Survey On Semantic Information Retrieval Techniques In Bigdata

M.Geetha, N.Vimala

Abstract: Big data reflects the exponential increase of everyday data generated by devices connected to a network. Big data is not only enormous amount of data, but also the ways with which it could be stored, processed and analyzed. It is called the 3 Vs of big data, Volume, Variety and Velocity of data that is beyond the compute capacity of conventional data processing facilities. It requires schema-less data processing facilities, which is one way of providing solution to manage such data. This solution also is not complete without imbibing the semantic information inherent in natural language data. Hence, there is a need for extracting information, which is concerned with semantic information instead of considering only the structural information. This brings an interesting concept called semantic relevancy. The semantic information retrieval systems have to adopt according the domain knowledge involved. When dealing with semantic information, it is highly important to construct queries that could fetch semantically relevant documents than a syntactic retrieval. This would help to build systems that are not affected by polysemy and granularity mismatch. Semantic matching achieves exact interlinking of concepts among documents thereby providing a holistic view of the domain. The linking problem can be solved with the help of knowledge graph obtained from semantic interlinking. Therefore, this work reviews the various semantic information retrieval techniques with respect to big data. The pros and cons of various techniques are analysed the suggestions are made to future researchers.

Index Terms: big data, Information Retrieval, Knowledge graph, Semantic Matching, Schema-less processes

1. INTRODUCTION

Big data is defined as the data that has enormous volume, with/without a concrete structure or in some cases semi-structured. Big data is characterized as 3Vs such as, enormous volume of data, multiple varieties of data and the lightning velocity of data being processed. The conventional data analysis frameworks cannot manage all these terabytes and petabytes data, easily. It takes polynomial time to process such data and setting up infrastructures to handle them is costly. The alternate approach to tackle big data depends on schema-less architectures and uncompromising data quality. Normally raw data is tagged with extended metadata and can be used in machine learning models. Artificial Intelligence applications look for patterns among data elements that are frequent to some level. [1] Automation of all the above process requires huge change to the existing architectures and human effort to tackle the large volume of data. Information Retrieval (IR) systems focus on knowledge filtering. It works on the principle of retrieving what the user exactly needs. It completely represents the domain knowledge to satisfy those needs. The efficiency of any IR system is its ability to translate queries into meaningful search operations. The document retrieved should save both the time and energy of the end user. The amount of time taken to fulfil the query shows the performance of the system. In addition, it is the next major issue in IR systems followed by relevancy of retrieved documents. Therefore, in any IR system the user becomes the prime element [2].

With the advent of social media, IR systems have become more advanced. The amount of natural language data in the World Wide Web is increasing exponentially. To represent the domain knowledge we need semantic Information systems. It helps to connect similar data points and form a network of sensible data. This becomes the knowledge graph. It presents the big picture of state of affairs of the domain of consideration. In order to filter the information overload resulting from explicitly portraying a domain, semantic IR systems are required. Semantic IR systems take the query entered and approach them along with the context. The machine understands the needs of the user from the users' point of view. The end user will prefer this kind of system as it not only reduces the information overload but also provides relevant guidelines for achieving the desired information. This survey reviews the various IR approaches with respect to semantic information systems. The reminder of the paper is organized as Section 2 deals with various existing research works related to IR systems and common approaches to information filtering. Section 3 deals with core survey of semantic IR approaches and their various advancements. Section 4 discusses the pros and cons of techniques reviewed in the previous section. And Section 5 concludes the review with findings of the survey.

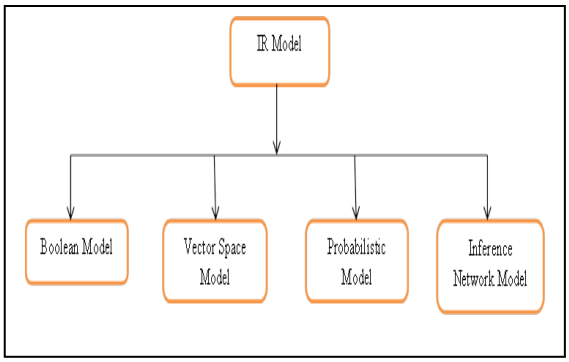

Fig 1: *IR Models*

- *M.Geetha is currently pursuing Doctoral degree program in Computer Science in Bharathiar University, Coimbatore, Tamilnadu, India.*
- *Dr.N.Vimala is currently a Assistant professor in the Department of Computer Science, LRG Government Arts College for Women, Tirupur, Tamilnadu, India.*

