A Survey on Multimodal Information Retrieval Approach

1st Enang Rusnandi
Department of Informatics
Universitas Majalengka
Jawa Barat, Indonesia
enang.rusnandi@mail.ugm.ac.id

2nd Edi Winarko
Dept. of *Computer Sciences and Electronics*
Universitas Gadjah Mada
Yogyakarta, Indonesia
ewinarko@ugm.ac.id

3rd Azhari SN
Dept. of *Computer Sciences and Electronics*
Universitas Gadjah Mada
Yogyakarta, Indonesia
arisn@ugm.ac.id

2020 International Conference on Smart Technology and Applications (ICoSTA) 978-1-7281-3083-5/20/$31.00 ©2020 IEEE 10.1109/ICoSTA48221.2020.1570611095

***Abstract*—There has been more content with various types needed by users. It comes from specific websites or databases. It causes a need to access the content through a retrieval system that can produce content that meets the demand. The existing searching applications still provide services aimed at only one particular modality, such as text, audio, or video. Therefore, research aimed at developing multimodal searching systems is still a significant opportunity in this field. In this paper, we review the multimodal information retrieval systems that have proposed in the previous research. And they discussed the components and mechanisms of the multimodal retrieval used by them. In this way, it will be clear on how they apply their approach to problem-solving on multimodality content searching. At the end of the discussion, there are some ideas or issues related to the problem.**

Keywords—Multimodal, Retrieval, Searching, Application, Approach, System

I. INTRODUCTION

Abundant content from various types of data in it, ranging from text, audio, video, guidance documents about the use of a particular machine, and database content that comes from various information systems, the number is increasing [1], [2], [3]. This condition has impacts on the user's need to be able to retrieve content through a retrieval system that can produce content that meets the requirements. The system can search and present content in the form of text, audio or video files. It can also provide further information about any entity related to the audio or video, for example, information about songwriters, singers, and their history as well as other details [4].

Content retrieval, which consists of one type of content, is called a single capital retrieval and content retrieval that can obtain at least two types of content or more is called a multimodal retrieval. The services provided by search engines that are currently widely used by the public are still related to single modal content searching. It means that if the process of searching for particular video content through the search application page, the results of searching will be displayed in the form of video-only. It does not produce multimodal information, which contains other information related to the video, starting from the time the video is making, its maker, and the meaning of the content contained in the video.

In this connection, a lot of research to try to provide convenience for users by developing approaches and algorithms aimed at this interest. For example, I-Search framework, which is part of European project activities able to retrieve various types of modalities. It is ranging from two-dimensional, three-dimensional, audio, and video through an interface designed to enable users to enter queries which are a combination of these multimodal. However, there is still a need for an assessment for this framework, including problems with its industrial use [5], [6].

On the other hand, there are still many search engines that mostly provide services for only one particular modality. It is a big opportunity for researchers to participate in conducting studies in this field, either to make improvements from existing methods or to develop new techniques and algorithms. The point is that in multimodal retrieval, the collaboration between various parties in various fields such as computer scientists, social scientists, and others. It is because talking about multimodal means that it involves the problem of human-computer interaction [7]. Further, there are two challenges in information retrieval, mainly related to effectivity and convenience, in which users prefer to choose multimodal from their search results. Therefore a query using various methods such as keywords and image examples is necessary [8]. The use of a multimodal retrieval approach to search for information on the type of images, audio, text, and video that is on a web page or in a particular database by using a combination of two or more retrieval models [9].

This article aims to obtain an overview of the mechanisms undertaken by several parties who have already conducted research, which discusses the multimodal retrieval problem by using their respective approaches. Therefore, they obtained clarity on how the system applied to solve the issues related to multimodality content searching, make analysis, comparison and bring up some ideas related to this problem so that they can be useful and contribute to providing a picture of the existence of this multimodal retrieval. The steps taken include exploring the results of previous studies, analyzing and interpreting multimodal retrieval systems. Several studies on multimodal retrieval problems aim mainly to review previous studies such as multimodal video indexing [10] and multimodal interaction [7].

The discussion in this article begins by providing an overview of the single capital model approach, which in content search activities only involves one type of content — followed by a discussion of multimodal retrieval, which provides an overview of the components involved in retrieval activities. Give examples of multimodal retrieval conducted to give clarity on research activities aimed at finding solutions to the problem of the increasing number of multimodal-based applications used in various fields of life.

