Delineating Disciplines Using Visual Information in Scientific Literature

Sean T. Yang[1], Po-shen Lee[1], Jevin D. West[2], and Bill Howe[2]

[1] Department of Electrical Engineering, University of Washington, Seattle WA 98195, USA {tyyang38, sephonlee}@uw.edu
[2] Information School, University of Washington, Seattle WA 98195, USA {billhowe@cs.washington.edu, jevinw@uw.edu}

Abstract. Figures are an important channel in scientific communication. These visual aids help the researchers to express their ideas and results in a structured and visualized way. However, this visual information is barely touched, especially in recommendation system. In this paper, we demonstrate the practicability of using scientific figures to delineate disciplines and compare it to a text-based and a citation-based method. We also show that visual distance conveys different but useful messages when it contrasts with citation distance. We aim to build an scientific recommendation tool based on visual information to facilitate the searching process.

Keywords: Viziometrics · Bibilometrics · Data Science.

1 Introduction

Visual communication is a significant channel for dissemination of scientific results. Over 60 % of authors choose figures that suggest the results of the study as the central figure of the paper[1]. These visual aids help readers absorb complex ideas and models better than text alone, and sometimes outperform text on presenting experimental results and complex scientific concepts. Not only does a single figure carry rich information, but a collection of figures also suggest messages that are worthy of attention. For example, Mounce et al. [2] extracted information from phylogenetic tree diagrams and merged the phylogenies to build a gigantic tree of life, and Kembhavi et al. [3] devised an LSTM-based method for syntactic parsing of diagrams and introduced an attention model for diagram question answering.

Viziometrics, first presented in [4], is an emerging field related to the analysis of visual information in the scientific literature. The term was adopted to distinguish this analysis from bibilometrics and scientometrics for different targets, while still conveying the common objectives of understanding and optimizing patterns of scientific influence and communication. Viziometrics has been contributing to several different topics, such as data extraction [5], machine learning [6], computer vision [7], and informatics [8] [1]. However, few studies leveraged

the figures for designing a recommendation system since, unlike text and citations, figures are not directly machine-readable. As a result, the value and the reach of the visual communication in information retrieval and recommendation system for scientific literature rarely been touced.

Analyzing similarity between academic articles has been a study area in the domain of information retrieval. Most of the studies are based on text and citation graph [9]. Therefore, we consider citation distance as gold standard on characterizing the differences between disciplines and documents. We want to demonstrate the practicability of making use of visual information as a measure of similarity among scholarly publications by answering the following questions in this paper:

- Does visual-based model agree or disagree with citation-based and text-based model?
- Does visual-based model suggest valuable information when it disagrees with citation-based and text-based model?

To answer the questions above, we characterize the distances between scientific disciplines using visual-based, text-based, and citation-based model. We further compare the three distance matrices using Mantel test [10], a well known statistical test of the correlation between two distance matrices, to provide quantitative result. We also perform hierarchical clustering on all distance matrices to offer qualitative comparison among three matrices. In addition, we examine the inconsistency between visual-based and citation-based matrices by case studies.

We demonstrate that visual information should be served as another measure of similarity in scientific literature. We design a simple visual-based model by characterizing fields of study with unsupervised learning to cluster the figures in scientific literature and grouping them by their disciplines. It shows that visual distance is moderate-high correlated to citation-based matric ($r = 0.706$, $p = 0.0001$, z score $= 5.103$) and is also moderate correlated to text-based metric ($r = 0.531$, $p=0.0002$, z score $= 5.019$). Besides, we also show that even when visual distance has disagreement with other methods, it suggests important messages. We conduct a case study on *Computation and Language*, specifically, which is visually far from other *Computer Science* disciplines while it is closer based on citations. We find that the isolation of *Computation and Language* due to heavy use of tables which is indicated that *Computation and Language* is more experimental and more reproducible by performing topic modeling technique on each cluster.

In this paper, we make the following contributions:

- We develop and present a pipeline to aggregate visual information to characterize and further delineate disciplines. We evaluate visual distance to text-based and citation-based measures and show that visual information should be also considered as a measure of similarity in scientific world.
- We examine the disagreements between visual distance and citation distance under *Computer Science* and reveal that *Computation and Language*

has been utilizing tables with increasing frequency in order to display the results of their research. We also demonstrate that the reason more tables are being used in these fields relates to an increase in data-driven, reproducible experimentation.

2 Related Work

Citations have been extensively studied and utilized as a measure of similarity among scientific publications. Marshakova proposed co-citation analysis [9] which uses the frequency that papers are cited together as a measure of similarity. Citations are also utilized to delineate the emerging nanoscience field in [11, 12] and are applied to design recommendation systems [13]. However, citations only reveal the structural information with the scholarly literature and ignore the rich content in the articles.

