IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 2, 2019, accepted January 21, 2019, date of publication February 6, 2019, date of current version February 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897849

The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval

MUHAMMAD NABEEL ASIM[1], MUHAMMAD WASIM[1], MUHAMMAD USMAN GHANI KHAN[2], NASIR MAHMOOD[2], AND WAQAR MAHMOOD[1]

[1]Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore 54890, Pakistan
[2]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore 54890, Pakistan

Corresponding author: Muhammad Nabeel Asim (nabeel.asim@kics.edu.pk)

This work was supported by the Higher Education Commission Pakistan under Grant TDF-077.

ABSTRACT Web contains a vast amount of data, which are accumulated, studied, and utilized by a huge number of users on a daily basis. A substantial amount of data on the Web is available in an unstructured format, such as Web pages, books, journals, and files. Acquiring appropriate information from such humongous data has become quite challenging and a time-consuming task. Trivial keyword-based information retrieval systems highly depend on the statistics of data, thus facing word mismatch problem due to inevitable semantic and context variations of a certain word. Therefore, this marks the desperate need to organize such massive data into a structured format so that information can be easily processed in a large context by taking data semantics into account. Ontologies are not only being extensively employed in the semantic Web to store unstructured information in an organized and structured way but it has also raised the performance of diverse information retrieval approaches to a great extent. Ontological information retrieval systems retrieve data based on the similarity of semantics between the user query and the indexed data. This paper reviews modern ontology-based information retrieval methods for textual, multimedia, and cross-lingual data types. Furthermore, we compare and categorize the most recent approaches used in the above-mentioned information retrieval methods along with their major drawbacks and advantages.

INDEX TERMS Ontology, text retrieval, multimedia retrieval, cross lingual retrieval.

I. INTRODUCTION

Since the commencement of written languages, people have been constantly striving to develop an efficient way to store, search and retrieve certain information. Initially, the concept of information retrieval remained limited to the library sciences. In 1945, Johnston and Webber [1] presented an idea that machine based information retrieval approaches will be common around the world in near future. In early 1970's people started using machines for instant information retrieval. But those systems were mostly designed for targeted groups such as medical practitioners, academic institutions, and government agencies. Whereas, With the advancement of technology and invention of social media platforms such as facebook, twitter and whatsapp, user started sharing huge amount of data in form of text, images, audio and video (also known as multimedia data). Likewise the data of multifarious domains such as E-Learning, E-Government, and E-Commerce started increasing exponentially due to the influx of technological devices and users. Thus, it was necessary to have a general purpose information retrieval system rather than several special purpose systems which were initially designed for a particular group [2]. The huge volume of information and its unstructured nature have made the information retrieval a quite tedious and time consuming task.

Numerous information retrieval techniques were developed to tackle aforementioned domain specific problems and many general purpose search engines were introduced that attracted many users but unfortunately they were not capable enough to semantically understand the user defined query and give the most accurate answer. Most search engines just processed the queries and provided approximate results and often returned bundle of unnecessary web pages to the user. A query might be a sentence or keyword feed by user

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Zhuang.

and search engines, like Google, provide a ranked list of relevant documents as a response to the user query. It is usually troublesome for the users to convey their complete information requirements in a precise way. User mostly uses keywords that are different from the keywords used by indexed documents in Database. For example, user poses a query, what precautions are suggested by practitioners for diabetes. In this case, the most relevant answer lies in only those documents which either contain the exact term *practitioner* or synonym of it such as *doctor* and *specialists*, thus in order to find those documents, it shall be resolved that both *practitioners* and *doctors* belong to same concept. To handle this problem, numerous methods were designed that use conceptual knowledge to help users in formulating efficient queries. Incorporation of a thesaurus-like component in IR systems is the most widely used method of conceptual knowledge. It represents various concepts of a domain along with their semantic relationships. Another method is to utilize conceptual knowledge as an intrinsic feature of IR systems. These methods have been proved very useful in domain of information retrieval, and the paradigm of IR from simple keywords based approaches to concepts based approaches. Finally, in 1998, Berners-Lee *et al.* [3] presented the idea of semantic web, in which knowledge is coupled with all information with specific format which is understandable by machine. This machine understandable semantic knowledge is integrated with the help of ontologies. Semantic web uses semantic languages such as RDF, RDFS, OWL, SPARQL etc. The proposed ontology based approach illustrated that conceptual mapping of information, such as ontologies, proved very beneficial in the retrieval of most relevant information [4].

In this paper, we have discussed state of the art and contemporary ontology based semantic information retrieval methodologies (Textual IR, Multimedia IR, Cross-lingual IR). We compare and classify most current diverse approaches used to carry out any one of the mentioned ontology based information retrieval. In the first section, we briefly describe the significance of semantics in domain of natural language processing and effectiveness in context of information retrieval. We also discuss modern knowledge representation models such as semantic maps, and ontologies which are employed to raise the performance of semantic based information retrieval. In next section, we give an overview of multifarious approaches proposed in context of textual information retrieval followed by multimedia and cross-lingual information retrieval. Then we compare the performance of several approaches proposed in textual, multimedia, and cross-lingual information retrieval. Finally, in last section, we give future directions for the task of ontology based information retrieval.

A. ROLE OF SEMANTICS IN NATURAL LANGUAGE PROCESSING

The ultimate purpose of natural language processing is to understand, and exchange facts and figures represented in certain language. In order to exploit gigantic amount of data more effectively in large context, semantics of the data are extremely significant. Semantics of the data do not only reveal the true meaning of the content but also help to discover the context of words present in the content. There are several researchers who have exploited data semantics to improve the performance of certain machine learning task. For instance, Roth and Small *et al.* [5] augmented the questions with appropriate syntactic and semantic category information and found a significant rise in the performance of question classification. Li *et al.* [6] utilized multiple knowledge resources to compute the semantic similarity among words. They investigated how knowledge sources comprising of organized semantic information and information content could be used in order to find semantic closeness of words. Moreover, Varelas *et al.* [7] focused on various WordNet based semantic similarity measures along with their applications in domain of web based information retrieval.

Exponential growth of textual data is causing word mismatch problem due to numerous form variations of particular word representing the same concept. This is why semantic information has proved an indispensable resource in domain of IR as it is extensively being employed in context of query expansion which raises the search results up to great extent. Information retrieval based on semantics, can be exploited in order to expand semantics of user's query just to make it more meaningful. Semantic extension improves the performance of a particular query in terms of retrieval results by enriching the specified query with more semantic features. Semantic based information retrieval is evolving rapidly and becoming the buzzword like "think out of the box" for all researchers and practitioners as it goes beyond the trivial information retrieval by utilizing the content semantics in order to assist the retrieval process. Semantic based information retrieval is typically performed in quest of following objectives [8]: i) To analyze and discover the semantics of content ii) To build a specific semantic pattern that shall depict the semantic features of content iii) To extend the user's query by utilizing semantic extension iv) All documents are clustered in order to produce semantic features before matching against user's query. Semantic based information retrieval can be considered as the most active area of research in domain of information retrieval.

The work of Raphael [9] is considered one of the earliest work in the context of semantic based information retrieval. They constructed Semantic Information Retrieval (SIR) system which was responsible to answer the queries represented in a confined form of English. SIR was developed in LISP programming language and possessed the ability to apprehend semantic information. The capability of this system was based on an internal model which used word associations along with property lists to parse relational information expressed in a certain conversational statement. At first, semantic content from input queries was extracted using a format-matching function. Then, system analyzed the queries to determine their types and processing them according to their types. If a given sentence was a declarative query, the system used to add most appropriate information in the model. On the other hand, if a given sentence was a question, system either used to return the answer of the question or determined the reason of not finding it in the underlay model. SIR was able to resolve semantic ambiguities present in user query and modify the structure of the model just to save computer's memory.

Moreover, semantic information has not only revolutionized the domain of IR but it is also being extensively used in several other domains such as semantic web. Semantic information plays a crucial in the construction of ontologies which are considered the backbone of semantic web. The work of some researchers who have expolited semantic information in context of semantic web is discussed below. Shah *et al.* [10] illustrated an approach for information retrieval by exploiting semantic web. They created a prototype that allowed the annotation of user's queries in terms of semantic information by utilizing a couple of existing ontologies. They were able to increase the precision over trivial text based information retrieval methods by using annotated semantic information. Similarly, Mukherjea *et al.* [11] exploited semantic web for knowledge discovery and information retrieval in context of biomedical patents. Yu *et al.* [12] focused on bringing the modern semantic web to personal and custom information retrieval by using web services. One of the major problems of semantic web is the necessity of annotation to recognize semantics. In this context many researchers have attempted to automate the semantic annotation process like Murthy *et al.* [13], Uricchio *et al.* [14], and Tosi *et al.* [15]. Dingli *et al.* [16] worked on creating seed documents through unsupervised information extraction techniques and in order to boost entire learning process. Dill *et al.* [17] developed the ''SemTag System'' that was able to tag semantic information to large corpora automatically.

In addition, researchers are also making conscious efforts to develop ontologies using deep learning [18]–[21], [22].

B. SEMANTIC NETWORK AND MAPS

Semantic networks or maps are considered as the most common knowledge representation methods [23] for structured data. Semantic networks employ directed graphs to represent concepts and their relations in form of nodes and edges. Nodes represent the concepts present in the documents and edges reveal the semantic relatedness between extracted concepts. They have been exploited widely in the domain of semantic based information retrieval.

When user poses a query through semantic networks, its relatedness against a given set of documents is determined by computing semantic relatedness among concepts extracted from these documents. As the related concepts are linked to each other through edges, it can be seen that all semantically related concepts are co-located, thus it, reduces search space and give better retrieval results. This section will briefly describe state of the art work on semantic network and maps. Ensan and Bagheri [24] proposed a semantic linking based document retrieval model named as semantic-enabled language model (SELM). Their proposed model represented the documents and queries in form of a graph in which nodes represented the set of concepts extracted from the documents and edges indicated their semantic relatedness. Their proposed model was based on the probabilistic reasoning in which conditional probability of query concepts was calculated by comparing against document concepts. For the evaluation of proposed approach, they used state of the art TREC dataset and illustrated that their semantic linking based approach outperformed keyword based approaches.

Lin [25] presented a detailed article on map displays for information retrieval. They highlighted that semantics networks and maps provide a large amount of information in a limited space and reveal the relationships between semantically related terms and documents. Cohen and Kjeldsen [26] created ''Grant System'' for the retrieval of funding sources by using constrained spreading activation. They observed a significant rise in the level of user satisfaction and a great boost in recall and precision values over previously built systems. Tang *et al.* [27] presented a decentralized peer to peer network based system for information retrieval. In their proposed network, indices of documents were distributed across the network based on the semantic of documents which were obtained using latent semantic indexing (LSI). In this way, search cost was reduced as all the semantically related indices were found at the same place within the network. Likewise, Lin *et al.* [28] assessed self-formulating semantic maps. They developed a semantic map using Kohonen's self formulating map algorithm and applied on a set of documents. The information obtained from maps enabled an uncomplicated navigation and processing of bibliographic data.

Hence, state of the art work on semantic networks and maps proves that semantics based document retrieval is way better than keyword based document retrieval.

C. ONTOLOGIES

Ontology is one of the most common knowledge representation model used extensively in information retrieval as it represents the knowledge in terms of machine readable, understandable and processable information hierarchies [29]. Typically, a computational ontology mainly consists of high level concepts or classes formed through the aggregation of domain specific terms, along with their attributes and relations. Ontology can be employed for semantic based information retrieval which is all about retrieving accurate results through query expansion, terms disambiguation resolution, document classification and enhancing IR model. All mentioned ways of ontology to assist semantic based information retrieval are briefly discussed below:

i) Query Expansion: User query is expanded with the semantically similar terms found in underly domain specific ontology [30] ii) Term Disambiguation Resolution: Multiple terms referring to same concept are resolved [31]. iii) Classifying Documents: Inducting ontological topics to classify documents and to assist semantic search [32]. iv) Enhanced

IR Model: Embedding ontology into existing IR model to get modified and enhanced information retrieval model due to the effects of semantic indexing [33].

Ontology has been employed in various information retrieval projects for different purposes, which are briefly discussed below: i) Semantic Digital Library (SDL): SDL has utilized ontology as knowledge base. The content and metadata of all documents are inserted in ontology, to enable quick search and retrieval. ii) Crime News Retrieval (CNR): CNR has mapped named entities into an ontology of news for sake of supporting semantic based information retrieval [34]. iii) Multi-Modality Ontology based Image Retrieval (MMOBIR): MMOBIR has proposed multiple ontologies such as textual ontology, visual ontology and domain ontology to describe the images in terms of textual, visual and domain specific semantic features. MMOBIR also reveals that how such ontologies can be embedded in DBpedia (An open source knowledge based) to assist a comprehensive search.

Ontology provides a structural and formal representation of data just to make it processable, shareable and reusable in large contexts. For instance, Khan *et al.* [35] employed an ontology model in quest of generating metadata for audio media type. They marked a significant rise in performance over trivial keyword based information retrieval approaches. Likewise, Soo *et al.* [36] exploited an ontology as a knowledge source of domain specific information just to raise the performance of particular image retrieval system. Moreover, Gómez-Pérez *et al.* [37] improved the performance of a legal based information retrieval system through ontology. They discovered that ontology helps the users significantly by suggesting better query words. Cesarano *et al.* [38] used an ontology to help categorize web pages on the fly in their semantic IR system. Likewise, at times, user query contains terms which belong to multiple domains, thus, it make hard for the information retrieval system to select accurate domain ontology. Undertaking this problem, Mannan and Sundarambal [39] proposed a methodology of query expansion. Each user query was expanded with the semantic terms found by aggregating several semantically related ontologies. If no semantic ontology was found then the WordNet ontology was used to expand the user query with semantic terms. Various ontology based IR domains are discussed in the following subsections.

II. TEXTUAL INFORMATION RETRIEVAL

Ontology based textual information retrieval has been cumbersome task for the researchers due to required semantic similarity between terms or concepts of query and corpus. Subsequent subsections will provide a birds eye view of state of the art ontology based textual information retrieval methodologies.

A. VECTOR SPACE BASED INFORMATION RETRIEVAL

Vector space model is also known as term vector model in which concepts of documents and queries are represented as vectors and similarity between them is calculated using cosine measure. Cosine similarity measure provides the level of closeness between textual documents vector and query vector. Besides, other prominent similarity measures like tf-idf and Okapi BM25 are also used in vector space based IR, however, they have not been exploited in ontological IR up to this time. Besides, ontologies play significant role for the extraction of concepts from documents and queries.

Paralic and Kostial [40] presented an ontology based IR methodology, by utilizing ontology for the extraction of a set of concepts from query. Their approach assumed that for a given query relevant concepts were already present in state of the art developed ontologies such as WordNet. For each document, set of related concepts were also extracted with the incorporation of ontology. Those extracted set of concepts were then compared with query concepts and a score was calculated in order to rank the documents based on similarity with user query. To evaluate the integrity of proposed system, they used 1239 files obtained from MEDLINE corpus where keyword Cystic Fibrosis was utilized for documents collection.

Castells *et al.* [41] augmented the previously mentioned ontology based IR system with semantic-based users personalization. For content retrieval, they used ontology based retrieval framework [42] and each domain concept was labeled with user personalization score. The similarity score between query terms and concepts of documents was obtained by following the above mentioned approach and then user preference score (personalized score) was integrated with it. The final score for each document was obtained and documents were ranked accordingly. Moreover, they also provided the way to adjust the degree of personalization automatically. For the evaluation of their proposed methodology, they used 145,316 documents obtained from CNN website.[1]

Ahmed-Ouamer and Hammache [43] presented another ontology based information retrieval approach for e-learning. They used vector space model for both query and documents and calculated their similarity through cosine similarity measure. Their proposed IR system, known as OBIREX, used an ontology to index a collection of documents and semantic links to enable inferences over all relevant documents. Ngo and Cao [44] proposed a generic vector space model in which Named entities (NEs) and keywords were combined to get the semantics of document text. They utilized NEs ontological features such as identifiers, aliases and classes. Entity classes were used for representing the latent or semantic information of all interrogative words present in Wh-queries, which are usually neglected in conventional keyword based searching. They performed experimentation on TREC dataset.

