A Multi-source Heterogeneous Data Storage and Retrieval System for Intelligent Manufacturing

Yaning Kong
School of Computer Science and Technology
Harbin Institute of Technology
Weihai, China
20s130405@stu.hit.edu.cn

Dongmei Li
Network Center
Shanxi Medical University
Taiyuan, China
lidongmei@sxmu.edu.cn

Chunshan Li, Dianhui Chu, Zekun Yao
School of Computer Science and Technology
Harbin Institute of Technology
Weihai, China
{lics, chudh20s130449, 20s130449}@hit.edu.cn

***Abstract*—The manufacturing industry produces massive multi-source heterogeneous data such as text, images, audio, and video in the process of design, production, sales, and service. The major problem facing manufacturing companies is how to efficiently manage and use these data resources to create value for manufacturing reproduction. Traditional data storage and retrieval systems classify heterogeneous data according to different forms or modalities and process them separately, resulting in a lack of correlation between cross-modal data (text, image, audio, and video data cannot be checked each other). It cannot support the problem of manufacturing business processes. In this article, we designed and implemented an efficient and fast cross-modal retrieval system for multi-modal industrial data such as text and pictures to realize efficient management and retrieval of multi-source heterogeneous data. Specifically, the system calculates the multi-modal content of manufacturing design, product, service, and other data as a set of unified semantic expressions and stores it in the index structure. When the user makes a query, the index system will return all the modal data related to the retrieved content. This article conducted experiments on the Flick30k data set. The experimental results show that: (1) This system can support millions of data storage and retrieval. (2) With millions of data, the system retrieval rate is in milliseconds. (3) The retrieval accuracy is higher than traditional vector retrieval methods.**

***Index Terms*—Multi-source heterogeneous data, cross-modal retrieval, similarity search framework, hybrid retrival**

I. Introduction

The manufacturing industry produces massive multi-source heterogeneous data such as text, images, audio and video in the design, production, sales and service links. In the manufacturing process, a product may face many query problems from conceptual design to manufacturing and sales services. For example, when designing a product, the designer wants to know "Can you get a yellow, cross-linked symbol in the existing component library? screwdriver"A yellow, Phillips screwdriver".Not only can it return the database records that meet the query conditions, but also the picture information that meets the description. Furthermore, designers often take or hand-paint a picture, hoping to find related or similar product components in the database. Furthermore, product engineers also expect to obtain structured production process details through mixed input of fuzzy text and images, as shown in Figure 1. In order to achieve the above query requirements, manufacturing enterprises need to efficiently integrate and utilize the massive multi-source heterogeneous data in their internals, construct a storage structure that correctly stores multiple forms and modal data, and realize effective cross-modal retrieval technology. In recent years, across text modalities and image modalities, efficient retrieval in a unified semantic space is a very active research field [1]–[3]. Thanks to the rapid progress of deep neural network technologies such as image recognition and text understanding [4], [5]. Zhou et al [6]. tried to use adversarial learning to achieve cross-modal retrieval, but the computational complexity was too high, which affected the retrieval rate; Jin et al. [7] improved the retrieval rate by optimizing the arrangement structure of features, but did not consider the semantics. The sparsity thus produces textual and visual misleading. Cross-modal indexing is extremely efficient for enterprises, especially the manufacturing industry, to manage and integrate their own multi-source heterogeneous data (multi-source mainly refers to the diversification of data sources; heterogeneity mainly refers to the difference in data structure. For example, structured data: fixed attributes Tables, etc.; unstructured data: text, images, audio and other multi-modal data, etc.). In response to the above requirements, this paper designs and implements a storage and retrieval system that supports massive multi-source heterogeneous data based on the virtual search (VisualSearch, Vearch) framework. The system mainly completes the following tasks:

- Projecting massive multi-source and multi-modal data into a unified data space for representation
- Realizing cross-modal graphic mutual search database
- Realizing the efficient retrieval of the three-level structure + HNSW algorithm
- A cross-modal index structure of forward index + inverted index + clustering is realized
- The efficiency and correctness of the data management system designed in this paper are verified on the flick30k data set

The rest of this article is as follows: Section 2 introduces the related work of the research. Section 3 presents the design and implementation of a storage and retrieval system that supports millions of multi-source heterogeneous data. Section 4 introduces the analysis of experimental results. Section 5 summarizes this research and looks forward to a retrieval

978-1-6654-4418-7/21/$31.00 ©2021 IEEE
DOI 10.1109/ICEBE52470.2021.00032

system that supports multiple query forms and builds indexes efficiently and quickly.