Text has also received significant attention on analyzing the connection within scientific disciplines and documents, especially in citation recommendations [14, 15]. Vilhena et al. [16] proposed a text-based metric to characterize the jargon distance between disciplines. However, ambiguity and synonymity of text makes text-based model less ideal[17].

Research has been conducted exploring the use of different components from research paper to measure the distance between disciplines. The frequency of mathematical symbols in papers are used to delineate fields in [18], but mathematical symbols are not as ubiquitous as other components. Visual communication is a significant channel in convey scientific knowledge, while is rarely explored to served as a measure of similarity among scientific articles.

A number of studies have focused on mining the scientific figures. Chart classification was well-studied by Futrelle et al. [19], Shao et al. [20], and Lee et al. [5]. Recent studies have been focusing on extraction of quantitative data from scientific visualizations, including line charts [21, 22], bar charts [23], and tables [24]. Researchers have also investigated the techniques to understand the semantic messages of the scientific figures. Kembhavi et al. [3] utilized a convolution neural network (CNN) to study the problem of diagram interpretation and reasoning. Elzer et al. [25] studied the intended messages in bar charts. Besides, several visualization-based search engines have been presented. DiagramFlyer [26], introduced by Chen et al., is a search engine for data-driven diagrams. VizioMetrix[27] and NOA[28] are both scientific figures search engines with big scholar data, while they both work by examining the captions around the figures. Therefore, a visual-based model to measure the distance among scientific disciplines is worth to explore.

3 Method

3.1 Data

We conducted the disciplinary analysis on the figures from arXiv, an open access to e-prints in physics, mathematics, computer science, quantitative biology,

quantitative finance, statistics, electrical engineering, systems science, and economics. We consider arXiv as a reliable and reasonable data source because of the following three reasons: (1) It is an open-access platform which provides low-cost and convenient bulk download service. (2) The disciplines of the articles are selected by the authors, so the disciplinary information is reliable to contact related research. (3) As listed above, there are a wide variety of research papers in the arXiv corpus. It provides a more generalized view for our method. There are 1,343,669 research papers which include 5,009,523 figures on arXiv through December 31st 2017.

3.2 Pipeline to Characterize Fields Using Visual Information

Fig. 1 shows the pipeline we develop to characterize scientific disciplines using visual information. Each step will be explained in the corresponding numbered paragraph.

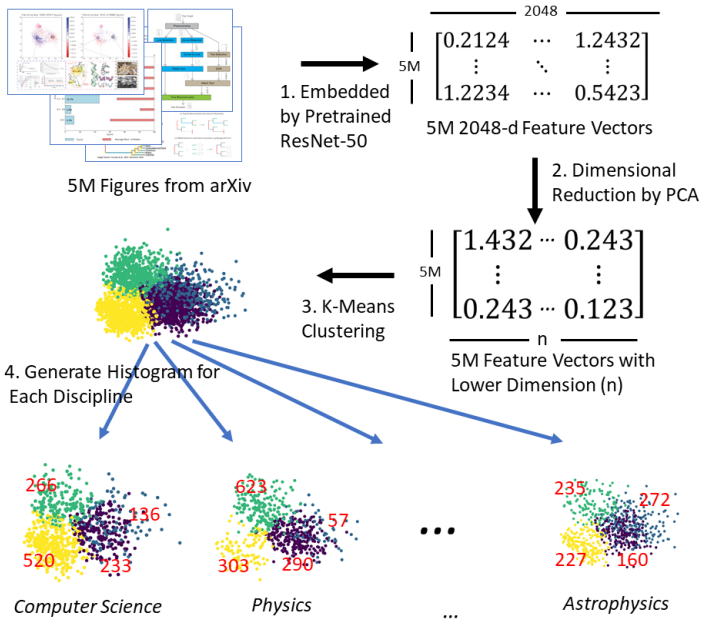

Fig. 1: The diagram illustrates the pipeline of our visual-based model to measure the distance within scientific disciplines.

1. Convert Figures Into Feature Vectors The first step in our pipeline is to embed each figure into a 2048-d feature vector using the pre-trained ResNet-50

model. The figures are re-sized and padded with white pixels to be 224 x 224 before being embedded by pre-trained ResNet-50. ResNet-50 was trained on the ImageNet corpus of 1.2M natural images. Even though the model was trained on natural images, we find that the early layers of the network identify simple patterns (lines, edges, corners, curves) that are sufficiently general for the overall network to represent the combinations of edges and shapes that comprise artificial images as well. Although we posit that a custom neural network architecture could be designed to incrementally improve performance on artificial images, we do not further consider that direction in this paper.

