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

20BCE204 Dhyan Patel, 20BCE261 Dhruvi Shah, and 20BCE315 Vedant Patel

Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

Abstract. On a daily basis, a large amount of data is acquired, studied, and utilized by a large number of people on the internet. Unstructured data, such as website content, books, journals, and files, make up a sizable portion of the data available on the Internet. Acquiring relevant information from such massive data has become difficult and time-consuming. Simple keyword-based retrieval techniques rely heavily on data statistics, resulting in word mismatch issues due to a given word's unavoidable semantic and context variations. This highlights the urgent necessity to arrange such vast amounts of data so that information may be quickly processed in a broad context while taking the semantics of data into account. In addition to being widely used on the semantic Web to store raw data in an orderly and organized manner, ontologies have also significantly improved the performance of various information retrieval techniques. Data is retrieved by ontological IR systems based on how semantically comparable the user's search query and the indexed data are. In this article, ontology-based information retrieval techniques for text, multimedia, and multilingual data types are reviewed. Additionally, we contrast and group the most recent techniques applied in the IR methods mentioned above, as well as their disadvantages and benefits.

Keywords: Ontology · cross lingual retrieval · text retrieval · multimedia retrieval.

1 Introduction

People have been working to create an effective method of information storage, search, and retrieval ever since written languages first appeared. Information retrieval was first only considered relevant to library sciences. The notion that machine-based retrieval of information techniques will become widely used in the near future was first put forth in 1945 by Johnston and Webber.[7] People began utilizing technology for quick information retrieval in the early 1970s. However, the majority of those systems were created for specific audiences like government organizations, academic institutions, and medical professionals.

However, as technology developed and social media sites like Twitter, WhatsApp, and Facebook were created, users began to share vast amounts of multimedia data, which includes audio, video, text, and images. The proliferation of technical tools and users also caused data in numerous fields, including e-learning,

[⋆] Supported by organization x.

e-government, and e-commerce, to start growing exponentially. In order to avoid having many systems that were initially created for different groups, it was vital to develop an all-purpose information retrieval system.[9] Information retrieval is a laborious and time-consuming operation due to the vast amount of information as well as its unstructured nature. To address the aforementioned domain-specific issues, many information retrieval methods were created, and numerous general-purpose search engine technologies were introduced. Unfortunately, despite their popularity, these search engines lacked the ability to semantically comprehend user-defined queries and provide the most accurate response. Most of the search engines just handled the user's requests, gave them approximative answers, and frequently returned a large number of pointless websites. Search engines, like Google, respond to user queries with a prioritized list of pertinent content in the form of a sentence or keyword feed.

It is typically difficult for users to express their complete information needs in a precise manner. The majority of keywords used by users differ from those used by documents in a database's index. For instance, a user asks what precautions medical professionals advise for diabetes. In this instance, the only papers that contain the term "practitioner" in its exact form or a synonym, such as "doctor" or "specialist," would have the most relevant answer. Therefore, in order to locate those documents, it must be decided whether or not both practitioners and doctors fall under the same notion. Numerous techniques have been developed to address this issue, using conceptual knowledge to assist users in creating effective queries. The most popular approach to conceptual knowledge in IR systems is the use of a thesaurus-like component. It displays several concepts from a domain together with the semantic connections between them.[9] Utilizing conceptual understanding as an inherent component of IR systems is another strategy. These techniques have shown to be quite helpful in the field of information retrieving, and the paradigm of IR has changed from straightforward keyword-based approaches to concepts-based ones. In this paper, we have covered the most recent and cutting-edge ontology-based semantic information retrieval approaches (Textual IR, Multimedia IR, and Cross-lingual IR) in this work. We evaluate and categorize the most recent and varied methods applied to the aforementioned ontology-based information retrieval.

2 Text Information Retrieval Methodologies

2.1 Vector Space Model

In the model of vector space, also known as the vector model, the idea of articles and queries for searching are represented as vectors, and the cosine measure is used to quantify how similar they are to one another. The cosine resemblance measure quantifies the degree of resemblance between the query vector and an array of textual texts. Aside from that, vector space-based IR also makes use of additional well-known similarity measures like tf-idf and Okapi BM25; however, ontological IR has not yet made use of these metrics. Additionally, ontologies are essential for concept retrieval from text and queries.

The survey of some studies conducted in this area by many eminent researchers on this subject will be examined in the next section.

By Paralic and Kostial,[2] the first method for leveraging ontology-based retrieval was presented. To extract a set of pertinent concepts from the queries, they employed ontology. According to their methodology, relevant ideas for a particular query will already be present in cutting-edge built ontologies like WordNet. For each article, an ontology was used to extract a collection of related concepts. In order to rank the research results as to how close they corresponded to the user query, the returned set of ideas was contrasted with the ideas included in the user query. A score was then computed. They evaluated the integrity of the proposed system using 1239 documents taken from the MEDLINE corpus, wherein the search term ”Cystic Fibrosis” was utilized to gather papers.

Castells put out the second strategy, which is based on the IR system that was first proposed. However, this system includes a few extra capabilities that are user-specific and based on semantics. For content retrieval, they employed an ontology-based retrieval framework, and each domain notion was given a user personalization score. The aforementioned technique was used to generate an assessment of the similarity between keywords and document concepts, which was then added to a customer preference score (personalized score). Each document was given a final grade, after which they were ranked. They also provided a way to instantly alter the level of personalization. They evaluated the effectiveness of their suggested methods using 145,316 documents from the internet.