The rest of this paper is organized as follows: Section II describes the single multimodal retrieval. Section III describes multimodal retrieval. Section IV provides examples of multimodal retrieval, and Part V concludes the paper.

II. SINGLE MULTIMODAL INFORMATION RETRIEVAL

In single-modal retrieval, there are several components involved, such as query, indexing, retrieving, and performance evaluation. As shown in Figure 1. The retrieval process on one side starts with a question from the user. Who has different thoughts and approaches when giving meaning to a query, and on the other hand, there is a large amount of content being crawled by search engines for further indexing using a variety of approaches made. After that, will be a synchronization process between the query and the indexing document in the form of retrieval, which is then evaluated and produces the desired information.

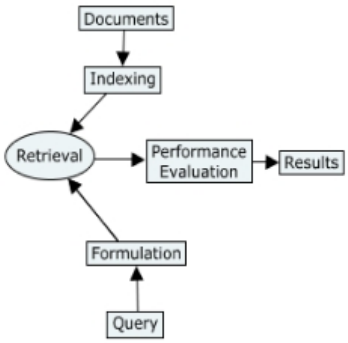

Fig. 1. Single-modal retrieval model

The single-modal stage of information retrieval is, on the one hand, carried out by the user by providing requests for a type of content, be it text, images, audio, or video. The query is then formulating using an appropriate retrieval approach. On the other hand, there is a process of crawling content from various sources, both from web pages and from specific databases. The results of the crawl are indexed to obtain data. That is well-ordering by the program that has made. The search process is presenting in the form of a query index of content. So that content obtained following the wishes of the user, but only includes one type of content.

In the single-modal retrieval query process, from indexing to the adjustment of information needs with the query given is only single. So the consequences from the interface query to the operation of adjusting the results of the retrieval result from only concern with one type of data, such as text, audio, or video, this goes along with what has been done by Yarmohammadi et al. [11], Tahyna et al. [12], dan Bayle et al. [13] and other articles. When querying to a structured data, the SQL language for selecting fields from a database is using. When dealing with unstructured queries, it is performing using queries that are estimating according to the information desired. When indexing is doing, each approach or model uses its algorithmic model. Similarly, it is using as well when adjusting between retrieval results and queries submitted [14].

In single-modal retrieval, the query processing desired by the user, content crawling, indexing, and synchronizing the information obtained does not involve the analysis of two types of objects. Even in image retrieval, data is not adjusting between the types of image objects with the description keywords attached to the image or video. In a single modal, the information presented to the user as a form of feedback from the requested query is also single-modal.

From the explanation above, the list of components in a single capital retrieval can look in table 1.

TABLE 1 COMPONENT OF SINGLE-MODAL RETRIEVAL

No	Component	Description
1	Users and information needed	The user enters the required information through an interface
2	Formulation of the query	The wording of this query is the algorithm
3	Crawling	The process of taking data from a group of content
4	Indexing	The method of sorting content according to specific criteria
5	Retrieval results	The information displayed after going through the synchronization process

III. MULTIMODAL INFORMATION RETRIEVAL

Multimodal information retrieval is an extensive research area. So the discussion is focused on several issues related to the fundamental technical problems of indexing and retrieval models, adaptation and reconfiguration of information retrieval, the combination of text and image on multimodal retrieval, collateral relations, and picture [15].
In multimodal, the problem of complexity of the document model, which includes the use of combination methods, feedback relevance, and image features used in the system, is of particular concern. It is a component of multimodal retrieval documents, as shown in Figure 2. Several studies that also discuss various aspects of the elements in multimodal retrieval, among others, are conducted by Datta et al., [3], Pinho et al. [16], Calumby et al. [17] and others. Meanwhile, multimodal processing and retrieval in terms of music, the modalities in it including motion and gestural data, music scores, audio recordings, video recordings, and others [18].

The multimodal information retrieval anatomy is related to the existence of modalities inherent in documents and queries, as exemplified that in searching of the book documents contain patterns in the form of book titles, reviews, pictures, tables, and other modalities [1]. Meanwhile, in the query, some modalities are stated explicitly or implicitly, where the textual description of the question is explicit. At the same time, other information such as acceptable language and book ranking are implicit. The researchers try to combine this multimodal retrieval problem by using machine learning. , neural networks, and supporting vector machines. To develop a retrieval system approach, which was able to meet user information needs [9], even trying to discuss this multimodal taking using extreme machine learning by exploiting modalities and text images [19].