2. Dimensional Reduction for Feature Vectors We reduce the dimension of each figure vector using Principal Component Analysis (PCA). The high-dimensional vectors produced by ResNet-50 contain more information than is necessary for our application of computing the visual similarity between fields, and all else being equal, we seek to make the pipeline as efficient as possible. Our original hypothesis was that a very low number of dimensions (10) would be sufficient to capture the differences between fields, but in our evaluation that higher values produced stronger correlations with other methods. We considered different values of this parameter using a sample of 1.5M figures from the 5M figure corpus. Our results appear in Table 1. "The Explained Variance Ratio" shows the percentage of variance explained by the selected components. The variance explained grows insignificantly after 256 components. "Average of Correlation to Citation Distance" shows the average of the correlations between visual distance of vectors with specified dimension and all the numbers of centroid k (from 2 to 30). We evaluate our method by conducting Mantel test [10] to compare the correlation between visual distance and citation distance, and it will be elaborated in Section 4. It confirms our hypothesis that the correlation increases when more components are used, but it converges after sufficient information is preserved. "Maximum Correlation to Citation Distance" shows the maximum correlation of the specified dimension among different options of number of centroid k, and the k contributing the maximum correlation is shown in "Maximum at k = ?". It surprises us that the maximum correlation happens at larger number of centroid with low dimension of figure vector. Our interpretation is that there is not sufficient information preserved by low dimensional space, and that it just produces random results.

3. Cluster the Figure Corpus The distribution of different types of figures carries significant information about how the visual communication is different in each discipline and could further represent each category. We cluster our figure corpus with K-Means clustering to aggregate similar figures.

We run an experiment on the number of cetroid k. Initially, we expected to acquire higher correlation with larger number of centroid because each cluster would only include similar figures. We experimented with k = 100, 200, and 400. However, larger numbers of k does not perform as expected on our task (correlation coefficient around 0.4) due to sparse feature vector of each discipline. We

lower the number of centroid to the range of 2 to 30 and conduct the experiment. The results appear in Table 1 and we explain the table briefly in the previous paragraph. The maximum correlation happens at k = 4 in most of the dimension experiments. We further investigate the reason and the interpretation in Section 5.2.

Table 1

Dimension	Explained Variance Ratio	Average of Correlations to Citation Distance	Maximum Correlation to Citation Distance	Maximum at k=?
16	52.0%	0.661	0.737	15
32	63.7%	0.631	0.768	3
64	73.9%	0.660	0.769	4
128	82.3%	0.662	0.770	4
256	88.9%	0.672	0.793	4
320	90.7%	0.674	0.793	4

4. Generate Normalized Histogram for Each Discipline We cluster the figures with number of centroid $k = 4$ and generate the normalized histogram for each discipline to acquire visual signature of each discipline.

After visual signature of each discipline is generated, we calculate the euclidean distance between each discipline. We evaluate the visual similarity between disciplines by comparing to citation-based and text-based metrics, which are explained in next section.

4 Comparison with Previous Methods

We use Mantel test [10], a popular statistical test of the correlation between two matrices, to evaluate visual distance by comparing visual distance and the distance matrices created by (1) Average shortest citation distance and (2) Natural language jargon distance. Citations and text have been extensively analyzed and employed to measure the similarity among research articles, and both of the measures have had success on information retrieval and recommendation systems among scholarly documents. Therefore, we consider citation distance as our benchmark of the task and text distance as another comparing method.

4.1 Average Shortest Citation Path

We compute the average shortest path between each pair of fields as a measure of similarity. Average shortest path [29] is one of the three most robust measures [30] of network topology, in addition to its clustering coefficient and its degree distribution. Vilhena et al [16] used this method to measure distance in the citation network to compare with their text-based metric.

Average shortest path is computed as follows:

$$D_{ij} = \frac{1}{n_i n_j} \sum_{n_i} \sum_{n_j} d(v_i, v_j)$$

where n_i is the number of vertices in field i and n_j is the number of vertices in field j. The average shortest path between field i and field j, D_{ij}, is the average of all paths between all vertex pairs, v_i and v_j.

Our citation graph is obtained from the SAO/NASA Astrophysics Data System (ADS)[31], a digital library portal maintaining three bibliographic databases containing more than 13.6 million records covering publications in Astronomy and Astrophysics, Physics, and the arXiv e-prints. The creation of the citations in ADS [32] is started by scanning the full-text of the paper to retrieve bibcode for each reference string in the article, followed by computing the similarity score between the ADS record and the bibcode. The citation pairs are generated if the similarity is higher than the threshold. This data has been extensively used on several bibliographic studies [33, 34]. There are 14,555,820 citation edges within our arXiv data corpus .

