VizioMetrix: A Platform for Analyzing the Visual Information in Big Scholarly Data

Po-shen Lee
University of Washington
185 Stevens Way
Seattle, Washington 98105
sephon@uw.edu

Jevin D. West
University of Washington
Box 352840
Seattle, Washington 98195
jevinw@uw.edu

Bill Howe
University of Washington
185 Stevens Way
Seattle, Washington 98105
billhowe@cs.washington.edu

ABSTRACT

We present VizioMetrix, a platform that extracts visual information from the scientific literature and makes it available for use in new information retrieval applications and for studies that look at patterns of visual information across millions of papers. New ideas are conveyed visually in the scientific literature through figures — diagrams, photos, visualizations, tables — but these visual elements remain ensconced in the surrounding paper and difficult to use directly to facilitate information discovery tasks or longitudinal analytics. Very few applications in information retrieval, academic search, or bibliometrics make direct use of the figures, and none attempt to recognize and exploit the *type* of figure, which can be used to augment interactions with a large corpus of scholarly literature.

The VizioMetrix platform processes a corpus of documents, classifies the figures, organizes the results into a cloud-hosted databases, and drives three distinct applications to support bibliometric analysis and information retrieval. The first application supports information retrieval tasks by allowing rapid browsing of classified figures. The second application supports longitudinal analysis of visual patterns in the literature and facilitates data mining of these figures. The third application supports crowdsourced tagging of figures to improve classification, augment search, and facilitate new kinds of analyses. Our initial corpus is the entirety of PubMed Central (PMC), and will be released to the public alongside this paper; we welcome other researchers to make use of these resources.

Keywords

Figure Retrieval, Information Retrieval, Crowdsourcing, Open-data, Bibliometrics, Scientometrics, Viziometrics

1. INTRODUCTION

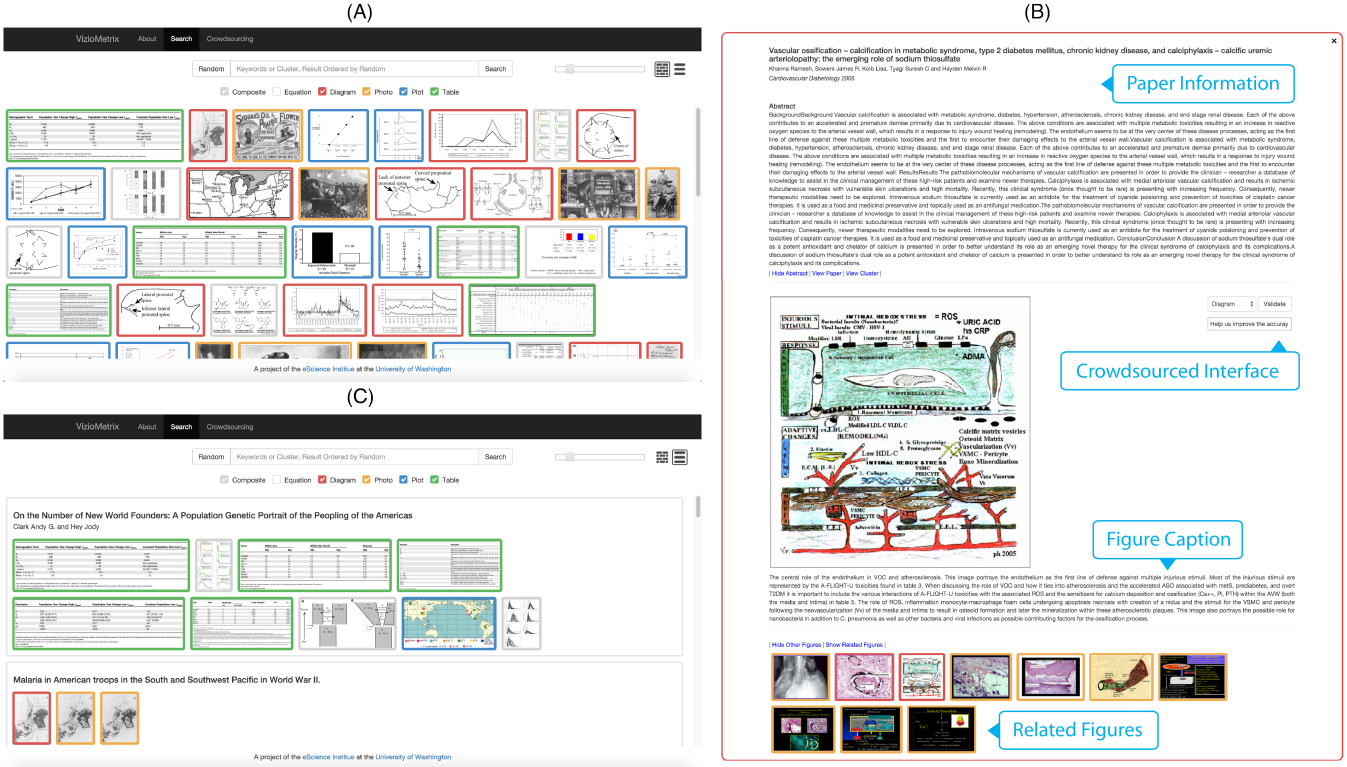
Scientific results are communicated visually in the scientific literature, but visual information is underused in academic search tools. We explore how visual information can be used to facilitate search tasks and design a new interface to support this exploration. Search tasks typically involve accessing a superset of relevant papers and manually scanning their contents for relevance. The text-based layout is not necessarily conducive for this ”scan for relevance” task, and as a result the anecdotal algorithm is to ”skim the figures.” In all fields, key experimental results are presented in plots, complex scientific concepts are visualized by graphics with text, and photographs provide evidence and insight. Reviewing these figures can help users quickly understand the style of article or the methods used (e.g. statistical analysis, experimental demonstration, survey). In many fields like Cell Biology, a figure can be worth a thousand words that summarizes the entirety of the paper [6]. Using a particular style of schematics, plots, or photos can be indicative of a particular type of paper. For example, a phylogenetic tree indicates a phylogenetic analysis has been performed. The visual information can carry details that are insufficiently described in the text. Our hypothesis is that reviewing these figures as first class artifacts during search can help users rapidly identify relevant articles, draw associations between related articles, and focus attention on key results rather than overarching topics.