Ahmed-Ouamer and Hammache[1] presented yet another information retrieval method for e-learning that is based on ontologies. In a vector space model, the query and the documents were both used, and the similarity between them was assessed using a cosine similarity measure. Their proposed IR system, termed OBIREX, used an ontology for storing a collection of documents and semantic links to enable inferences over all pertinent texts.

Mestrovic[12] and others suggested using an ontology-based approach with vector space-based IR. Their approach relies on the fundamental taxonomy produced from a linguistic database or linked data set for query extension and document generation. They developed a mapping tool for placing those ontological levels onto the basic taxonomy. The vector space structure was weighted using a system that was developed. They took advantage of the conceptual and lexical connections between ideas and terms in order to reduce the vector's size in IR and avoid problems with vocabulary mismatch.

2.2 Probability Based Information Retrieval Methodologies

Instead of utilizing a cosine metric, this method uses a probability distribution to assess how similar the pages are to the user's query. The consumer's original query is improved and refined using domain ontologies. This expanded query is then used to compute the total score of the documents. The relevance of a document is determined by assessing the probability of each relevant (expanded query-based phrase) and non-relevant term inside the same text. Prior to this, several efforts had been made by scholars to include semantics in possibility-based IR systems using domain ontologies. Stojanovic suggested a logic-based search refinement to support ontology-based IR systems. They used domain ontology to add to the user's first inquiry. Then, to examine larger question phrases, they conducted a test of assessment known as an acceptance test. Their recommended approach used hazy probability and progressed gradually to focus on the stated question.

Zhai[23] built on the research of Stojanovic and also proposed the concept of imprecise ontologies for information retrieval in the setting of e-commerce. They proposed three components for the system: concepts, conceptual characteristics, and values reflecting those qualities. Both the common data types and the linguistic interpretations of fuzzy notions are valid values for properties. They reasoned that questions might be expanded by combining domain ontologies and fuzzy language variables. They also provided a Probabilistic Latent Semantic Indexing, which determined document scores through query expansion.

2.3 Context Based Information Retrieval Methodologies

The context of concepts like location, moment, date, and data about users, among others, is the foundation of this methodology. As soon as academics discovered how crucial these criteria are, information retrieval algorithms started using them. The search query is improved with the use of spatiotemporal domain-specific ontologies, and the context, such as user profiles and geographical regions, is incorporated into the query in order to optimize the retrieval of results. User-defined queries can occasionally be converted into RDF triples using ontologies, and document metadata can occasionally be saved as RDF multiplies as well.

The degree of RDF similarity between the documents and the query is then calculated to determine which documents are the most pertinent. utilizing precise and unambiguous search phrases, pertinent documents are located utilizing this technique, leading to effective information retrieval outcomes.

Liaqaut[10] offered a query language and a role ontology-based IR system. The user's description, setting, history, ontologies, and user comments were among the context parameters they advised modeling. Additionally, they expanded the query phrases by using word extension techniques based on online resources like WordNet, subject ontologies, and various thesaurus. To highlight the contextual problems with existing IR systems, Wang and Zhu [5] proposed an ontology-based tasking-agent IR framework. They developed command-line options according to the user's query and distributed user privileges so that the user could select whether to search using a domain ontology or their own custom-built ontology.

2.4 Semantic Similarity Based Information Retrieval Methodologies

The aforementioned techniques only consider a user question's environment or concept-based ontology. However, they do not consider the concepts and semantics of the publications. Consequently, the document itself always has a semantic gap. The semantic aspect of IR systems won't be finished until the creation of the semantic comparable measure. Query and document concepts can be conceptually closer to one another when employing semantic similarity-based models.

With this technique, the semantics of both the question and the files are enhanced for better document retrieval results. Ontologies serve to annotate texts semantically, whereas domain-specific ontologies are utilized to focus and broaden the user's query. After that, a similarity grade is created by contrasting these semantic annotations based on ontologies with semantically enhanced questions. Based on that score, the top relevant publications with the highest similarity are obtained.

In the innovative technique presented by Ozcan[15] and Aslangdogan, ontologies were used throughout the IR process. Their proposed architecture improved queries and generated information based on ontologies. In order to obtain results that were comparable to previously supplied data inside that same word area, the requested query was expanded.

A different model created by Nagypál [14] enhanced ontology-based inquiries and generated semantic data. A simple user inquiry was converted into a semantic query using a variety of ontology tools, which was subsequently used to perform the information retrieval procedure. Gu and Yu enhanced the previous research by choosing composite terms over single words.

2.5 Semantic Association Based Information Retrieval Methodologies

Principles of association and ontologies are combined in this complex subset of information retrieval to extend the query, a key element of IR systems. Bypassing the user's query to the IR engine, documents are retrieved using this manner. These documents go through preprocessing after retrieval. Stop and useful phrases are taken out after preprocessing. The remaining words in the text are handled as items, and each document that is retrieved is considered a transaction. This results in the development of transactional databases that are used to find significant correlations in the files, which are subsequently assigned to ideas using domain ontologies. Before being sent to the data retrieval engine to find relevant material, the query is enhanced using associations and ontologies. Association rules also provide algorithms that can be utilized to give documents different weights.