4.2 Jargon Distance

We also compare our results to text metrics based on cultural information as represented by patterns of discipline-specific jargon. Jargon distance was first proposed by Vilhena et al. [16], where the authors quantitatively measure the communication barrier between fields using n-grams from full text. The jargon distance (E_{ij}) between field i and field j is defined as the ratio of (1) the entropy H of a random variable X_i with a probability distribution of the jargon or mathematical symbols within field i and (2) the cross entropy Q between the probability distributions in field i and field j:

$$E_{ij} = \frac{H(X_i)}{Q(p_i||p_j)} = \frac{-\sum_{x \in X} p_i(x) \log_2 p_i(x)}{-\sum_{x \in X} p_i(x) \log_2 p_j(x)}$$

Imagine a writer from field i trying to communicate with a reader from field j. The writer has a codebook P_i that maps the natural language or mathematical symbols to codewords that the reader has to decode using the codebook P_j from field j. A small jargon distance means high communication efficiency between two fields and are closely related. This metric could be easily applied to natural language jargon to explore how the communication varies through these two channels across disciplines. We compute the jargon distance between two different disciplines by applying the metrics on unigram from abstracts.

5 Results

To demonstrate the practicality and the value of visual distance, we prove that visual distance is able to (1) determine the overall relationships between fields

and (2) to expose exclusive and valuable information that is not able to be revealed by citation and text alone. In Section 5.1 we show the capacity of visual distance to reveal the relationships across scientific disciplines by displaying high correlation between visual distance and citation distance, a well-studied and reliable measure of similarity between research articles. In Section 5.2, we examine each cluster to understand the visual composition and uncover that each cluster is dominated by a certain type of visualization, which is connected to our previous work [4], where we divided the figure corpus into 5 types (Diagram, Plot, Table, Photo, and Equation) and studied the relationships between the use of the type of images and the impact of the article. Last but not the least, we discover that citation distance and visual distance have the disagreement on *Computation and Language*, which is close to other *Computer Science* areas while is visually farther. The reason is that a abnormally high amount of table is used in *Computation and Language* due to an increase of more reproducible experimentation.

5.1 Delineating Disciplines

In this section, we demonstrate the ability of visual distance to characterize the relationships between fields, quantitatively and qualitatively. Quantitatively, we conduct Mantel test [10] with Spearman rank correlation method to compare two different distance matrices to reveal the similarity between two structures. We also perform hierarchical clustering using UPGMA algorithm [35] to visualize the hierarchical relationships across disciplines, qualitatively. Vilhena et al.[16] used similar technique to qualitatively visualize how disciplines are delineated, but the data they used was from JSTOR, which focuses on biological science and social science so that it is not comparable with our task.

Table 2: The correlation results between distance matrices.

		Results
Visual Distance	Citation Distance	r = 0.706 p = 0.0001 z = 5.103
Visual Distance	Jargon Distance	r = 0.531 p = 0.0002 z = 5.019
Jargon Distance	Citation Distance	r = 0.697 p = 0.0001 z = 5.989

Table 2 shows the correlation results between different distances. The first two columns indicate the comparing methods and the "Results" column shows the correlation results. Visual distance is slightly more correlated to citation distance ($r = 0.706$, p value = 0.0001, z score = 5.103) than jargon distance to

citation distance ($r = 0.697$, p value = 0.0001, z score = 5.989). Visual distance is also moderately correlated to jargon distance with $r = 0.531$, p value = 0.0002, and z score = 5.019. This result is expected. Correlation between visual distance and citation distance is sufficient enough to show that visual distance is capable to characterize general relationships between disciplines, but it also reveals that there are still differences between citation distances and visual distance. We will elaborate the different connections visual distance expose in 5.2.

We then perform hierarchical clustering, using UPGMA algorithm, to qualitatively visualize how different methods group similar disciplines together and separate dissimilar disciplines. The hierarchical clustering results for visual distance, citation distance, and jargon distance are shown in Fig.2, Fig.3, and Fig.4, respectively. We observe similar patterns between visual distance and citation distance where *Computer Science*, *Statistics*, *Math*, and *Mathematical Physics* are isolated from other physics-related fields of study. There is inconsistency between visual distance and citation distance in the field of *Quantitative Biology*, which is the outlier in citation distance, but is assigned to the physics-related cluster in visual distance.

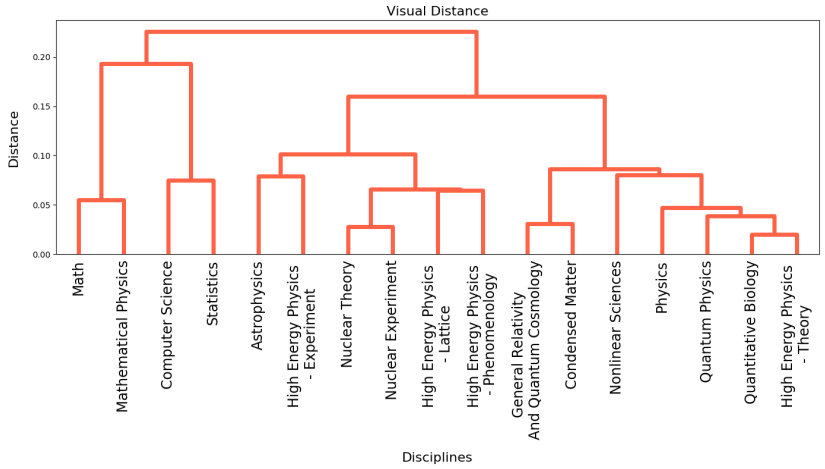

Fig. 2: Qualitative result of how visual information delineates the fields. *Math*, *Mathematical Physics*, *Computer Science* and *Statistics* are separated from other physics-related fields. It confirms our hypothesis that visual information is able to separate distinct fields and group similar fields.