II. Related Work

This section will introduce standard text search algorithms, image search algorithms, deep learning-based cross-modal retrieval algorithms, and their applications in the unified representation of multi-source heterogeneous data.

A. Text Retrieval Algorithm

The primary task of text retrieval is to find a subset of documents related to the user query in a given document collection for any user query [8]. Converting the target text into a vector is a common task in the retrieval step. Related methods include the Vector Space Model (VSM). The VSM model can integrate the weight characteristics of the vocabulary into the model; the word2vec/doc2vec distribution representation [9], using neural networks (so-called deep learning) to analyze and process massive texts, and express the text as relatively low-dimensional dense Vector format; LSA/LDA topic model [10], the topic model assumes that there is a K-dimensional topic space, and the document is represented as a numerical representation on K topics.

B. Image Search Algorithm

At present, there are two most commonly used image search algorithms based on image content feature description [11]: This is a semantic level matching. It is necessary to manually describe and classify the image's content (such as objects, background, composition, color features, etc.), give descriptive words and sentences, and query by matching relevant descriptive words and sentences. The query effect of this method is good, and generally speaking, a better precision rate can be obtained. Extraction based on image form features [12]: The image analysis software automatically extracts the color, shape, texture, and other features of the image and establishes a feature index library. Users only need to describe the general features of the image to be searched, and then they can be found. Images with similar characteristics. This is a kind of mechanical matching based on the image feature hierarchy. It is especially suitable for clear retrieval goals, significant differences between query requirements (such as retrieval of everyday things), and the retrieval results can meet users' needs.

C. Image Text Cross-modal Matching Algorithm

Because the type of input information is different from the type of information obtained, we call this task "cross-modality" [13]. There are many cross-modal matching methods for image and text. Global-based cross-modal matching method for image and text: generally extracts the global features of the image and the text and matches the worldwide image and the global text to improve the performance of the model. Specifically, to solve the problem that global image and global text features cannot fully express their global semantic information, Faghri et al. [14] introduced indivisible samples in the triple loss function, which can learn a better mapping matrix and improve it. A good measure of the relevance of images and text. Zheng et al. [15] proposed fine-tunable visual and textual representations, using the fine-tuned global image features and global text features for matching learning to improve the effect of image text cross-modal retrieval. Huang et al. [16] proposed a semantically enhanced image and text matching model, which can improve the representation of images by learning semantic concepts and then organize them in the correct semantic order. Partial-based cross-modal matching of image and text is also the current mainstream method. They usually extract the local features of the image and the text and match the local image and the local text to improve the performance of the model. Specifically, to solve the problem that the local features of images and text cannot be fully optimized, Lee et al. [17] proposed a stacked cross-attention to capture the potential alignment between image regions and text words. In summary, the above retrieval algorithms rarely consider the characteristics of multi-source heterogeneous data but simply assess its generality. When implemented in a specific field, the effect is not good. At the same time, an algorithm that is only aimed at one direction cannot consider other aspects.

III. System Design and Implementation

A. Overview

This section will describe the Multi-source heterogeneous data storage and retrieval system (MHSRS) we designed. The system is mainly divided into two parts: storage and retrieval. Figure 2 shows the system architecture diagram. First, in the storage part, preprocessing is required before storing multi-source heterogeneous data, and operations such as model training are converted into corresponding vectors. Then the data is normalized and stored in the database, and indexed. The retrieval part uses the high-performance retrieval framework Vearch's three-level architecture model, and the hierarchical connectivity naive composition algorithm [18] (Hierarchical Navigable Small World graphs, HNSW) to achieve high-performance similar retrieval.

B. Storage

In order to achieve the retrieval requirements mentioned in the previous article, our MHSRS has designed four storage modes, namely, text search and text storage mode, image search and image storage mode, image and text mutual search storage mode, and event storage mode. The event storage mode is mainly for efficiently managing the production process plan in the manufacturing process. The image search data storage process is similar, and the difference is that we use vgg16 to convert the image data into a 512-dimensional vector and then normalize it and store it in the database. As a classic network model, vgg16 can adapt to most graph data sets and has good classification performance. At the same time, its network structure is relatively straightforward and organized, and it is easy to modify and optimize it. The image-text mutual search data mode combines the two, storing images and text