Zhang *et al.* [45] proposed another ontology based information retrieval system. At first, Semantics of the query were analyzed and transformed into domain related

[1] $http://dmoz.org/News/Online_Archives/CNN.com$

ontological concepts. To understand the semantics of query, they performed query expansion using domain ontology. The whole corpus was labeled with ontological concepts and then clustering was performed to find a bunch of those documents which were semantically related to the user query. Documents were then sorted and returned to the user. Results of their experiments depicted that proposed approach outperformed other state of the art keyword based approaches. Meštrović *et al.* [46] proposed ontology based approach using vector space based IR. In their approach, query expansion and document depend on the base taxonomy which was produced from Linked Data set or lexical database. They introduced a mapping function to map the ontological layers onto the base taxonomy. They introduced a weighting schema which was used for vector space model. They utilized semantic and lexical relations between concepts and terms to reduce the vector size in IR and avoid vocabulary mismatch problems.

B. PROBABILITY BASED INFORMATION RETRIEVAL

This technique is based on probability distribution rather than a cosine measure to calculate the similarity between the documents and the query posed by the user. The query posed by the user is refined and expanded with the help of domain ontologies. This expanded query is then utilized in calculating the score of documents. The Relevance of document is determined by calculating the ratio between the probability of all relevant (expanded query based terms) and irrelevant terms within the same document. Previously, researchers have put great efforts in order to incorporate semantics in probability based IR systems using domain ontologies.

Such as, Stojanovic *et al.* [47] proposed two methods for query refinement in order to support the IR system semantically. One of the proposed method suggested an equivalent feature or attribute that directed towards correlated content while the other one referred to retrieve those resources that were relevant to the user query. Finally, the query was represented based on the probability of both methods. Stojanovic and Stojanovic [48] proposed a logic based query refinement to support ontology based IR systems. They expanded the original user query using domain ontology. Afterwards, they performed a test known as acceptance test in which expanded query terms were examined. Their proposed approach was based on fuzzy probability and proceeded step-by-step for the refinement of the posed query. Zhai *et al.* [49] extended the work of Stojanovic and Stojanovic [48] and presented a concept of fuzzy ontology for information retrieval in the domain of e-commerce. Their proposed system had three components: concepts, conceptual properties, and values of those properties. The values of properties can be either linguistic values of fuzzy concepts or standard data types. They considered that queries can be expanded by combining the fuzzy linguistic variables with domain ontologies. Moreover, they presented Probabilistic latent Semantic Index (PLSI) [50] in which document scores were calculated based on query expansion.

C. CONTEXT BASED INFORMATION RETRIEVAL

This technique is based on the context of concepts such as time, location, date, and details of users etc. Recently, researchers identified the importance of such parameters and started utilizing them in information retrieval systems [51]. The query is expanded with the help of spatio-temporal domain ontologies and the context such as user profile details and geographical locations are incorporated into the query inquest of better retrieval results. In some cases, user defined query is converted into RDF triples with the help of ontologies and the metadata of documents is also stored in form of RDF triples. Then, the similarity between the RDF of query and documents are computed in order to find the most relevant documents. In such a way, precise and clear query terms are used to find the relevant documents leading to quality information retrieval results.

Liaqaut *et al.* [52] presented a context based IR system using query schema and role ontology. Their proposed approach modeled various context parameters such as user's profile, context, history, ontologies and user feedback. Moreover, they also employed term expansion techniques to expand the query terms based on online available resources such as WordNet, domain ontology, and various thesauruses. Wang and Zhu [53] focused on highlighting the contextual problems of existing IR systems and proposed ontology based multi-agent IR framework. They developed command line parameters from query posed by the user and provided privileges to the user in such a way that user can select either to use their self-built customized ontology or domain ontology for searching purpose.

A methodology for semantic based information retrieval by taking into account the context of query concepts was elaborated by Mustafa *et al.* [54]. In order to capture the context of concepts present in query, thematic similarity technique was employed. They stored the metadata of sources in form of RDF triples and used existing RDF triple to search user query. Tuominen *et al.* [55] proposed another context based technique in which they expanded query based on spatio-temporal Ontology. They considered the time and geographical location of the user as a context for the construction of IR system.

Mule and Waghmare [56] proposed a technique namely ontological indexing in context of textual information retrieval and made comparison with text based search. The proposed technique exploited the context of query words in order to improve search results. They created a movie ontology using Protege editor. They employed WordNet to find synonyms of query words, part of speech tagger for word sense disambiguation, and stored all web pages in an xml database.

D. SEMANTIC BASED INFORMATION RETRIEVAL

These techniques are based on semantics in which meaning and concepts of a query are focused instead of simple query terms, which helps in retrieving relevant information with reduced search space. In this approach, the queries are augmented semantically before they are used in the process of information retrieval. This semantic augmentation of query requires clear conceptual knowledge which is obtained from domain specific ontologies and the redundant or irrelevant information is removed. The query can be automatically augmented or the user may give privileges to semantically augment the query by using thesaurus and domain ontologies. These semantically enhanced queries are then send to the IR engines for relevant information retrieval. Fernández *et al.* [57] proposed a semantic search model that integrated and utilized the advantages of keyword and semantic based search. They contributed a novel rank fusion technique which was used to minimize unwanted effects of knowledge sparseness on the semantic web. A large scale evaluation benchmark based on TREC IR evaluation standard, was presented to allow an exhaustive comparison between SW approaches and IR. Results of their proposed methodology was higher than TREC automatic system.

Xiao and Cruz [58] presented a query processing approach for IR systems, in which they built a semantic query based on the user query using various domain ontologies. Guan *et al.* [59] implemented a similar technique in which they provided the user certain privilege to draw the input query which was later semantically enhanced by query processing module. Then, this semantically augmented query was used to retrieve documents with high relevancy. Hu and Ju *et al.* [60] presented a model to combine IR systems with ontology which was automatically learned from searched results. The constructed ontology had the facets of domain knowledge, semantic ontology, and the hierarchy of fuzzy concepts. To assess the quality of system, they followed the evaluation approach of ImageCLEF. Then, Gu and Yu [61] presented a 3-layer semantic based indexing approach to surpass the performance of conventional index based approach. They exploited semantic markups and transformed the content of documents into an effective collection of ontology terms. Their proposed approach raised the precision of IR due to improved matching of query terms with converted terms.

Jimeno-Yepes *et al.* [62] also shared his contribution by generating a query model in which concepts were set into a query relevant domain. It provided privilege to the user to visualize and manually select those concepts that were highly correlated to the user query. Dong *et al.* [63] used metathesaurus for query expansion and assigned weights to the related terms in order to accurately understand the semantics of the query. Yadav and Singh [64] presented an ontology based IR methodology in quest of document retrieval. Their defined ontology was flexible as it gave total freedom for the selection of docuemnts, evidently assisting to produce more effective retrieval results. They compared the performance of proposed approach with state of the art technique using F-measure.

E. SEMANTIC SIMILARITY BASED INFORMATION RETRIEVAL

The above mentioned methods consider only the context or concept based ontology for particular user query. However, they do not consider concepts and semantics within the documents. Therefore, a semantic gap from document's side always remains. Until the development of semantic similarity measure, the semantic IR systems will remain incomplete. Semantic similarity based models can be of great importance to reduce the gap between the concepts of query and documents. For this purpose, researchers put great effort in order to build such semantic models [65], [66]. In this approach, both query and documents are semantically enhanced to have better retrieval results. The query posed by the user is expanded and refined with the help of domain ontologies whereas documents are annotated semantically with the help of ontologies. Then, these ontology based semantic annotations are evaluated against semantically enhanced queries and a similarity score is calculated. On the basis of that score, top relevant documents having highest similarity are obtained.

Ozcan and Aslangdogan [67] proposed an innovative approach in which they exploited ontologies during IR process. Their proposed framework generated metadata and expanded query based on ontologies. Given query was expanded and results were generated and compared against already generated metadata within the same word space.

Nagypál [68] presented another model which generated semantic metadata and expanded the query based on ontologies. A simple user query was converted into a semantic query by processing it from various ontology tools and then it was used for the retrieval of data. Gu and Yu [61] enhanced the previously mentioned work by employing composite words selection instead of individual words.

F. SEMANTIC ASSOCIATION BASED INFORMATION RETRIEVAL

This is an advanced branch of information retrieval system in which association rules and ontologies are combined to expand the query which plays a key role in IR systems. In this approach, the user provided query is fed to the IR engine and documents are retrieved. These retrieved documents are then preprocessed. Functional and stop words are removed after preprocessing. The remaining content words are considered as items and each document is considered as a transaction thus, it leads to the creation of transactioned database which is used to find important associations present in the retrieved documents and then these associations are mapped to the concepts using domain ontologies. Both the associations and ontologies are combined to expand the query which is then fed to the IR engines for the retrieval of relevant information. Moreover, association rules also provide heuristics on the basis of which documents can be assigned weights.

Song *et al.* [69] presented a semantic query expansion approach that combined IR techniques and association rules along with ontologies. Their proposed approach exploited linguistic and semantic properties of totally unstructured textual corpus and utilized the context of significant terms determined by association rules. The proposed approach was evaluated on a small set of TREC collections.

Later, Hu *et al.* [70] proposed a rough ontology based IR method for effective retrieval in the non certain information space. The proposed method could exploit the keywords of query to compute query properties and individuals through a search procedure. Then it used to build approximation space for ontology driven information systems by the help of equivalence relation. Finally, it employed approximation space in order to determine the similarity among individual and document set. They compared their proposed approach against two other state of the art approaches namely OntoSCORM, and Lucene using F-measure.

G. SEMANTIC ANNOTATION BASED INFORMATION RETRIEVAL

This is an evolving field of IR systems in which documents are annotated with respect to the keywords of query which helps IR engines to find relevant documents. The query terms give us information about the documents which need to be searched. Therefore, annotating documents with key terms present in the query provide great ease [71], [72]. In this approach, user defined query is converted into RDF triple form and sent to the knowledge bases and ontologies to retrieve relevant instances. These retrieved instances are then used to label the documents. These semantic concepts present in the instances are searched in the whole text of the document. If those concepts are found in the document then the particular document is labeled with that instance. In such a way, documents are semantically labeled. After the labeling of a document, weights are assigned to the documents based on these annotations and highest weighting documents are returned to the user.

Rodríguez-García *et al.* [73] proposed a system for desired retrieval of cloud services to facilitate the user. They have performed this task in two main modules: firstly, they created a repository for could services in which each service was stored in the form of semantic vector after semantic annotation of the cloud services description. Secondly, they used Information and Communication Technology (ICT) for the user required cloud services retrieval task and to get best results from similar broad domains. Vallet *et al.* [42] utilized this technique and presented an algorithm of annotation weighting. Their proposed approach utilized an ontology based scheme to annotate the documents semi-automatically and built a retrieval system. First of all, Resource description query language (RDQL) was generated directly using input query terms. Then this RDQL query was sent to the system and list of instances were retrieved from knowledge bases which were used to annotate the documents. Finally, based on these annotations documents were assigned weights and topmost relevant documents were returned to the user.

All above mentioned ontology based IR methods can be summarized in terms of precision, recall, retrieval and processing time. Vector space model is efficiently applied to real time applications but its retrieval time is not satisfactory. However, it gives an impressive value of precision and recall. On the other hand, probabilistic models give very efficient retrieval and processing time but their precision and recall values are quite low due to which they can not be applied in real time applications where accuracy has great importance. When semantics based approaches were introduced, the performance was very impressive in terms of all parameters with efficient processing and retrieval time. Therefore, they proved optimal for many real time applications. Later, semantic similarity, associations, and annotations were introduced [74], [75] and analyzed in which similarity based approaches gain a lot of fame due to their precise results and impressive recall and retrieval time values but other approaches are still struggling [76]. Semantic association based approaches give efficient results in terms of recall and retrieval time but they neither give precise results nor satisfactory processing time. Overall we can say that semantic similarity based approaches offer best results and opportunities to bring advancement in the field of semantic web.

H. FUTURE WORK

Semantic based information retrieval is yet to overcome few problems before being adopted widely by numerous researchers and practitioners for different languages. The first problem is the availability of semantic knowledge sources. Typically, it is easy to find semantic information sources for English, however semantic resources are still deficient for other languages such as Urdu. Another problem is that, generally, information retrieval algorithms dealing with the semantics are slower as compared to trivial information retrieval methods. These problems will alleviate once researchers and practitioners will excel in their research work on semantics. Semantic based information retrieval approach will start getting used extensively once semantic web achieved its goals and automatic annotation methods become feasible.

Heterogeneous nature of web environment poses a great challenge to semantic based information retrieval. Internet contains information from almost all domains of the world. Hence, it is not feasible to use semantic based information retrieval on all types of semantic resources. Therefore, we predict that this problem will be mitigated by first identifying concepts in query and performing query classification to determine its domain. For instance, if query lies in medical domain, then only medical ontologies will be used for information retrieval and if it is from political domain, then only politics related ontologies will be selected.

Another side effect of semantic based information retrieval is incompleteness of knowledge bases. We believe that this can be catered by opting technologies such as ontology learning. We anticipate that users will be able tackle this problem by using tools that take unstructured documents as input and return fully developed ontology as output. So far, only Text-to-Onto is available as off-the-shelf tool to carry out ontology learning from text [77]. Currently, researchers are focusing to develop ontologies using deep learning [18]–[22]. Development of such tools looks mandatory in future to tackle the problem of incomplete knowledge base. Besides this, domain evolution poses another threat to the semantic information retrieval. There are domains that evolve very quickly because new concepts and terms are discovered at frequent pace. For such domains, there exists a desperate to update existing ontologies in order to keep them upto date and this task is very costly, if ontologies are manually updated. However, advanced research in the domain of ontology enrichment and extension can help in this regard.

Unavailability of evaluation benchmarks and datasets is another major hurdle in semantic based information retrieval. To tackle this problem, large scale challenges need to be encouraged. For instance, information retrieval domain flourished after TREC challenge. Therefore, such a challenge needs to be established which shall not only provide datasets and evaluation benchmarks, but shall also motivate researchers in a competitive environment to improve semantic based information retrieval.

III. MULTIMEDIA INFORMATION RETRIEVAL (MIR)

Searching, indexing, storing and sharing of multimedia data such as images, videos, sounds, 3D graphics, or their combination is a common practice nowadays. Data is being stored in personal computers and on the Internet in form of audio, video, and images. The storage rate of media data like pictures, music, and videos has been increased with the rapid growth of social media platforms such as Facebook, and twitter. Multimedia is considered more simplest way of sharing information. In order to illustrate, lets take an example of whatsapp which allows instant messaging. People share audio messages, and textual images more than trivial text messages as they find multimedia resources to be more convenient. This is why the use of multimedia resources is increasing as the speed of light, thus, making multimedia retrieval an indispensable task in domain of IR. But the question arises, how to interpret the features of images and utilize them better for sake of content analysis and retrieval purposes. To answer these questions, multimedia retrieval systems were introduced in which descriptive features from images and audio data were extracted and then mapped to high level features of the query. But the continued growth of web data has made processing and management of massive scale multimedia information quite tedious. According to a survey, in 1999, there was approximately 9 terabytes of data on web and size of Internet data doubled in every two years [78]. At the beginning of 2016, around 7.7 zeta bytes of data was reported on the web and in 2020, it is expected to reach around 40 zeta bytes.[2] Moreover, MIR is a heterogeneous field with an extensive range of research issues, methodologies and supported data types like audio, graphics, images, animation, videos, rich text, hypertext, combination of all these data types [79]. The storage rate of media data like pictures, music, and videos has increased massively over time. This makes a growing need of efficient multimedia search and retrieved methodology for the web user.

