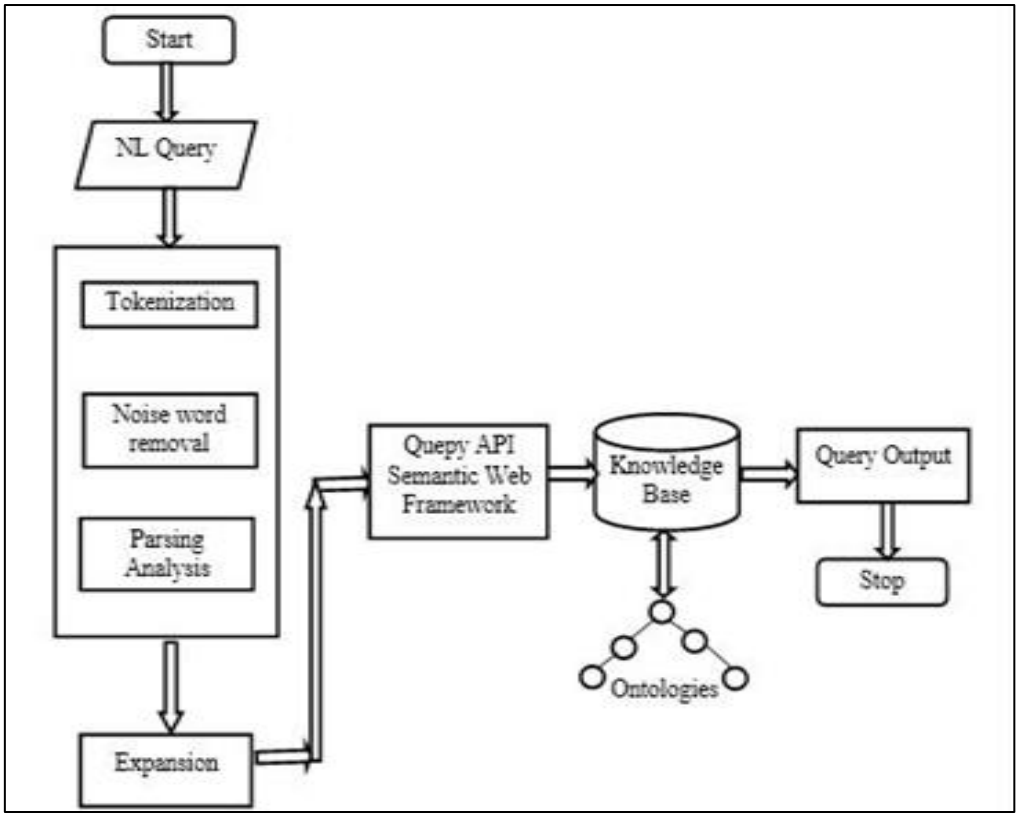

Fig. 1: System description of existing system

manner. If output of question is exist in ontology than it also available in KB so, question's reply is got quickly. Information retrieval search engine retrieves result and display in the structured form [11].

D. Ontology

When we need to present objects of classes and relation between them then ontology used. To match conceptual information between two knowledge base on the web a code must require for common meaning and a solution is collecting that information at a common place called ontology, ontology creation process has 3 sub-steps: Ontology capture, Ontology coding, Ontology integration of existing ontologies. Ontology has also life cycle as every system which includes steps like Specification, Conceptualization, Integration, Implementation and Maintenance [18].

E. RDF

When we need to interchange data on the web than RDF is used. It provides facility of merging of data even if the schemas different. It allowed structured or semi-structured data to be mixed and shareable over the different application. It uses linking structured and in form of directed labelled graph in which edge present named link between two resources. In RDF graph view is the easy format for understanding RDF also utilize in knowledge management application [12], [2].

F. SPARQL

SPARQL stand for SPARQL protocol and RDF query language is semantic query language for database and retrieves data, manipulation data which is stored in RDF format [19]. It is directed, labelled graph data format to represent context on the web which defines the syntax and semantics of the SPARQL query language for RDF. SPARQL query in triplet form <Subject, Object, and Predicate>. It also provides query of triplet form with conjunctions, disjunctions and optimal patterns [12].

G. DBPedia

DBPedia extracts structured information from DBPedia's ontology. Generally, in Wikipedia, all the information available and in the form of unstructured which is more trouble full for getting simple data so DBPedia have their own ontologies and it fired queries on such ontologies and provide result in the structured form [4].