Later, a rough ontology-based information retrieval (IR) strategy was proposed for effective retrieval in the ambiguous information space. The suggested method might leverage the keywords from the query to find persons and properties through a search procedure. It is then used to build an approximate field for ontology-driven systems of information with the use of an equivalence relation. Finally, it computed the degree of similarity among a person and a group of documents using approximation space.

2.6 Semantic Annotation-Based Information Retrieval Methodologies

In this emerging field of IR systems, materials are annotated in relation to the words of the query, making it simpler for IR engines to find related content. The query terms give information about the files that require to be searched. As a result, it is relatively easy to annotate documents with crucial query terms.

In order to retrieve relevant instances, this approach sends user defined queries in RDF triple form to information databases and ontologies. Following that, labels are assigned to the documents depending on the situations that were found. These semantic ideas that appear in the instances are looked for across the whole text of the document. If certain ideas are found in that particular document, it is indicated by that instance. This is how documents are semantically labeled. After labeling, the papers are assigned weights based on these notes, and the ones with the highest weights are sent back to the user.

Rodrguez-Garca[18] created a structure for desired recovery of cloud services to aid the user. These efforts were achieved in two main modules: first, an archive for cloud services was developed, where each service was preserved as a semantic vector after being meaningfully annotated in the cloud's service description. Second, they used information and communication technology, or ICT, to extract the best data from broadly linked domains and get cloud services as needed by consumers. Vallet suggested a weighting system for annotations using this technique. They proposed a technique that built a retrieval system and semi-automatically annotated the texts using an ontology-based methodology.

3 Multimedia Information Retrieval

Information exchange using multimedia is seen to be more straightforward or simplistic. Data in the form of music, video, and graphics is saved on personal computers and on the Internet. The fast expansion of social media platforms like Facebook and Twitter has raised the storage rate for media data including images, music, and videos. Let's use WhatsApp, which supports instant messaging, as an example to show. Because they perceive multimedia resources to be handier, people tend to share audio messages and textual pictures more frequently than they do simple text messages. Because of this, the utilisation of multimedia resources is expanding at the speed of light, making multimedia retrieval a necessary job in the field of IR.

Descriptive characteristics from audio and visual data were collected and then translated to high level query features in the development of multimedia retrieval systems. Processing and managing enormous amounts of multimedia data, however, has become extremely laborious due to the ongoing expansion of online data. Every two years, the Internet's size roughly doubles. Around 7.7 Zettabytes of data are expected to be distributed across Internet servers at the start of 2016 according to the Counter. Data volume is anticipated to reach 40 zettabytes by the year 2020. The number of devices linked to the Internet is predicted to reach 50 billion by that time.

Additionally, MIR is a heterogeneous field with a wide variety of methodologies, supported data types, and research issues. These data types include audio, graphics, images, animation, videos, rich text, hypertext, and a combination of all these data types. If online users are thoroughly aware with the representation of multimedia contents and its structure, they may be able to access pertinent multimedia data. To do this, more techniques to bridge the semantic gap between objects and lexical libraries like FrameNet, WordNet, VerbNet, ConceptNet, etc. are required. Additionally, algorithms like SOR are also employed in Hadoop and MapReduce.[6]

Multimedia information retrieval (IR) systems are less well established than text retrieval systems, such as online search engines, which are widely established and open to the public. The overview of multimedia retrieval provided in this part will give researchers a fantastic opportunity to make ground-breaking discoveries by systematically obtaining important domain expertise.

3.1 Image Retrieval

Since the invention of Digital camera, images are taken from every corner of the Earth and also uploaded on the web in various fields. 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 easier to grasp. The huge information has made the search query more complex and with irrelevant fields. For example, if the user enters the query 'Red Ferrari', then we have to manage huge database and an efficient retrieval method. Thus the images are retrieved by two methods: 1) Text Based 2) Content Based.

Text Based TIBR are used to retrieve images from web portals in which it uses the text connected with the images itself in the form of image file name, links or hyperlinks, captions with the images,etc. When the user enters a query, different methods are applied to the query to preprocess it or solve the polysemy problem which for example like people have two ears but also the animals have. Then keywords and annotation are extracted from the query. Then the similarity is calculated from the indexed or tagged images present in the database with the annotated query. Thus, the images corresponding to the query are retrieved. The Architecture of TBIR is shown below.

Popular websites or search engine like Google, Yahoo, Bing, etc are almost refreshed 2 billion times in an hour. Thus, have more than 200 billion indexed images in their servers. Thus, they have to be fast and robust in retrieving the images which have text as a description in their files. This could lead to low precision rate in TIBR. And there would arise a problem of polysemy in which one word can be linked to many different meanings. Thus Soo[20] presented an ontological solution to extract the image of Chinese culture. In this system the text was sent to algorithm after using ontologies. The images and query both were transformed into a RDF structure and similarity is calculated between the triples of query and image. Thus, the highest matched was returned to the user.

Content-Based Image Retrieval Content Based Image retrieval is a framework that can overcome problems as it is based on the visual analysis of the contents that are part of the query. It is established on the premises of method of indexing and extracting of the low-level characteristics like color, shape and texture which automatically catalogues the images with the description given to the content. Overall architecture of the CBIR is given in the figure. The method of content-based retrieval is given below.