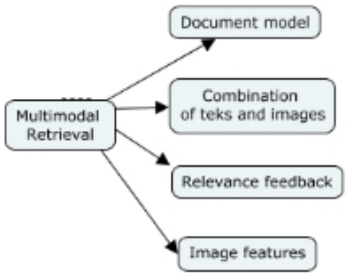

Fig. 2. Components of a multimodal retrieval document

Meanwhile, dealing with the problem of abundant content through internet networks originating from users, it must be intelligent, three-dimensional, interactive, multimodal, etc. Therefore it provides a natural state that enables the communication between one type of modality and others [5]. During the process of searching and retrieving multimodal content, it is easier to provide various types of media into a container and give labels on each box with the same semantic variety.

Multimodal is not only limited to discussing multimedia products but also how to utilize various modalities to be used when doing the query process [20]. Therefore, the query activities can include keyword-based queries, using graphics, using resources, and using standard languages such as XQuery. Meanwhile, a form of a platform that can be expanding to retrieve multimodal health images in open-source software then detailed the architectural, technical approach, and query techniques that are creating as well as the implementation of the search engine itself [16]. The information stored in the form of documents consists of several modalities, as we can find in an article in which there are titles, layouts, pictures, tables, and others, which are called multimodal documents. Therefore, with the development of the internet and digital library, efforts are necessary to develop information retrieval for the benefit of indexing and retrieval intelligence [15].

In a multimodal information system, some activities or components that determine the success of the information retrieval process, which is the query process that must take into account the formulation and analysis of two or more different modalities, for example, text with images. The freebase used as a source of knowledge to develop the method, so information can obtain and summarized to prepare each request [21]. Likewise, with the use of MeSH medical ontology, by using extended queries with medical terms for the benefit of multimodal retrieval [22]. In the case of indexing, which also has to make representations of two or more modalities, the researchers try to provide solutions, for example, developing retrieval and indexing systems based on multimodal geospatial content [23]. Likewise, Pallavi dan N. Pathan [24], who proposes to combine multimodal queries with propagation annotations into an integrated system, also uses techniques that are suitable for indexing.

In terms of data retrieval, we integrated two mining techniques to take pictures that are related to semantics using tree pattern-based algorithms [25]. Furthermore, several studies that developed methods on the side of relevance feedback are conducting by Aksoy and O. Cavus [26]. To be able to implement a multimodal approach, other articles discuss semantic network issues and performance evaluations. With this discussion, the actual need for the presence of a multimodal information retrieval model is still urgent. Further, when the final application produced can be shared to be an open-source that can be easily used by users without having to do complicated configuration and, of course, provide information that meets the user's needs. This condition is an opportunity for researchers to get into more complex fields, such as relating to how multimodal information systems can enter various areas of public life and can realize in the form of adaptive applications. The challenges that exist in multimodal information retrieval relating to semantic models for combining each media model, new interfaces for managing input and presentation of various media data, and a retrieval engine to cross-media boundaries during searching [9].

The list of single and multimodal information retrieval articles surveyed are:

TABLE 2 THE LIST OF ARTICLES SURVEYED

Scope	Reference
Experiments or surveys related to models, methods, frameworks related to single-capital information retrieval	[11], [12], [13], [14], [21], [23]
Experiments related to models, methods, frameworks related to multimodal information retrieval	[1], [2], [3], [5], [6], [16], [17], [19], [20], [22], [24], [25], [26], [27]
Review of frameworks, issues, challenges, models and methods in multimodal information retrieval	[7], [9], [10], [15], [18]

IV. EXAMPLES OF MULTIMODAL RETRIEVAL SYSTEM

To clarify this multimodal retrieval, an example of some research results that discuss multimodal is giving. The example given discusses the background of the problem to the stage carried out for each multimodal retrieval approach. Some examples of these approaches include the Octopus framework model, Multimodal retrieval based on Extreme Learning Machine (ELM), mutual information-based textual query reformulation, and the I-Search framework.

A. Octopus

Octopus is a multimodal information system to answer two challenges related to the convenience and effectiveness of an information retrieval approach. It proposes an aggressive searching mechanism to answer multimodal user queries by investigating various kinds of knowledge [8]. The cooperative interface is designing to facilitate user queries and presentations of retrieval results and collaboration with users and active learning of users' behavior to obtain better retrieval results. It means by the term aggressiveness of the search mechanism on Octopus is because it can get multimodal information, explore diverse knowledge, and learn from user interactions to improve retrieval results. The components of this aggressive searching mechanism include a multifaceted knowledge base, a link analysis based retrieval approach, and a learning-from-interaction strategy, as shown in Figure 3.