5.2 Analyzing Clusters

We classify the figures in each cluster to understand the visual composition of the each cluster. We use the convolutional neural network classifier in [5] to

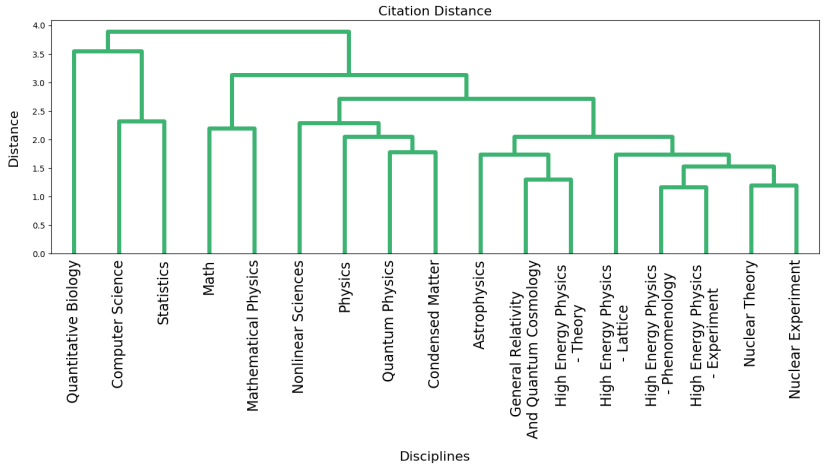

Fig. 3: The hierarchical clustering dendrogram of citation distance. Citation distance is a benchmark in our task. It shows similar pattern as visual distance where *Computer Science*, *Statistics*, *Math*, and *Mathematical Physics* are separated from the rest of the disciplines. The inconsistency between citation distance and visual distance is *Quantitative Biology*, which is clustered with physics-related disciplines in visual distance while it is isolated in citation distance.

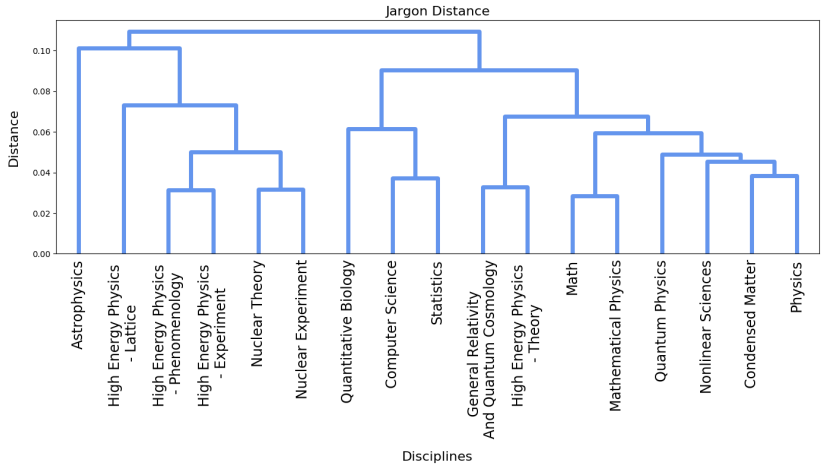

Fig. 4: The hierarchical clustering dendrogram based on jargon distance. Jargon distance segregates disciplines differently from visual distance and citation distance in the high level. *High Energy Physics* and *Nuclear* are separated from the rest where *Quantitative Biology*, *Computer Science* and *Statistics* are isolated in the sub-cluster.

categorize figures into five categories: (1) Diagrams (2) Plots (3) Table (4) Photo and (5) Equation. The classification results are shown in Fig. 5. Surprisingly, each cluster is dominated by a certain type of visualization, where Cluster#0 (We will refer as **Cluster Diagram**) is primary composed of diagrams, Cluster#1 (We will refer as **Cluster Table** consists most of the tables, Cluster#2 (We will refer as **Cluster Plot**) features plots, and Cluster#3 (We will refer as **Cluster Photo**) has large amount of photos. This results notably corroborate to our previous work in [4], where we categorized figures into 5 classes (Diagram, Plot, Table, Photo, and Equation). Our interpretation of the connection is that each type of figures has distinctive purposes. Tables, for example, are excel to display comparison, and plots, on the other hand, are largely used to present quantitative result. The distribution of figures helps to reveal the property of the disciplines. For instance, Cluster Plot is dominant figure cluster in *Quantitative Biology* (48%) and *Nuclear Experiment* (60%), which shows the characteristic of the two disciplines being more experimental and data-driven. The distribution could further be used to group similar disciplines and separate the dissimilar fields as we show in the previous section.