1. Initially the features are extricated from the images in the database.
2. Then the features are categorized or indexed and are stored in the database.
3. As the same way, processing is on the query and features are extricated.
4. Finally, the distance between both the query and features of the image stored in database is calculated. On the basis of this computation, top ranked results are retrieved.

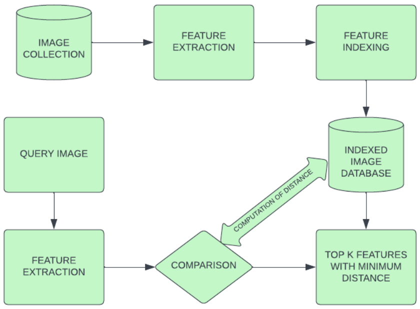

Fig. 1. Content based Image Retrieval

When the image is taken into picture, there would a number of objects or content in the image which would have a different semantic sense. To overcome this problem, as it can contain garbage values or other unnecessary things, features are extracted which form the meta data of the image. And CBIR need to have the full image. So, Computer vision technique methods are employed to extract the images from the database. Various researchers have developed their own Ontological models or system for CBIR

1. LSH (Locality sensitive Hashing): It was a bag of words model used to determine nearest neighbours on a heritage image dataset.[17]
2. OAV(Object Attribute and Values): The keywords of the user queries where rectified in form of object, attribute and values. However it had semantic issues.[22]
3. HCI (Human Computer Interaction): It was a form of relevance feedback which used to increase retrieval performance. It assigned the weight to each user and calculated the similarity which gave higher accuracy.[19]

4. QVE (Query by Visual Expansion): It is based on sketch similarity in which a user gives a sketch and an image is retrieved using the semantic techniques.[3]
5. Radlex Technology: To solve the semantic problem they used automatic annotation for sentences to high level attributes.[8]

In CBIR, there has been a huge gap or difference between the high level and low-level features. This is also called semantic gap. For example, if the user enters a query "Yellow Lamborghini", Then high level concept that Lamborghini and low-level features such as yellow. This is the semantic gap which the system should understand that There is a Lamborghini which is yellow, not every yellow object is a Lamborghini. Thus, this problem is solved by Relevance Feedback and Automatic Image Annotation. Relevance Feedback is an iterative process in which the users will specify the results as relevant and irrelevant and the system will learn from the users and iterate all over again. The solution would not be irrelevant if only small number of results are only retrieved. Thus, it might fail to provide a better or relevant results in form of large number of images. In Automatic Automation, the images are directly classified into predefined classes or keywords. In Ontological approach, Direct relationship is formed between the high level and low level content which is worthy as it reduces the semantic gap between the results and provides accurate results.

3.2 Video Retrieval

The advancement of digital devices used to capture various forms of digital data has led to a tremendous increase in its volume. This digital data can include documents, images, sounds, videos, and more. Among these, video is a particularly valuable form of digital media as it can contain images, sound, and text. Retrieving specific information from a video database based on a user's needs is known as video retrieval. Due to the large amount of data involved, efficient information retrieval approaches are necessary to manage such a vast database. This section briefly outlines the latest methods for video retrieval, including text-based and content-based approaches.

Text Based Video Retrieval (TBVR) In Text based video retrieval, videos are being retrieved on the basis of the text or captions or any other information associated with the video. The words, characters or piece of sentences are being examined in the video frames. As TBIR, the textual query is processed and polysemy problem is solved and keywords are extracted. Then annotations are made. And in the video, using OCR techniques, key words are extracted and annotated videos are stored in Database. After which similarity is computed and relevant results are retrieved. Various researchers have used ontological concepts in order close the semantic gap which proved effective.

However, Automatic Annotation proved very space and time consuming using ontological concepts. Thus, ontological text video retrieval methods are useful in finding a song of artist, an album, a speech by a political leader and other events happening which are tagged their location in the video. These methods are also useful in educational videos or speeches in which the text is easily classified. But, the choice of OCR tool is still a problem for the TBVR. Furthermore, these techniques do not perform semantic evaluation of the videos. As a result, researchers[16] are now focusing on the content of the videos themselves, rather than just the text, to improve retrieval accuracy.

Content Based Video Retrieval (CBVR) This method involves analyzing the real content of video frames to extract semantic information about the video. Instead of just text, the term ”content” in this context may refer to characteristics such as colors, textures, and shapes of objects. In Content Based Video Retrieval, initially there is segmentation of moving objects into frames which is basically the image obtained at a particular time from the video. Once The keyframes are classified or obtained, the features are divided into high and low level features. High level features deal with the semantic concepts while the low level features are color, shape, bandwidth, frame of the videos and loudness which are directly accessible from the database. For example, for a query “find images of Modiji waving”. To answer this query, the technique should have some information about Modiji that there exists a man who is a special man rather than ordinary man.