Others have observed this need for figure-oriented retrieval and built tools to support the corresponding tasks[1]. Structured Literature Image Finder system [1] proposed by Carnegie Melon University researchers focus their interest in microscope images. DiagramFlyer [3], proposed by University of Michigan researchers, facilitates search over figures in the literature, but focuses on keyword search over extracted text rather than classification of the figure type itself. A research team from Pennsylvania State University engages in mining data-driven visualizations [10, 2]. Our approach is to classify figures based on their *type*, then organize a search and analysis interface that uses these classified images as the primary unit of interaction.

VizioMetrix is designed to be an initial suite of tools in support of *Viziometrics*, an emerging field pertaining to the analysis of visual information in the scientific literature [8]. The term is intended to convey shared goals with bibilometrics and scientometrics while indicating our focus on the visual information. VizioMetrix (the platform) includes functionality for three groups of users: (1) academic users performing search tasks, (2) researchers who specialize in computer vision for document understanding, and (3) scientometricians interested in understanding general communication patterns across the literature. For group (1), VizioMetrix

[1] Zanran, D8taplex

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2890523 .



A project of the eScience Institue at the University of Washington

Vascular ossification – calcification in metabolic syndrome, type 2 diabetes mellitus, chronic kidney disease, and calciphylaxis – calcific uremic arteriolopathy: the emerging role of sodium thiosulfate