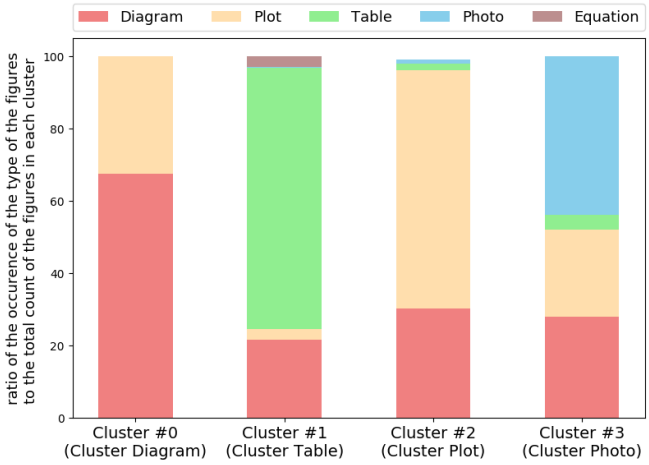

Fig. 5: The visual composition of each cluster. It appears that each cluster has one dominant visualization.

5.3 Different information delivered by figures

In this section, we focus on the cases where visual distance and citation distance have disagreements and answer the following questions: (1) What causes the disagreements? and (2) What is revealed by visual distance in these cases? We

focus on our own field of computer science to be able to interpret nuanced messages.

We normalize visual distance and citation distance and subtract visual distance from citation distance to expose the discrepancies. Fig. 6 uncovers strong disagreement on *Computation and Language* between visual distance and citation distance. Red cells show the disagreements where fields are visually farther and close based on citation, and green cells, on the other hand, point out the disciplines are close visually with high citation distance. We observe that *Computation and Language* are general close to all other categories in *Computer Science*, while it is visually far from others. We further examine the visual profile of *Computation and Language* to investigate the reasons for the divergence between these two distances.

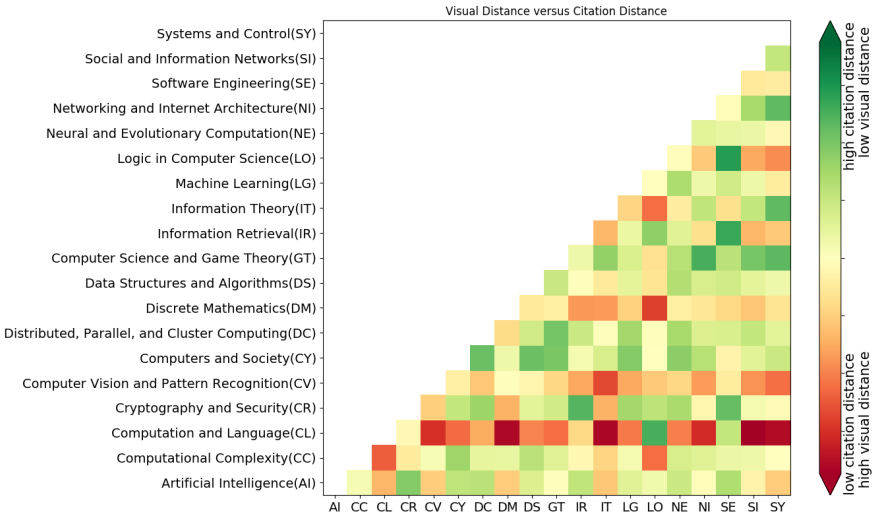

Fig. 6: The heat map to visualize the difference between visual distance and citation distance. It reveals the distinctions on *Computation and Language* and *Computer Vision* between citation distances and visual distance.

Computation and Language Fig. 7 shows the distribution of the figure usage in *Computation and Language* (CL) and *Computer Science* (CS) over the past ten years. We make two observations from this stacked bar chart: (1) Cluster Table dominates the visual communication with over 50% in *Computation and Language* in 2017, compared to approximately 30% in *Computer Science*, and it has been growing over the past few years. (2) The researchers in *Computation and Language* use very little figures in Cluster Photo. It is pretty intuitive that photos are barely used in *Computation and Language*, because the research focus