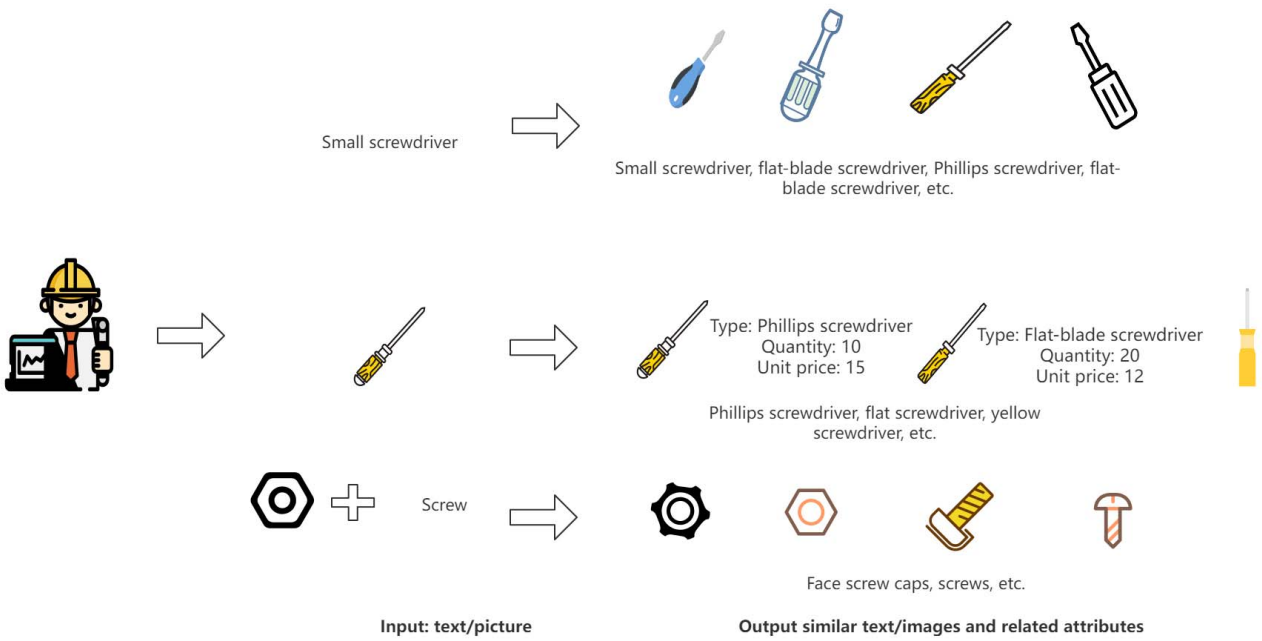

Fig. 1. Manufacturing scene

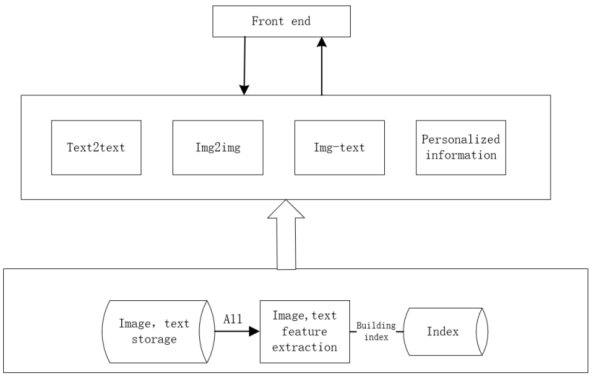

Fig. 2. System structure

in the same data block. The customized database is designed according to the characteristics of multi-source heterogeneous data, which can better meet the needs of enterprises and users. In this system, we crawled the news data of the news section of the official website of Harbin Institute of Technology. According to the characteristics of the news data, we designed a database of news topics. It contains basic text and picture information and adds corresponding news attributes, such as The title of the news, the time and place it happened, the people involved, etc. The storage attributes of the text-to-text search database, the image-to-image search database, the image-to-text search database, and the news custom database are shown in table I. Data storage needs to be normalized. At present, researchers often use min-max normalization, z-score normalization, a tan function conversion, and log function conversion. The minimum-maximum normalization is used here, which can linearly change the original data without changing its distribution, so the result falls within the interval [0,1].