2. LITERATURE REVIEW

Initially the general IR techniques in big data frameworks will be analyzed. IR models as given in fig 1, are divided into four types such as, Boolean model, Vector space model, Probabilistic model, and Inference model. In general, Information retrieval (IR) is to filter document information from the storage domain in response to user query. Many IR models in commercial setups use boolean techniques. However, it has some drawbacks that need to be rectified. These models do not have the ability to deal with precision errors, subjectivity of query, which are natural issues in IR systems. [4]The fuzzy IR approaches are emerging research area. In natural language searching, Boolean operators and relevance ranking are used. The evaluation method called, batch retrieval where, documents, queries and relevance judgments are analyzed to find out the efficiency of various systems. This method performed better than human searchers. Query expansion can be used to tackle the problem of relevance scores. [5] Fuzzy ordinal linguistic IRs uses extended Boolean queries. It automates learning criteria using multi-objective evolutionary algorithms such as, non-dominated sorting genetic algorithm, and strength Pareto evolutionary algorithm. The vector space model (VSM) [7] technique always represents both documents and relevant queries as vectors. Vectors along with weighted term frequencies are obtained and cosine similarity is used to obtain similarity between query vector and documents Word based similarity alone is studied this will be an issue when one word has multiple meanings. [6]Term weighting is used in semantic space and query building, stripping suffix and using more than monosyllable words. VSM with Latent semantic Indexing improves performance of retrieval techniques. LSM can adopt to synonymy with ease therefore performs better than key-word based similarity techniques. In addition, performance can be further enhanced by combining semantic structure analysis and raw VSM. To avoid words with different parts of speech, priori collapsing is carried out in syntactic categories. It improves secondary performance gain factors. It also allows for using user feedback after query processing for long term domain knowledge. Computer Aided Design systems for commercial establishments [8] resulting in high volume databases of assemblies. The partial matching technique of VSM model is used to improve this assembly retrieval. In addition to this, the documents are ranked with cosine ranking formula according to its query similarity. Assembly retrieval is extended with parts based matching and greedy matching algorithm to solve bipartite graph matching problems. However, it doesn't consider sequence parts of query. [9] In another work, prosodic information is embedded in VSM models to facilitate query expansion. [10] Linked dependence assumption can be improved using term dependency based probabilistic retrieval model. The query was expanded with chow expansion and dependency structured index system. Dependency parser was used to derive dependency between terms in query terms. Estimation of relevant and irrelevant classes is an issue in IR systems [11] Probabilistic IR model with DCM document modelling and estimation techniques were proposed. The two components were responsible for efficient information retrieval and exact ranking. DCM helps in relevance feedback mechanism by using negative feedback to improve retrieval algorithm. This considers only the burstiness of repetitive terms and not related but different terms. [12] Okapi BM25, logical model, language model, latent semantic indexing are used for document ranking in IR systems. Term proximity information is found to improve performance of retrieval function. It is used for XML document retrieval that has an interesting internal structure in the form of tree. To model this type of information retrieval system, term proximity is calculated in a distributed environment with different resources. [13] Query dependent features and query independent features are applied to solve the problem of relevance and granularity. [14] In natural language processing based retrieval system the query in the first step is fed in to question processing stage. It extracts entity information from the query terms and other semantic data. Then the question is classified according to the domain, topic and the entity of focus. Case-Based Reasoning [15] is becoming more and more difficult due to the presence of complex case retrieval data types that hinder similarity computation. Therefore, Bayesian Network calculates probabilistic inference instead of calculating similarity, this is to avoid data type transformation. The BN technique is used for systems using big data, with WC algorithm to assign computation task in a parallel manner to achieve parameter independence test. Then probability learning is measured with WICDD algorithm that improves the efficiency of case based retrieval. [16] Knowledge graphs are vital for any IR system in aiding query expansion and corresponding result ranking. It also has its applications in recommender systems, relation extraction and question answering systems. Here, inference is concerned with estimating the missing relation between various entities to facilitate query construction in knowledge retrieval systems.