Khanna Ramesh, Sowers James R, Kolb Lisa, Tyagi Suresh C and Hayden Melvin R

Cardiovascular Diabetology 2005

Abstract

BackgroundBackground Vascular calcification is associated with metabolic syndrome, diabetes, hypertension, atherosclerosis, chronic kidney disease, and end stage renal disease. Each of the above contributes to an accelerated and premature demise primarily due to cardiovascular disease. The above conditions are associated with multiple metabolic toxicities resulting in an increase in reactive oxygen species to the arterial vessel wall, which results in a response to injury wound healing (remodeling). The endothelium seems to be at the very center of these disease processes, acting as the first line of defense against these multiple metabolic toxicities and the first to encounter their damaging effects to the arterial vessel wall.Vascular calcification is associated with metabolic syndrome, diabetes, hypertension, atherosclerosis, chronic kidney disease, and end stage renal disease. Each of the above contributes to an accelerated and premature demise primarily due to cardiovascular disease. The above conditions are associated with multiple metabolic toxicities resulting in an increase in reactive oxygen species to the arterial vessel wall, which results in a response to injury wound healing (remodeling). The endothelium seems to be at the very center of these disease processes, acting as the first line of defense against these multiple metabolic toxicities and the first to encounter their damaging effects to the arterial vessel wall. ResultsResults The pathobiomolecular mechanisms of vascular calcification are presented in order to provide the clinician – researcher a database of knowledge to assist in the clinical management of these high-risk patients and examine newer therapies. Calciphylaxis is associated with medial arteriolar vascular calcification and results in ischemic subcutaneous necrosis with vulnerable skin ulcerations and high mortality. Recently, this clinical syndrome (once thought to be rare) is presenting with increasing frequency. Consequently, newer therapeutic modalities need to be explored. Intravenous sodium thiosulfate is currently used as an antidote for the treatment of cyanide poisioning and prevention of toxicities of cisplatin cancer therapies. It is used as a food and medicinal preservative and topically used as an antifungal medication.The pathobiomolecular mechanisms of vascular calcification are presented in order to provide the clinician – researcher a database of knowledge to assist in the clinical management of these high-risk patients and examine newer therapies. Calciphylaxis is associated with medial arteriolar vascular calcification and results in ischemic subcutaneous necrosis with vulnerable skin ulcerations and high mortality. Recently, this clinical syndrome (once thought to be rare) is presenting with increasing frequency. Consequently, newer therapeutic modalities need to be explored. Intravenous sodium thiosulfate is currently used as an antidote for the treatment of cyanide poisioning and prevention of toxicities of cisplatin cancer therapies. It is used as a food and medicinal preservative and topically used as an antifungal medication. ConclusionConclusion A discussion of sodium thiosulfate's dual role as a potent antioxidant and chelator of calcium is presented in order to better understand its role as an emerging novel therapy for the clinical syndrome of calciphylaxis and its complications.A discussion of sodium thiosulfate's dual role as a potent antioxidant and chelator of calcium is presented in order to better understand its role as an emerging novel therapy for the clinical syndrome of calciphylaxis and its complications.

| Hide Abstract | View Paper | View Cluster |

The central role of the endothelium in VOC and atherosclerosis. This image portrays the endothelium as the first line of defense against multiple injurious stimuli. Most of the injurious stimuli are represented by the A-FLIGHT-U toxicities found in table 3. When discussing the role of VOC and how it ties into atherosclerosis and the accelerated ASO associated with metS, prediabetes, and overt T2DM it is important to include the various interactions of A-FLIGHT-U toxicities with the associated ROS and the sensitizers for calcium deposition and ossification (Ca++, Pi, PTH) within the AVW (both the media and intima) in table 5. The role of ROS, inflammation monocyte-macrophage foam cells undergoing apoptosis necrosis with creation of a nidus and the stimuli for the VSMC and pericyte following the neovascularization (Vv) of the media and intima to result in osteoid formation and later the mineralization within these atherosclerotic plaques. This image also portrays the possible role for nanobacteria in addition to C. pneumonia as well as other bacteria and viral infections as possible contributing factors for the ossification process.

| Hide Other Figures | Show Related Figures |

On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas

Clark Andy G. and Hey Jody

Malaria in American troops in the South and Southwest Pacific in World War II.

A project of the eScience Institue at the University of Washington

Figure 1: Screenshot of search engine interface (viziometrics.org). We use different colors to highlight different figure types (e.g., red indicates diagrams) (A) shows the grid layout, which is designed for reviewing many images (B) shows the alternative layout, which bundles figures from the same paper and related papers. Related papers are selected based on out and in-citations and then ranked by the ALEF score. This is made to look more like a paper, whereas the grid layout provides a general overview on a particular topic. (c) is the page showing figure and paper details. A simple crowdsourced labelling interface is embedded in the page to gather human labels.

provides a figure-centric search service that allows browsing and filtering by figure types. The figure labels are determined by our classification process, but the labels can be edited directly through the interface to improve accuracy over time. For group (2), VizioMetrix provides an efficient bulk-labeling interface to produce better training data for figure type classification and related viziometric analysis tasks. For group (3), VizioMetrix provides a longitudinal analysis interface that allows aggregate analysis of the patterns of visual information in the literature.