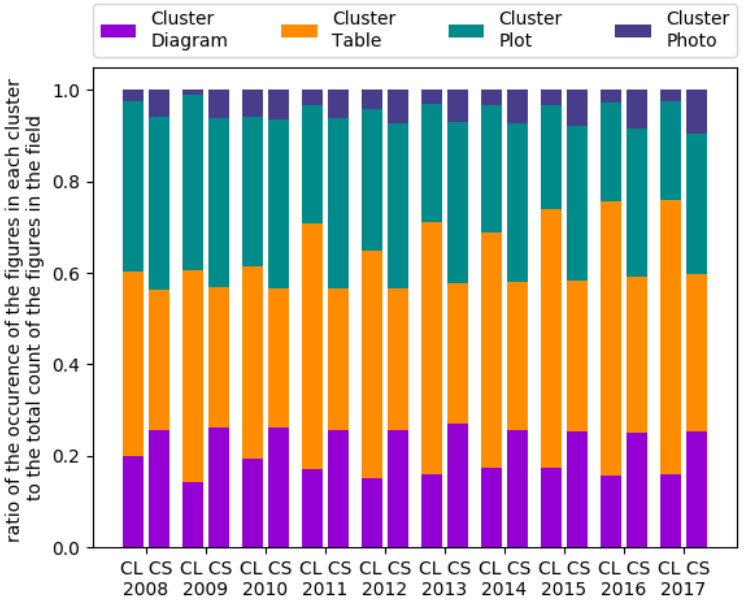

Fig. 7: The stacked bar chart shows how the distribution of the clusters evolves in *Computation and Language* and *Computer Science* over the past ten years. From this plot, we could observe that Cluster Table has been growing in *Computation and Language* and researchers in *Computation and Language* use relative low number of figures in Cluster Photo

of *Computation and Language* is language, which is rarely visualized with photos. We further investigate the reason that tables are largely used in *Computation and Language* by analyzing the cluster textually. We conduct topic modeling on the captions of the figures of Cluster Table using Non-negative Matrix Factorization (NMF) [36] with five topic numbers. In Table 3, we display the top 10 keywords of each topic along with the ratio of the count of the figures in each topic to the total count in the cluster over the past 10 years. We also look at the images in each topic to help us understand the purpose of each topic. Based on the keywords and the images, we can infer that Topic 0 mostly contains table with comparison data to other models, Topic 1 includes the examples of the language and words, Topic 2, which is similar to Topic 0, also involves comparing results with different models, Topic 3 consists of the statistics of the dataset, and Topic 4 is a mix of the tables and diagrams which mostly are used to illustrate the architecture of LSTM models. It appears that tables to compare the accuracy of different models have been growing drastically (from 46.4% (28.6% + 17.8%) in 2008 to 60% (47.6% + 12.4%) in 2017). This could be explained by the fact that reproducing experiments becomes easier due to the advance of computation power, easy assess to the data and code, and the rapid growth of the field itself.

Table 3: Top 10 keywords for each topic in Cluster Table along with the ratio of the figure in each topic over time

	Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
Cluster Table	results	words	et	set	model
	table	figure	al	test	language
	models	word	2015	training	trained
	different	number	2016	data	baseline
	performance	example	2014	development	lstm
	best	table	2017	sets	proposed
	scores	sentence	2013	table	models
	dataset	example	results	dev	attention
	comparison	sentences	2011	used	layer
	accuracy	used	taken	statistics	performance
year	ratio	ratio	ratio	ratio	ratio
2008	28.6%	25.0%	17.8%	22.9%	5.7%
2009	31.1%	26.8%	16.1%	18.6%	7.4%
2010	31.2%	24.2%	16.9%	21.0%	6.7%
2011	34.2%	25.1%	17.2%	16.3%	7.2%
2012	39.1%	22.7%	16.0%	15.5%	6.7%
2013	37.3%	21.7%	17.3%	15.9%	7.8%
2014	39.4%	19.8%	16.0%	14.9%	9.9%
2015	43.9%	18.7%	14.2%	12.2%	11.0%
2016	45.3%	18.5%	13.4%	10.4%	12.4%
2017	47.6%	17.4%	12.4%	10.1%	12.5%

6 Conclusion

In this study, we demonstrate the feasibility of visual information being used as a measure of similarity. We show that visual distance is able to determine the overall relationships between fields by acquiring moderate high correlation (0.706) between visual distance and citation distance. In addition, we show that visual distance still delivers valuable information when it disagree with citation distance. We plan to extend our study to investigate the techniques of using visual information in information retrieval and recommendation system tasks and further integrate text and citations to facilitate the search of scholarly documents. Besides, we will explore clustering techniques to obtain more fine-grained clusters of scientific figures.

7 Acknowledgement

This research has made use of NASA's Astrophysics Data System Bibliographic Services.