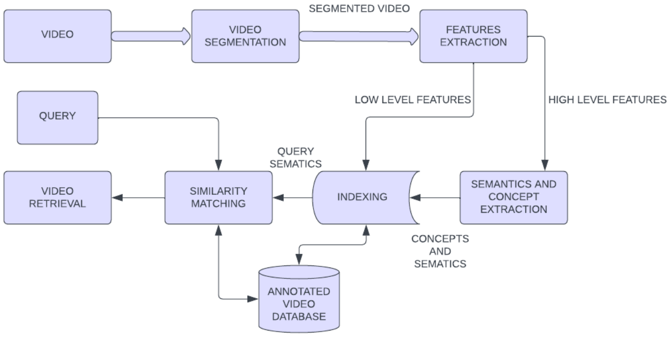

Fig. 2. Content Based Video Retrieva

Video retrieval is in active state of research as lots of different videos are being created and upload on the web. Different researchers have developed different techniques for CVBR in which indexing is done the video and then it analysed and was interpretated on automatic reasoning. In another study SVM[13] was used to minimize the gap between the high level and low-level features. Then another researcher used Bayesian Classifier[4] that dealt with the problem of semantic gap from the hug databases. Researchers have proposed two methods to address the issue of semantic gap. The first approach involves automatically generating metadata for videos, which requires the use of semantic concepts and different schemas. The second approach suggests the use of relevance feedback to learn and comprehend the semantic context of a query.

3.3 Audio Retrieval

Audio retrieval is also important as video and image as a lot of information is shared on the web in form of audio which can be recordings, speeches, songs and confidential clips also. It is available with different kind of quality which contain noises or some are of pure form. As a result, there is a need to automatically extract information from the data rather than manually searching through all of it. Therefore, it is being retrieved in text based and content based.

Text Based Audio Retrieval (TBAR) In Text based Audio retrieval, Sound information is extracted on the basis of meta data provided with the audio. In the audio retrieval process, the user's query is first pre-processed to extract relevant features such as artist name and audio type. The audio database containing various audio documents is then processed, and its features are extracted and tagged. These tags are then compared with the features obtained from the query, and their similarity is computed[21]. Finally, the audio document with the highest similarity to the query features is retrieved and returned to the user.

One clear disadvantage of these approaches is that audio descriptions can be subjective. As a result, there has been a shift in the paradigm from relying solely on text-based approaches to adopting content-based approaches.

Content Base Audio retrieval Content based audio retrieval techniques were introduced to answer queries which do not support a semantic approach. And also there were inadequacy of file name and more subjectivity in the text description, it was challenging to retrieve the audio. In CBAR, they index audio files using file names and tags given by the user, then is being compared with the query.The audio signals can be variable so the CBAR is based on set of retrieved audio features like frequency distribution.[11]

The general approach for:

1. To begin with, sound information is sorted into predetermined categories such as speech, noise, or music.
2. Next, each type of sound is handled and organized using distinct methods. For instance, if the sound is identified as speech, speech recognition techniques are utilized to recognize the words and the sound is indexed accordingly.
3. Likewise, when a query is made using an audio sample, it is processed, categorized and indexed.
4. Finally, the likeness between the query index and the audio indices in the database is assessed. Based on this similarity, the most relevant audio samples are retrieved.

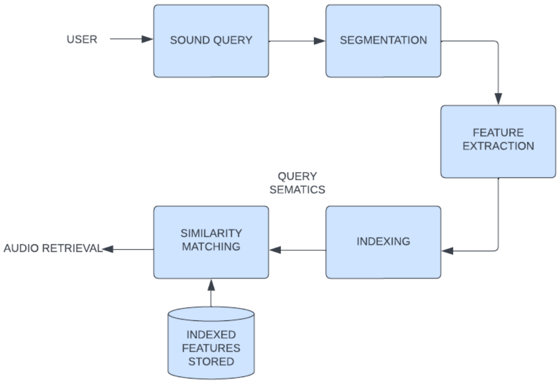

Fig. 3. Content Base Audio retrieval

4 Cross-Lingual Information Retrieval

The capacity to acquire relevant data in language other than the one used in the inquiry is a vital necessity in a field whose worldwide scope has expanded. The process of conducting a search in one language and providing results in a variety of languages is known as cross-lingual retrieval of data.

In order for the user to grasp the data that was typically retrieved, the created papers are then translated into query language. While some CLIR systems employ parallel corpora or bilingual dictionaries to translate user queries to the target language, others use machine translation to convert content from corpora in various languages into the original language. Foreign language corpus documents can now be located using a specific user's query thanks to machine translation. As an illustration, if a user types "flower arrangements" into the search bar in English, they will find that "ikebana," the Japanese word for flower arrangements, is what they are looking for. The recovery of cross-lingual knowledge is a topic that many researchers are interested in. . More seminars and programmes supporting cross-lingual information retrieval research have become accessible in recent years. For instance, the Cross-Language Evaluation Forum (CLEF), which was established in 2000, focuses on European languages. Although there are certain non-translation techniques, such as latent semantics searching, relevancy models, semantics indexing, and kindred pairing, most cross-lingual systems for retrieving data use a kind of translating mechanism, which is regarded to be the more prevalent and efficient technique. The three main topics investigated in this field are the material that needs to be interpreted, how it's supposed to be interpreted, and how to prevent inaccurate translations. Language translation is the main challenge in cross-lingual information retrieval. The most effective way to collect an enormous amount of translational data. The paper's conclusion talks about current developments in cross-lingual information retrieval, which could aid researchers in solving all of the aforementioned issues.