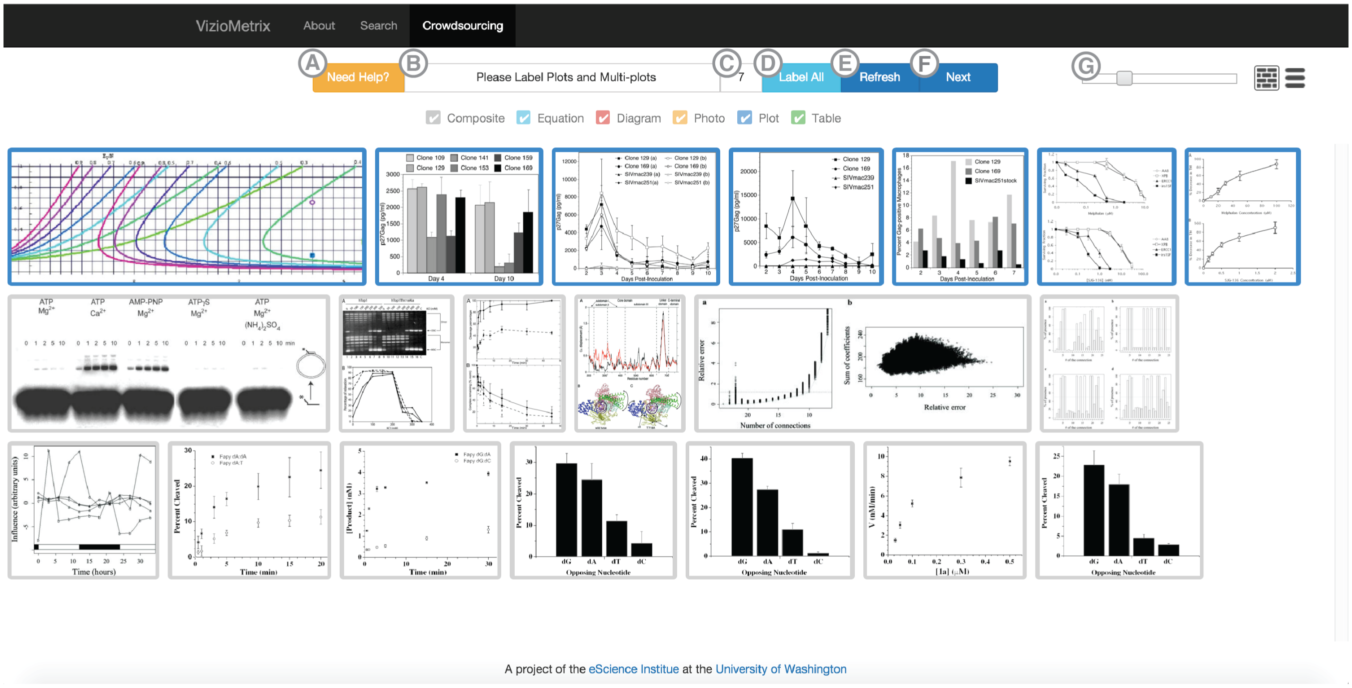
These three interfaces are powered by a figure-processing pipeline that extracts features and classifies the figures using patch-oriented techniques in computer vision and machine learning [12]. Our current corpus consists of 8.9 million images extracted from 680K papers freely accessed from PubMed Central (PMC) (see [8] for more details). In the remainder of the paper, we will describe how the user interacts with the applications (Section 2), briefly walk through the system architecture (Section 3), report the future work (Section 4), and give a brief summary (Section 5).

2. APPLICATION INTERFACE

VizioMetrix includes a suite of three web applications targeting three different categories of users: a visual search application for researchers, an analytics interface for viziometricians, and a crowdsourced interface for efficient bulk figure-labelling by those contributing to the project.

Visual Search Engine Figure 1 shows the user interface for the visual search application. At a basic level, users search for figures associated with a given keyword query. Each figure can be clicked to see metadata about the paper that contains it. This "inversion" of the search to emphasize figures before papers represents an important shift, even on its own, one that is shared by other recent tools, such as DiagramFlyer. In particular, the figures are more closely related to specific results, and are therefore, we hypothesize, more closely related to the intent of the user's search.

At a more advanced level, users can search for figures from a scientific domain or subdomain. To identify the fields, we use the mapequation [11], and to identify the important articles, we use the article-level Eigenfactor (ALEF) score, a variant of PageRank designed specifically for time-directed, article-level networks. The benefits of the ALEF score for ranking was recently shown in a competition against other article-level ranking algorithms in the WSDM data challenge [14]. We find (through informal conversations with users) that the images in this view are useful for gathering field-level information — particularly useful for new scholars to a field or for reminding experts of the common findings and models resident in the field.



