Figure Retrieval from Collections of Research Articles

Saar Kuzi(✉) and ChengXiang Zhai

University of Illinois at Urbana-Champaign, Urbana, IL, USA
{skuzi2,czhai}@illinois.edu

Abstract. In this paper, we introduce and study a new task of figure retrieval in which the retrieval units are figures of research articles and the task is to rank figures with response to a query. As a first step toward addressing this task, we focus on textual queries and represent a figure using text extracted from its article. We suggest and study the effectiveness of several retrieval methods for the task. We build a test collection by using research articles from the ACL Anthology corpus and treating figure captions as queries. While having some limitations, using this data set we were able to obtain some interesting preliminary results on the relative effectiveness of different representations of a figure and different retrieval methods, which also shed some light regarding possible types of information need, and potential challenges in figure retrieval.

1 Introduction

Devising intelligent systems to assist researchers and improve their productivity is crucial for accelerating research and scientific discovery. Tools for literature search such as Google Scholar and many digital library systems are essential for researchers; their effectiveness directly affects the productivity of researchers. Conventional literature search systems often treat a literature article as a retrieval unit (i.e., a document) and the retrieval task is to rank articles in response to a query. In this paper, we introduce and study a novel retrieval task where we would treat a figure in a literature article as a retrieval unit and the retrieval task is to return a ranked list of figures from all the literature articles in a collection in response to a query.

An effective figure retrieval system is useful in many ways. First, major scientific research results (e.g., precision-recall curves in information retrieval research) are often summarized in figures and key ideas of technical approaches (e.g., neural networks and graphical models in machine learning research) are often illustrated with figures, making figures important "information objects" in research articles that researchers often want to locate and pay special attention to. While one can also navigate into relevant figures after finding a relevant article, it would be much more efficient if a researcher can directly retrieve relevant figures by using a figure retrieval system. Second, a figure search system may supply useful features for improving the ranking of literature articles in

L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11437, pp. 696–710, 2019.
https://doi.org/10.1007/978-3-030-15712-8_45

a conventional literature search system by rewarding an article whose figures also match well with a query. Third, a figure search system can be very useful for finding examples of illustrations of a concept, thus potentially having broad applications beyond supporting researchers to also generate benefit in education. For example, a figure search engine operating on a collection of research articles in the natural language processing domain can conveniently allow anyone to find some examples of parse trees, which would be useful for learning about a parse tree or just citing an example in a tutorial of natural language processing.

As a retrieval problem, figure retrieval is different from conventional retrieval tasks in many ways, making it an interesting new problem for research. First, the types of information need of users in figure retrieval are expected to be different than in document retrieval, thus potentially requiring the development of novel approaches to satisfy those needs. Another challenge in figure retrieval is how to effectively represent a figure in the collection. One way to represent figures is to treat them as independent units (i.e., image files). However, such a representation does not benefit from the rich context of a figure in the research article that contains the figure. For example, text in the article that explicitly describes the figure as well as other related parts of the article can be used to represent a figure. Finally, it would be important to study models for measuring the relevance between a figure and a query.

In this work, as a first step, we focus on textual queries (i.e., keywords) and represent figures using text extracted from their articles. We propose multiple ways to represent figures and study their effectiveness when using different retrieval methods. Specifically, we propose to represent a figure using multiple textual fields, generated using text in the article that explicitly mentions the figure and also other text in the article that might be related. We then use existing retrieval models, based on lexical similarity and semantic similarity, to measure the relevance between a figure field and a query. Finally, a learning-to-rank approach is used in order to combine different figure fields and retrieval models.

We perform experiments using research articles from the natural language processing domain (ACL Anthology). Since no data sets of queries for figure retrieval are publicly available, we created an initial test collection for evaluation in which figure captions are used to simulate queries (thus, the task is to retrieve a single figure using its caption). While having some limitations, using this data set we were able to obtain some interesting preliminary results. Specifically, our experimental results show that it is beneficial to use a rich textual representation for a figure and to combine different retrieval models. We also gain some initial understanding of the figure retrieval problem, including some illustration of potential types of information need and possible difficulties and challenges. We conclude the paper by suggesting a road map for future research on the task.