The users of the web might be able to retrieve relevant multimedia data if they are deeply familiar with the representation of multimedia contents and its structure. The user can fruitfully retrieve contents of multimedia from the web of data(SW) due to a proper and structured representation of contents. For this purpose, there is a need of some extra mechanisms to overcome the semantic gap of objects by using lexical libraries, such as FrameNet, WordNet, VerbNet, ConceptNet, etc. The premature research work on MIR was based on the research of computer vision [80]. Guo *et al.* [81] proposed an efficient algorithm namely Semantic Ontology Retrieval (SOR) which was used to retrieve multimedia ontology from diverse big data domains using a processing tool known as Hadoop. They presented a framework called MapReduce-based retrieval for parallel processing of SOR in distributed nodes. They exploited user feedback scheme for good precision and user experience. More recently, researchers have started moving from feature based retrieval to content based retrieval.

To make the MIR system more human-centered, the response of the system must be accurate for the user's satisfaction. There are several users who are using different MIR systems like Google for image and video search, Altavista for audio search, and many others in quest of different multimedia resources. Moreover, there are lot of workshops and conferences on MIR such as ACM SIGMM. Typically, MIR systems are used to fulfill two major user needs: (i) searching, and (ii) browsing with the summarization of media gathering [80]. In order to achieve these two needs, there are mainly two types of methodologies namely, featured-based and category-based. In recent times, category-based methods have got the popularity due to being able to express media semantically which is considered a fruitful facet in IR.

In recent times, researchers have moved towards content-based approaches [82] and technology is improving at a phenomenal pace. Moreover, non-textual data is becoming more common as compared to textual data and in near future, it will become a common format for sharing information. By considering these trends, it is necessary to have a detailed review of all the state of the art retrieval approaches of non-textual, multimedia data. Text retrieval systems such as web search engines are well established and publicly available whereas multimedia IR systems are less established. This section presents an overview of multimedia retrieval that will serve as a great opportunity for the researchers to bring remarkable breakthroughs by gaining significant domain knowledge in a systematic way.

A. IMAGE RETRIEVAL

Since last decade, the invention of digital cameras has increased the trend of photography in different fields like

[2] http://www.live-counter.com/how-big-is-the-internet/

fashion, medicine, publishing, architecture and remote sensing etc [83]. A huge amount of digital information is shared on the web on daily basis which is mostly in the form of images because visual information is more effective and easy to grasp [84]. This huge volume of information has made the search of the required image very hard especially for the queries where the user just asks, "show me images of yellow "Lamborghini". To answer such queries and manage a huge database for effective and efficient retrieval, there is need to develop methodologies. The image can be retrieved by using three methodologies: text-based image retrieval(TBIR), Content-based image retrieval(CBIR) and hybrid image retrieval method, which are discussed in following subsections.

1) TEXT-BASED IMAGE RETRIEVAL (TBIR)

TBIR systems are mostly used for image retrieval in web services [85]. This technique uses text which is associated with the image as a file name, hyperlink or annotation. This text usually demonstrates what exactly image contains. When a user enters a textual query, first of all, various techniques are applied to it to resolve polysemy problem and then keywords are extracted from it. These keywords are then used to tag the query. The database contains indexed or tagged images which are compared against annotated query and maximum matching images are retrieved as a result. The overall architecture of text based image retrieval is shown in figure 1.

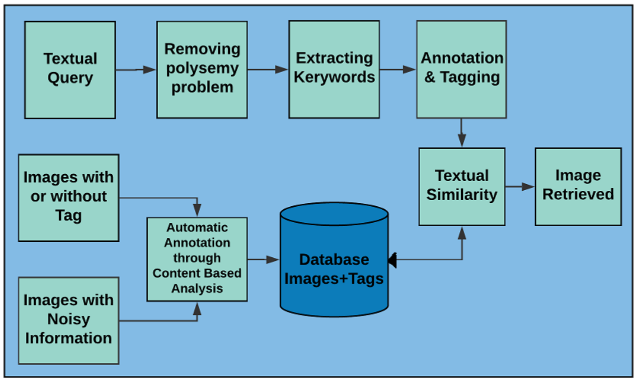

Textual Query
Removing polysemy problem
Extracting Kerywords
Annotation & Tagging
Images with or without Tag
Textual Similarity
Image Retrieved
Automatic Annotation through Content Based Analysis
Database Images+Tags
Images with Noisy Information

FIGURE 1. Overall architecture of text base image retrieval.

Yahoo and Google are famous example of search engines for images having almost one billion indexed images [86]. These engines are robust and fast but sometimes they produce an irrelevant result due to many reasons, such as redundant words in the textual description eventually leading to a low precision rate in image search. Another reason of irrelevant result production is polysemy problem [87] in which one word is used in order to reference of multiple objects. Soo *et al.* [36] presented an ontology based system to retrieve the image of Chinese culture. It was a text based approach in which textual queries were sent to the systems using ontologies. In their proposed approach, they converted both query and images into RDF structure and then similarity was calculated between the triples of query and images. The image with the highest similarity was returned to the user.

kara *et al.* [88] proposed an ontology based IR system with its application in soccer domain. They mainly focused on three hurdles in semantic search like as scalability, retrieval performance and usability. They improved the retrieval performance using inferencing, rules and domain specific information extractor. A model for semantic indexing which is based on Apache Lucene, was presented by them to enhance the ability of keyword based search by providing extracted and inferred information with the help of domain ontology. They used this indexing model for solving the problem of structural ambiguities. This system achieved maximum aspects of the semantic web due to efficient semantic indexing model.

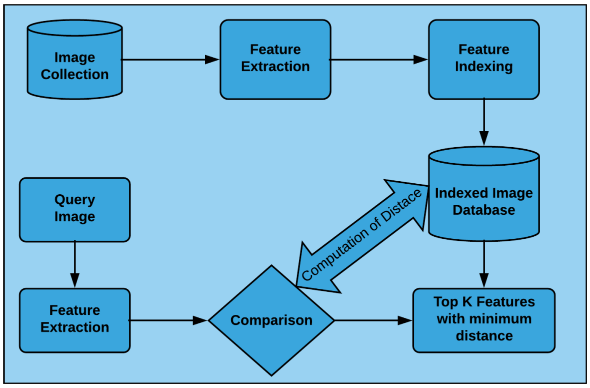


FIGURE 2. Overall architecture of content base image retrieval.

2) CONTENT-BASED IMAGE RETRIEVAL (CBIR)

CBIR is a picture retrieval methodology. It is based on the process of indexing and extracting of low-level characteristics of images. These low-level characteristics are like shape, color and texture for supporting visual queries in an instinctive way and automatically index images with content descriptors [89]. Figure 2 illustrates the general methodology of content-based image retrieval approaches which is summarized in this section.

1) Initially image collection is preprocessed and features are extracted.
2) Then feature indexing is performed and stored database is obtained.
3) Similarly, image query is processed and features are extracted.
4) Finally, distance is computed between features obtained by query and features stored in the image database. On the basis of this distance, images are ranked and topmost images are retrieved.

Most state of the art systems take into account every image as a whole; however, a picture can have numerous objects or regions with separate semantic sense. A user is mostly interested to search one region of the image rather than whole image. Hence, instead of observing every image completely, it is very logical and decent to view every image as a set of regions. In CBIR the word *content* might deal with texture, color, shapes or spatial orientation that can be acquired within the image. The images related to web search engine depend on meta-data and produce garbage results, so CBIR is useful in this case because CBIR search an image by the help of image contents and these contents are similar like meta-data of image. The features employed by most IR systems include spatial layout, texture, color, and shape [83]. If these features are extracted from the whole image then these are not much fruitful for CBIR. The CBIR methodology uses the methods of computer vision to retrieve digital images from database. Content based retrieval performs the inspection with actual content and does not take into account annotated tags or keywords [90]. Podder *et al.* [91] proposed an ontology-driven CBIR system which was responsible to retrieve semantically and structurally similar images (using bag of words model) from heritage image dataset. For this work, a technique Locality-sensitive hashing (LSH) was used to determine nearest neighbor images. They used a manually build hindu heritage ontology using protege.

Vijayarajan *et al.* [92] proposed a web model for information extraction in the form of object, attribute and value (OAV) to refine the keywords of user query. Different image search engines like Google use content based approach for image extraction, however, content base image approaches face semantic gap problem, thus, might not be able to extract most relevant images. To overcome the problem of semantic gap, they built a domain ontology using images description. Moreover, they proposed a model which examined the subject, predicate and object (SPO) from user query using a NLP parser. Then the acquired SPO information was used generate the SPARQL query automatically which was deployed in ontology based image retrieval approach

A technique based on HCI (Human Computer Interaction) was used to take relevance feedback from the user for guiding the retrieval system causing an increase in retrieval performance [93]. A technique has been used for searching the image using content based image retrieval approach after annotating the image with appropriate content [94]. Several researchers have done significant work in the field of MIR. For instance, Smeulders [95] categorized the queries of content based image retrieval into three classes, (i) when user does not decide which kind of image he/she want to retrieve, called retrieval by association, (ii) when the user has a specific goal for specific object or image, called retrieval by targets and (iii) when the user wants to retrieve a picture of an object like ‘‘a picture of car’’. They utilized image texture and color features. In this approach user allocated the weight to each feature and then computed the similarity with combined texture and color features. This approach got higher retrieval accuracy than conventional methods.

The work of Corridoni *et al.* [96] for image retrieval was based on semantics of color using low level properties like numerical characteristics of color, texture features, and high level properties of color such as the ability of color to describe something like sensations. Kato *et al.* [97] introduced a retrieval system based on sketch similarity. In their proposed system, query by visual example (QVE) accepted a user drawn sketch and through semantic techniques, they retrieved a relevant image.

Natsev *et al.* [98] worked for computing similarity between database and query image using several signatures. In Conventional approaches, only signature of the image based on color was used, but Natsev introduced an algorithm namely WALRUS in which query image was used to match with database image and this algorithm was very efficient to scale and translate the objects present in an image. In this technique, images of both query and database are decomposed into particular regions and then the similarity between them is computed using the signatures of both images.

Camille Kurtz *et al.* [99] proposed a system for semantic retrieval of biomedical images using a strategy in which they combined the features of image content and ontological semantic dissimilarities. For this purpose, Multi scale Riesz wavelet was used for the extraction of low level image features. They used Radlex ontology for automatic semantic annotation and achieved high level features to overcome the semantic gap problem. The similarity between annotations of images was calculated in order to retrieve similar images using a measure called dissimilarity measure. Finally, after retrieving similar images, they combined both strategies (Computing features of image content and Image annotation) in quest of accurate image retrieval. They claimed that presented work will be helpful for the radiologists to get similar biomedical images with their associated diagnoses and responses to therapies. Chang *et al.* [100] and Rui *et al.* [101] improved the performance of the visual information incorporating human efforts in retrieval process. The CBIR system substantially faces the problems of ‘‘semantic gap’’, and computational load.

Mostly, CBIR system performance is intrinsically constrained by the visual low level features(which are used for describing an image) because they are not able to express the concept of high level(which human recognize). This is called the problem of semantic gap. In the retrieval process of images, user feeds the system images as an example, these example images are converted into a feature vector and compared with those database feature vector [102]. while the visual extracted features are objective and natural, there is remarkable space between low-level features relating to images [103] and high-level concepts of queries. For example, a user enters a query, show me images of black Limozin. This query contains high level concepts such as Limozin is a car and database contains various images annotated with low level features such as cars. Thus, a semantic gap gets introduced automatically between high level and low level features, for which system must be capable enough to understand that Limozin is a special type of car, not every car is Limozin.

The conventional CBIR merely depends on comparison and extraction of primitive features. It does not understand the semantic content of the image and can not fulfill the needs of users due to the rich human semantics and semantic space of features [104], [105]. The reduction of the semantic gap has been active research topic in CBIR [106]. Current research provides some approaches to reduce the semantic gap which are mainly categorized into (1) relevance feedback (2) automatically image annotation.

Relevance Feedback (RF): is a user response based on-line process in which semantic space between semantic concepts and features is reduced. The query is refined and re-evaluated in every iteration by using feedback of the user against particular image result concerning its relevance to the specific target image. Primarily image retrieval procedure with RF consist of four steps [107].

1) Images are shown to the user after retrieval
2) User indicates about non-relevant and relevant images
3) The system learn from user feedback about user needs
4) Based on user feedback, new set of images are retrieved and furnished

This iterative process will continue until the desired image result is obtained. The small set of labeled samples by the user is the main reason for poor classification of image database [108]. The performance surety of RF is fully dependent on the proficient standard of the top-N images which are used by RF. This process might be insignificant if there are only a small number of images returned [109]. It might fail to provide a large number of relevant images even if the user provide a lot of feedback. This is because user's target object is impossible to represent by the composition of obtainable features [110].

Automatic Image Annotation: Manual annotation of large size data is very hard and time consuming. In manual annotation, user supervision is needed, but some tools [111] are available to accomplish this task easily without user supervision. Images are automatically organized into predefined classes (keyword). Low-level features extracted from training images and classifiers are assembled with them for providing to the concept decision. Finally, the instructed classifiers, classify the new occurrence and automatically annotate the ignored images [112]. A lot of methods are available now for automatic annotation which can be grouped into two categories [113]. **(i) Ontology based**, the hierarchical representation of concepts along with relations is called ontology. It is same as classification with a keyword, but in reality, the keywords being a part of the taxonomy enriches the automatic annotation. **(ii) Keyword based**, arbitrarily selected keywords from supervised vocabularies are used for image description. Since our survey on information retrieval is ontology based, we will further describe image annotation based on ontology.

Ontology is a hierarchical structure of concepts in the form of taxonomy. The components of a concept are terms, which are used to refer to a concept, attributes for detail description of concepts and relationships of concepts. The terms have attributes and relations, these terms are represented in ''is a'' relationship. These relations are helpful to determine high level conceptual relations. Image retrieval through ontology is fruitful because it undertakes semantic relations and nor only overcomes the semantic gap and also provides accurate retrieval. Ontology building and annotation in detail are tedious tasks in this field. The rich ontology with semantics refines the precision for retrieval. The deficiency of textual information influences the approach of keyword based retrieval. These approaches stuck in the problem of mismatching between the terms and not fruitful at all because they do not examine the image content features. For this reason, ontology is good with the collaboration of visual features of an image.

3) HYBRID IMAGE RETRIEVAL

In order to take advantage of both textual and content based approaches, researchers are now combining the two modalities of web images, visual features and textual context for retrieval [114]. The joint use of visual features and textual context can provide good results [115]. In this hybrid approach, frequency of words is used for indexing purpose, however, in order to treat text and image as same data, extra weights could be assigned to all those words which occurred in the src and alt tage of the image. These approaches [116] generally learn the correlation of keywords by analyzing the occurrence of keywords in the lexicon for keywords semantic meanings like as synonyms [117] or web images annotations. As a result, perfect composition of traditional content and text based approaches are not adequate to tackle image retrieval problem on the web.

B. VIDEO RETRIEVAL

Digital data is growing tremendously due to the remarkable advancement of digital gadgets which are used to capture them. Digital data contains documents, images, sounds, and videos etc. Video is an important digital media which contains a rich source of information as it contains all the other digital data such as images, sound, and text. The retrieval of information from video database according to the need of the user is called a video retrieval process. In order to manage such an immense database, efficient information retrieval approaches are required. This section briefly describes state of the art video retrieval (Text base, Content base) methodologies.

1) TEXT BASED VIDEO RETRIEVAL (TBVR)

This technique retrieves videos on the basis of text present in videos in the form of textual captions, names of performers and location of events etc [118]. In the text based video retrieval, textual regions such as characters, words, sentences, and blocks of textual information are analyzed within the video frames. These video frames require labeling on the basis of textual information present in them [119]. Conventionally, Optical Character Recognition(OCR) tools are utilized to extract textual information from videos. Information is extracted in the form of text strings which are then used as keywords for indexing purpose. The overall architecture of text based video retrieval system is shown in Figure 3.