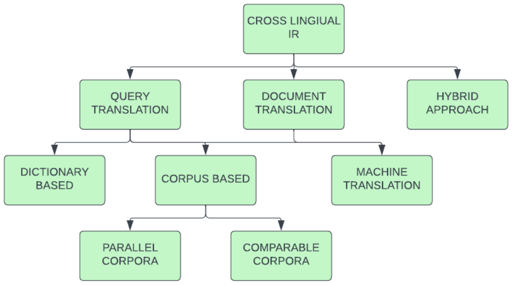

Fig. 4. Cross-Lingual Information Retrieva

4.1 Firstly, what should be translated?

The entire document, a particular query, or both of the aforementioned options must be translated. Unlike query translation, which converts an input question into the target language, document translation requires translating the complete document into the query language.

4.2 Question Translation

The user's request may be converted into the necessary language using particular multilingual dictionaries, corpora, or automated translation. In this section, some of the most significant query translation techniques are illustrated with examples.

A dictionary based approach Using MRDs (machine-readable dictionaries), the user's query is linguistically analysed to identify its keywords, which are then translated into the target language. This technique is called "dictionary-based query translation approach". MRDs are just the digital equivalents of printed dictionaries, which may be general or niche dictionaries. The texts of both general and specialised dictionaries can also be included in MRDs. A list of words from the source language and their translations into the target language make up a typical bilingual dictionary. Instead of trying to translate entire paragraphs into the target language, which would seem to require more time and effort, using dictionaries to translate the query is far simpler and faster. Cross-lingual information retrieval (CLIR) systems have reportedly been used to improve query translation performance, according to Mustafa Abusalah et al. To create a better search engine for the travel and tourist industry, the multilingual ontology

(MO) took on the function of the machine-readable dictionary (MRD). When mean average precision was used as the criterion, their suggested ontology-based solution performed better than the benchmark (MRD technique).

Considering the fact that there are numerous bilingual dictionaries in various literary works, the primary issues with a bilingual glossary-based method are translation unpredictability, phrase inflection issues, interpreting phrases, phrase compounds, specialist terminology, and typographical differences. The fixation of word inflection is made possible through lemmatization and stemming. In order to improve dictionary-based query translation and address the problem of translation ambiguity, Yahya et al. suggested a method in CLIR. Using an identical multilingual Malay and English corpus, the dictionary was produced . They assess the efficacy of three IR systems according to natural language queries, changed queries based on natural language using a corpus of words (baseline), and changed natural speech searches using the Quran ontology using a mean average accuracy and a median accuracy determined at 11 recall points. Their suggested methodology produced better retrieval outcomes for the gathering of English documents as compared to Malay resources. They showed how query modifications can improve the retrieval performance of the suggested system in a specific language.

Another factor that could be to blame for performance degradation is the absence of vocabulary coverage. It typically occurs when a query contains underlined words that aren't found in dictionaries. Even the best dictionaries can suffer from the Out-of-Vocabulary issue, also referred to as the dictionary problem. Even when search extension has been employed, it has been challenging to obtain the crucial missing terms due to the precision of user inquiries. The majority of terms and proper names employed in OOV nomenclature are completely unique. For example, to learn more about the outbreak of influenza A (HINI) viruses in Malaysia, one could type "HINI Malaysia" into an internet search engine. The lexicon from a few years ago surely does not contain the word HINI because it is brand-new. Additionally, if someone does not include the word HINI in their search, they could not come across any relevant documents.

Corpuse-Based Methods A corpus is, in essence, an accumulation of language-based elements, such as words, chapters, and text, that may have been created in a variety of languages. For query translations, two separate bilingual database management systems—parallel and similar corpora—have been deployed. Here is a brief description of each corpus.

Parallel corpora Sets of identical texts that have been translated into numerous target languages are known as parallel corpora. To clearly illustrate the exact connection between words in the language of the source and the target language, annotated paralleled corpus may be employed. They can be used to study a wide range of operations, including the translation of concepts, ideas, and data between languages. They are evolving into informational resources that both humans and machines can understand. The search query does not need to be converted into the target language in order to retrieve a specific text from an aligned corpus because it can easily match the corpus's part from that language, and the corresponding component in the desired language can then be swiftly retrieved. Typically, parallel corpora are populated via human and machine translation as well as multilingual websites. For instance, many scholars are developing multilingual corpora utilizing ”Spider” buildings, which can collect online articles with similar translations. After that, texts produced in a unique language and texts written in the target language are synchronized, either using specialized technology or by comparing texts with indicators. All information—including content date, author, special title or quantity, and acronyms—that precisely suits both the source and a particular target language document is taken into account when doing alignment utilizing indicators analysis on documents. But using tools like PTMiner, it is also common to practice synchronizing parallel corpora. Prior to choosing crucial pages from each website that are generally obtained by a search engine on the internet (like Google), PTMiner first chooses potential websites. By analyzing similarities in the URLs (default.phpvsdefaultf.php), the program then generates pairs of web pages. The system then filters the prospective parallel web page. For multilingual reports generated from election results, Braschler and Scäuble created an alternative alignment approach.

Comparable Corpora Comparable Corpora are collections of texts that have been written in a variety of languages but aren't a perfect translation of one another. Comparable corpora of multilingual writings encompass related subject matter, and as a result, their lexicon is similar. CNN, BBC, Reuters, Xinhua News, and BERNAMA are a few examples of news organisations that create multilingual news feeds. These feeds are easily accessible by web users with a variety of site and language combinations. They frequently contain many sentence pairings that are excellent translations of each other. It is feasible to create a bilingual topic-specific dictionary from aligned corpora using a variety of statistical techniques.