2 Related Work

In most retrieval tasks, the retrieval units are documents, though the retrieval of other units, notably entities (e.g., [1,6,20,21]) and passages (e.g., [11,23,26]) has also been studied. Our work adds to this line of research a new retrieval task where the retrieval units are figures in scientific research articles.

As an effective way to communicate research results, figures are especially useful in domains such as the biomedical domain. As a result, how to support biologists to search for figures has attracted a significant amount of attention, and multiple systems were developed [10,13,24]. These previous works have focused on the development of a figure search engine system from the application perspective, but none of those systems or algorithms used in those systems has been evaluated in terms of retrieval accuracy.

Some works [14,31] studied the ranking of figures within a given article based on the assumption that figures in an article have different levels of importance. These works suggested a set of features for ranking so as to measure the centrality of a figure in the article. The suggested features, however, have not been used for figure retrieval. In this paper, we analyze the performance of our approach as a function of the figure centrality in the article, which serves as a first step toward utilizing such features for figure retrieval in the future.

In another line of works, methods for extraction of text from figures in the biomedical domain were studied (e.g., [12,19,29]). Using the text inside a figure can potentially improve retrieval effectiveness by enriching the figure representation. Yet, these works focused mainly on testing the text extraction accuracy, and not the retrieval effectiveness. In our work, we focus on studying the effectiveness of general figure retrieval models, which we believe is required in order to establish a solid foundation for research in figure retrieval; naturally, the general retrieval models can be enhanced by using many additional techniques to enrich figure representation to further improve accuracy as happens in many other applications such as Web search, which we leave as an interesting future work.

Finally, our work is also related to the large body of work on image search. As an effort for improving image search, the ImageCLEF Track was established. In one task, for example, participants were asked to devise approaches for ranking images in the medical domain using visual and textual data [18]. Content-based Image Retrieval (CBIR) was also explored in some works [9,25]. In CBIR, the idea is to extract visual features from the image (e.g., color, texture, and shape) and use them for ranking with respect to an image query. Other works focused on combining visual and textual data for image representation and retrieval (e.g., [2,7,27]). Figures in research articles can also be viewed as images, but we study the problem from the perspective of textual representation of figures. An interesting future work would be to try to incorporate some of the approaches for image search in figure retrieval.

3 Figure Retrieval

In this section, we introduce and define the new problem of figure retrieval, discuss strategies for solving this problem, and present specific retrieval methods that we will later experiment with.

3.1 Problem Formulation

As a retrieval problem, figure retrieval treats each figure in a research article as a retrieval unit. As those figures do not naturally exist as well separated units, the notion of a collection in figure retrieval is defined based on a collection of research articles D, which can be used to build a collection of figures F_D as follows. For every article $d \in D$, k_d figures are extracted; each figure can be uniquely identified in its article by a number $i \in \{1, .., k_d\}$. Then, all figures, extracted from all articles in D, constitute the figure collection F_D.

The goal of the figure retrieval task is to rank figures in F_D according to their relevance to a user query q, where q can be a set of keywords (i.e., textual), an image, or a combination of the two. In general, a user may use keywords to describe what kind of figures he/she wants to find and may also (optionally) use one or multiple example images to define what kind of figures should be retrieved. As a first step in studying this problem, we only consider keyword queries, though we should note that a full treatment of the figure retrieval problem should also include matching any user-provided examples of images with the figure collection, which would be a very interesting direction for future work.

With a keyword query, the figure retrieval problem is quite challenging because it requires matching a keyword query with a figure, which does not necessarily have any readily available text description. Fortunately, we can extract relevant text information from the article with a figure to represent the figure; indeed, all figures have captions, which we can conveniently use to represent them. We can also extract any sentences discussing a figure in an article as an additional text description of the figure. This way, we would obtain a pseudo text document to represent each figure, which we refer to as a *figure document*. Thus, our figure collection contains a set of figures where each figure is associated with a figure document, and the main task for retrieval now is to match a query with those figure documents. This transformation of problem formulation allows us to leverage existing text retrieval models to solve the problem. There are two key technical challenges that we need to study in order to solve the problem effectively: (1) How to derive effective text representations of the figures. (2) How to measure the relevance between a figure and a query. We discuss each next in detail.