C. Search

The retrieval is based on Vearch's three-level structure, and the retrieval structure is shown in Figure 3. The workflow is rough that when receiving a query from a user, the front end forwards the query request to a specific mixer. The mixer then sends the query to all agents, and each agent requires a subset of the searchers to perform searches in parallel. At the same time, each agent has multiple identical instances to achieve load balancing and fault tolerance. Each searcher is responsible for searching for similar data from a partition of the entire data set. The searcher returns the top k most similar data to the requesting agent. The agent then merges the results from its subset of searchers and sends them back to the mixer. The mixer ranks the results and returns them to the user. Each index data partition has a searcher. The searcher is responsible for searching the corresponding index partition. Each searcher node identifies clusters most similar to the queried image/document based on its characteristics. Then, it scans the inverted list of clusters and calculates the similarity to each image/document in the inverted list. Return the top N

TABLE I
DATABASE DEFINITION AND ITS RELATED ATTRIBUTES

Type	Definition	Attributes
Text-to-text	Text information	Id,vector,text_name,text_url
Image-to-image	Image information	Id,vector,image_name,image_url
Image-to-text	Image-text information	Id,vector,custom_name,text,image_url
Event(news)	Customized information	Id,vector,time,place,people,title,text,url

most similar images/documents. The most similar items are identified by traversing the inverted list and calculating the Euclidean distance to each image/document in the inverted list, and finally, the results are ranked. The searcher is also responsible for processing messages from the message queue. The three-level structure can ensure the system's scalability through massive data and a large number of searcher nodes. At the same time, the retrieval also uses the HNSW algorithm,

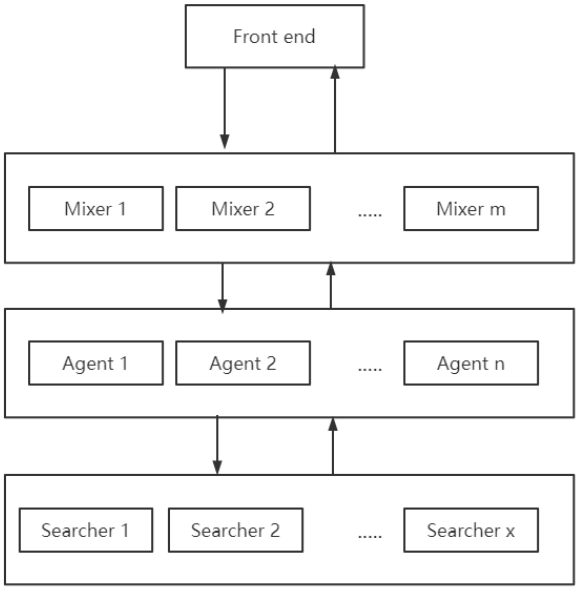

Fig. 3. System retrieval architecture

which is a graph-based algorithm in the ANN search field. HNSW first constructs all the vectors in the high-dimensional space into a connected graph and searches for the K nearest neighbors of a vertex based on this graph. First, introduce the NSW algorithm. Insert points into the graph one by one. Whenever a new point is inserted, the naive search method in the naive idea (by calculating the distance between the adjacent point and the point to be inserted is used to determine whether the next entry point is Which point) Find the nearest n points to this brand new point (the user sets n), and connect the brand new point to n points. Therefore, the HNSW principle is shown in Figure 4. It can be seen that each layer in the figure uses the NSW algorithm and then connects the layers through a jump table to find the target vector.

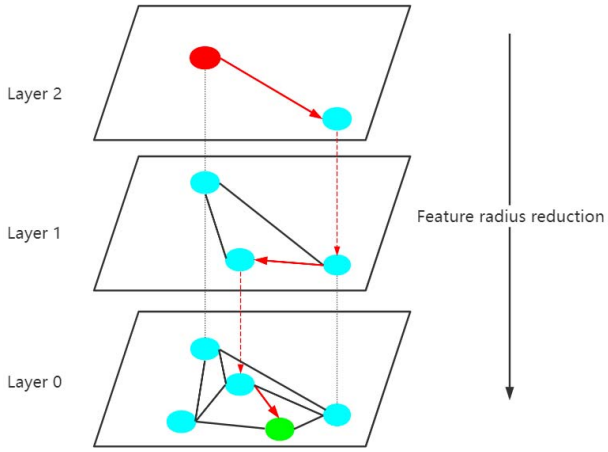

Fig. 4. HNSW algorithm

D. Index

This system index adopts a combination of forwarding index and inverted index. The forward index table is shown in Figure 5. Each image or text is numbered in sequence, and the attributes are stored in a forward index. The index is a self Define array, and each element in the array contains the corresponding attribute information. ID, vector, URL, and other attributes are stored in it. The inverted index consists of N inverted lists. Each inverted list represents a class of images with similar high-dimensional features. The inverted