3. SEMANTIC INFORMATION RETRIEVAL TECHNIQUES

This portion deals with the core area of the paper, the Semantic Information Retrieval (SIR) techniques. According to the Fig.2, the SIR techniques can be grouped into two types, Indexing Techniques and Ontology Search Techniques. Each type of techniques will be reviewed in detail in the following paragraphs.

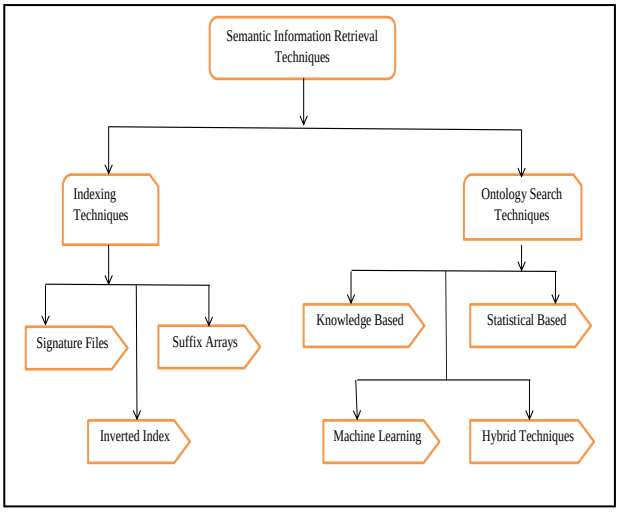

Fig 2: *Semantic Information Retrieval Techniques*

In general, Semantic analysis [17] is defined as the vector representation of words or in some cases documents that has a document corpus as the knowledge base. [18] This comes under the field of information extraction. The major problem with information extraction, especially from natural language texts is the presence of words that are polysemic and synonymic. The words will often have syntactic dependencies among neighbouring terms and acquire different meanings in different contexts. This word sense disambiguation plays a major role in effective implementation of SIR systems. These systems should be able to extract relevant natural language information and also support both the keyword and natural language queries. [19] The application of data mining algorithms in information extraction such as named entity extraction is an interesting research direction in SIR.

3.1 Indexing Techniques

The search engines used nowadays matches the keywords in the query with the words in the documents. It retrieves the matching documents and ranks them using a ranking algorithm [20]. When the query becomes complicated, a semantic search will first deduce the similarity between the query elements and entities in the documents. To achieve this, a special type of indexing is required, that could model entities, its types and keywords. Using this indexing a suitable data structure is customized to accommodate Boolean retrieval and ranking. This type of framework can retrieve semantic information with keywords and annotated queries. The speed of constructing such index, amount of storage requirements and time taken for processing queries is taken as the measurement factor for optimality. [21] SemIndex+ is a semantic indexing and retrieval framework that facilitates, result selection and ranking for any type of data, structured, unstructured and partly structured data as in NoSQL.

3.1.1 Signature Files

These are abstract documents to filter lengthy documents and to facilitate number of page access counts for a given query [22]. Signature files are constructed from the bit strings generated from the words contained in a document that are again used for building the index. In large databases, the applications require this type of files to enable efficient information filtering. To achieve a dynamic storage, 2-dimensional dynamic signature file is constructed with the help of multilevel extendible hashing and frame slicing techniques. It supports effective insertions, deletions and updations. If the signature files are segregated into number of frames, the filtering efficiency is increased. [23] The signatures are in fact properties of documents stored sequentially as a separate file in the database. When querying, only the signature files are searched and other files are discarded. When the query signature weights are not up to the threshold levels it results in performance failure. To avoid this, frame slice approach is applied to organize signature files, the resultant file is a hierarchical signature file. This kind of file organization increases retrieval time and reduces storage overhead, even when the data distributions are uniform, normal or exponential.[24]These files can also be made as parallel files to retrieve huge data in big data scenario. Vertically partitioned parallel signature file has extendable hashing and frame-sliced signature file technique designed for dynamic retrieval environments. It works without the disadvantage of data skew and frame by frame processing. This technique performs better than Hamming partitioned parallel signature in space and time complexities.