3.2 Figure Representation

While figures can be treated just as independent images (i.e., sets of pixels), they appear in the context of research articles, which offers opportunities to build a rich representation for them. For example, text in the article that explicitly

mentions the figure can be utilized. Such text can be the figure caption or other parts of the article that describe or discuss the figure. Other text in the article may not explicitly mention the figure but can still be useful. The abstract of the article, for instance, may serve as a textual representation of the figure since both are in the topic of the article. Finally, other information can be derived from the context of the article which is not necessarily textual. The "authority" of the article (e.g., the number of citations) can serve as a prior for the figure relevance. Our approach to the computation of figure representation is to generate a set of textual fields for each figure, using text that explicitly mentions the figure, as well as other parts of the article.

Explicit Figure Mentions: We generate textual fields using text in the article that explicitly mentions the figure. The caption of the figure, for example, can be regarded as such text. Nevertheless, since figure captions serve as queries in our experiments, we were not able to use them for figure representation at this point. Thus, we only utilize text in the article that discusses or describes the figure (e.g., "The results for the experiment are depicted in Figure 1 ..."). While the general location of such text can be detected easily (since the figure number is explicitly mentioned), it might be challenging to determine its boundaries. That is, automatically detecting at what point in the text the discussion about the figure begins, and at what point the subject changes. A similar problem has been studied in the context of identifying the text that describes a cited article [8]. Yet, it was not studied, to the best of our knowledge, for figure retrieval. In this paper, we take the following approach for extracting this type of text. Given an explicit mention of a figure (i.e., the string "Figure i"), we include w words that precede the figure mention and w words that follow it; w is a free parameter. We denote these textual fields as **FigText** fields and generate three such fields for $w \in \{10, 20, 50\}$. In the case where a figure is mentioned several times in the text, we concatenate all of the text segments that correspond to the different mentions to form a single textual field for a given value of w; overlapping texts are merged so as to avoid textual redundancy.

General Article Text: Other parts of the article that do not explicitly mention the figure can also be useful for figure representation. This might be the case since a figure is usually related to some of the topics of the article, and these topics may also be discussed in some other parts of the article. Using this type of text can be potentially advantageous when the text that explicitly mentions the figure is very short or not highly informative. In such a case, other parts of the article can help to bridge the lexical gap between the query and the figure when measuring the relevance between them. We denote this type of fields **FigArticle** fields. We use the title, abstract, and introduction of the article to generate three separate fields, denoted **Title**, **Abs**, and **Intro**, respectively. By using these sections of the article we can obtain textual fields with different levels of length and generality. We do not use other parts of the article as these may be too general (e.g., using the entire text), or too narrow (e.g., using sections that describe the model). Furthermore, these three sections appear in almost every research article and are easy to detect automatically.

An alternative approach for using the text of an entire article section would be to select only parts of it that are presumably more related to the figure. Motivated by a previous work [30], we select a single sentence from the abstract to represent a figure. This sentence serves as an additional field and is denoted **Abs-sen**. We select a single sentence from the abstract in the following way. We measure the similarity between a sentence in the abstract and the figure using the cosine similarity between their $tf.idf$ representations; a figure is represented using the FigText field ($w = 50$). Then, we choose a single sentence with the highest similarity. If the scores for all abstract sentences with respect to a figure are zeros, we do not represent the figure with a sentence from the abstract. In that sense, using this field we can somehow measure the centrality of the figure in the article (i.e., if the similarity with all abstract sentences is zero then the figure is not likely to be central). The importance in considering the figure centrality was discussed in previous works [14,31].

3.3 Retrieval Models

As each figure is represented by a figure document which consists of multiple text segments, conventional retrieval models are applicable to measure relevance. Our study thus focuses on understanding how effective the basic standard retrieval models are for this new retrieval task, and what kind of representation of figures is the most effective. Specifically, we generate a set of features for each figure where each feature corresponds to a combination of a textual field and a retrieval model and use these features to learn a ranking function using a learning-to-rank (LTR) algorithm [15]. We use LTR so as to effectively combine the different retrieval models and textual fields. Furthermore, LTR offers a flexible framework for adding more features in the future that are not necessarily generated using text data.