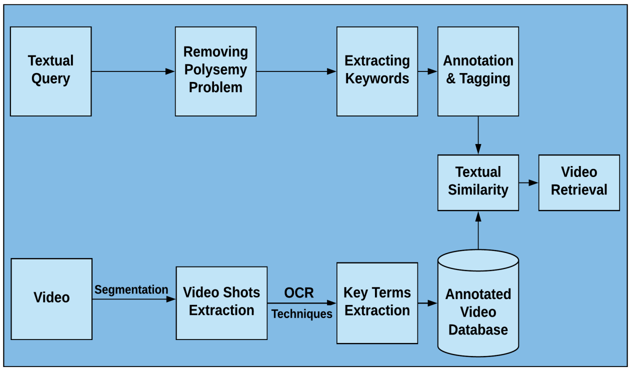


FIGURE 3. **Overall architecture of text base video retrieval.**

This section provides a birds eye view of ontological textual video retrieval methodologies. Antoniou and Van Harmelen [120] used ontology to disambiguate polysemous words for multimedia information retrieval. They also proposed the usage of ontology based reasoning in order to fixed associations between terms and concepts. It proved very beneficial for establishing relationships between classes and concepts which made information retrieval quite effective.

In 2005, Jawahar *et al.* [119] introduced a technique to search based on textual information present in a video. At first, regions that contain textual information were identified. Videos were annotated based on the extracted text which helped in quick search and classification of videos. Yanai *et al.* [121] presented an innovative text based approach for the automatic extraction of video shots. Their proposed methodology ranks relevant videos by analyzing the relationships of video tags and query words. However, their proposed approach was very time and space consuming.

Ontological textual video retrieval techniques are helpful in searching an album of an artist, a genre of songs and happenings of events according to geographical locations mentioned in the videos. These techniques are also helpful in digital libraries of video lectures for the applications of distant learning. The selection of optimal OCR tool is a constraint for such techniques and these techniques do not semantically evaluate the videos. Therefore, the interest of researchers has been shifted towards the content of video instead of just text in order to have better retrieval accuracy [122].

2) CONTENT BASED VIDEO RETRIEVAL (CBVR)

In this technique, actual content of the video frames is analyzed to extract semantics about the video. The term 'content' may refer to the colors, textures, shapes of objects instead of just text [122].

The first step of *'Content based Video Retrieval'* is the segmentation of moving objects into shots. A shot is basically an image sequence which is used as an indexing unit. Out of these sequence of images, key-frames are extracted which provide the abstract idea of very informative visual content available in video and helps in faster browsing of videos.

Once keyframes are extracted, the next step is the extraction of features. Features of video content are classified into following two classes: *low level features* and *high level features*. Low level features include color, shape, the motion of objects, pitch, bandwidth and loudness which are directly accessed from video databases. They help in answering the queries such as "Extract images with squared shaped objects." Whereas high level features are known as semantic features in which concepts are detected. Such features deal with semantic queries such as "find the images with the smile of Mona Lisa". This is quite challenging as it needs some semantics to be known. To answer such queries the system must have some prior knowledge e.g., Mona Lisa is a woman, who is a special character in a painting rather than an ordinary woman. These are semantics which are related to the identification of concepts present in the frame. The overall process of content based video extraction is shown in Figure 4.

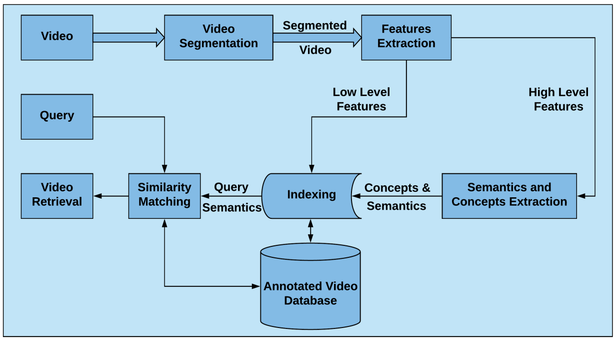


FIGURE 4. **Overall architecture of content base video retrieval.**

Video retrieval is an active research area as the amount of video data is increasing rapidly [123], [124]. The traditional approaches for video indexing are based on the content present in the videos by utilizing a computational feature set [125]. Jiang *et al.* [126] proposed a modified text-based inverted index approach for indexing large-scale videos. The initial step they took was the adjustment of video concepts that can be easily indexed using proposed approach. This indexing approach achieved remarkable results in content based video retrieval. Sikos [127] proposed a system which represented video events and scenes in an ontological format for content based video retrieval. He presented a logic based architecture for representation of video scenes and their interpretation which was based on automated reasoning.

Color and texture are considered as basic visual features of index frames [128]. For object based video retrieval, representative features obtained from index frames are utilized, which are stored in feature database [129]. Another important property of index frame is texture. Mittal and Gupta [130] presented a learning framework for high-level video indexing to minimize the gap between low-level features and semantic video classes. Their proposed approach supported support vector machine (SVM) parameters which result in better classification. This framework was only designed for video shots.

Huayong [131] in his paper, presented a content-based video retrieval approach which was based on the semantics of videos. Chen *et al.* [132] presented a video retrieval system, in which they proposed a statistics-based algorithm for the extraction of those videos that possess requested object motion from video database. Their proposed system deal with a crucial problem of quick semantic information retrieval from huge databases. Ying Dai presented a semantic tolerance relation model [133]. This model presented videos and images by their semantics and low level features. The semantics of each index frame or key-frame was provided by a class weighted vector. These weights were assigned using Bayesian classifier.

A lot of research has been conducted on the retrieval of textual resources using ontology but for visual sources, it is relatively new area [134]. Hoogs *et al.* [136] and Stein *et al.* [137] proposed a system in which, at the very first step video shots were analyzed and both low level and high level features were extracted such as color, shape and luminous etc. which were then used to search in the extended Wordnet for relevant annotation. These annotated videos were then used in IR system to obtain efficient retrieval results. In another study, the authors presented a keyframe based approach in which they extracted a set of key frames from videos which were then used for indexing and retrieval purposes [137]. They computed block matching of keyframes and linked similar shots together. Semantic gap problem also exist in video retrieval approaches. High level queries require external or prior knowledge, which needs to be embedded with low level features. But it is difficult to cover user's requirements with the low level features of video as query contains many high level concepts. This is known as semantic gap [138]. To minimize the semantic gap, high level features of a query must be mapped with the low level features of video. Two approaches have been proposed by researchers to reduce the semantic gap. The first one suggests to automatically generate metadata for videos [139]. It would still need semantic concepts and diverse schemas. The second suggests relevance feedback in order to learn and understand the semantic context of a query [122], [140].

Mezaris *et al.* [141] presented an approach which integrated a thesaurus with relevance feedback. They reduced the problem of semantic gap by allowing users to describe high-level keywords. This was done with the help of terms present in the ontology using medium level descriptors such as luminescence etc. The comparison was conducted between descriptions using keywords and extracted visual tags of images. This eventually returned matching images. Finally, relevance feedback was applied to refine the retrieval results.

C. AUDIO RETRIEVAL

In this era of digital media, audio information is very useful among the users due to its expressiveness. Both online and off-line audio content is present either in an isolated form such as sound recordings or in combination with other data such as movie soundtracks in which audio is combined with video. A huge amount of sound data is available on Youtube and similar websites with different captions and sound quality. Therefore, it is requested to automatically retrieve information out of it instead of manually probing the whole data. For the effective and efficient retrieval of sound data, following subsections present the research efforts proposed by various researchers.

1) TEXT BASED AUDIO RETRIEVAL (TBAR)

In this approach, sound data is retrieved on the basis of metadata such as artist name, file name, and tags etc [142]. When user poses a query, it is preprocessed and features such as the name of an artist, type of audio are extracted. The database which contains a collection of audio documents is processed and features are extracted and tagged. These tags are then compared with the features extracted from the query and similarity is calculated. Then an audio document with the highest similarity is returned to the user. The figure 5 illustrates different phases of text base audio retrieval.

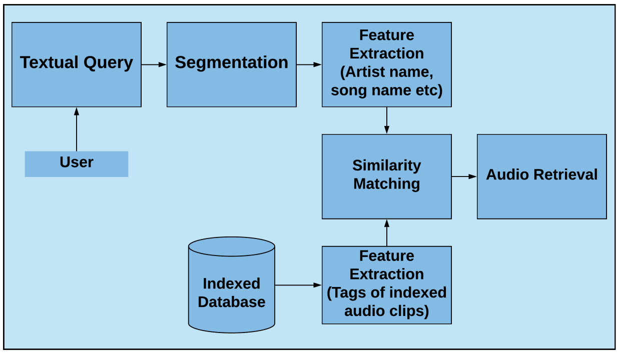

Textual Query
Segmentation
Feature Extraction (Artist name, song name etc)
User
Similarity Matching
Audio Retrieval
Indexed Database
Feature Extraction (Tags of indexed audio clips)

FIGURE 5. **Overall architecture of TBAR.**

For many cases where the sensitivity of information is not so crucial,the technique of retrieving audio through text is still used. For example, if people want to find examples of sounds but do not have a recorded sample at hand or preparing a presentation on the life of jungle and one want to add some sound effects of animal's roaring and rain etc. In all these cases, a natural way to define the desired sound is by a textual name, label, or description which is fed to the search engine and required sound data is retrieved [143]. Khan and McLeod presented an ontology based audio retrieval system [144]. Their proposed model had several tasks such as segmentation, acquisition of metadata and scheduling etc. In segmentation phase, boundaries of audio objects were identified based on five-tuple: identifier, description, starting time, ending time and audio data [145].

Turnbull *et al.* [146] presented a text based audio retrieval system in which a textual query was sent to the system and the most relevant audio document was retrieved using Gaussian mixture model. Their proposed system experimented on musical data. They utilized a small vocabulary of emotions, instrument names and pre-defined semantic tags. Later on, they extended their system to extract sound effects from a library, using a vocabulary of 348 words [147]. A group of Technische Universitat Berlin also used text based approach for sound data in which they mapped a sound clip to a text description. Then they retrieved relevant audio clip based on the description provided by the users in their textual query [148]. An obvious drawback of such approaches is the subjective nature of audio descriptions. Therefore, the paradigm shifted from text-based to content-based approaches.

2) CONTENT BASED AUDIO RETRIEVAL

Human beings can easily differentiate among different types of audios. Provided with an audio sample, humans can instantly identify: the type of audio as music, human voice, noise, speed of utterance, mood of speaker, and the topic of audio.

For machine, each audio sample is just a sequence of sample values. Since a long time, accessing of audio pieces based on their titles or file names remained very common. But due to the incompleteness of file name and subjectiveness of text description, it was quite challenging to retrieve audio pieces according to particular requirements of applications [149]. Moreover, this retrieval approach could not support semantic queries such as "retrieve audio samples similar to the one being played". To answer such queries, by avoiding above mentioned problems, content based audio retrieval techniques were introduced.

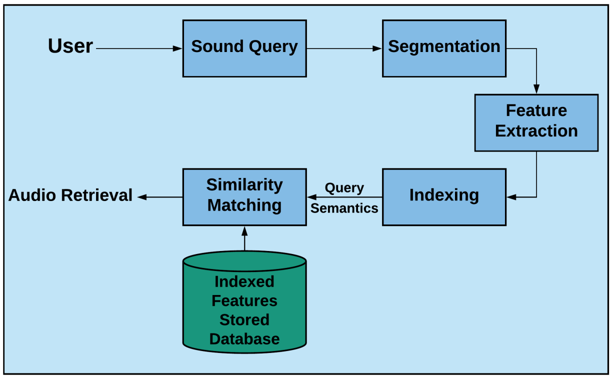


FIGURE 6. Content based audio retrieval system.

Content-based approach differs from text based retrieval approaches as they index media files using file names and user tags etc [143]. In simple content based audio retrieval approaches, each and every sample of stored audio pieces is compared with the query, but this approach is not practical as audio signals are usually variable. In addition, different snippets of audio are represented by different sampling rates and utilize a different number of bits per sample. Therefore, content based audio retrieval is based on a set of extracted audio features e.g., frequency distribution etc [149]. The general approach for content based audio indexing and retrieval is shown in the figure 6 which is as follows:

- Initially, sound data is classified into some predefined types of audio such as speech, noise, and music etc.
- After that, different types of audios are processed and indexed using different ways. For example, if an audio is of speech type, speech recognition techniques are applied and then indexed based on the recognized words.
- Similarly, audio samples of a query are processed, classified and indexed.
- Finally, the similarity between query index and audio indices in a database is calculated. On the basis of this similarity, topmost related audio samples are retrieved.

This section briefly describes state of the art work on text based audio retrieval. Barrington *et al.* [150] proposed a content based audio retrieval system. In this particular system, an audio sample was provided as a query instead of textual description and a list of audio snippets was returned similar to the query.

In content based audio retrieval system, two kinds of approaches were generally used. The first approach was the query-by-semantic example (QBSE) in which audio was retrieved on the basis of semantic information [150] and second approach was query-by-acoustic example (QBAE) in which audio was retrieved on the basis of acoustic similarity to the query [145].

3) HYBRID APPROACHES

In this technique, both text and content based approaches are combined to retrieve sound data. Kannan *et al.* [145] showed that hybrid approaches possess the ability to retrieve most relevant audio resources in large numbers as they improve the effectivenss of audio retrieval up to great extent. Wichern *et al.* [151] presented an integrated system, which can be utilized for both text based retrieval and content-based query-by-example. This hybrid network linked the sounds based on perceptual similarity and semantic tags based on user provided weights which helped in improving the audio retrieval. A group of researchers in Berlin [148] utilized both text and audio content for the retrieval by mapping a sound clip to a text description based on the content of sound data. This text description was later used in retrieval of audio data.

D. FUTURE WORK

The need of multimedia information retrieval system is continuously growing with exponentially growth of multimedia data. Currently, most multimedia information retrieval systems perform content based retrieval which exploits the features of video, image and audio resources. An improved version of underlying algorithms along with features that better depict semantics would ultimately raise the usefulness and precision of multimedia information retrieval system. The future of multimedia information retrieval approaches heavily depends on effective and high performance computing. Since multimedia data sources are humongous, the processing of such gigantic resources is only feasible through high performance computers.

Query plays a key role for efficient information retrieval. However, in multimedia information retrieval systems, semantic gap between user's query keywords and features of multimedia resources is very huge. It is difficult to cater this semantic gap just by using traditional query reformulation techniques. Therefore, this problem needs to be tackled by designing special query languages for multimedia information retrieval. Such query language will not only contain the user query, but will also possess characteristics that corresponds to semantics of multimedia resources.

Another major bottleneck in multimedia information retrieval is the high dimensionality of multimedia features. For such a humongous feature space, traditional indexing algorithms are not suitable. To tackle this challenge, high dimensional indexing algorithms are required. However, there exists a gap as state of the art indexing algorithms were designed for generic indexing without considering the properties and characteristics of multimedia feature space. Therefore, there is a need to improve these high dimensionality indexing algorithms in context of multimedia information retrieval such that they perform indexing while considering each type of multimedia content.

In domain of multimedia information retrieval, more research needs to be done on high level semantic multimedia retrieval. Most of the multimedia search engine either involve very little semantics or do not involve them at all. For example, almost all search engines are able to cater queries like ''find picture like this'' or ''find picture of flower''. However, if semantics are get involved in query that asks for a ''picture of some brown colored wet mud in the center of field which is of similar size as vegetation beside it'', they would not be able to do so. Therefore, trends need to be shifted to high level semantic based multimedia retrieval.