3.1.2 Inverted Index

This works in the same way as bag of words models, it represents every word in the document as a keyword in inverted index [25]. Every word is recorded with the respective document and its corresponding location. Therefore, when the query contains a keyword present in the inverted index, it is translated to the specific location of the document. The major difficulty in maintaining an inverted index in a dynamic space such as search engines is that the number of pages related to a particular key word changes often and also the web document content gets changed. It is highly desirable in situations that are more concerned about response time for each query, as the required document is retrieved in one search itself. Related inverted files are constructed as an extension to the inverted file structure. They are built after appending semantically related words in the inverted index. The semantic relation is determined using pagerank values that improve system efficiency. [26] Inverted index can be further extended by incorporating full text information by forming semantic knowledge base and constructing a hybrid retrieval model. SemIndex is a hybrid of generic semantic network and conventional inverted index characterized as multi-graphs. [27] There is an interesting research gap in this kind of IR systems in the form of partial queries in large dictionaries. To solve this n-gram indexing and bit-sliced signature file compression is proposed. The smaller signature file improves the search efficiency of the model to enhance flexibility in memory saving.

3.1.3 Suffix Array

This is a tree based data type to solve string related problems, called suffix trees [28]. A suffix array is similar to a suffix tree, instead it consumes very less space. It is called a suffix array since, it contains pointers to the suffix terms of the text arranged in alphabetical order. Suffix is the starting point of a string and searching is done with the help of binary search. There is a trade off between higher word segmentation and information retrieval performance but it do not guarantee improvement of efficiency. Suffix arrays are used to solve this issue. Suffix arrays calculate term and document frequency for every n-gram in the data corpus. In the next step, a filter algorithm is implemented to reduce the size of n-grams used in n-gram indexes. Previously uni- and bi- grams were used without word segmentation. Now it can be extended with n-grams. Suffix trees also eliminate the need for computing term and document frequency. It is also found that longer n-grams improves precision score of the IR system. It could be seen in future, that if length weighting can further improve IR performance along with longer n-grams. [29] Selfindex and Compressed Suffix arrays are another major research direction in semantic information retrieval techniques. It is applied in semantic web, where data is in Resource Description Format (RDF). The file structure is in triples, composed of a entity, its property and its value. For faster retrieval, some compact storage schemes facilitate the

presence of entire dataset in the main memory itself. When adding index structure with the data, the compactness gets affected when SPARQL queries are given. Therefore, a Selfindex is built that could attach index with the data in a single representation and also consumes limited space when compared to raw triples. Selfindex also supports SPARQL queries. The compressed suffix arrays works in an interesting manner by arranging the triples in cyclic string format and optimize them according to their domain. Using compressed suffix arrays reduced the space requirement by half. It is also useful to solve graph patterns by providing stable response times.

3.2. Ontology based Search Techniques

Web page interlinking has resulted in enormous amount of data in the form of n-number of links for a given query and is also expected to produce only the most relevant search results to the end user [30]. Ontology is the methodology to describe entities and their relationships of a particular domain in a machine-readable format. It is also known as domain ontology. It can be thought of as a method to provide meaning, rules and constraints to words used in the semantic web. When searching is executed in semantically, the results are plenty, such as, gathering relevant pages, semantic analysis, accurate ranking in the descending order and respective similarity to the given query. It is widely applied in expert systems, decision support, prediction, control and repair.

3.2.1. Knowledge based Techniques

The application of semantic retrieval in mechanical domain is considered very important while designing a product [31]. This is a challenging task given the vastness of the domain knowledge, also, the conventional keyword search or other semantic methods are not sufficient. As an alternative, ontology based search system was proposed. It directly holds the entire knowledge base, thereby improving the IR performance. It derives the original intent of a given query using domain ontology, which is a novelty of this framework. The normal keyword oriented query is transformed into a Boolean retrieval model that has weights assigned on each keyword. The mechanical domain ontology is used to assess the inherent knowledge for end user satisfaction. The conventional semantic keys could be extended with sophisticated semantic keys with the help of domain ontology. However, excessive query extension might lead to incoherent keywords and error prone retrieval. [32] IR systems are also useful in achieving search engine optimization. In question answering system, the user questions are subjected to multiple levels of transformation to convert them to relevant queries to retrieve exact response. Therefore, the ability to produce exact responses instead of list of responses is seen as desirable trait in IR systems. Triple extraction algorithm is considered efficient for such type of problems in RDF databases. [33] The structural complexity of enormous databases and semantically inter-related data pose a major challenge to lay users expecting to retrieve required information. In order to bring results expected by a user, interactive query construction is explored. It can be made possible using ontologies and by introducing a specialized interface to cater to complex search requests. Ontology based approaches are concerned with ontology modelling, ontology processing and converting knowledge base of the ontology into appropriate search queries. When working with ontologies, the things to be considered are, loss of data while transforming ontology to database, mapping between different ontologies or within the ontology and usefulness of domain knowledge. To construct ontology for big data, the data model has to be semantic. It should elaborately represent the domain knowledge and all its linked entities. Even meta data can be incorporated into existing ontological knowledge base to extend the functionality of the domain ontology.