In our experiments, we considered two retrieval models in order to measure the relevance between a query and a textual field. The first model we use is **BM25** [22]. This model can also be viewed as a model that measures the lexical similarity between the query and some text as it heavily relies on exact keyword matching. The second model that we use is based on word embeddings (e.g., Word2Vec [17]). Specifically, word embeddings can be used to measure the semantic similarity between the query and a textual field, thus this approach is expected to be complementary to BM25. We learn an embeddings model using the entire collection of research articles. Then, we represent the query and a textual field using the idf weighted average of their term vectors. Finally, the similarity between them is measured using the cosine function. This retrieval approach is denoted **W2V** in our analysis of experimental results.

4 Evaluation

Our main goal is to study the effectiveness of the various approaches we proposed for computing figure representation and ranking figures. Unfortunately, as figure

retrieval is a new task, there does not exist any test collection that we can use for evaluation. Thus, we first need to address the challenge of creating a test collection.

4.1 Test Collection Creation

A test collection for figure retrieval generally consists of three components: (1) a collection of figures; (2) a set of queries; (3) a set of relevance judgments. We now discuss how we construct each of them and create the very first test collection for figure retrieval (available at figuredata.web.illinois.edu).

Figure Collection: To construct a figure collection, we leveraged the ACL Anthology reference corpus [3]. This is one of the very few publicly available full-text article collections. This corpus consists of 22,878 articles whose copyright belongs to ACL. Figures and their captions were extracted from all articles in the corpus using the PdfFigures toolkit [5], resulting in a collection of 42,530 figures; figures that were not mentioned in the text of the article at least one time were excluded from the collection. In order to extract the full text from the PDF files of the articles, we used the Grobid toolkit (github.com/kermitt2/grobid).

Queries Data Set and Relevance Judgments: Ideally, we should create our query set based on real queries from users. Unfortunately, there are no such queries available to us. To address this challenge, we opt to use figure captions as queries with the assumption that if a user would like to search for figures, it is conceivable that the user would use a sentence similar to a caption sentence of a figure. One additional benefit of this is that we can then assume that the figure whose caption has been taken as the query is relevant to the query and thus should be ranked on the top of other figures by an effective figure retrieval algorithm. Of course, we have to exclude the caption sentences from the representation of the figure, or otherwise, the relevant figure would be trivially ranked on the top of other figures by every ranking method. The other figures are assumed to be non-relevant. We note that this assumption is clearly invalid as some of those figures may also be relevant. However, it is still quite reasonable to assume that the figure whose caption has been used as a query should be regarded as more relevant than any other figures, thus measuring to what extent a method can rank this target figure on top of all others is still quite meaningful and can be used to make relative comparisons of different methods. To further improve the quality of the queries, we use only captions that have between 2 and 5 words (not including stopwords), resulting in 16,829 queries; 17%, 33%, 30%, and 20% of the queries in the data set are of length 2, 3, 4, and 5, respectively. The data set of queries was split at random such that one half was used for training the LTR algorithm and the other half was used for evaluation.

4.2 Implementation Details

The Lucene toolkit (lucene.apache.org) was used for experiments. Krovetz stemming and stopword removal were applied to both queries and figure fields. For our word embeddings-based retrieval model, we trained a CBOW Word2Vec model [17] with a window size of 5 and 100 dimensions (radimrehurek.com/gensim/models/word2vec). We used the LambdaMart algorithm [28] in order to learn an LTR model (sourceforge.net/p/lemur/wiki/RankLib). Using the LTR model for ranking the entire collection of figures is not practical as several features are quite expensive to compute for all figures (e.g., word embeddings). We address this issue by adopting a 2-phase retrieval paradigm as follows. We perform an initial retrieval of 100 figures using the FigText field with $w = 50$ (and the BM25 retrieval model). Then, we re-rank the result list using the LTR model with the entire set of features. We use the mean reciprocal rank (MRR@100) and the $success@k$ ($k \in \{1, 3, 5, 10\}$) as our evaluation measures. $success@k$ is the fraction of queries for which the relevant figure is among the top k results.