IV. CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR)

A desperate need in a rapidly growing globalized economy is the ability to find significant information in the languages which are different from query language. Cross-lingual information retrieval is all about making a query in one particular language and retrieving documents in diverse languages. The yielded documents are then translated into query language to permit the user to get the gist and basic idea about retrieved information. Few CLIR systems exploit bilingual dictionaries or parallel corpora to translate user's query into a target language, whereas, others translate foreign language corpora documents into source language beforehand by the use of machine translation. Machine translation makes foreign language corpora documents retrievable by particular user's query. For instance, the user formulates a query in English regarding ''Arrangement of flower'' and get the documents in Japanese related to ''Ikebana'' which is actually an arrangement of flower in Japanese. Cross-lingual information retrieval has literally proved an area of great interest for numerous researchers [152]. Recently, a significant number of workshops and tracks have come into the picture to aid research in Cross-lingual information retrieval. For example, Cross Language Evaluation Forum (CLEF) is working on European languages and it has been active since 2000. Similarly, The NII Test Collection for Information Retrieval Systems (NTCIR) is a project competition which is held in Japan on yearly basis. It covers several topics such as Cross-lingual Information Retrieval specifically dealing with the languages like Chinese, Japanese, Korean and English. Moreover, Text Retrieval Conference (TRec) was running a multi-language information retrieval track for few years which ended in 2002. However, it is still being extensively used in both NTCIR and CLEF. Mostly cross-lingual based information retrieval systems employ some translation mechanism which is considered more dominant and effective approach, although, there exists some non-translation methods like semantic indexing, cognate matching [153], relevance models [154] and latent semantic indexing [155]. We have illustrated the current tree of diverse approaches used in the context of cross-lingual information retrieval through pictorial representation in Figure 7. The approaches of query translation are also discussed in [156].

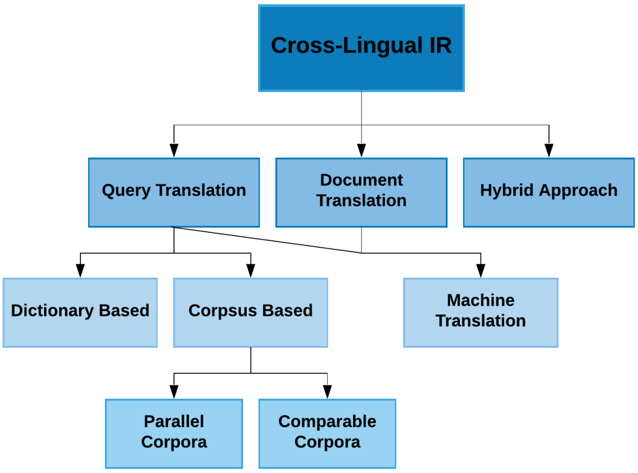


FIGURE 7. Cross-lingual information retrieval diverse approach tree.

The major problem of Cross-lingual information retrieval is language translation that is why dominant areas of research in this domain are as follows: i) What needs to be translated, ii) how it must be translated and iii) how to truncate bad and faulty translations iv) How to gain a large amount of translation data. The remaining section is devoted to the discussion of recent advancements in the field of cross lingual information retrieval which may help the researchers to alleviate all mentioned problems. This section will progress as follows. First, we will address what needs to be translated followed by the methods usually used for sake of translation. Next, we will discuss the methods proposed by researchers to automatically achieve resources in quest of translation. Finally, we will take a look at the future of Cross-lingual Information Retrieval.

A. WHAT NEEDS TO BE TRANSALATED

The major choices in what needs to be translated are a query, entire document or both. Query translation refers to the process of translating input query into a target language, whereas, document translation is the process of translating an entire document into the query language.

B. QUERY TRANSLATION

User's query is translated into the target language using certain bilingual dictionaries, corpora or machine translation. This section illustrates highly significant current approaches used for query translation.

1) DICTIONARY BASED APPROACH

Dictionary based query translation approach refers to the process of first processing the user's query linguistically to find the keywords and then translating the keywords into target language through the use of machine readable dictionaries (MRD). MRD's are basically electronic version of either general or domain specific printed dictionaries. MRD's can also combine the texts of general and domain specific dictionaries. Generally, a bilingual dictionary consist of list of words expressed in source language and their equivalent translations in target language [157]. These dictionaries may have translation probabilities that enable weighting and disambiguation. The use of existing significant linguistic resources like MRD's for query translation is a more natural and convenient approach in domain of cross-lingual information retrieval. Translating the query through the use of dictionaries is way simpler and faster approach as compared to translating entire documents into source language which seems more costly in terms of time and effort [158]. Mustafa Abusalah *et al.* [159] proposed an ontology based approach to raise the performance of query translation using a cross lingual information retrieval (CLIR) system. They replaced machine readable dictionary (MRD) with multilingual ontology (MO) in order to provide better search engine for tourism domain. Their proposed ontology based approach surpassed the performance of baseline (MRD approach) using mean average precision as evaluation metric.

Although, there are many bilingual dictionaries in several literatures but the major problems in bilingual dictionary based approach are translation ambiguity, word inflection problems, translating phrases, word compounds, special terms and spelling variants [160]–[162]. Word inflection can be solved through stemming and lemmatization [163]. Undertaking the problem of translation ambiguity, Yahya *et al.* [164] proposed a method in CLIR to improve dictionary based query translation using a domain ontology of Quran concepts. The dictionary was written as a bilingual parallel corpora in English and Malay language. They assess the performance of three IR systems based on natural language query, transformed natural language query using dictionary (baseline), and transfigured natural language query using Quran ontology respectively through mean average precision and an average precision computed at 11 recall points. Their proposed methodology produced better retrieval results for the collection of English documents as compared to Malay documents. They proved that proposed system can acquire effects of query expansion and improve the performance of retrieval in certain language

Another problem which cause a reasonable degradation in the performance is the lack of vocabulary coverage. It usually occurs when query contain some phrases or words which are not present in the underline dictionary. This problem is also known as Out-of-Vocabulary (OOV) problem [165] and it exist even in some of the best dictionaries. The context of user's query is not always clear because of their preciseness, hence, even query expansion has fails to recover the significant missing terms. Typically, OOV terms are totally new words or proper names. For instance, a user is looking for the information regarding a disease namely Influenza A (HINI) in Malaysia through formulating the query as ''HINI Malaysia''. HINI is a totally new term and most probably it may not be present in underlying dictionary which was created couple of years ago.

Furthermore, if the user omits the term HINI from the query, he may not get any relevant documents. OOV terms usually include proper nouns, compound words and technical terms [166]. In large cases, proper names and technical terms are considered as significant assets of documents but usually all dictionaries only contain most common technical terms and proper nouns such as countries and major cities. Translating all terms perfectly is one of the biggest challenge for cross-language information retrieval system. One of the most trivial approach used to cater untranslatable keywords is to simply add untranslatable content in target language query. If target language does not have these particular terms, the query will less likely retrieve relevant documents. Moreover, phrase translation is also pretty complex as a phrase can not be translated directly by translating all of its individual words [165]. For example, translating an idiom word by word will change its actual meaning as expressed in source language. Likewise, named entities (NEs) acquisition and translation are immensely significant in almost all tasks of natural language processing such as construction of bilingual lexicon, cross-lingual information retrieval and machine translation [167]. NEs are considered the integral component in news texts [168]. NEs like person, organization and location are pretty tough to handle through a fixed and non-dynamic set of rules because new entities are continually being created. This marks the growing need of advanced NEs acquisition and translation techniques. Usually bilingual dictionaries have few entries regarding NEs [169], however, when NEs are wrongly tagged as simple words and translated by exploiting a bilingual dictionary, poor results are produced often.

2) CORPUS BASED METHODS

Corpus is basically a collection of natural language resources comprising of text, sentences, and paragraphs which may exist in one or different languages. Query translation has exploited two various bilingual corpora named *parallel and comparable* corpora [170]. Both corpora are briefly discussed below:

Parallel Corpora:

Parallel corpora comprise exactly same documents but in different target languages. Generally, in order to reveal the exact correspondence between source language sentence and target language sentence, an annotated parallel corpus is used. They can be exploited for sake of analyzing many processes like transferring ideas, concepts, and information from one language to other. They are becoming the translation equivalent information sources for machine translation based applications or humans. The query does not require to be transformed into target language in case of retrieving particular text from aligned corpus because the query can easily match the corpus component of the same language and then its corresponding component in the target language can be easily retrieved. Typically, parallel corpora are populated using multi-lingual websites, machine translation, and human translation. For instance, many researchers are generating multi-lingual corpora through the use of ''Spider'' systems which are able to collect documents of equivalent translation from web. The alignment between a source language and target language documents is then performed either by comparing documents through indicators or using special tools. Alignment through indicators compares documents considering anything which exactly corresponds to both source and particular target language documents such as document date, author name, special name or numbers, and acronyms. On the other hand, tools like PTMiner [171] are also being extensively used to align parallel corpora. In PTMiner, the system first decides candidate sites and then filters significant web pages from each web site which are generally indexed by a search engine (e.g Google). In next step, the system builds pairs of web pages after matching pattern among URL's ($default.phpvsdefault_f.php$). Finally, the system filters the candidate parallel web page [172]. Moreover, Braschler and Scäuble [173] developed another alignment approach for bilingual reports generated from election results.

Comparable Corpora:

Comparable corpora consist of sets of documents present in multiple languages but documents are not the translations of each other [174]. In comparable corpora, multi lingual documents cover the same topic, therefore they contain an equivalent vocabulary [163]. For example, different news agencies like CNN, Xinhua news, BBC, Reuters, and BERNAMA produce multilingual news feeds. Such feeds are easily available on Web for several domains and language pairs. They often have numerous sentence pairs that are pretty good translations of one another [175]. Several statistical methodologies can be exploited to construct bilingual topic-specific dictionaries from aligned corpora.

Corpus based approach is way better than dictionary based translation because it marks greater performance, as found by McNamee and Mayfield [176]. However, it is extremely difficult to find such a large significant parallel corpora for some languages. In addition, the creation of a parallel corpus is extremely daunting and expensive task. Corpus based translation approaches also face the challenges of coverage and quality. McNamee and Mayfield [176] have concluded that poor quality corpora can immensely affect the performance of a particular system. Likewise, coverage, which is all about vastness and enormousness of vocabulary words has a similar effect on a system, because if a number of query words would not find any translation in an underline corpus, it will eventually decrease the performance of an entire system. Although, in some languages, coverage is not a big deal but in other languages like English and Chinese, it is considered a big problem [177]. Considering all the mentioned challenges, researchers have done a reasonable work to acquire bilingual lexicons or parallel corpora semi-automatically or automatically.

C. DOCUMENT TRANSLATION

Typically, document translation is performed by exploiting machine translation system like SYSTRAN [178], AppTek [179] and PROMPT [180]. Systems based on machine translation usually yield translations of different natural languages either semi-automatically or automatically. Major tasks of any machine translation system are source text and language analysis, source-target conversion and generation of the target language by exploiting either bilingual or multilingual dictionaries [181]. Syntactic, morphological and semantic information is captured and stored through the entire process. One of the earliest machine translation company namely SYSTRAN was funded by US government in 1995 for the sake of developing an efficient cross-language information retrieval system having abilities of parsing and translating any natural language. SYSTRAN developed a software [182] which combined statistical and rule based machine translation approaches to produce high quality results. The success of SYSTRAN's software has been proved a great landmark in the field of cross-lingual information retrieval because it is able to deliver a tremendous quality translation of natural language for all domains. Likewise, PROMPT [180] provides machine translation services, standalone data mining, translation and translation memory systems and dictionaries. PROMPT provides efficient software tools such as linguistic editor, post editing tool, building and editing dictionaries and user oriented interface to resolve the problems of translation modules and dictionary volume. Thenmozhi and Aravindan [183] proposed a methodology for Tamil-English cross lingual information retrieval system. They translated the Tamil query into English query in order to retrieve documents written in English language. They exploited word sense disambiguation module to remove ambiguity from Tamil query and an automatically generated English ontology was used to resolve ambiguity from equivalent English query. They developed named entity meta-thesaurus, morphological analyzer and bilingual dictionary for Tamil-English in quest of Tamil-English query translation. Moreover, they utilized ontology in order to reformulate translated query before embedding into search engine for document retrieval. They performed experiments in agriculture domain and outshined other techniques in terms of precision.

McCarley [158] marks various advantages of document translation. The biggest engaging advantage of document translation is a significant increase in the chances for either translating a particular word correctly or determining the synonymous from for a query. Researchers have invested a lot of effort to compare document translation with query translation and they have found that document translation is way better than query translation [184], [185]. For instance, Oard [185] performed experiments on the TREC-6 dataset and concluded that document translation convincingly produced better results as compared to query translation. Likewise, Chen and Gey [186] observed that document translation produced better results on test collection CLEF 2003. However, document translation through machine translation system is pretty computationally expensive and sometimes it seems infeasible for a large set of documents. Although, modern computers have allowed the problems of computational cost to some extent, the high cost of available translation systems and non-availability of translation systems for certain language pairs are still the major challenges in document translation. Undoubtedly, translating the user's query seems more practical approach as document collections are either too large or outdated most of the time [169]. However, translation ambiguity is the biggest challenge in query translation approach as already discussed.

1) MACHINE TRANSLATION (MT)

Machine translation approach is used for both query and document translation. It exploits machine translation system for the sake of translating either query or document into a target language. Machine translation is implemented in two different methods [187]. The first method is all about using an off-line machine translation system to transform foreign language corpora documents into the equivalent language of user's query beforehand. However, this methodology is not pretty efficient as it has been proved computationally expensive for large corpora or for multi-lingual documents collection [188]. For instance, Braschler [189] were unable to find a direct German/Spanish machine translation in their experiments on German-Spanish cross-lingual information retrieval. Hence, they used German/English machine translation leading to English/Spanish machine translation but still all terms of German documents were not translated. Furthermore, off-line machine translation approach becomes infeasible if there is a need to search required resources or documents on the web against a particular user query. The second way of using machine translation in cross-lingual information retrieval is by transforming the user's query into a target language (the language of corpora documents). The transformed query can then be used to acquire documents of target language through classical information retrieval methodologies. Machine translation and retrieval work separately in both mentioned methodologies. The major problem of machine learning in the context of query translation is translation disambiguation which often rooted by homonymy and polysemy as transformed query may not represent the true sense of original query [190]. Homonymy generally refers to a certain word that possesses at least two diverse meanings. For example, translating an English word like ''Bark'' into some other language may affect the true meaning of a word since its context is not clear at all. ''Bark'' may either refer to woof of a dog or tree skin. Whereas, polysemy typically refers to a specific word which is used to express a couple of different but related meanings like ''Head'' of particular family or humans head. Therefore, usually machine translation systems perform context analysis [189] to determine the true word sense and resolve ambiguity for translation. Typically, most of the queries lack context because of its preciseness in terms of a small number of used keywords. Therefore, machine learning is considerably more effective and efficient in document translation rather than query translation where context is not clear at all.

V. FUTURE WORK

Cross-lingual information retrieval has marked significant advancements in last couple of years but it has failed to surpass the results of trivial monolingual information retrieval. Moreover, acquiring parallel corpora and lexicons for certain minority languages is still considered as the highly challenging task. Translation disambiguation is one of the major problem in cross-lingual information retrieval as one word can have multiple translations depending upon the context in which it is used. This problem arises because of linguistic properties of homonymy and polysemy. For example, the word ''head'' can be considered as a human body part and can also be the leader of some organization. As user queries are small and contain very little contextual information, therefore they do not provide enough information that could be used to disambiguate the meaning of key terms. That is why this problem is more common in query translation task. To cater this problem, we can look into development of interactive cross lingual platforms that allow users to select the correct translation using graphical visualization to resolve the disambiguity problem. This user feedback can be saved and used as training corpus to train different types of machine learning algorithms in order to automate the process of translation disambiguation.

Also, there are many non-popular languages like Urdu and Hungarian that lack resources in the form of corpus, ontologies and lexicons. To perform cross-lingual information retrieval between such languages, there is a need to either design ontologies and resources for them, or use machine translation. For example, it is difficult to perform Urdu to Hungarian cross-lingual information retrieval as there are no machine readable ontologies of these languages. Designing domain specific ontologies is a costly task and an extensive research domain. Therefore, such cases can be resolved by involving a third common language as the intermediary language. If we involve English as intermediate language, the system will first perform Urdu to English translation and retrieve the information in English. At the end, it will return the output to user after performing English to Hungarian.

There is a need to involve high level of semantic understanding in query or document translation stage, especially for phrasal translation. There exists a gap in traditional translation approaches as they are unable to cater the semantics behind various idioms and phrases. Translation of idioms using dictionary based approach changes the meaning behind the idiom completely. For this, we need to involve more phrase specific entries in language resources and ontologies.