V. CONCLUSION

Semantic web promotes common data format and exchange protocol on the web. So, the limitation related to knowledge base searching is solved and get proper required structured output of the natural language or user query. Knowledge base stored the ontologies and the related term, Quepy API mapped SPARQL with RDF knowledge base. Semantic web is also applied on documents, email, books, etc… to get semantic meaning of term or question.

REFERENCES

[1] B. T. Berners-lee and J. Hendler, "The Semantic Web," IEEE intelligent systems, vol. 21, no. 1, pp. 53–59, 2006.
[2] Yadav Usha, Gagandeep Singh Narula, Neelam Duhan and Vishal Jain, "A Novel Approach for Precise Search Results Retrieval Based on Semantic Web Technologies," Computing for Sustainable Global Development , 3rd International Conference on. IEEE, pp. 1357–1362, 2016.
[3] Y. Lei, V. Uren, and E. Motta, "SemSearch : A Search Engine for the Semantic Web," International Conference on Knowledge Engineering and Knowledge Management, pp. 238–245, 2006.
[4] M. Jishma Mohan, C. Sunitha, A. Ganesh, and A. Jaya, "A Study on Ontology Based Abstractive Summarization," Procedia Computer Science, pp. 32–37, 2016.
[5] J. S. Kallimani, "Information Extraction by an Abstractive Text Summarization for an Indian Regional Language," Natural Language Processing and Knowledge Engineering, 7th International Conference on. IEEE, pp. 319–322, 2011.
[6] C. Nowak, "On ontologies for high-level information fusion," Information Fusion 2003 Proceedings of the Sixth International Conference pp. 657–664, 2003.
[7] M. Batet and D. Sanchez, "Improving Semantic Relatedness Assessments: Ontologies Meet Textual Corpora," Procedia Computer Science, pp. 365–374, 2016.
[8] Finkelstein, Lev, et al., "Placing Search in Context : The Concept Revisited," Proceedings of the 10th international conference on World Wide Web, pp. 406–414, 2001.
[9] Gorskis, Henrihs, Ludmila Aleksejeva, and Inese Polaka, "Database Analysis for Ontology Learning," Procedia Computer Science, vol. 102, no. 12, pp. 113–120, 2016.
[10] Al-Nazer, Ahmed, Saeed Albukhitan and Tarek Helmy, "Cross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case," Procedia Computer Science, pp. 607–614, 2016.
[11] Liu Haishan, "Towards semantic data mining" 9th International Semantic Web Conference, pp. 7–11, 2010.
[12] S. Decker, F. Van Harmelen and J. Broekstra, "The Semantic Web - on the respective Roles of XML and RDF," IEEE Internet computing, vol. 4, no. 5, pp. 63–73, 2000.
[13] Chen Hsin-Hsi, Ming-Shun Lin, and Yu-Chuan Wei, "Novel Association Measures Using Web Search with Double Checking," Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1009–1016, 2006.
[14] Gabrilovich, Evgeniy and Shaul Markovitch, "Wikipedia-based Semantic Interpretation for Natural Language Processing," Journal of Artificial Intelligence Research, pp. 443–498, 2009.

[15] Cheng Tao, Hady W. Lauw and Stelios Paparizos, "Entity Synonyms for Structured Web Search," IEEE transactions on knowledge and data engineering, vol. 24, no. 10, pp. 1862–1875, 2012.
[16] G.A. Beckwith, and R. Fellbaum, "Introduction to WordNet: An on-line lexical database," vol. 3, no. 4, pp. 235–244, 1990.
[17] Abirami A. M. and A. Askarunisa, "A Semantic Based Approach For Knowledge Discovery And Acquisition From Multiple," International Journal of Web & Semantic Technology, vol. 4, no. 3, pp. 57–63, 2013.
[18] Lee, Chang-Shing, Zhi-Wei Jian and Lin-Kai Huang, "A Fuzzy Ontology and Its Application to News Summarization," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 35, no. 5, pp. 859–880, 2005.
[19] Zhang Xiang, Gong Cheng and Yuzhong Qu, "Ontology Summarization Based on RDF Sentence Graph," Proceedings of the 16th international conference on World Wide Web, pp. 707–715, 2007.
[20] https://en.wikipedia.org/.