Table 1. Main result. Figure retrieval performance when different figure fields and different retrieval models are used. The differences in MRR between all LTR models and the initial retrieval are statistically significant (two-tailed paired t-test, $p < 1.0e-7$).

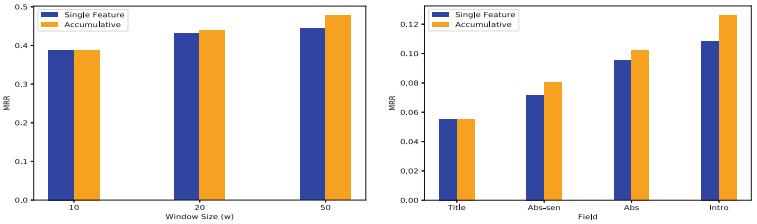
		MRR	$success@1$	$success@3$	$success@5$	$success@10$
Initial retrieval		.443	.353	.497	.547	.607
LTR						
BM25	FigText	.478	.391	.531	.577	.639
	FigArticle	.126	.079	.142	.172	.218
	FigText+FigArticle	.483	.394	.538	.586	.648
W2V	FigText	.212	.129	.233	.291	.377
	FigArticle	.070	.026	.064	.096	.154
	FigText+FigArticle	.212	.127	.230	.289	.380
BM25+W2V	FigText+FigArticle	.487	.398	.541	.592	.649

4.3 Experimental Results

Main Result: The performance of our suggested approach for the figure retrieval task is presented in Table 1. We compare the effectiveness of the initial retrieval with that of the re-ranking approach in which LTR was used. In the case of LTR, we report the performance of using different figure fields and different retrieval models. The LTR performance when the BM25 retrieval model is used is reported in the upper block of the table. According to the results, this approach outperforms the initial retrieval by a very large margin when FigText fields are used. This result attests to the benefit of using different sizes of window for the FigText fields (recall that only a single window size of 50 was used for the initial retrieval). Using the FigArticle fields, on the other hand, results in

an ineffective LTR model compared to the initial retrieval. Yet, according to the results, there is clear merit in combining FigText and FigArticle fields. When W2V is used as a retrieval model, we can see that it is not effective with respect to the initial retrieval. Furthermore, as in the case of BM25, FigText fields are more effective than FigArticle fields when W2V is used. Finally, when all figure fields and all retrieval models are combined, the highest performance is achieved for all evaluation measures. We conclude, based on Table 1, that the most useful figure fields are the FigText fields and the most effective retrieval model is BM25. The W2V retrieval model and the FigArticle fields, on the other hand, are not very effective when used alone and only improve performance when added on top of the other features.

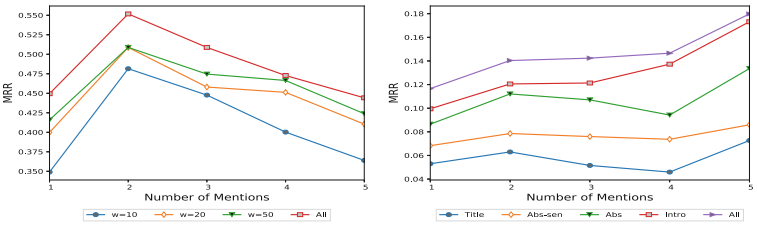
Analysis of Individual Fields: The performance of using individual FigText fields and FigArticle fields for re-ranking the initial result list is reported in Fig. 1(a) and 1(b), respectively. In each graph, the performance (MRR) when a single field is used is reported (blue bar) as well as when a single field is used together with all the fields presented to its left (i.e., accumulative performance; orange bar); BM25 was used as a retrieval model. According to Fig. 1(a), all FigText fields are quite effective and the re-ranking performance increases with the size of the window. Moreover, there is a clear benefit in combining different sizes of the window as the accumulative performance also increases as a function of the window size. Indeed, the length of the text which describes a figure can often vary. In this paper, we address this issue by using different values for the text length. In future work, we plan to explore automatic approaches for setting this value dynamically on a per-figure basis. As for the FigArticle fields, the performance increases as a function of the average field length. That is, the lowest performance is achieved for the title and the highest performance is achieved for the introduction. As in the case of the FigText fields, we can see that there is always an added value when using multiple fields.