Finally, when cross-lingual information retrieval would match the level of trivial monolingual information retrieval, there would still exist a problem of information representation with respect to the user. All users might not have the capability to read the yielded documents. Considering the information representation problem, we expect a reasonable increase in the research of reliable and efficient machine translation techniques. All in all, we are optimistic to see more advanced research in the domain of cross lingual information retrieval that exploits world wide web, ontologies, and modern machine translation algorithms more effectively.

TABLE 1. **Performance of several techniques proposed in context of textual, multimedia, and cross lingual information retrieval.**

Textual Information Retrieval			
Author	**Dataset**	**Algorithm**	**Results**
Jibran Mustafa et al.	R. Publication Ontology	Context Based IR	0.82 (correlation)
BEEN-CHIAN CHIEN	General semantic Ontology	Semantic Based IR	55.84 (MAP)
Zheng et al.	General Ontology generated from documents	Semantic Based IR	0.813 (Recall) 0.729(Precision)
POONAM YADAV et al.	Mobile communication dataset and emerging concepts in Ontology	Semantic Based IR	0. 771836 (F-measure)
Min Song et al.	WordNet Ontology and TREC5	Semantic Association Based IR	0.458 (F-measure)
HU Jun et al.	Rough Ontology	Semantic Association Based IR	0.77(F-measure)
Faezeh Ensan et al.	TREC Robust04	Semantic Networks and Maps	0.2858 (MAP)
Latifur Khan et al.	Sport News Ontology	Ontologies	88.7% (F-score)
Miriam FernÃąndez	TREC dataset	Semantically enhanced IR	0.1641 (MAP)
Multimedia Information Retrieval			
R. Uma et al.	General Ontology	MIR/Semantic Annotation Based IR	0.9915(F-measure)
V. Vijayarajan et al.	Domain ontology using image descriptions	Content-based image retrieval in MIR	(0.4618) precision
Latifur Khan et al.	Sports News Ontology	Text-based audio retrieval in MIR	Conflict (98% & 97%) Recall & Precision
Camille Kurtz et al.	RadLex ontology	Image contents and semantic dissimilarity	AUC (0.77) & NDCG (0.929)
Soner Kara et al.	Soccer domain Ontology	Semantic Indexing	0.8434% (MAP)
Lamberto Ballan et al.	Airplane events (LSCOM) Ontology Broadcast news and surveillance	Video Annotation	0.797 (Average Precision)
Miguel ÃĄngel RodrÃ■guez-GarcÃ■a et al.	ICT Ontology	Ontology based annotation	0.89 (Avg. F-measure)
Kehua Guo et al.	Culture, geography and technology	Semantic Ontology Retrieval Algorithm	0.82% (Avg. Precision Rate)
Cross-Lingual Information Retrieval			
Mustafa Abusalah et al.	Travel domain Ontology	Query Translation	0.63 (MAP)
D THENMOZHI et al.	Agriculture domain Ontology	Ontology based query translation	95.36 (MAP)
Zulaini Yahya et al.	Quran Ontology in English	Dictionary based and concept similarity based	English/Malay doc 0.098/0.072 (MAP)

VI. DISCUSSION

This section summarizes and compares the performance of multifarious approaches proposed in quest of textual, multilingual, and multimedia retrieval with respective domains and evaluation metrics. The table 1 highlights the results produced by several proposed techniques for the task of textual, multimedia, and cross-lingual information retrieval.

As the table suggests, in textual information retrieval, Mustafa *et al.* [54] proposed framework of using context based semantic information retrieval based on thematic similarity measure raises the precision of search results. RDF triples are used to store metadata of sources and query concepts are matched against existing RDF triples instead of keywords. This enables the search framework to focus on concept combination and the similarity between their relations at the same point of time. User query is expanded with semantic neighborhood and synonym before passing it to Semantic matcher. Then triple searching along with semantic matching are performed. Finally, all yielded results are ranked according to the relevancy with user's query. They make the comparison of correlation among precision and recall for keyword, simple semantic, and semantic neighborhood based information retrieval systems. They report that keyword based information retrieval system manages to produce the upper and lower bounds of 0.73, and 0.15 compared to 0.82, and 0.39 of simple semantic based information retrieval system. However, semantic IR with semantic neighborhood marks the correlation figures of 0.79, and 0.49 which reveals a strong correlation among precision and recall. Moreover, Mule and Waghmare [56] proposed approach of ontological indexing considers the context of query words in order to find accurate results from database compared to trivial text base search which extract the results on the basis of keyword match. Hu and Ju [60] proposed ontology based IR raises the effectiveness of IR up to great extent. They embed a fuzzy ontology constructed automatically from documents. They compare the effectiveness of TF-IDF based IR, TFIDF based IR using query expansion (WordNet), and IR using combinations of multifarious ontologies. They evaluate top 1000 retrieved documents of all aforementioned approaches. They extrapolate that IR performance gets increased 2-5% in terms of average precision and 1-6% in terms of mean average precision. They report the mean average precision figures of 52.33%, 47.41%, and highest 55.84% produced by TFIDF, TFIDF with query expansion, and ontology based IR respectively. Gu and Yu [61] proposed 3-layer semantic based indexing approach produces the precision figures of 72.9%, which is 11.6% better than conventional index approach based IR.

Likewise, Yadav and Singh [64] proposed ontology based IR methodology improve the retrieval of relevant documents. They evaluate their proposed approach on three datasets and outperform benchmark approach on every dataset. Ontology based IR approach reveals the highest F-score of 77% compared to 70% of benchmark approach on one of three datasets. In addition, Song *et al.* [69] proposed semantic query expansion technique utilize association rules along with ontologies and IR techniques to raise the performance of search results. Thier proposed technique outshines all three state of the art approaches namely SLIPPER, semantic query expansion with ontologies and consine similarity, and semantic query expansion without ontologies by producing the F-score of 11.90% compared to 9.2% on TREC5 data. Moreover, in term of top 20 ranks(P@20), their approach manages to mark the figure of 20.98% than 15.49% of other approaches. Hu *et al.* [70] presented hybrid information retrieval method integrates rough ontology with efficient search and association search methodology in order to improve the performance of IR. They compare rough ontology based IR approach with two other approaches namely OntoSCORM, and Lucene using F-measure over number of queries. Rough ontology based IR approach manages to produce the F-score of 0.77% compared to 0.74% and 0.68% of OntoSCORM and Lucene respectively.

Likewise, Ensan and Bagheri [24] proposed semantic based language model (SELM) represents the documents and user queries in form of graph concepts where the closeness and similarity of a query with certain document is computed by determining the semantic relatedness of their corresponding concepts. They perform experiments over three corpora namely TREC Robust04, ClueWeb09-B, and ClueWeb12. They compare their proposed language model with state of the art information retrieval model namely Sequential Dependence Model (SDM) and with couple of query expansion models namely Relevance Model (RM3), and Entity Query Feature Expansion (EQFE). They report that their proposed language model and EQFE collectively surpass the performance of state of the art models by producing the mean average precision of 0.33, 0.11, 0.049 over all three corpora respectively. Selvalakshmi and Subramaniam [33] propose preprocessing and document classification methodologies in quest of semantic information retrieval through ontology matching. The proposed methodologies raise the relevancy of retrieved documents and possess the ability to manage and retrieve massive data. The proposed methodologies assist semantic information retrieval to produce the average relevancy score of 98 and 99 for web documents and tweets respectively. Fernández *et al.* [57] proposes semantic search model which embeds semantic knowledge in form of ontologies and utilizes state of the art IR ranking methodologies. They compare semantic retrieval approach with Lucene and TREC automatic engine in terms of mean average precision. Their proposed approach outperforms other two approaches over 60% and 65% of queries. On average, semantic retrieval approach marks the mean average precision of 0.16 compared to 0.1 and 0.2 of Lucene and TREC automatic engine.

Turning to multimedia information retrieval, Uma and Muneeswaran [72] present collaborative tagging model which exploits block acquiring page segmentation technique in order to retrieve the tagged information which gets added when multimedia resources are semantically annotated. They exploit single ontology in order to extract multimedia information. For experimentation, they use the multimedia corpus formed by collecting 20,000 images, 10,000 audios, 10,000 videos, and 10,000 texts from Youtube, Wikipedia and organize them into different categories such as Agriculture, Geography, etc. They evaluated their proposed ontology based multimedia retrieval approach in terms of precision, recall, and F-measure. They report the f-score of 0.99 over aforementioned dataset. Kurtz *et al.* [99] presented semantic framework retrieves highly similar radiological images by utilizing the high level semantic annotations of certain image. This is because annotations are actually semantic terms belonging to domain specific ontology. They obtain NDCG score of greater than 0.92 over a dataset having 25 images, and AUC score of greater than 0.77 over a dataset having 72 images. Moreover, with EMD distance metric, the proposed method manages to produce the AUC score of 0.64 and NDCG score of 0.87 on respective datasets.

Moreover, Kara *et al.* [88] proposed keyword based semantic retrieval framework performance gets improved by using semantic indexing, ontology based information extraction, and inferencing. For experimentation, they crawled 10 UEFA matches to get 1182 narrations, and 902 events. They built an index through information extraction and inferred knowledge to evaluate over ten queries related to soccer domain. This index manages to produce the mean average precision of more than 98% over six queries. Rodríguez-García *et al.* [73] presented ontology based semantically enhanced platform aids in discovering cloud services according to the needs of multifarious users. Their proposed platform creates a repository containing descriptions of cloud services in order to support the search of user oriented services. An ICT ontology of 500 cloud services is used and system is evaluated over five different queries written by experts for ten predefined cloud services. The propose system manages to produces the average F-score of 0.82 and 0.89 for single and multi topic queries. In addition, Guo *et al.* [81] present optimized semantic ontology retrieval exploits processing tools related to big data in order to save and retrieve ontologies out of heterogeneous multimedia data sources. They perform experiment over multimedia dataset formed by collecting 40,0000 documents, 10,000 audios, 10,000 text, and 10,000 videos from Bing and Google. Their proposed approach produces the average precision rate of 80% for the retrieval task of multimedia.

On the other hand, in cross-lingual information retrieval, Thenmozhi and Aravindan [183] proposed ontology based cross lingual information retrieval assists Tamil farmers to feed query in Tamil and to acquire documents in English language. They resolve the ambiguities in Tamil and English query using WSD and agricultural ontology, which has been learnt automatically from underly textual documents. They compare ontology based CLIR system with trivial keyword search approach by the help of multifarious Tamil search engines. Google translation, and CLIR system. Their proposed system manage to retrieve top 20 pages and outperforms other techniques by producing the mean average precision of 95.36% computed by aggregating the performance value of several queries posed by farmers.

Likewise, Abusalah *et al.* [159] proposed approach of using ontology rather than trivial machine readable dictionary has improved the results of query translation by the figure of 21% as baseline reports the mean average precision of 0.42% in travel domain. Although this sort of approach was degraded in earlier monolingual work, however, the benefit of exploiting ontology is being widely accepted now. Yahya *et al.* [164] compare the effect of query translation using bilingual ontology with bilingual dictionary for the task of cross-lingual information retrieval. They extrapolate that query translation using first translation provided by dictionary produce good results than using all translation candidates enlisted in dictionary either for Malay or English set of documents. Likewise, using ontology rather than conventional dictionary receives the effects of query expansion and surpass the retrieval results of dictionary bases translation for English documents but not for Malay language documents because of concept matching problem. They reported the mean average precision figures of 0.098%, and 0.072% for English and Malay document collections respectively.

VII. CONCLUSION

Semantic based information retrieval is facing multifarious problems such as unavailability of semantic knowledge sources, evaluation benchmarks, datasets, fast IR methods, and inevitable evolution of domain. Likewise, multimedia information retrieval is yet to overcome the challenges of semantic gap which does exist between the keywords of user query and features of multimedia resources. Another major bottleneck in multimedia IR is the lack of high dimensional indexing algorithms which are indispensable techniques for highly dimensional multimedia features. On the other hand, cross-lingual information retrieval is also lacking significant resources such as corpus, ontologies, and lexicons for several renowned languages (e.g Urdu). In addition, cross-lingual information retrieval is still unable to surpass the problem of knowledge representation which would prove a major hurdle for several researchers and practitioners. Considering all aforementioned problems, a reasonable research in machine translation, automatic ontology learning from unstructured text, and the semantic information annotation and extraction is required to excel in the field of information retrieval.

REFERENCES

[1] B. Johnston and S. Webber, "As we may think: Information literacy as a discipline for the information age," *Res. Strategies*, vol. 20, no. 3, pp. 108–121, 2005.

[2] H. B. Styltsvig, "Ontology-based information retrieval," Ph.D. dissertation, Dept. Comput. Sci., Roskilde Univ., Roskilde, Kingdom of Denmark, 2006.

[3] T. B. Lee, "The semantic Web roadmap," 1998.

[4] F. Ramli, S. A. Noah, and T. B. Kurniawan, "Ontology-based information retrieval for historical documents," in *Proc. 3rd Int. Conf. Inf. Retr. Knowl. Manage. (CAMP)*, Aug. 2016, pp. 55–59.

[5] D. Roth and K. Small, "The role of semantic information in learning question classifiers," in *Proc. 1st Int. Joint Conf. Natural Lang. Process.*, 2004, pp. 184–187.

[6] Y. Li, Z. A. Bandar, and D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, Jul. 2003.

[7] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the Web," in *Proc. 7th Annu. ACM Int. Workshop Web Inf. Data Manage.*, 2005, pp. 10–16.

[8] K. Krishnan, R. Krishnan, and A. Muthumari, "A semantic-based ontology mapping–information retrieval for mobile learning resources," *Int. J. Comput. Appl.*, vol. 39, no. 3, pp. 169–178, 2017.

[9] B. Raphael, "Sir: A computer program for semantic information retrieval," 1964.

[10] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield, "Information retrieval on the semantic Web," in *Proc. 11th Int. Conf. Inf. Knowl. Manage.*, 2002, pp. 461–468.

[11] S. Mukherjea, B. Bamba, and P. Kankar, "Information retrieval and knowledge discovery utilizing a biomedical patent semantic Web," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 8, pp. 1099–1110, Aug. 2005.

[12] H. Yu, T. Mine, and M. Amamiya, "An architecture for personal semantic Web information retrieval system ntegrating Web services and Web contents," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, Jul. 2005, pp. 329–336.

[13] V. N. Murthy, S. Maji, and R. Manmatha, "Automatic image annotation using deep learning representations," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 603–606.

[14] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognit.*, vol. 71, pp. 144–157, Nov. 2017.

[15] D. Tosi and S. Morasca, "Supporting the semi-automatic semantic annotation of Web services: A systematic literature review," *Inf. Softw. Technol.*, vol. 61, pp. 16–32, May 2015.

[16] A. Dingli, F. Ciravegna, and Y. Wilks, "Automatic semantic annotation using unsupervised information extraction and integration," in *Proc. SemAnnot Workshop*, 2003, pp. 1–8.

[17] S. Dill *et al.*, "SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation," in *Proc. 12th Int. Conf. World Wide Web*, 2003, pp. 178–186.

[18] G. Petrucci, C. Ghidini, and M. Rospocher, "Ontology learning in the deep," in *Proc. Eur. Knowl. Acquisition Workshop*, 2016, pp. 480–495.

[19] M. A. Casteleiro *et al.*, "Ontology learning with deep learning: A case study on patient safety using pubmed," in *Proc. SWAT4LS*, 2016, pp. 1–10.

[20] P. Hohenecker and T. Lukasiewicz. (2017). "Deep learning for ontology reasoning." [Online]. Available: https://arxiv.org/abs/1705.10342

[21] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for gene ontology annotation predictions," in *Proc. 5th ACM Conf. Bioinf., Comput. Biol., Health Inform.*, 2014, pp. 533–540.