Please Label Plots and Multi-plots

A project of the eScience Institue at the University of Washington

Figure 2: Screenshot of the bulk-labelling interface. (A) Instructions for using the interface. (B) Indicator for the type of figure the user is asked to label (e.g., photos). (C) The total number figures that have been labelled by the user. (D) Controls to allow users to label all figures directly. (E) Controls to allow users to refresh the pool of figures. (F) submits the result. (G) Zoom control to afford fluid inspection of figures and groups of figures.

The search performs a free-text match on the caption text extracted from the pdf (and eventually, like DiagramFlyer, the extracted text from the image itself) and returns relevant figures via a free-text index on the underlying database that incorporates stemming and tolerance for spelling errors. The returned figures are ordered by the ALEF score as an estimate of impact. To facilitate browsing and prevent certain figures from dominating the experience, the user can shift to random order by clicking a button near the search box.

The returned figures are arranged in a grid layout (Figure 1(a)) to make better use of screen real estate and account for the widely varying shapes and sizes of figures. We color the figure border to indicate the type of the figure identified by the classifier; the legend for these colors appears at the top of the screen (described in Section 3.1). Users can retrieve additional figures by scrolling down to the bottom of the page.

The most important feature in our approach is that users can restrict the search to figures of specific types by using the checkboxes just under the search box: photographs, tables, visualizations, diagrams, or equation. This faceted search feature is simple. In order to support figure type filtering, we developed a new classifier to properly label the figures. We posit that this categorization of figures based on their semantics (as opposed to just their embedded or surrounding text) fundamentally changes the way users can interact with the literature. For instance, a clinical software engineer in a cancer lab may search for papers describing a cloud architecture for electronic health records that they can use to inform the design of their own system. Searching for "cloud" and "EHR" may return figures from relevant papers, but the precision will be low since the task is to find *specific architectures*. With this interface, filtering figure types to diagrams only can improve precision. An alternative layout is bundling the figures from the same paper together and listing the papers (Figure 1(b)). This mode is designed for users who are looking for particular papers, but who may recall a memorable figure from the paper if not the title or author. Viewing article titles together with figures may help them narrow the scope. For figures with dense information such as composite figures, users can shift the slider to zoom in for close inspection or click the figure to review the figure caption, the source paper details, and other figures from related papers associated by the citation network. We bring the crowdsourced labelling function in this page showing the figure and paper details. We simplified the labelling interface (Figure 1(c)) to a quick click if the machine-label type is correct. For an incorrect case, it needs an extra move of modifying the figure-type in the drop-down menu. For future work, we plan to allow users to tag keywords of a figure.

Crowdsourced Labelling The viziometrics analysis tasks we aim to support rely strongly on the availability of human-labeled data. We hand-labeled thousands of images when training our dismantling algorithm for separating composite images [9] and our classification algorithm [8]. Going forward, additional labels for figure sub-types, content tags, and information extraction techniques will require even more human-labeled data for training. For example, a line chart in oceanography may be a "depth chart," while a line chart in machine learning could be an ROC curve. The level of expertise needed now and in the future motivated an efficient, intuitive, and expert-oriented interface for bulk-labeling of figures.

Due to the expertise required, these labelling tasks are not directly appropriate for Mechanical Turk. We therefore include the labelling interface as part of the core system in the interest of enticing our (expert) users to contribute to the labelling task directly. Figure 2 shows the interface for bulk-labelling. Labelers are shown an instructional page as a first step, and can access the instructions at any time by clicking the "Need Help?" button. In each round, a user is asked to select figures of a certain type (e.g. Diagram) chosen randomly. Figure can be included or excluded in the selection by clicking.

We hypothesize that an approach of "one label, many images" will allow the user to quickly become efficient at recognizing a specific type rather than spending too much time considering a specific figure. The decision to make on each figure is binary, and "difficult" figures can simply be ignored. We suspect that this approach will produce high accuracy labels, with a possible downside that difficult-to-classify images will be consistently ignored, perhaps complicating training tasks.