(a) FigText (b) FigArticle

Fig. 1. Performance of using individual figure fields. The performance of the FigText and FigArticle fields is depicted in Figure (a) and (b), respectively. (Color figure online)

Figure Centrality Analysis: A figure in a research article can be mentioned in the text several times. We define the number of figure mentions as the number of times the figure number was explicitly mentioned in the article (i.e., the number



(a) FigText (b) FigArticle

Fig. 2. Performance of using different figure fields as a function of the number of mentions of the figure in the article. "All" refers to using all fields. The value of '5' in the x-axis refers to figures with *at least* five mentions. (Color figure online)

of mentions of figure i is the number of appearances of the string "Figure i" in the text). We examine the performance of using different figure fields (using BM25) for re-ranking the initial result list as a function of the number of figure mentions in Fig. 2. Figures with 1, 2, 3, 4, and 5 (or more) mentions constitute 65%, 23%, 7%, 3%, and 2% of the entire figures in the test set, respectively. The performance of using the FigText fields is depicted in Fig. 2(a). According to the graph, the poorest performance is achieved when the figure has only one mention and the highest performance is achieved for two mentions. Furthermore, increasing the number of mentions to more than two almost always results in a performance decrease. A possible explanation for that can be that when the figure is mentioned many times, there are high chances for the window of text to include irrelevant text. The results for the FigArticle fields are presented in Fig. 2(b). According to the graph, the performance almost always increases with the number of mentions for all fields. A possible explanation for that can be that once the figure is mentioned many times in the article, there are high chances that it describes a central topic in the article. Consequently, the text that does not explicitly describes the figure is expected to serve as a more reliable

Table 2. Representative queries and the rank of the relevant figure.

Query	Rank	Query	Rank
(1) Dialog strategy architecture	6	(6) word gloss algorithm	2
(2) Dependency tree english sentence	2	(7) precision recall graph query	32
(3) Performance official runs	1	(8) example graphic tree	1
(4) Full simulation naive bayes f1	9	(9) graphical model sdtm	1
(5) Hierarchical recurrent neural network	1	(10) example dependency tree	0

representation of the figure. Further exploration revealed that adding the number of mentions as an additional feature in the LTR algorithm does not result in further performance gains. An interesting future work would be to explore the effectiveness of more features that capture the centrality of a figure in an article as suggested in previous works [14,31].

Query Analysis: In Table 2, we provide ten representative examples of queries with variable performance and information needs and the corresponding rank of the relevant figure when all features are used for re-ranking the initial result list. (Rank=0 means that the relevant figure did not appear in the top 100 results.) The queries in Table 2 help to illustrate the different information needs that can be addressed by figure retrieval. For example, queries 4 and 7 describe a need for experimental results, while queries 5 and 9 describe a need for some model. Table 2 also helps to illustrate the variance in performance of different queries. For example, query 10 fails to retrieve the relevant figure presumably since this query is very general, resulting in many other figures that match those keywords. Other queries are well specified (e.g., query 4) and thus result in a much better performance. As we already mentioned, one limitation of our experiments is that only one figure is considered relevant for a query. Thus, it is plausible that in a more realistic scenario we would be able to see much better performance for these queries. Nevertheless, these examples help illustrate the potential information needs in figure retrieval and the difficulty of some queries.

We perform an analysis of the query topics in order to gain further understanding about the types of information need in figure retrieval and the effectiveness of their corresponding queries. Specifically, we learn an LDA topic model [4] using all queries in both training and test set. (We use the MeTA toolkit to learn the topic model [16].) Ten words with the highest probabilities in five topics are presented in Table 3. We also present the performance of each topic, which is calculated as follows. We first assign a topic for each query. This topic is the one with the highest probability in the multinomial distribution over topics for this query. Then, we report the average MRR of the queries in each topic. (Each topic ended up containing about 20% of the queries.) The results in Table 3 illustrate potentially five types of information need. For example, Topic 1 contains words that are frequently used in figures that describe examples in the ACL corpus (e.g., "example", "tree", and "parse"). Words that describe a model or an algorithm, on the other hand, can be seen in Topic 2. Finally, Topic 3 contains words that are related to the description of experimental results (e.g., "accuracy" and "performance"). Examining the performance of the different topics, we can see that it can be very different. For example, the worst performance is achieved for Topic 1 (potentially queries for retrieving examples), and the best performance is achieved for Topic 4 which presumably describes an information need for an experimental setup (e.g., "corpus", "annotation", and "text").