3.2.2. Statistical Techniques

Statistical techniques are used for extracting terms related to domain, their concepts and the implicit associations [34]. Natural language processing techniques are used to achieve ontology learning from the initial stage to final. Ontology, after it is constructed, is to be presented in a formal manner to aid in better retrieval. In order to achieve this, Inductive Logic Programming is used. It simplifies the domain logic and offers other algorithms to formal representation. Ontology learning can be grouped into three categories, such as, social, statistical and logical. The statistical techniques consider only the statistical information present in the knowledge base, instead of semantic information. Mostly these techniques are probabilistic and applied right after the data pre-processing stage. Some of the application areas are entity and concept extraction and taxonomical relation extraction. The of the techniques are named as, contrastive analysis, term subsumption, co-occurrence analysis, ARM and C/NC value. C/NC value, Co-occurrence analysis and Contrastive analysis are used for term and concept extraction in ontology. When C value is used in the place of pure frequency, the precision value is increased. While using contextual information, real terms get to the top of the list and facilitate extraction of multi-word concepts. [35] Ontology is useful in so many situations like, vocabulary representation, concept reusability, semantic IR, multi-agent interaction and structuring the unstructured domain knowledge. One of the challenging aspect with respect to ontology based search techniques is learning automation. Constructing a full-fledged ontology automatically using a machine-learning algorithm is still a research gap. However, two algorithms are considered, that could achieve auto-learning to some extent. Singular Value Decomposition and Latent Drichlet Allocation are the two algorithms available to model topic information in the form of ontology. It first extracts the statistical relationship among terms and document containing the terms to build topic ontology. Then ontology graph is generated with lesser human involvement. Initially, terminology ontology is constructed to aid in semantic query optimization in effective knowledge management. Some of the other algorithms employed in topic ontology learning are, Pachinko allocation model and Hierarchical Drichlet allocation. This type of learning has its applications in Topic tracking, E-commerce, knowledge engineering, natural language processing, artificial intelligence and education.

3.2.3. Machine Learning (ML) Techniques

In order to figure out the mapping weight between two different entities in ontology, machine-learning classification algorithms are used [36]. To achieve semantic search in a distributed manner three steps are required, selection of apt resource, query alteration according to ontology alignment, ranking retrieved data and fusion. Ontological mapping can be achieved with the use of machine-learning classifiers and terminology based classifiers. Automatic resource selection among multiple resources is a field of concern. For which, query based sampling of ontological knowledge base is tried and applied. A rule based sophisticated language and ontology web language for individual domain modelling is very important nowadays. Multi-level classifiers can be employed to attach semantic information through various terminological resources. The efficiency of ML based IR system can be further improved with measures like, KL-distance. However, selecting one particular classifier for similar tasks is difficult than finding out a suitable combo of multiple classifiers for the same.