According to McNamee and Mayfield, corpus-based translation is substantially more effective than dictionary-based translation. For some languages, it can be extremely challenging to locate such a huge, sizable parallel corpus. Additionally, building an analogous corpus is a very difficult and expensive task. In corpus-based translation techniques, the issues of breadth and quality must also be taken into consideration. According to McNamee and Mayfield, low-quality corpora can greatly impact how well a particular system performs. Coverage, that is all on the number and breadth of language phrases, has similar impacts. The speed of the overall system will suffer if many query words in the underlying corpus lack any translations.

4.3 Document Translation

Usually, machine translation software such as SYSTRAN, AppTek, and PROMPT is used to translate documents. Machine translation technologies commonly translate a number of natural languages fully or automatically.

Any machine translation system's primary tasks include analyzing the source text and language, translating between the source and target languages, and creating the target language using bilingual or multilingual dictionaries. During the procedure, syntactic, morphological, and conceptual information is collected and preserved. In order to produce high-quality results, SYSTRAN created technology that combines statistics and rule-based translation techniques. Due to its ability to provide remarkable high-quality translations of natural language for all domains, SYSTRAN's software has been shown to be a significant advancement in the area of cross-lingual information retrieval. PROMPT offers comparable services in addition to independent data mining, dictionary entries, translations and translation storage systems, and machine translation. PROMPT provides useful software tools, such as a language editor, a post-editing tool, generating and modifying dictionaries, and a user-oriented interface, to handle the problems with translating modules and dictionary volume.

Machine translation Both query and translation of documents use machine translation. A query or document is translated into a target language using machine translation software. Two alternative methodologies are used to implement machine translation. The first technique entails using an offline automated translation system to convert content from corpora of other languages into the user's chosen language. This methodology has been shown to be computationally expensive for big corpora or collections of multilingual documents, however, it is not particularly efficient. For instance, in their cross-lingual retrieval of data research on German and Spanish, Braschler was unable to locate a straight machine translation. Although not all terms in the German papers were translated, they employed machine translation to transfer German to English before converting English to Spanish.

Furthermore, if an internet search is required to find pertinent documents or publications in answer to a specific user request, the offline automated translation method is rendered useless.

The user's query is translated into the desired language in the second method for using machine translation in cross-lingual data retrieval. The query can be modified and then used with conventional information retrieval techniques to find articles in the target language. In both of the aforementioned approaches, machine translation and retrieval operate independently. The main issue with artificial intelligence when it comes to search translational is translational disambiguation, which is typically triggered by homonymy and polysemy since converted queries may not accurately represent the original queries' purpose. A word that has at least two distinct meanings is frequently described as homonomous. Because the setting of an expression such as "Bark" in English isn't at all evident, translating it into a different language may change its underlying meaning. Bark may refer to a dog's woof or the outer layer of a tree. A term is considered to be polysemic if it has several interrelated significances, like the "Head" of a particular clan or "Head" of a person. Therefore, when interpreting documents as opposed to inquiries where the context is utterly ambiguous, machine learning is far more effective and efficient.

5 Conclusion

Semantic-based information retrieval is plagued by a variety of issues, including the absence of semantic sources of knowledge, benchmarks for evaluation, datasets, quick IR techniques, and the inescapable growth of the domain. In a similar vein, the semantic gap between the properties of multimedia resources and the keywords of user queries is still a barrier to multimedia information retrieval. The lack of techniques for high dimensional indexing algorithms, which are essential methods for high dimensional multi-media features, is another significant barrier to multimedia IR.

On the other hand, substantial resources like corpora, ontologies, and lexicons for a number of well-known languages (like Urdu) are also deficient in cross-lingual information retrieval. Additionally, the issue of knowledge representation remains unsolved in cross-lingual information retrieval, which presents a significant challenge for many researchers and practitioners. To succeed in the field of information retrieval, it is necessary to do appropriate research in the fields of machine translation, automated ontology learning from text that is unstructured, and semantics of annotation and extraction.

6 Research Opportunities

Since OBIE is an emerging field with plenty of future potential growth in various directions can be anticipated. Here, we make an effort to pinpoint several key directions. In order to cover the numerous technologies which can be applied to the advancement of the sector, these are detailed at a higher level.

6.1 Increasing the IE process's efficiency:

Instead of being restricted to OBIE, research on this dimension is tied to the larger subject of information extraction. In general, this can be viewed as targeting to increase recall and precision. This procedure would greatly benefit from the development of fresh information extraction methods as well as the integration of current ones. To employ such strategies for OBIE, it is necessary to study the ways in which ontologies might serve as a guide. Additionally, some well-known issues must be resolved in order to increase the efficacy of IE and OBIE. Reference reconciliation, also known as object reconciliation, is one of these issues. It is the process of figuring out whether the two instances are the same thing.