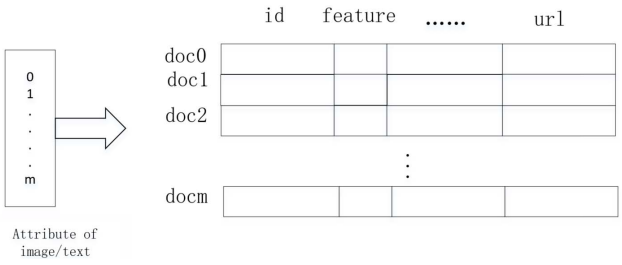

Fig. 5. Forward index

index table is shown in Figure 6. The k-means algorithm for a given training data set (i.e., image or text features) is used to generate the classification. Each Center represents a type of image. In the indexing process, the nearest neighbor algorithm is used to calculate the category to which the image belongs

TABLE II
PERFORMANCE COMPARISON BETWEEN MHSRS AND CNN BASED ON THE MATCONVNET FRAMEWORK

Method	Accuracy	Recall	Rate/s
MHSRS	92.32%	90.75%	0.025
CNN	90.16%	86.78%	0.53

based on the similarity, and the image or text ID is appended to the corresponding inverted list.

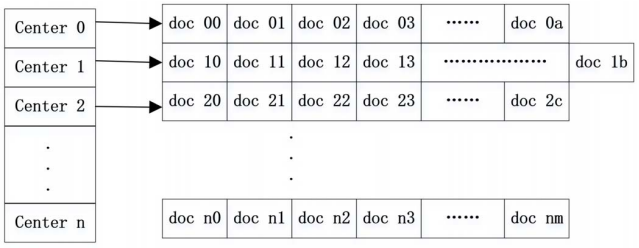

Fig. 6. Inverted index

IV. RESULT

A. Experimental Data

The experiment uses the public graphic data set flick30k, which contains 158915 text descriptions and 31783 images; in addition, to verify the event model, we crawled about 1000 groups of news events in the news section of the official website of Harbin Institute of Technology (each group of information contains pictures and news content), News headlines, people, locations, etc.). We tested the query effects of the multi-source heterogeneous data storage and retrieval system in the image search mode, the image-text search mode, and the event mode. The results are shown in Figure 7. The specific content of the news is shown in Figure 8. As you can see in the figure, the query results are ranked according to the degree of similarity, and users can manually adjust the number of returned results they want; at the same time, the system also has a sharing function, and users can choose the returned results they want to package.

B. MHSRS Test on the Data Set

We conducted the comprehensive performance test of the four databases on a multi-source heterogeneous data storage and retrieval system. At the same time, we compared the CNN based on the MatConvNet framework. We filled the flick30k data set and obtained about one million pictures and One million text descriptions. To test the speed and accuracy of the system, as shown in table **??**. It can be seen that the single retrieval speed can reach the millisecond level, and the precision and recall rate is stable above 90%. Both the speed and accuracy are better than CNN. For the mutual inspection of images and texts, we use a single flick30k data set for comparison experiments. The experimental results are shown in Table 3, and the CNN based on the MatConvNet framework is shown in the tableIII. Among them, Recall@ K represents the proportion of the samples whose correct answer appears in the first K returned results to the total samples; Median Rank represents the median of the position where the first real sample appears in the result sorting is also such that Recall@K¿ = 50% of the minimum K value. MR of 0 means that the first return result is the target result. Similarly, the Average Rank represents the average number of positions of actual samples. This experiment verified the cases of k=1, k=3, and k=5. In the case of a single flick30k data set, as the number of tests increases, the recall rate remains stable; simultaneously, both the accuracy and the speed are ours is better.

In order to verify whether the system can have a good effect on the new data outside the database, we set 8 categories from noun pictures and behavior pictures respectively. And then select a number of each category. Pictures for testing. The experiment is shown in table IV. It can be seen that for the new type of data, the Recall@k of the two are stable at about 90%; due to the large number and types of new data tested, the performance of the two is slightly reduced.

It can be seen that for the new type of data, the Recall@k of the two are stable at about 90%; due to the large number and types of new data tested, the performance of the two is slightly reduced.