In connection with this Octopus problem, the challenge in the information retrieval is how to reinforce the retrieval results to the specific needs of each user, which requires the

synergy of both user modeling techniques and information retrieval techniques. A new user modeling and adaptation techniques used in the Octopus that enable user adaptation (personalization) in connection with the process of taking and presenting results [27].

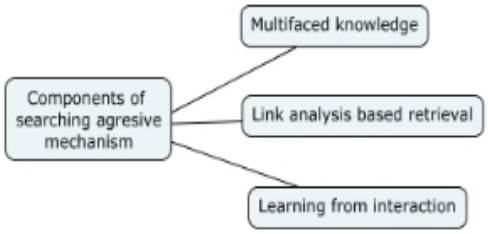

Fig. 3. Components of an aggressive searching mechanism on Octopus

In the existing model in octopus, the suitability of the information is base on each user's view of relevance. It relates to the object with crucial descriptions that exist on that object, individually or in general. While in other information retrieval models, the presentation of information is a match between knowledge and the interests of the user. Evaluation activities are, of course, carried out on this adaptive retrieval model to measure the performance of the proposed model, namely by involving primitive experimental methods involving the user.

B. Multimodal retrieval model based on ELM

In this approach, Extreme Learning Machine is used to make a description of the process of composing text and images, as well as obtaining the desired images and text representations, and then use the probabilistic Latent Semantic Analysis for semantic analysis. Further, some vector features of models are grouped and treated as one word, where one image can have several visual words. The next process is to synchronize the photos with the text in the training set to get the same picture and meaning of words, t. The connection between text and image from the semantic side is making by analyzing it using single-hidden layer feedforward neural networks SLFNs [19]. The general sequence of this approach is looking in Figure 4.

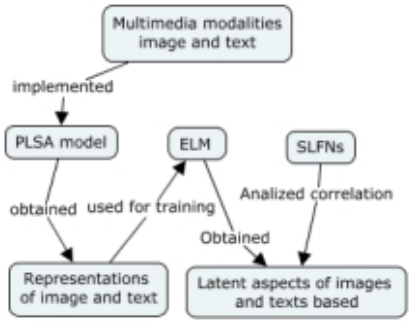

Fig. 4. The general process of multimodal based on ELM

a. Multimedia modality modalities consisting of images and text are analyzed using the PLSA model approach.
b. Obtained the results of the representation of images and text that have been through the analysis process.
c. The ELM approach using to obtain aspects of the images and text for training.
d. On the other hand, SLFNs analyze correlations on aspects of the image and text.

The emergence of a multimodal approach based on Extreme Learning Machine (ELM) is present by the increasing number of multimedia applications that are used in various fields of life, ranging from social, education, sports, and commercial activities. This condition raises new challenges regarding how to obtain the right multimodal product using a fundamental approach.

Experiments are making out through the stages of environmental recognition and measures of performance to measure the model. Furthermore, a retrieval performance evaluation is carried out by implementing image examples and performing image search processes using text. The results of the experiments carried out show that this model is effective and efficient [19].

C. Mutual Information based Textual Query Reformulation

Various researches are carrying out because there are sub-optimal results from the search process using text, and do not directly bring up images related to the version. Even looking for images that are not optimally can bring up a book that matches the desired image. The query text expansion method is carried out through keyphrase extraction so that it can produce better results on text search or image and text retrieval. Query reformulation with query expansion can implement for text exploration. The use of query expansion can improve image capture capabilities with using of optimal combination parameters studied with Fisher Linear Discriminant Analysis (Fisher-LDA) and combined queries [2]. The use of two approaches, namely the highest tf-idf score of the top-k document and the KEA approach, is carried out to expand textual queries. There are several steps undertaken in this framework, namely preprocessing activities for text and images. Preprocessing on the document is done to eliminate unnecessary and redundant data. The next step is to calculate indexes and match scores, learn the parameters of combination and ranking.

The steps taken in this multimodal retrieval method include preprocessing, matching score calculation, indexing, learning combination parameters, and ranking, as can be seen in Figure 5.