6.2 Combining Ontology Based Information Extraction systems with the Semantic Web:

As was already said, one key aspect that makes OBIE a fascinating research area is its ability to provide semantic content for the Semantic Web. The best way to connect these items with the semantic web hasn't been decided upon, though. For this objective, a number of options are available, such as the construction of web-based services that respond to ontology-based queries. The placement of the OBIE systems and the semantic web's interfaces must also be decided. They can be made available for every website or implemented independently of individual websites, most likely producing semantic content for a specific domain.

6.3 Enhancing the use of ontologies:

Having "good" ontologies is crucial to the success of an OBIE system because they are used to direct the procedure for extracting data and present the findings in OBIE. Ontology construction systems that are automatic or semi-automatic will be crucial to this process. Furthermore, ontologies' quality can be assessed using OBIE. The use of OBIE systems that automatically improve ontologies via the information retrieval procedure is also anticipated to grow in the future. It's also noteworthy that the majority of OBIE systems only employ one ontology. However, using multiple ontologies in a system is not prohibited by any rules.

References

1. Rachid Ahmed-Ouamer and Arezki Hammache. Ontology-based information retrieval for e-learning of computer science. In *2010 International Conference on Machine and Web Intelligence*, pages 250–257. IEEE, 2010.
2. Pablo Castells, Miriam Fernández, David Vallet, Phivos Mylonas, and Yannis Avrithis. Self-tuning personalized information retrieval in an ontology-based framework. In *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops: OTM Confederated Internationl Workshops and Posters, AWeSOMe, CAMS, GADA, MIOS+ INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005, Agia Napa, Cyprus, October 31-November 4, 2005. Proceedings*, pages 977–986. Springer, 2005.
3. Jacopo M Corridoni, Alberto Del Bimbo, and Pietro Pala. Image retrieval by color semantics. *Multimedia systems*, 7:175–183, 1999.
4. Ying Dai. Semantic tolerance-based image representation for large image/video retrieval. In *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pages 1005–1012. IEEE, 2007.
5. Liang Dong, Pradip K Srimani, and James Z Wang. Ontology graph based query expansion for biomedical information retrieval. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 488–493. IEEE, 2011.
6. Kehua Guo, Zhonghe Liang, Yayuan Tang, and Tao Chi. Sor: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. *Journal of computational science*, 28:455–465, 2018.
7. Bill Johnston and Sheila Webber. As we may think: Information literacy as a discipline for the information age. *Research strategies*, 20(3):108–121, 2005.

8. Camille Kurtz, Adrien Depeursinge, Sandy Napel, Christopher F Beaulieu, and Daniel L Rubin. On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. *Medical image analysis*, 18(7):1082–1100, 2014.
9. Yuhua Li, Zuhair A Bandar, and David McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882, 2003.
10. Humaira Liaqaut, Nadeem Iftikhar, and Muhammad Abdul Qadir. Context aware information retrieval using role ontology and query schemas. In *2006 IEEE International Multitopic Conference*, pages 244–249. IEEE, 2006.
11. Goujun Lu. Indexing and retrieval of audio: A survey. *Multimedia Tools and Applications*, 15:269–290, 2001.
12. Ana Meštrović. Collaboration networks analysis: Combining structural and keyword-based approaches. In *Semantic Keyword-Based Search on Structured Data Sources: Third International KEYSTONE Conference, IKC 2017, Gdańsk, Poland, September 11-12, 2017, Revised Selected Papers and COST Action IC1302 Reports 3*, pages 111–122. Springer, 2018.
13. Ankush Mittal and Sumit Gupta. Automatic content-based retrieval and semantic classification of video content. *International Journal on Digital Libraries*, 6:30–38, 2006.
14. Gábor Nagypál. Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops: OTM Confederated Internationl Workshops and Posters, AWeSOMe, CAMS, GADA, MIOS+ INTEROP, ORM, PhDS, SeBGIS, SWWS, and WOSE 2005, Agia Napa, Cyprus, October 31-November 4, 2005. Proceedings*, pages 780–789. Springer, 2005.
15. Rifat Ozcan and YA Aslangdogan. Concept based information access using ontologies and latent semantic analysis. *Dept. of Computer Science and Engineering*, 8:2004, 2004.
16. BV Patel and BB Meshram. Content based video retrieval systems. *arXiv preprint arXiv:1205.1641*, 2012.
17. Dipannita Podder, Jit Mukherjee, Shashaank Mattur Aswatha, Jayanta Mukherjee, and Shamik Sural. Ontology-driven content-based retrieval of heritage images. *Heritage Preservation: A Computational Approach*, pages 143–160, 2018.
18. Miguel Ángel Rodríguez-García, Rafael Valencia-García, Francisco García-Sánchez, and J Javier Samper-Zapater. Ontology-based annotation and retrieval of services in the cloud. *Knowledge-based systems*, 56:15–25, 2014.
19. Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.
20. Von-Wun Soo, Chen-Yu Lee, Chung-Cheng Li, Shu Lei Chen, and Ching-chih Chen. Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, pages 61–72. IEEE, 2003.
21. Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446, 2007.
22. V Vijayarajan, M Dinakaran, Priyam Tejaswin, and Mayank Lohani. A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-centric Computing and Information Sciences*, 6(1):1–30, 2016.

23. Jun Zhai, Yiduo Liang, Yi Yu, and Jiatao Jiang. Semantic information retrieval based on fuzzy ontology for electronic commerce. *J. Softw.*, 3(9):20–27, 2008.