[22] S. Albukhitan, T. Helmy, and A. Alnazer, "Arabic ontology learning using deep learning," in *Proc. Int. Conf. Web Intell.*, 2017, pp. 1138–1142.

[23] S. Peters and H. E. Shrobe, "Using semantic networks for knowledge representation in an intelligent environment," in *Proc. 1st IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2003, pp. 323–329.

[24] F. Ensan and E. Bagheri, "Document retrieval model through semantic linking," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 181–190.

[25] X. Lin, "Map displays for information retrieval," *J. Amer. Soc. Inf. Sci.*, vol. 48, no. 1, pp. 40–54, 1997.

[26] P. R. Cohen and R. Kjeldsen, "Information retrieval by constrained spreading activation in semantic networks," *Inf. Process. Manage.*, vol. 23, no. 4, pp. 255–268, 1987.

[27] C. Tang, Z. Xu, and S. Dwarkadas, "Peer-to-peer information retrieval using self-organizing semantic overlay networks," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2003, pp. 175–186.

[28] X. Lin, D. Soergel, and G. Marchionini, "A self-organizing semantic map for information retrieval," in *Proc. 14th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1991, pp. 262–269.

[29] K. Gai, M. Qiu, S. Jayaraman, and L. Tao, "Ontology-based knowledge representation for secure self-diagnosis in patient-centered teleheath with cloud systems," in *Proc. IEEE 2nd Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)*, Nov. 2015, pp. 98–103.

[30] K. Glocker *et al.*, "Optimizing a query by transformation and expansion," *Stud. Health Technol. Inform.*, vol. 243, pp. 197–201, Sep. 2017.

[31] M. Sheth, S. Popat, and T. Vyas, "Word sense disambiguation for Indian languages," in *Proc. Int. Conf. Emerg. Res. Comput., Inf., Commun. Appl.*, 2016, pp. 583–593.

[32] M. K. Elhadad, K. M. Badran, and G. I. Salama, "A novel approach for ontology-based feature vector generation for Web text document classification," *Int. J. Softw. Innov.*, vol. 6, no. 1, pp. 1–10, 2018.

[33] B. Selvalakshmi and M. Subramaniam, "Intelligent ontology based semantic information retrieval using feature selection and classification," *Cluster Comput.*, vol. 21, pp. 1–11, 2018.

[34] Y. Liu, Y. Huang, S. Zhang, D. Zhang, and N. Ling, "Integrating object ontology and region semantic template for crime scene investigation image retrieval," in *Proc. 12th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2017, pp. 149–153.

[35] L. Khan, D. McLeod, and E. Hovy, "Retrieval effectiveness of an ontology-based model for information selection," *VLDB J.-Int. J. Very Large Data Bases*, vol. 13, no. 1, pp. 71–85, 2004.

[36] V.-W. Soo, C.-Y. Lee, C.-C. Li, S. L. Chen, and C.-C. Chen, "Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques," in *Proc. Joint Conf. Digit. Libraries*, May 2003, pp. 61–72.

[37] A. Gómez-Pérez, F. Ortiz-Rodríguez, and B. Villazón-Terrazas, "Ontology-based legal information retrieval to improve the information access in e-government," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 1007–1008.

[38] C. Cesarano, A. d'Acierno, and A. Picariello, "An intelligent search agent system for semantic information retrieval on the Internet," in *Proc. 5th ACM Int. Workshop Web Inf. Data Manage.*, 2003, pp. 111–117.

[39] J. M. Mannan and M. Sundarambal, "Semantic term based information retrieval using ontology."

[40] J. Paralic and I. Kostial, "Ontology-based information retrieval," in *Proc. Inf. Intell. Syst.*, 2003, pp. 23–28.

[41] P. Castells, M. Fernández, D. Vallet, P. Mylonas, and Y. Avrithis, "Self-tuning personalized information retrieval in an ontology-based framework," in *Proc. Int. Conf. Move Meaningful Internet Syst.*, 2005, pp. 977–986.

[42] D. Vallet, M. Fernández, and P. Castells, "An ontology-based information retrieval model," in *Proc. Eur. Semantic Web Conf.*, 2005, pp. 455–470.

[43] R. Ahmed-Ouamer and A. Hammache, "Ontology-based information retrieval for e-Learning of computer science," in *Proc. Int. Conf. Mach. Web Intell. (ICMWI)*, Oct. 2010, pp. 250–257.

[44] V. M. Ngo and T. H. Cao. (2018). "A generalized vector space model for ontology-based information retrieval." [Online]. Available: https://arxiv.org/abs/1807.07779

[45] X. Zhang, X. Chen, X. Hou, and T. Zhuang, "A semantic retrieval framework for engineering domain knowledge," in *Proc. Int. Conf. Intell. Comput.*, 2011, pp. 488–493.

[46] A. Meštrović, "Collaboration networks analysis: Combining structural and keyword-based approaches," in *Proc. Int. Keystone Conf. Semantic Keyword-Based Search Struct. Data Sources*, 2017, pp. 111–122.

[47] N. Stojanovic, R. Studer, and L. Stojanovic, "An approach for step-by-step query refinement in the ontology-based information retrieval," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Sep. 2004, pp. 36–43.

[48] N. Stojanovic and L. Stojanovic, "A logic-based approach for query refinement in ontology-based information retrieval systems," in *Proc. 16th IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2004, pp. 450–457.

[49] J. Zhai, Y. Liang, Y. Yu, and J. Jiang, "Semantic information retrieval based on fuzzy ontology for electronic commerce," *JSW*, vol. 3, no. 9, pp. 20–27, 2008.

[50] L. A. F. Park and K. Ramamohanarao, "Efficient storage and retrieval of probabilistic latent semantic information for information retrieval," *VLDB J.*, vol. 18, no. 1, pp. 141–155, 2009.

[51] I. Ruthven, "Information retrieval in context," in *Advanced Topics in Information Retrieval*. Berlin, Germany: Springer, 2011, pp. 187–207.

[52] H. Liaqaut, N. Iftikhar, and M. A. Qadir, "Context aware information retrieval using role ontology and query schemas," in *Proc. IEEE Multitopic Conf. (INMIC)*, Dec. 2006, pp. 244–249.

[53] J.-Y. Wang and Z. Zhu, "Framework of multi-agent information retrieval system based on ontology and its application," in *Proc. IEEE Int. Conf. Mach. Learn.*, vol. 3, Jul. 2008, pp. 1615–1620.

[54] J. Mustafa, S. Khan, and K. Latif, "Ontology based semantic information retrieval," in *Proc. 4th Int. IEEE Conf. Intell. Syst. (IS)*, vol. 3, Sep. 2008, pp. 14–22.

[55] J. Tuominen, T. Kauppinen, K. Viljanen, and E. Hyvönen, "Ontology-based query expansion widget for information retrieval," in *Proc. 5th Workshop Scripting Develop. Semantic Web (SFSW) 6th Eur. Semantic Web Conf. (ESWC)*, vol. 449, 2009, pp. 52–57.

[56] K. S. Mule and A. Waghmare, "Context based information retrieval based on ontological concepts," in *Proc. Int. Conf. Inf. Process. (ICIP)*, Dec. 2015, pp. 491–495.

[57] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: An ontology-based approach," *J. Web Semantics*, vol. 9, no. 4, pp. 434–452, 2011.

[58] H. Xiao and I. F. Cruz, "A multi-ontology approach for personal information management," in *Proc. Int. Conf. Semantic Desktop Workshop, Next Gener. Inf. Manage. D, Collaboration Infrastruct.*, vol. 175, 2005, pp. 1–15.

[59] J. Guan, X. Zhang, J. Deng, and Y. Qu, "An ontology-driven information retrieval mechanism for semantic information portals," in *Proc. 1st Int. Conf. Semantics, Knowl. Grid (SKG)*, 2005, p. 63.

[60] B.-C. C. C.-H. Hu and M.-Y. Ju, "Intelligent information retrieval applying automatic constructed fuzzy ontology," in *Proc. 6th Int. Conf. Mach. Learn. Cybern.*, 2007, pp. 2239–2244.

[61] Z. Gu and S.-N. Yu, "Ontology-based inverted tables in information retrieval system," in *Proc. 3rd Int. Conf. Semantics, Knowl. Grid*, Oct. 2007, pp. 354–357.

[62] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann, "Ontology refinement for improved information retrieval," *Inf. Process. Manage.*, vol. 46, no. 4, pp. 426–435, 2010.

[63] L. Dong, P. K. Srimani, and J. Z. Wang, "Ontology graph based query expansion for biomedical information retrieval," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2011, pp. 488–493.

[64] P. Yadav and R. P. Singh, "An ontology-based intelligent information retrieval method for document retrieval," *Int. J. Eng. Sci. Technol.*, vol. 4, no. 9, pp. 3970–3974, 2012.

[65] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 101–110.

[66] T. Hofmann, "Probabilistic latent semantic indexing," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 211–218, 2017.

[67] R. Ozcan and Y. Aslangdogan, "Concept based information access using ontologies and latent semantic analysis," Dept. Comput. Sci. Eng., Univ. Texas Arlington, Arlington, TX, USA, Tech. Rep. CSE-2004-8, 2004.

[68] G. Nagypál, "Improving information retrieval effectiveness by using domain knowledge stored in ontologies," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.*, 2005, pp. 780–789.

[69] M. Song, I.-Y. Song, X. Hu, and R. Allen, "Semantic query expansion combining association rules with ontologies and information retrieval techniques," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, 2005, pp. 326–335.

[70] J. Hu, Z.-L. Li, and C. Guan, "A method of rough ontology-based information retrieval," in *Proc. IEEE Int. Conf. Granular Comput. (GrC)*, Aug. 2008, pp. 296–299.

[71] X. Lv and N. M. El-Gohary, "Semantic annotation for supporting context-aware information retrieval in the transportation project environmental review domain," *J. Comput. Civil Eng.*, vol. 30, no. 6, 2016, Art. no. 04016033.

[72] R. Uma and K. Muneeswaran, "OMIR: Ontology-based multimedia information retrieval system for Web usage mining," *Cybern. Syst.*, vol. 48, no. 4, pp. 393–414, 2017.

[73] M. Á. Rodríguez-García, R. Valencia-García, F. García-Sánchez, and J. J. Samper-Zapater, "Ontology-based annotation and retrieval of services in the cloud," *Knowl.-Based Syst.*, vol. 56, pp. 15–25, Jan. 2014.

[74] T. Slimani. (2013). "Description and evaluation of semantic similarity measures approaches." [Online]. Available: https://arxiv.org/abs/1310.8059

[75] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. Petrakis, and E. Milios, "Information retrieval by semantic similarity," in *Medical Informatics: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2009, pp. 647–665.

[76] T. Vrbanec and A. Meštrović, "The struggle with academic plagiarism: Approaches based on semantic similarity," in *Proc. 40th Jubilee Int. ICT Convention–MIPRO*, 2017, pp. 1–6.

[77] I. Lukšová, "Ontology enrichment based on unstructured text data," 2013.

[78] S. Lawrence and C. L. Giles, "Accessibility of information on the Web," *Intell.*, vol. 11, no. 1, pp. 32–39, 2000.

[79] J. Z. Wang, N. Boujemaa, A. Del Bimbo, D. Geman, A. G. Hauptmann, and J. Tesić, "Diversity in multimedia information retrieval research," in *Proc. 8th ACM Int. Workshop Multimedia Inf. Retr.*, 2006, pp. 5–12.

[80] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[81] K. Guo, Z. Liang, Y. Tang, and T. Chi, "SOR: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data," *J. Comput. Sci.*, vol. 28, pp. 455–465, Sep. 2018.

[82] E. Y. Chen, D. C. Gibbon, L. W. Ruedisueli, and B. Shahraray, "Method for content-based non-linear control of multimedia playback," U.S. Patent 15/296 182, Feb. 9, 2017.

[83] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.

[84] R. R. Saritha, V. Paul, and P. G. Kumar, "Content based image retrieval using deep learning process," *Cluster Comput.*, vol. 21, pp. 1–14, 2018.

[85] H. Müller, "Text-based (image) retrieval," Tech. Rep., 2010.

[86] A. M. Riad, H. K. Elminir, and S. Abd-Elghany, "A literature review of image retrieval based on semantic concept," *Int. J. Comput. Appl.*, vol. 40, no. 11, pp. 12–19, 2012.

[87] K. Saenko and T. Darrell, "Unsupervised learning of visual sense models for polysemous words," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1393–1400.

[88] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Inf. Syst.*, vol. 37, no. 4, pp. 294–305, 2012.

[89] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[90] V. Vijayarajan and M. Dinakaran, "A review on ontology based document and image retrieval methods," *Indian J. Sci. Technol.*, vol. 9, no. 47, pp. 1–13, 2016.

[91] D. Podder, J. Mukherjee, S. M. Aswatha, J. Mukherjee, and S. Sural, "Ontology-driven content-based retrieval of heritage images," in *Heritage Preservation*. Singapore: Springer, 2018, pp. 143–160.

[92] V. Vijayarajan, M. Dinakaran, P. Tejaswin, and M. Lohani, "A generic framework for ontology-based information retrieval and image retrieval in Web data," *Human-Centric Comput. Inf. Sci.*, vol. 6, no. 1, p. 18, 2016.

[93] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[94] S.-K. Chang and T. L. Kunil, "Pictorial data-base systems," *Computer*, vol. 14, no. 11, pp. 13–21, Nov. 1981.

[95] Z.-C. Huang, P. P. Chan, W. W. Ng, and D. S. Yeung, "Content-based image retrieval using color moment and gabor texture feature," in *Proc. Int. Conf. Mach. Learn. (ICMLC)*, vol. 2, Jul. 2010, pp. 719–724.

[96] J. M. Corridoni, A. Del Bimbo, and P. Pala, "Image retrieval by color semantics," *Multimedia Syst.*, vol. 7, no. 3, pp. 175–183, 1999.

[97] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database-query by visual example," in *Proc. 11th IAPR Int. Conf. IEEE Pattern Recognit. Conf. Comput. Vis. Appl.*, 1992, pp. 530–533.

[98] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 3, pp. 301–316, Mar. 2004.

[99] C. Kurtz, A. Depeursinge, S. Napel, C. F. Beaulieu, and D. L. Rubin, "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications," *Med. Image Anal.*, vol. 18, no. 7, pp. 1082–1100, 2014.

[100] E. Y. Chang, B. Li, G. Wu, and K. Goh, "Statistical learning for effective visual information retrieval," in *Proc. Int. Conf. Image Process. (ICIP)*, vol. 3, Sep. 2003, p. III-609.

[101] Y. R. Y. Rui, T. S. Huang, S. Mehrotra, and M. Ortega, "A relevance feedback architecture for content-based multimedia information retrieval systems," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Jun. 1997, pp. 82–89.

[102] Q.-F. Zheng and W. Gao, "Constructing visual phrases for effective and efficient object-based image retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 5, no. 1, p. 7, 2008.

[103] P. Chandrika and C. V. Jawahar, "Multi modal semantic indexing for image retrieval," in *Proc. ACM Int. Conf. Image Video Retr.*, 2010, pp. 342–349.

[104] R.-B. Huang, S.-L. Dong, and M.-H. Du, "A semantic retrieval approach by color and spatial location of image regions," in *Proc. Congr. Image Signal Process. (CISP)*, vol. 2, May 2008, pp. 466–470.

[105] N. Van Nguyen, A. Boucher, J.-M. Ogier, and S. Tabbone, "Region-based semi-automatic annotation using the bag of words representation of the keywords," in *Proc. 5th Int. Conf. Image Graph. (ICIG)*, Sep. 2009, pp. 422–427.

[106] W.-C. Lin, M. Oakes, and J. Tait, "Improving image annotation via representative feature vector selection," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1774–1782, 2010.

[107] J. A. dos Santos, C. D. Ferreira, R. D. S. Torres, M. A. Gonçalves, and R. A. Lamparelli, "A relevance feedback method based on genetic programming for classification of remote sensing images," *Inf. Sci.*, vol. 181, no. 13, pp. 2671–2684, 2011.