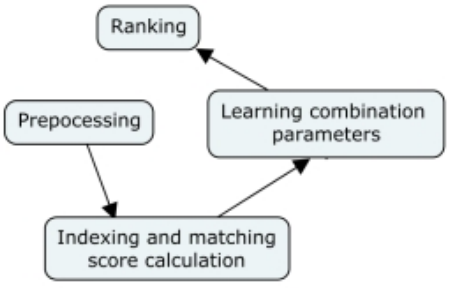

Fig. 5. Steps on using mutual information-based textual query reformulation

a. Preprocessing on text is done to delete unnecessary data by deleting all text, which does not use English, removing stopwords, deleting characters that are not understood, extracting titles, and combining titles with positive narration. But the image is not preprocessed.

b. Indexing of the text is done using IR Terrier and not on image indexing
c. A learning combination made by dividing the entire set of topics with a ratio of 80 percent for training purposes and 20 percent for testing purposes.
d. The ranking by taking the final score of the combination weights at the testing stage is doing.

D. I-Search framework

This framework shows the retrieval process of multimodal information content on the searching and retrieval side. This framework is adaptive to all types of devices owned by users, ranging from mobile to personal computers that have a high capacity [5], [6]. In general, the I-Search framework can be explaining, as shown in Figure 6.

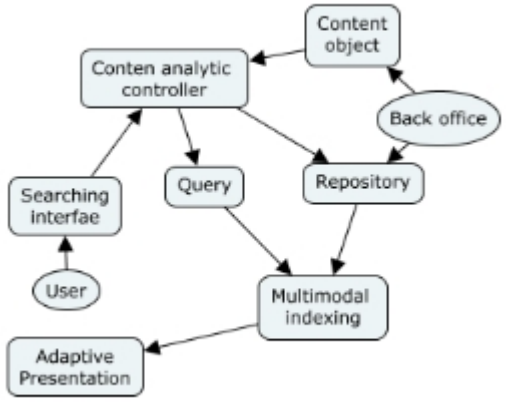

Fig. 6. I-Search framework

a. Through an existing interface, the user searches for the desired information.
b. On the other side of the back office, content objects consisting of various types of information carry out the process of controlling and analyzing the content of existing repositories and queries.
c. Indexing was done after obtaining the appropriate information and displayed as adaptive information to the user.

In the I-Search, there is a well-known term such as content object (CO), which can consist of various types of modalities, ranging from text, images, audio, and video. Then the CO that is behind the activities carried out by the back office enables people to attach metadata to be included in the media container and sent to the content analytic controller, which will do the extraction process for the low-level descriptor in the offline phase. On the other hand, in the online stage, users process the query through the interface that has been prepared and then retrieves the multimodal index. By using a combination of various types of content when querying, I-Search can take multimodal content together. Even I-search provides a proposed solution in the form of an approach to extract and multimodal index descriptors, dynamic interfaces, ways to analyze behavior, and adaptive presentation of search results [5].

V. CONCLUSION

Retrieval information systems that can produce appropriate content are users' expectations to face an abundance of content on the website. The research aimed at just one modality ultimately leads to multimodal analysis; this is in line with the demands of obtaining adequate information by combining various types of data. Through this article, a description of the importance of developing a multimodal retrieval model, the components involved in it, and the methods employed to improve performance indexing and retrieval. When a searching application on a website or a particular information system still provides a single-modal interface, then this is an open opportunity for research and a future research trend.

REFERENCES

[1] M. Imhof and M. Braschler, “A study of untrained models for multimodal information retrieval,” Inf. Retr. J., vol. 21, no. 1, pp. 81–106, 2018.

[2] D. Datta, S. Varma, C. R. Chowdary, and S. K. Singh, “Multimodal Retrieval using Mutual Information based Textual Query Reformulation,” Expert Syst. Appl., vol. 68, pp. 81–92, 2017.

[3] D. Datta, S. K. Singh, and C. R. Chowdary, “Bridging the gap: effect of text query reformulation in multimodal retrieval,” Multimed. Tools Appl., vol. 76, no. 21, pp. 22871–22888, 2017.

[4] L. I. Qing, J. Yang, and Y. Zhuang, “Multi-modal information retrieval with a semantic view mechanism,” Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA, vol. 1, pp. 133–138, 2005.

[5] A. Axenopoulos et al., “I-SEARCH: A unified framework for multimodal search and retrieval,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7281 LNCS, pp. 130–141, 2012.