3.2.4 Hybrid Techniques

Searching techniques for hypermedia application domains is a field gaining importance [37]. A hybrid spread activation technique can be applied in a graph of hybrid instances. The interesting notion about this technique is every relation will have two aspects, one is the label from the domain ontology definition and a weight assigned from weight mapping processes. It first investigates the concept graph by aggregating the inter-related concepts by traversing through the links in the graph. Insights are obtained naturally, since, there exist some nodes, which are not explicitly connected but derived from connections in the initial nodes. One major problem associated with hybrid techniques are, they do not consider the semantic value of concepts in the ontology or knowledge graphs. Not all the derived inferences or insights have to be true, some may provide false conclusions. Relevance feedback algorithm can be used to update the system based on dynamic performance analysis. So that, after testing, the system will become capable of learning the good configuration and preferred paths for a specific domain.[38] In question answering systems, when user query is ambiguous it is difficult for the system to run query translation.. Mostly, the system retrieves syntactic responses. A hybrid technique for semantic response retrieval was proposed using semantic fuzzy ontology. Initially a hierarchical ontology is built and fuzzy co-clustering algorithm is used to extract desired response that matches the query. Along with this, a fuzzy scale arranges the answers based on their priorities using fuzzy co-clustering. This kind of system achieves deep search and surface search as well. The increasing number of ontology repositories calls for effective search mechanism of ontology with appropriate keywords. Swoogle works based on page rank algorithm. However, it has a drawback due to the poor inter-reference between ontologies that lack quality. There is an interesting improvement of the retrieval technique by introducing a measure called semantic closeness measure. [40] Hybrid ontology is constructed to identify query uncertainty using fuzzy logic. A fuzzy ontology is constructed from the predefined concept hierarchy. A clustering technique is used to auto-create fuzzy ontology with the help of Formal Concept Analysis (FCA). It enriches the existing ontology with resembling reasoning technique.

4. DISCUSSION

When dealing with semantic information, it is highly important to construct queries that could fetch semantically relevant documents than a syntactic retrieval. The efficiency of any IR system is its ability to translate queries into meaningful search operations. Query expansion can be used to tackle the problem of relevance scores. Fuzzy ordinal linguistic IRs uses extended Boolean queries. It automates learning criteria using multi-objective evolutionary algorithms such as, non-dominated sorting genetic algorithm, and strength Pareto evolutionary algorithm. VSM with Latent semantic Indexing improves performance of retrieval techniques. LSM can adapt to synonymy with ease therefore performs better than key-word based similarity techniques. Probabilistic models have query dependent features and query independent features that are simultaneously applied to solve the problem of relevance and granularity. Knowledge graphs are vital for any IR system in aiding query expansion and corresponding result ranking. Word sense disambiguation plays a major role in effective implementation of SIR systems. These systems should be able to extract relevant natural language information and support both the keyword and natural language queries. Indexing techniques are concerned with the speed of constructing an index, amount of storage requirements and time taken for processing queries to measure optimality.To avoid insufficient query weights in signature files, frame slice approach is applied and the resultant file is a hierarchical signature file. This kind of file organization increases retrieval time and reduces storage overhead. The major difficulty in maintaining an inverted index in a dynamic space such as search engines is that the number of pages related to a particular key word changes. Partial queries in large dictionaries problem can be solved by n-gram indexing and bit-sliced signature file compression. Interactive query construction is explored to achieve exact retrieval in one search itself. It can be made possible using ontologies and by introducing a specialized interface to cater to complex search requests. Constructing a full-fledged ontology automatically using a machine-learning algorithm is still a research gap. Singular Value Decomposition and Latent Drichlet Allocation are the two algorithms available to model topic information in the form of ontology, though with some human intervention. Semantic closeness measure can improve ontology search with precise keywords.

5. CONCLUSION

Big data has resulted from the data explosion. Information retrieval and information filtering are used interchangeably, due to the nature of work they perform. Instead of a conventional syntactic retrieval, semantic retrieval would be more meaningful. This survey aims to review various techniques related to modelling semantic relevancy. As stated in the paper, polysemy and granularity mismatch are major problems hindering an effective IR system. After analyzing the pros and cons of various techniques the different research gaps emerge and probable solutions are also noticeable. Inverted indexing techniques are best suited for structured and unstructured data and dynamic

SIR systems. Signature files solve the problem of time complexity while searching large databases. They also support parallel processing. Suffix arrays are best suited for RDF databases and works well with compact storage space. Ontology based SIR techniques helps to achieve accurate query expansion. Statistical techniques are mainly concerned with constructing ontologies that in turn will enable effective semantic search. Hybrid techniques are an emerging field with respect to semantic searching, where the scope for future research is plenty. Thus, this review will be useful for researchers and practitioners in deciding about the right semantic information retrieval technique for a particular domain. Moreover, make them aware of probable issues and their respective solutions. In future, information retrieval techniques for graph databases have to be probed. Since, highly inter-connected data will be the next stage of data evolution after NoSQL.