Clicking the "Next" button will submit the choice and go to the next round. The figures that were not chosen will stay in the new round. We expect the unchosen figures will eventually meet their categories if the user stays for enough rounds. We feed 20 figures in each round to help prevent unconscious mistakes caused by repetitive tasks. Since we group the candidate figures according to their machine-labels, the user's task is often confirmatory. In this case, using the "Label All" button can save time. Although this approach may lead to confirmation bias in the human labels, but we find anecdotally that mistakes in the machine labels stand out and attract attention. When the user is satisfied that each figure in the set is labeled correctly or ignored, clicking the "Fresh" button can retrieve a new pool of unlabeled images. We don't set the end of round so the user can stop in any round.

Open-data Platform — We have argued previously that the study of the visual information in the literature can lead to new insights into scientific communication [8]. For instance, we found the top 5% papers ranked by ALEF in PMC tend to have higher diagram density (total number of diagrams per total number of pages) and plot density (Figure 3), perhaps suggesting that important papers carry ideas that are easier to express visually. In related work, Early, Fawcett et al. [4] used their own data to show the heavy use of equations can impede communication among biologists. These two studies show a new direction in scientometrics by analyzing the use of scholarly figures. To aid in these investigations, we include a tool for aggregate, longitudinal, online analysis of the patterns of visual information for use by social scientists. In addition to the interface, the human-labeled figures will be released for research use. We hope it can incubate new findings of visual patterns from our audience and stimulate users to pursue innovative ideas in viziometrics.

3. ARCHITECTURE

VizioMetrix's architecture proceeds in two sub-systems. The offline data backend extracts visual information from a corpus and stores them in a database used to power the online system. The online system includes three applications for different audiences (researchers searching the literature using figures as the central facet and researchers mining figures from the literature). Figure 4 shows the system architecture. Our design goal is to offer a reusable platform to support viziometrics applications.

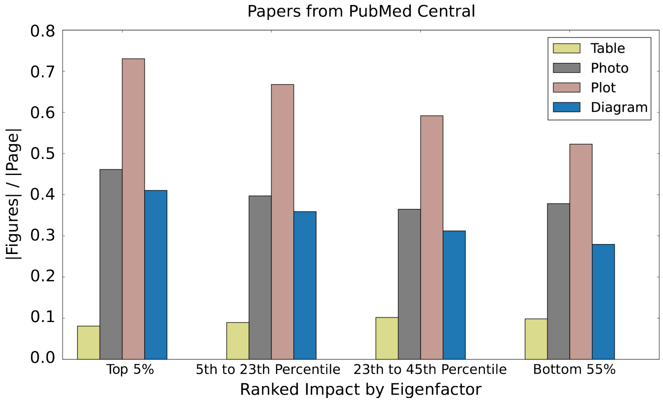


Figure 3: Scholarly impact versus figure density [8]. We rank PMC papers by ALEF and group them into 4 tiers: top 5%, 5th to 25th percentile, 25th to 50th percentile, and the bottom 50%. Any two papers with ALEF difference within 1E-12 are regarded as having equal impact and forced to be in the same bin. Hence, the boundary between 2nd tier and 3rd tier shifts to 23th percentile (papers in 23th to 25 percentile have equal ALEF) and the last bin increases to 55% (papers in bottom 55% have ALEF of zero), where are the closest positions to 25th and 50% respectively. For each tier, we average the densities (total number of figure per paper page) of 4 figure categories: table, photo, plot, and diagram. We ignore equations here because they are text. The top 5% paper tend to have statistically higher densities of plots and diagrams.

3.1 Data Backend

We extracted all papers from PubMed Central (PMC) repository, an archive of biomedical and life science literature. It offers free access to approximately 1 million articles. Every article is packaged with its PDF files, meta data file, and figure images. We extracted the figure images and dumped them in Amazon S3 server. We parsed the meta data to get paper titles, abstracts, citations, etc and stored this information in our database. The figures are then fed into figure processing pipeline (Figure 4) for classification. The citations are then used to calculate ALEF scores.

Image Processing We found that 66% of article files had image files and extracted 8.9 million source images from these files. The majority of images were in JPEG format. There are few images in GIF, TIF, TIFF, and PNG. We filtered all GIF images out since they are duplicate copies of images in other formats. Moreover, TIF and TIFF images can be extremely large and cannot be rendered by common browsers. Thus, we resized TIF and TIFF images to 1280 pixels corresponding to the longer side without modifying aspect ratios and then converted them into JPEF format.