Table 3. Query topics (LDA). The average performance of the queries in each topic in terms of MRR is reported in the parenthesis.

Topic 1 (.417)	Topic 2 (.506)	Topic 3 (.501)	Topic 4 (.541)	Topic 5 (.471)
Example	Example	Result	Example	System
Tree	Algorithm	Distribution	Sample	Architecture
Sentence	Model	Accuracy	Annotation	Overview
Parse	Rule	Different	Model	Result
Structure	Learning	Set	Corpus	Process
Dependency	Word	Score	Dialogue	Question
Derive	Alignment	Data	Interface	Framework
Sample	Base	Performance	Entry	Evaluate
Graph	Process	Comparison	Structure	Flow
Rule	Graph	Training	Text	Example

5 Conclusions and Future Work

In this paper, a novel task of figure retrieval from collections of research articles is suggested and studied. According to the new task, figures of research articles are treated as retrieval units and the goal is to rank them with response to a query. We propose and study different approaches for building a representation for a figure using the article text as well as different retrieval methods. Our empirical evaluation demonstrates the benefit of using a rich textual representation for a figure and of combining different retrieval models. Furthermore, an analysis of the queries in the data set sheds some light on the potential information needs in figure retrieval and their relative difficulty.

Figure retrieval is a very promising novel retrieval task; an effective figure search engine would enable researchers to increase productivity, thus accelerating scientific discovery. Our work is only a small initial step; there are many interesting novel research directions that can be further studied in the future which we briefly discuss below.

First, as there does not exist any test collection for figure retrieval, evaluation of figure retrieval is quite challenging. Although we created a test collection, which allowed us to make some interesting relative comparisons of different methods, the test collection we constructed has two limitations: (1) captions do not necessarily represent information needs of real users; (2) captions have only one relevant figure. This data set allowed us to gain some initial understanding of the problem and study the relative effectiveness of different approaches, but those findings have to be further verified with additional experiments. Thus, a very important future work is to build a more realistic data set using a query log and verify our findings. We are currently working on collecting such data by using a figure search engine which we developed (figuresearch.web.illinois.edu).

Second, related to the challenge of constructing a test collection is a better understanding of the information needs in figure retrieval. To that end, it is necessary to conduct a user study in order to obtain some realistic queries. It would also be interesting to study what kind of queries are harder to answer. Another interesting question would be whether there are some common types of information need shared among different research disciplines. A thorough understanding of the users' information needs is also crucial for devising effective retrieval methods that are optimized with respect to user needs.

Third, in this paper, we assumed that the user query is textual. However, in the most general case, the query can involve both textual and visual information. For example, the user would describe an information need using text and also provide figure examples. This raises the question of how to create an effective representation of the user query. To that end, it would make sense to leverage ideas from the area of computer vision, creating an interesting opportunity for interdisciplinary research of information retrieval and computer vision. Furthermore, different representations of the query may also necessitate the development of new ranking models that have to combine multiple ranking criteria.

Figure representation is another subject worth exploring in future work. In this work, we used only textual information for figure representation. In the general case, however, it might be useful to combine different types of information. For example: text data, visual information, article citation information, and figure centrality information. One line of works in this direction would be to identify useful sources of information. Another direction would be to combine heterogeneous information into an effective figure representation.

Finally, devising approaches for the extraction of relevant information for representing a figure is also important. For example, devising methods for automatically identifying the text in the article that discusses a figure, and devising computer vision methods for extraction of useful information from figures to enhance retrieval accuracy are all very interesting directions for future work.

Acknowledgments. We thank the reviewers for their useful comments. This material is based upon work supported by the National Science Foundation under Grant No. 1801652.