References

1. Olga Kazakova, Poshen Lee Lee, Bum Mook Oh, T. Sean Yang, Jevin West, and Bill Howe. Viziometrics: Identifying central figures in scientific papers. 2017.
2. Ross Mounce, Peter Murray-Rust, and Matthew Wills. A machine-compiled microbial supertree from figure-mining thousands of papers. *Research Ideas and Outcomes*, 3:e13589, 2017.
3. Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251. Springer, 2016.
4. Poshen Lee, Jevin West, and Bill Howe. Viziometrics: Analyzing visual patterns in the scientific literature. *IEEE Transactions on Big Data*, 2017.
5. Poshen Lee, T. Sean Yang, Jevin West, and Bill Howe. Phyloparser: A hybrid algorithm for extractingphylogenies from dendrograms. 2017.
6. Maxim Grechkin, Hoifung Poon, and Bill Howe. Ezlearn: Exploiting organic supervision in large-scale data annotation. *arXiv preprint arXiv:1709.08600*, 2017.
7. Satoshi Tsutsui and David Crandall. A data driven approach for compound figure separation using convolutional neural networks. *arXiv preprint arXiv:1703.05105*, 2017.
8. Akshat Dave. *Application of convolutional neural network models for personality prediction from social media images and citation prediction for academic papers*. University of California, San Diego, 2016.
9. IV Marshakova. Co-citation in scientific literature: A new measure of the relationship between publications.". *Scientific and Technical Information Serial of VINITI*, 6:3–8, 1973.
10. Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220, 1967.
11. Michel Zitt and Elise Bassecoulard. Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information processing & management*, 42(6):1513–1531, 2006.
12. Loet Leydesdorff and Ping Zhou. Nanotechnology as a field of science: Its delineation in terms of journals and patents. *Scientometrics*, 70(3):693–713, 2007.
13. J.D. West, I. Wesley-Smith, and C.T. Bergstrom. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, 2(2):113–123, June 2016.
14. Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C Lee Giles. Refseer: A citation recommendation system. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 371–374. IEEE Press, 2014.
15. Trevor Strohman, W Bruce Croft, and David Jensen. Recommending citations for academic papers. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 705–706. ACM, 2007.
16. D. Vilhena, J. Foster, M. Rosvall, J.D. West, J. Evans, and C. Bergstrom. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221–238, 2014.
17. Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. Direction awareness in citation recommendation. 2012.
18. J.D. West and J. Portenoy. Delineating fields using mathematical jargon. In *JCDL Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, 2016.

19. Robert P Futrelle, Mingyan Shao, Chris Cieslik, and Andrea Elaina Grimes. Extraction, layout analysis and classification of diagrams in pdf documents. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, pages 1007–1013. IEEE, 2003.
20. Mingyan Shao and Robert P Futrelle. Recognition and classification of figures in pdf documents. In *International Workshop on Graphics Recognition*, pages 231–242. Springer, 2005.
21. Xiaonan Lu, J Wang, Prasenjit Mitra, and C Lee Giles. Automatic extraction of data from 2-d plots in documents. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 188–192. IEEE, 2007.
22. Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision*, pages 664–680. Springer, 2016.
23. Rabah A Al-Zaidy and C Lee Giles. Automatic extraction of data from bar charts. In *Proceedings of the 8th International Conference on Knowledge Capture*, page 30. ACM, 2015.
24. Jing Fang, Prasenjit Mitra, Zhi Tang, and C Lee Giles. Table header detection and classification. In *AAAI*, pages 599–605, 2012.
25. Stephanie Elzer, Sandra Carberry, and Ingrid Zukerman. The automated understanding of simple bar charts. *Artificial Intelligence*, 175(2):526–555, 2011.
26. Zhe Chen, Michael Cafarella, and Eytan Adar. Diagramflyer: A search engine for data-driven diagrams. In *Proceedings of the 24th International Conference on World Wide Web*, pages 183–186. ACM, 2015.
27. Poshen Lee, Jevin West, and Bill Howe. Viziometrix: A platform for analyzing the visual information in big scholarly data. In *BigScholar Workshop (co-located at WWW)*, 2016.
28. Jean Charbonnier, Lucia Sohmen, John Rothman, Birte Rohden, and Christian Wartena. Noa: A search engine for reusable scientific images beyond the life sciences. In *European Conference on Information Retrieval*, pages 797–800. Springer, 2018.
29. Stuart E Dreyfus. An appraisal of some shortest-path algorithms. *Operations research*, 17(3):395–412, 1969.
30. Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.
31. Guenther Eichhorn. An overview of the astrophysics data system. *Experimental Astronomy*, 5(3-4):205–220, 1994.
32. Alberto Accomazzi, Gunther Eichhorn, Michael J Kurtz, Carolyn S Grant, Edwin Henneken, Markus Demleitner, Donna Thompson, Elizabeth Bohlen, and Stephen S Murray. Creation and use of citations in the ads. *arXiv preprint cs/0610011*, 2006.
33. Eugene Garfield. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93, 2006.
34. Michael J Kurtz and Edwin A Henneken. Measuring metrics-a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the Association for Information Science and Technology*, 68(3):695–708, 2017.
35. F James Rohlf and David R Fisher. Tests for hierarchical structure in random data sets. *Systematic Biology*, 17(4):407–412, 1968.
36. Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.