6. REFERENCES

[1] R. Priyadarshini, Latha Tamilselvan, T. Khuthbudin, S. Saravanan and S. Satish, (2015)" Semantic Retrieval of Relevant Sources for Large Scale Virtual Documents", Procedia Computer Science 54, 371 – 379.

[2] Antonio M. Rinaldi, Cristiano Russo, (2018)" User-centered Information Retrieval using Semantic Multimedia Big Data", DOI: 10.1109/BigData.2018.8622613.

[3] William Hersh, Andrew Turpin, Susan Price, Dale Kraemer, Daniel Olson, Benjamin Chan, Lynetta Sacherek, (2001)" Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations", Information Processing and Management, 37 ,383-402.

[4] Daniel Z. & Zanger, (2002)" Interpolation of the extended Boolean retrieval model", Information Processing and Management 38 743–748.

[5] A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, (2009)" Applying multi-objective evolutionary algorithms to the automatic learning of extended Boolean queries in fuzzy ordinal linguistic information retrieval systems", Fuzzy Sets and Systems 160 ,2192–2205.

[6] karen e. lochbaum and lynn a. streeter,(1989)" comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval" Information Processing & Management Vol. 25, No. 6, pp. 665-676.

[7] Xiaoying Tai , Fuji Ren, Kenji Kita, (2002)" An information retrieval model based on vector space method by supervised learning", Information Processing and Management 38 749–764.Elseiver.

[8] Kai-Mo Hu, Bin Wang,Jun-Hai Yong,Jean-Claude Paul, (2013)" Relaxed lightweight assembly retrieval using vector space model", Computer-Aided Design 45 739–750.Elseiver.

[9] Nigel G. Ward , Steven D. Werner , Fernando Garcia , Emilio Sanchis, (2015)" A prosody-based vector-space model of dialog activity for information retrieval", Speech Communication 68 85–96.Elseiver.

[10] Changki Lee , Gary Geunbae Lee, (2005)" Probabilistic information retrieval model for a dependency structured indexing system", Information Processing and Management 41 161–175.

[11] Zuobing Xu, Ram Akella, (2010)" Improving probabilistic information retrieval by modeling burstiness of words", Information Processing and Management 46 143–158. Elseiver.

[12] Ben He, Jimmy Xiangji Huang , Xiaofeng Zhou, (2011)" Modeling term proximity for probabilistic information retrieval models", Information Sciences 181 3017–3031.

[13] Fouad Dahak , Mohand Boughanem , Amar Balla , (2016)" A probabilistic model to exploit user expectations in XML information retrieval", Information Processing and Management 1–19.

[14] Mourad Sarrouti , Said Ouatik El Alaoui, (2017)" A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering", Journal of Biomedical Informatics 68 96–103. Elseiver.

[15] Yuan Guo , Yuan Guo , K. Wu, (2018)" Research on case retrieval of Bayesian network under big data", https://doi.org/10.1016/j.datak.2018.08.002, Data & Knowledge Engineering, Elseiver.

[16] Daifeng Li, Andrew Madden, (2019)" Cascade embedding model for knowledge graph inference and retrieval", Information Processing and Management 56 102093,Elseiver.

[17] Rabia Tehseen, (2018)" Semantic Information Retrieval: A Survey", Journal of Information Technology & Software Engineering, DOI: 10.4172/2165-7866.1000241.LongDom Publishing.

[18] Licia Sbattella, Roberto Tedesco, (2013)" A novel semantic information retrieval system based on a three-level Domain model", The Journal of Systems and Software 86 1426– 1452

[19] Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan,(2008)" An Intelligent System for Semantic Information Retrieval Information from Textual Web Documents", IWCF , LNCS 5158, pp. 135–146. Springer.

[20] Fatemeh Lashkari, Faezeh Ensan, Ebrahim Bagheri, Ali A. Ghorbani, (2016), "Efficient Indexing for Semantic Search", Expert Systems With Applications, doi:10.1016/j.eswa.2016.12.033.Elseiver

[21] J. Tekli, R. Chbeir, A.J.M. Traina (2018), "SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data", Knowledge-Based Systems https://doi.org/10.1016/j.knosys.2018.11.010.Elseiver.