[108] H. Modaghegh, H. Sadoghi Yazdi, and H. R. Pourreza, "Learning of relevance feedback using a novel kernel based neural network," *Austral. J. Basic Appl. Sci.*, vol. 4, no. 2, pp. 171–186, 2010.

[109] L. Shao, "An efficient local invariant region detector for image retrieval," in *Proc. Can. Conf. Comput. Robot Vis. (CRV)*, May 2009, pp. 208–212.

[110] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," in *Proc. 8th ACM Int. Conf. Multimedia*, 2000, pp. 31–37.

[111] Z. Muda, "Ontological description of image content using regions relationships," 2008.

[112] Y. Li, J. Lu, Y. Zhang, R. Li, and W. Xu, "Ensemble of two-class classifiers for image annotation," in *Proc. Int. Workshop Geosci. Remote Sens. Educ. Technol. Training (ETT GRS)*, vol. 1, Dec. 2008, pp. 763–767.

[113] S. Feng and D. Xu, "Transductive multi-instance multi-label learning algorithm with application to automatic image annotation," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 661–670, 2010.

[114] R. He, N. Xiong, L. T. Yang, and J. H. Park, "Using multi-modal semantic association rules to fuse keywords and visual features automatically for Web image retrieval," *Inf. Fusion*, vol. 12, no. 3, pp. 223–230, 2011.

[115] J. Hou, D. Zhang, Z. Chen, L. Jiang, H. Zhang, and X. Qin, "Web image search by automatic image annotation and translation," in *Proc. 17th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, 2010, pp. 105–108.

[116] R. C. F. Wong and C. H. C. Leung, "Automatic semantic annotation of real-world Web images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1933–1944, Nov. 2008.

[117] H. Xu, X. Zhou, L. Lin, Y. Xiang, and B. Shi, "Automatic Web image annotation via Web-scale image semantic space learning," in *Advances in Data and Web Management*. Berlin, Germany: Springer, 2009, pp. 211–222.

[118] V. Bante and A. Bhute, "A text based video retrieval using semantic and visual approach," *Int. Res. J. Eng. Technol.*, vol. 2, no. 7, pp. 1–6, 2015.

[119] C. Jawahar, B. Chennupati, B. Paluri, and N. Jammalamadaka, "Video retrieval based on textual queries," in *Proc. 13th Int. Conf. Adv. Comput. Commun.*, Coimbatore, Indian, 2005, pp. 1–6.

[120] G. Antoniou and F. Van Harmelen, *A Semantic Web Primer*. Cambridge, MA, USA: MIT Press, 2004.

[121] K. Yanai *et al.*, "Automatic extraction of relevant video shots of specific actions exploiting Web data," *Comput. Vis. Image Understand.*, vol. 118, pp. 2–15, Jan. 2014.

[122] B. V. Patel and B. B. Meshram. (2012). "Content based video retrieval systems." [Online]. Available: https://arxiv.org/abs/1205.1641

[123] S. Marchand-Maillet, "Content-based video retrieval: An overview," Tech. Rep., 2000.

[124] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 593–601, 2001.

[125] A. Girgensohn, J. Adcock, M. Cooper, and L. Wilcox, "Interactive search in large video collections," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, 2005, pp. 1395–1398.

[126] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann, "Fast and accurate content-based semantic search in 100 m Internet videos," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 49–58.

[127] L. F. Sikos, "Spatiotemporal reasoning for complex video event recognition in content-based video retrieval," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Springer, 2017, pp. 704–713.

[128] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, p. 5, 2008.

[129] P. Sinha and R. Jain, "Concept annotation and search space decrement of digital photos using optical context information," *Proc. SPIE*, vol. 6820, Jan. 2008, Art. no. 68200H.

[130] A. Mittal and S. Gupta, "Automatic content-based retrieval and semantic classification of video content," *Int. J. Digit. Libraries*, vol. 6, no. 1, pp. 30–38, 2006.

[131] L. Huayong, "Content-based TV sports video retrieval based on audio-visual features and text information," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Sep. 2004, pp. 481–484.

[132] J.-F. Chen, H.-Y. M. Liao, and C.-W. Lin, "Fast video retrieval via the statistics of motion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2005, p. ii-437.

[133] Y. Dai, "Semantic tolerance-based image representation for large image/video retrieval," in *Proc. 3rd Int. IEEE Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Dec. 2007, pp. 1005–1012.

[134] S. Handschuh and S. Staab, *Annotation for the Semantic Web*, vol. 96. Amsterdam, The Netherlands: IOS Press, 2003.

[135] A. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, "Video content annotation using visual analysis and a large semantic knowledgebase," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2003, p. II.

[136] G. C. Stein, J. Rittscher, and A. Hoogs, "Enabling video annotation using a semantic database extended with visual knowledge," in *Proc. Int. Conf. Multimedia Expo (ICME)*, vol. 1, Jul. 2003, p. I-161.

[137] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 9, no. 4, pp. 580–588, Jun. 1999.

[138] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua, "Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 33–42.

[139] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann, "Bridging the ultimate semantic gap: A semantic search engine for Internet videos," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 27–34.

[140] G. Y. Bobhate and U. A. Jogalekar, "Reduction of the semantic gap using pseudo relevance feedback algorithm," *Int. J. Adv. Comput. Eng. Netw.*, vol. 1, no. 2, pp. 56–60, 2013.

[141] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Region-based image retrieval using an object ontology and relevance feedback," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 6, 2004, Art. no. 231946.

[142] A. Swartz, "MusicBrainz: A semantic Web service," *IEEE Intell. Syst.*, vol. 17, no. 1, pp. 76–77, Jan. 2002.

[143] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 105–112.

[144] L. Khan and D. McLeod, "Audio structuring and personalized retrieval using ontologies," in *Proc. IEEE Adv. Digit. Libraries*, May 2000, pp. 116–126.

[145] P. Kannan, P. Bala, and G. Aghila, "Multimedia retrieval using ontology for semantic Web—A survey," *Int. J. Soft Comput. Eng.*, vol. 2, no. 1, pp. 47–51, 2012.

[146] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the cal500 data set," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 439–446.

[147] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[148] B. Feiten, R. Frank, and T. Ungvary, "Organization of sounds with neural nets," in *Proc. Int. Comput. Music Conf.*, 1991, p. 441.

[149] G. Lu, "Indexing and retrieval of audio: A survey," *Multimedia Tools Appl.*, vol. 15, no. 3, pp. 269–290, 2001.

[150] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio information retrieval using semantic similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 2, Apr. 2007, p. II-725.

[151] G. Wichern, H. Thornburg, and A. Spanias, "Unifying semantic and content-based approaches for retrieval of environmental sounds," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2009, pp. 13–16.

[152] S. K. Gupta, A. Sinha, and M. Jain, "Cross lingual information retrieval with SMT and query mining," *Adv. Comput.*, vol. 2, no. 5, p. 33, 2011.

[153] C. Buckley, M. Mitra, J. Walz, and C. Cardie, "Using clustering and superconcepts within SMART: TREC 6," *Inf. Process. Manage.*, vol. 36, no. 1, pp. 109–131, 2000.

[154] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2002, pp. 299–306.

[155] S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer, "Automatic cross-language retrieval using latent semantic indexing," in *Proc. AAAI Spring Symp. Cross-Lang. Text Speech Retr.*, vol. 15, 1997, p. 21.

[156] V. K. Sharma and N. Mittal, "Cross lingual information retrieval (CLIR): Review of tools, challenges and translation approaches," in *Information Systems Design and Intelligent Applications*. New Delhi, India: Springer, 2016, pp. 699–708.

[157] V. K. Sharma and N. Mittal, "Cross-lingual information retrieval: A dictionary-based query translation approach," in *Advances in Computer and Computational Sciences*. Singapore: Springer, 2018, pp. 611–618.

[158] J. S. McCarley, "Should we translate the documents or the queries in cross-language information retrieval?" in *Proc. 37th Annu. Meeting Assoc. Comput. Linguistics Comput. Linguistics*, 1999, pp. 208–214.

[159] M. Abusalah, J. Tait, and M. Oakes, "Cross language information retrieval using multilingual ontology as translation and query expansion base," *Polibits*, vol. 40, pp. 13–16, Dec. 2009.

[160] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," *ACM SIGIR Forum*, vol. 31, no. S1, pp. 84–91, 1997.

[161] L. Ballesteros and W. B. Croft, "Resolving ambiguity for cross-language retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1998, pp. 64–71.

[162] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and K. Järvelin, "Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002," *Inf. Retr.*, vol. 7, nos. 1–2, pp. 99–119, 2004.

[163] D. M. Christopher, R. Prabhakar, and S. Hinrich, "Introduction to information retrieval," *Introduction Inf. Retr.*, vol. 151, no. 177, p. 5, 2008.

[164] Z. Yahya, M. T. Abdullah, A. Azman, and R. A. Kadir, "Query translation using concepts similarity based on Quran ontology for cross-language information retrieval," *J. Comput. Sci.*, vol. 9, no. 7, p. 889, 2013.

[165] C. Lu, Y. Xu, and S. Geva, "Translation disambiguation in Web-based translation extraction for English-Chinese CLIR," in *Proc. ACM Symp. Appl. Comput.*, 2007, pp. 819–823.

[166] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin, "Dictionary-based cross-language information retrieval: Problems, methods, and research findings," *Inf. Retr.*, vol. 4, nos. 3–4, pp. 209–230, 2001.

[167] N. A. Nasharuddin and M. T. Abdullah, "Cross-lingual information retrieval: State-of-the-Art," *Electron. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 1, pp. 1–5, 2011.

[168] C.-J. Lee, J. S. Chang, and J.-S. R. Jang, "Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources," *ACM Trans. Asian Lang. Inf. Process.*, vol. 5, no. 2, pp. 121–145, 2006.

[169] C.-C. Lin, Y.-C. Wang, C.-H. Yeh, W.-C. Tsai, and R. T.-H. Tsai, "Learning weights for translation candidates in Japanese–Chinese information retrieval," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7695–7699, 2009.

[170] E. Picchi and C. Peters, "Cross-language information retrieval: A system for comparable corpus querying," in *Cross-Language Information Retrieval*. Boston, MA, USA: Springer, 1998, pp. 81–92.

[171] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 74–81.

[172] W. Kraaij, J.-Y. Nie, and M. Simard, "Embedding Web-based statistical translation models in cross-language information retrieval," *Comput. Linguistics*, vol. 29, no. 3, pp. 381–419, 2003.

[173] M. Braschler and P. Scäuble, "Multilingual information retrieval based on document alignment techniques," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, 1998, pp. 183–197.

[174] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2010, pp. 403–411.

[175] D. S. Munteanu and D. Marcu, "Extracting parallel sub-sentential fragments from non-parallel corpora," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 81–88.

[176] P. McNamee and J. Mayfield, "Comparing cross-language query expansion techniques by degrading translation resources," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2002, pp. 159–166.

[177] Y. Zhang and P. Vines, "Using the Web for automated translation extraction in cross-language information retrieval," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2004, pp. 162–169.

[178] Systran. *Systran Language Translation Technology*. Accessed: Apr. 30, 2018. [Online]. Available: http://www.systransoft.com

[179] A. Boretz and A. Adam. (Feb. 2009). *AppTek Launches Hybrid Machine Translation Software*. [Online]. Available: https://SpeechTechMag.com

[180] *A Brief Guide to Prompt Machine Translation Technology*, PROMT, Saint Petersburg, Russia, 2005.

[181] D. A. Gachot, E. Lange, and J. Yang, "The SYSTRAN NLP browser: An application of machine translation technology in cross-language information retrieval," in *Cross-Language Information Retrieval*. Boston, MA, USA: Springer, 1998, pp. 105–118.

[182] Knowledge, Multilingual Corporate, "The systran linguistics platform: A software solution to manage multilingual corporate knowledge," Systran, Seoul, South Korea, White Paper, 2002.

[183] D. Thenmozhi and C. Aravindan, "Ontology-based Tamil–English cross-lingual information retrieval system," *Sādhanā*, vol. 43, no. 10, p. 157, 2018.

[184] D. Farwell, L. Gerber, and E. H. Hovy, Eds., *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. London, U.K.: Springer-Verlag, 1998.

[185] D. W. Oard, "A comparative study of query and document translation for cross-language information retrieval," in *Proc. Conf. Assoc. Mach. Transl. Amer.*, 1998, pp. 472–483.

[186] A. Chen and F. C. Gey, "Combining query translation and document translation in cross-language retrieval," in *Proc. Workshop Cross-Lang. Eval. Forum Eur. Lang.*, 2003, pp. 108–121.

[187] S. Tripathi and J. K. Sarkhel, "Approaches to machine translation," 2010.

[188] M. Araujo, J. Reis, A. Pereira, and F. Benevenuto, "An evaluation of machine translation for multilingual sentence-level sentiment analysis," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, 2016, pp. 1140–1145.

[189] M. Braschler, "Combination approaches for multilingual text retrieval," *Inf. Retr.*, vol. 7, nos. 1–2, pp. 183–204, 2004.

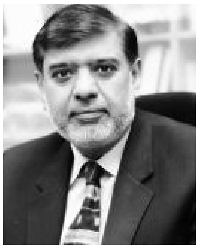
[190] S. Sloto *et al.*, "Leveraging data resources for cross-linguistic information retrieval using statistical machine translation," in *Proc. 13th Conf. Assoc. Mach. Transl. Amer.*, vol. 2, 2018, pp. 223–233.

MUHAMMAD NABEEL ASIM received the bachelor's degree from the University of Management and Technology and the master's degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, where he is currently a Research Assistant Manager with the Al-Khawarizmi Institute of Computer Science. His research interests include text classification, semantic Web, information retrieval, and deep learning.

MUHAMMAD WASIM is currently pursuing the Ph.D. degree in computer science with the University of Engineering and Technology, Lahore. His Ph.D. degree is concerned with improving biomedical question answering systems, which involves techniques from both information retrieval and natural language processing. He is currently an Experienced Research Associate with the University of Engineering and Technology. With strong industry background, he is committed to improve the state of the art in his area of interest and develop innovative research-oriented products. His research interests include biomedical information retrieval, information extraction, ontology learning, and deep learning.

MUHAMMAD USMAN GHANI KHAN received the Ph.D. degree from The University of Sheffield, U.K. His Ph.D. degree was concerned with statistical modeling for machine vision signals, specifically language descriptions of video streams. He is currently an Associate Professor with the Department of Computer Science, University of Engineering and Technology, Lahore. He has been studying on spoken language processing using statistical approaches with applications such as information extraction from speech and speech summarisation. His recent work is concerned with multimedia, incorporating text, and audio and visual processing into one frame work.

NASIR MAHMOOD is currently pursuing the Ph.D. degree with the University of Engineering and Technology, Lahore. He worked for United Nations on different IT positions and working on IT projects at the Police Department of Punjab. He is also an Experienced IT Professional having IT project planning, designing, management, and execution experience of both international and national levels. His research area is human behavior modeling and prediction from social media contents. His research focused on developing knowledge base of extremist human behavior and designing association matrix among entities using advanced ontology, link analysis, and activity charting algorithms.



WAQAR MAHMOOD received the bachelor's degree from the University of Engineering and Technology (UET), Lahore, in 1989, and the master's and Ph.D. degrees from Georgia Tech, USA, in 1992 and 1996, respectively, all in electrical engineering. He has served in industry as well as academia in North America and Pakistan. He served as the Director of process development at CIENA Corporation, MD, USA. He is currently serving as the Sultan Qaboos IT Chair at UET, where he is also the Director of the Al-Khwarizmi Institute of Computer Science, a center of excellence for ICT research and development. He has published over 80 reviewed papers and holds 12 U.S. patents to his credit. His research interests include optical communications, process technology development, discrete event systems, networking systems, wireless systems, energy optimization, power systems, and renewable energy. He was selected as a Voting Member and a U.S. Delegate to the IEC Committee on Optical Interconnect Standardization. He has consistently received research funding in a number of IT and renewable energy technologies.