[6] D. Rafailidis, S. Manolopoulou, and P. Daras, “A unified framework for multimodal retrieval,” Pattern Recognit., vol. 46, no. 12, pp. 3358–3370, 2013.

[7] M. Turk, “Multimodal interaction: A review,” Pattern Recognit. Lett., vol. 36, no. 1, pp. 189–195, 2014.

[8] J. Yang, Q. Li, and Y. Zhuang, “A Multimodal Information Retrieval System: Mechanism and Interface,” IEEE Trans. Multimed.

[9] M. UbaidullahBokhari and F. Hasan, “Multimodal Information Retrieval: Challenges and Future Trends,” Int. J. Comput. Appl., vol. 74, no. 14, pp. 9–12, 2013.

[10] 1C. G. M. Snoek and M. Worring, “Multimodal video indexing: A review of the state-of-the-art,” Multimed. Tools Appl., vol. 25, no. 1, pp. 5–35, 2005.

[11] H. Yarmohammadi, M. Rahmati, and S. Khadivi, "Content-based video retrieval using information theory," Iran. Conf. Mach. Vis. Image Process. MVIP, no. 3, pp. 214–218, 2013.

[12] B. Tahayna, S. Alhashmi, Y. Wang, and K. Abbas, “Combining content and context information fusion for video classification and retrieval,” ICSPS 2010 - Proc. 2010 2nd Int. Conf. Signal Process. Syst., vol. 2, pp. V2-600-V2-604, 2010.

[13] Y. Bayle, M. Robine, and P. Hanna, “SATIN: a persistent musical database for music information retrieval and a supporting deep learning experiment on song instrumental classification,” Multimed. Tools Appl., pp. 1–16, 2018.

[14] J. Foote, “Overview of audio information retrieval,” Multimed. Syst., vol. 7, no. 1, pp. 2–10, 1999.

[15] N. Chen, “A Survey of Indexing and Retrieval of Multimodal Documents : Text and Images,” no. February, p. 40, 2006.

[16] E. Pinho, T. Godinho, F. Valente, and C. Costa, “A Multimodal Search Engine for Medical Imaging Studies,” J. Digit. Imaging, vol. 30, no. 1, pp. 39–48, 2017.

[17] R. T. Calumby, R. Da Silva Torres, and M. A. Gonçalves, “Multimodal retrieval with relevance feedback based on genetic programming,” Multimed. Tools Appl., vol. 69, no. 3, pp. 991–1019, 2014.

[18] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal music information processing and retrieval: Survey and future challenges," Proc. - 2019 Int. Work. Multilayer Music Represent. Process. MMRP, 2019, pp. 10–18, 2019.

[19] Y. Zhang, Y. Yuan, Y. Wang, and G. Wang, “A novel multimodal retrieval model based on ELM,” Neurocomputing, vol. 277, pp. 65–77, 2018.

[20] M. Kharrat, A. Jedidi, and F. Gargouri, "A system proposal for multimodal retrieval of multimedia documents," Proc. - 9th IEEE Int. Symp. Parallel Distrib. Process. With Appl. Work. ISPAW 2011 - ICASE 2011, SGH 2011, GSDP 2011, pp. 177–182, 2011.

[21] C. Lv, R. Qiang, F. Fan, and J. Yang, "Knowledge-based query expansion in real-time microblog search," Lect. Notes Comput. Sci. (including Subset. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9460, pp. 43–55, 2015.

[22] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López, “Query expansion with a medical ontology to improve a multimodal information retrieval system,” Comput. Biol. Med., vol. 39, no. 4, pp. 396–403, 2009.

[23] C. R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial information retrieval and indexing system - Content mining, semantics modelling, and complex queries," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 4, pp. 839–852, 2007.

[24] M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras, “Multimedia search and retrieval using multimodal annotation propagation and indexing techniques,” Signal Process. Image Commun., vol. 28, no. 4, pp. 351–367, 2013.

[25] D. Pallavi and N. Pathan, “A New Approach for Information Retrieval in Multimodal Fusion using Association Rule Mining,” vol. 5, no. 8, pp. 1897–1901, 2016.

[26] S. Aksoy and O. Cavus, “A Relevance Feedback Technique for Multimodal Retrieval of News Videos,” vol. 00, pp. 139–142, 2006.

[27] J. Yang, “User-Adaptive Retrieval of Multimodal Information Using Relevance Network Model,” pp. 1–26.2002