V. SUMMARY

There are many image recognition and search software and image-text search software on the market, but most have their limitations. The results returned by Baidu Recognition and Sogou Recognition, which pay attention to versatility, are mixed; letter-taking is subject to database restrictions, and the results returned for behavioral pictures are very poor; Taobao's Polaroid is a typical customized image recognition software. It is only applicable to inquiries on commonly used commodity categories. How to efficiently integrate and manage multi-source heterogeneous data is a significant challenge facing the current manufacturing industry. This paper designs and implements a storage and retrieval system that supports millions of multi-source heterogeneous data. We used the flick30k data set to test the system. We also compared the CNN based on the MatConvNet framework and the query effect of the inspection system on the new data. The results show that our system is slightly better than its accuracy, recall, or rate. In the future, we will continue to optimize the system and design new functions to support more prosperous and diversified data types such as audio.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China (No.2018YFB1700400), in part by the National Natural Science Foundation of China (No. 61902090, 61772159, 61832004), and the Natural Science Foundation of Shandong Province (No. ZR2020KF019).

Fig. 7. MHSRS system demonstration

TABLE III
COMPARISON OF MHSRS AND CNN PERFORMANCE INDICATORS

Number	Method	Recall@1	Recall@3	Recall@5	AVE	MR	AR
100	MHSRS	90.91%	90.06%	90.59%	**90.52%**	0	0.10
100	CNN	80.13%	80.02%	80.23%	80.13	0	0.80
500	MHSRS	89.01%	90.42%	90.22%	**89.88%**	0	0.27
500	CNNN	80.72%	80.31%	80.21%	80.41%	0	0.92
1000	MHSRS	90.34%	90.59%	91.19%	**90.71%**	0	0.31
1000	CNN	80.42%	80.24%	80.41%	80.36%	0	1.21

Fig. 8. News presentation

TABLE IV
COMPARISON OF MHSRS AND CNN ON NEW DATA SET

Method	Recall@1	Recall@2	Recall@3
MHSRS	89.98%	89.18%	87.47
CNN	89.18%	89.12%	88.90

REFERENCES

[1] LIN X, GOKTURK B, SUMENGEN B, et al. Visual search engine for product images[J]. Proc Spie, 2008, 6820:22.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.

[3] GAO L, SONG J, ZOU F, et al. Scalable multimedia retrieval by deep learning hashing with relative similarity learning[C]//ACM. [S.l.: s.n.], 2015: 903-906.

[4] AHMED G F, BARSKAR R. A study on different image retrieval techniques in image processing[J]. International Journal of Soft Computing Engineering, 2011.

[5] FENG X Q, WANG Z W, LIU T C. Port container number recognition system based on improved yolo and crnn algorithm[C]//2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA). [S.l.: s.n.], 2020.

[6] ZHOU N, DU J, XUE Z, et al. Cross-modal search for social networks via adversarial learning[J]. Computational Intelligence and Neuroscience, 2020, 2020:1-12.

[7] L J, K L, Z. L. Cross-modal search for social networks via adversarial learning[J]. Neural Networks and Learning Systems, 2019, 2019:1429-1440.

[8] ZHAO J, JIN Q L, XU B. Semantic Computing for Text Retrieval[J]. Chinese Journal of Computers, 2005, 028(012):2068-2078.

[9] CHEN Q, SOKOLOVA M. Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries[J].2018.

[10] CHIRU C, REBEDEA T, CIOTEC S. Comparison between lsa-ldalexical chains[C]//WEBIST 2014. [S.l.: s.n.], 2014.

[11] RUI Y, HUANG T S. Relevance feedback: a power tool for interactive content-based image retrieval[J]. Circuits Systems for Video Technology IEEE Transactions on, 1998, 8(5):644-655.

[12] LOWE D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[13] Zhang T, Jin C, Tie Y, et al. Research on audio database content matching method for cross-modal retrieval[J]. Signal Processing, 36(6):11.

[14] FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visualsemantic embeddings with hard negatives[J]. arXiv, 2017.

[15] ZHENG Z, ZHENG L, GARRETT M, et al. Dual-path convolutional image-text embedding[J]. 2017.

[16] YAN H, QI W, LIANG W. Learning semantic concepts and order for image and sentence matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(3).

[17] LEE K H, XI C, GANG H, et al. Stacked cross attention for image-text matching[J]. 2018.

[18] MALKOV Y A, YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, PP(99).