[22] Jeong-ki kim Choon-hee lee, Jae-Woo Chang, (1997)." Two-Dimensional Dynamic Signature File Method Using Extendible Hashing and Frame-Slicing Techniques", INFORMATION SCIENCES 98, 1-26, Elseiver.

[23] Byoung-Mo IM , Myoung Ho Kim , Jae Soo Yoo , Kil Seong Choi, (1999)" Dynamic construction of signature ®les based on frame sliced approach", Data & Knowledge Engineering 30 101-120. Elseiver.

[24] Jeong-Ki Kim , Jae-Woo Chang, (2000)" Vertically-partitioned parallel signature file method", Journal of Systems Architecture 46 ,655-673.Elseiver.

[25] shaojun zhong, min shang and zhijuan deng, (2011)" Design of the Inverted Index based on Web Document

Comprehending", Journal of Computers, vol. 6, no. 4,Elseiver.

[26] Richard Chbeir, Yi Luo, Joe Tekli,_, Kokou Yetongnon, Carlos Raymundo Ibanez, Agma J. M. Traina, Caetano Traina Jr.,and Marc Al Assad,(2014)" SemIndex: Semantic-Aware Inverted Index", ADBIS, LNCS 8716, pp. 290–307, Springer International Publishing Switzerland.

[27] Ben Carterette , Fazli Can, (2005)" Comparing inverted files and signature files for searching a large lexicon", Information Processing and Management 41, 613–633,Elseiver.

[28] Jin Hu Huang and David Powers, (2008)" Suffix Tree Based Approach for Chinese Information Retrieval", Eighth International Conference on Intelligent Systems Design and Applications, DOI 10.1109/ISDA.2008.365,IEEE.

[29] Nieves R. Brisaboa, Ana Cerdeira-Pena, Antonio Farina, and Gonzalo Navarro,(2015)" A Compact RDF Store Using Suffix Arrays", SPIRE, LNCS 9309, pp. 103–115, 2015. DOI: 10.1007/978-3-319-23826-5 11.

[30] Prerna Parmeshwaran, Juilee Rege, Sindhu Nair,(2015)" The Use of Ontology in Semantic Search Techniques", International Journal of Computer Applications (0975 – 8887) Volume 127 – No.6.Elseiver.

[31] Songhua Ma and Ling Tian, (2015)" Ontology-based semantic retrieval for mechanical design knowledge", International Journal of Computer Integrated Manufacturing, Vol. 28, No. 2, 226–238, Elseiver.

[32] Amol N. Jamgade and Shivkumar J. Karale, (2015)" Ontology Based Information Retrieval System for Academic Library", 2nd International Conference on Innovations in Information, Embedded and Communication systems . IEEE.

[33] Kamran Munir, M. Sheraz Anjum, (2018)" The use of ontologies for effective knowledge modelling and information retrieval", Applied Computing and Informatics 14,116–126.

[34] Asim,M.-N.,Wasim,M., Khan,M.U.G. (2018) ,"A survey of ontology learning techniques and applications", Database, doi:10.1093/database/bay101.

[35] Monika Rani, Amit Kumar Dhar, O.P. Vyas,(2017)," Semi-automatic terminology ontology learning based on topic modeling", Engineering Applications of Artificial Intelligence, Volume 63, Pages 108-125,Elseiver.

[36] Umberto Straccia and Raphael Troncy,(2006)" Towards Distributed Information Retrieval in the Semantic Web: Query Reformulation Using the oMAP Framework", ESWC, LNCS 4011, pp. 378–392. Springer.

[37] Cristiano Rocha, Daniel Schwabe, Marcus Poggi de Aragao(2004)" A Hybrid Approach for Searching in the Semantic Web", The 2004 International World Wide Web Conference, May 17-22, New York, USA. ACM .

[38] Monika Rani, Maybin K. Muyeba, O.P. Vyas, (2014),"A hybrid approach using ontology similarity and fuzzy logic for semantic question answering." In Advanced Computing, Networking and Informatics-Volume 1, pp. 601-609. Springer.

[39] K.Sridevi ,R. Umarani,(2014),"A Novel and Hybrid Ontology Ranking Framework using Semantic Closeness Measure", International Journal of Computer Applications (0975 – 8887) Volume 87 – No.5. Taylor & Francis.

[40] Balasubramaniam K, (2015)" Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50 135 – 142, Elsevier.