References

1. Adafre, S.F., de Rijke, M., Sang, E.T.K.: Entity retrieval. In: Recent Advances in Natural Language Processing (RANLP 2007) (2007)
2. Ah-Pine, J., Csurka, G., Clinchant, S.: Unsupervised visual and textual information fusion in cbmir using graph-based methods. ACM Trans. Inf. Syst. (TOIS) **33**(2), 9 (2015)
3. Bird, S., et al.: The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics (2008)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

5. Clark, C., Divvala, S.: Looking beyond text: Extracting figures, tables, and captions from computer science papers (2015)
6. Demartini, G., Missen, M.M.S., Blanco, R., Zaragoza, H.: Entity summarization of news articles. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 795–796. ACM (2010)
7. Dey, S., Dutta, A., Ghosh, S.K., Valveny, E., Lladós, J., Pal, U.: Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. arXiv preprint arXiv:1804.10819 (2018)
8. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: the next generation of citation analysis. J. Assoc. Inf. Sci. Technol. **65**(9), 1820–1833 (2014)
9. Eakins, J., Graham, M.: Content-based image retrieval (1999)
10. Hearst, M.A., et al.: Biotext search engine: beyond abstract search. Bioinformatics **23**(16), 2196–2197 (2007)
11. Kaszkiel, M., Zobel, J.: Passage retrieval revisited. In: ACM SIGIR Forum, vol. 31, pp. 178–185. ACM (1997)
12. Kim, D., Yu, H.: Figure text extraction in biomedical literature. PloS one **6**(1), e15338 (2011)
13. Liu, F., Jenssen, T.K., Nygaard, V., Sack, J., Hovig, E.: Figsearch: a figure legend indexing and classification system. Bioinformatics **20**(16), 2880–2882 (2004)
14. Liu, F., Yu, H.: Learning to rank figures within a biomedical article. PloS one **9**(3), e61567 (2014)
15. Liu, T.Y.: Learning to rank for information retrieval. Found. Trends® Inf. Retr. **3**(3), 225–331 (2009)
16. Massung, S., Geigle, C., Zhai, C.: Meta: a unified toolkit for text retrieval and analysis. In: Proceedings of ACL-2016 System Demonstrations, pp. 91–96 (2016)
17. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)
18. Müller, H., Deselaers, T., Deserno, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: Peters, C., et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74999-8_72
19. Murphy, R.F., Kou, Z., Hua, J., Joffe, M., Cohen, W.W.: Extracting and structuring subcellular location information from on-line journal articles: the subcellular location image finder. In: Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering, pp. 109–114 (2004)
20. Petkova, D., Croft, W.B.: Proximity-based document representation for named entity retrieval. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 731–740. ACM (2007)
21. Raviv, H., Carmel, D., Kurland, O.: A ranking framework for entity oriented search using markov random fields. In: Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search, p. 1. ACM (2012)
22. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232–241. Springer-Verlag New York, Inc., London (1994)
23. Salton, G., Allan, J., Buckley, C.: Approaches to passage retrieval in full text information systems. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49–58. ACM (1993)

24. Sheikh, A.S., et al.: Structured literature image finder: Open source software for extracting and disseminating information from text and figures in biomedical literature. Technical report, Carnegie Mellon University School of Computer Science, Pittsburgh, USA, CMU-CB-09-101 (2009)
25. Shete, D.S., Chavan, M., Kolhapur, K.: Content based image retrieval. Int. J. Emerg. Technol. Adv. Eng. **2**(9), 85–90 (2012)
26. Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G.: Quantitative evaluation of passage retrieval algorithms for question answering. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 41–47. ACM (2003)
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
28. Wu, Q., Burges, C.J., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. Inf. Retr. **13**(3), 254–270 (2010)
29. Yin, X.C., et al.: Detext: a database for evaluating text extraction from biomedical literature figures. PLoS One **10**(5), e0126200 (2015)
30. Yu, H., Lee, M.: Accessing bioscience images from abstract sentences. Bioinformatics **22**(14), e547–e556 (2006)
31. Yu, H., Liu, F., Ramesh, B.P.: Automatic figure ranking and user interfacing for intelligent figure search. PLoS One **5**(10), e12983 (2010)