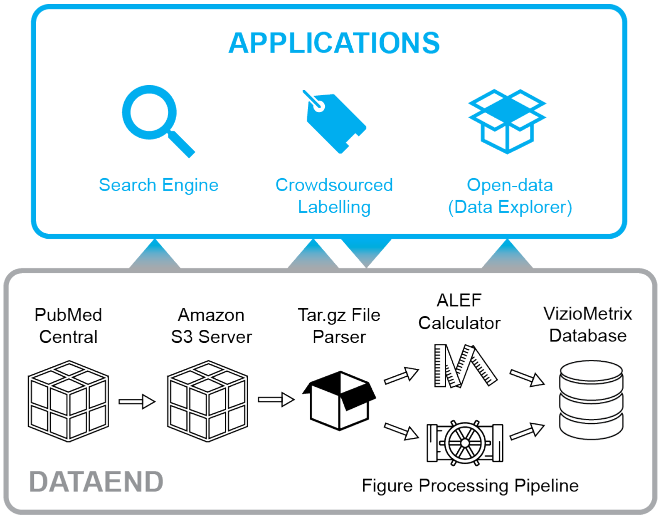


Figure 4: VizioMetrix's Architecture. The gray box illustrates the processing system on which VizioMetrix is based. This includes a data pipeline, which parses articles files, classifies figures, and calculates the article influence. The data pipeline then pushes the results into the VizioMetrix database. The blue box lists the three applications powered by the database.

We ended up with 4,986,302 images that are extracted from 680,494 papers. All file resources are stored in our Amazon S3 server so that all sub-systems can access them via Amazon AWS API. We also make these images available to other researchers interested in mining these figures.

We fed the images into the figure processing pipeline as shown in Figure 4. The composite figures (an image with multiple figures) are separated from singleton figures (an image with one figure). The composite figures were dismantled into several single figures using a method we developed in a previous paper [9]. Then all the single figures were classified into five categories: equation, diagram, photo, plot, and table. Each category can be described in terms of representative examples:

- Equation: embedded mathematics, Greek and Latin characters
- Diagram: schematics, conceptual diagrams, flow charts, model explanations, phylogenetic graphs
- Photo: microscopy images, diagnostic images, radiology images, fluorescence imaging, etc
- Table: tables
- Plot: numeric data visualization, bar charts, scatter plots, line charts, etc

We run figure processing in parallel. The final results are pushed into the VizioMetrix database (currently implemented in MySQL) together with figure information such as the figure id, the file path on S3, the image size, etc. More detail about the figure processing pipeline can be found in [8]. We join the publication table with the image table, and materialize the result for performance reasons. Due to the limitation of full-text search in MySQL, we integrate Solr with MySQL as a free-text indexing solution to achieve reliable performance.

3.2 Visual Search Engine

The application backend is responsible for data ingestion. Whenever a client sends a GET request with search keywords, the backend system will query the database to find the literature containing the keywords in title, abstract and figure captions. A list of figure information from matched literature will be return in JSON format. Since the figure files are stored in Amazon AWS S3, we store urls only and use S3 to deliver the images to the client. The frontend fetches the returned JSON and renders the layout in the user's desired mode (a grid or a conventional list). We return only the first 100 figures ordered by paper impact in descending order to achieve reliable performance. When the user scrolls to the end of the set of figures, a new GET request will be issued to retrieve the next 100 figures. When a user clicks a figure to get further information, another GET request is sent to retrieve relevant figures from associate papers. The current system returns the figures from the papers that cite and have been cited by the paper containing the selected figure. We plan to employ scholarly article recommendation techniques to provide more accurate results[13].

3.3 Backend for crowdsourced Labelling

The crowdsourced labelling backend responds to GET and POST requests. Whenever a GET request is received, it returns 20 figure image URLs. The 20 figures are in the same category randomly selected by the system. In the future, we plan on providing additional options for the user. For example, users may want images from the top conference papers or they may want to label papers from a particular time period. When the user returns the labelling result, the system will push the user's IP address, the figure id, and the given label into the database. The ground-truth type of a figure will be determined via voting. Since we have a very large image inventory and is growing continuously, one figure is likely to be labelled once and never be selected as a candidate again, which affects the credibility of the ground-truth data. Hence, we increase the probability of selecting images that have been labelled previously. The crowdsourced labelling data will be used offline to improve our machine learning models and will be open to the public for academic use via our open-data platform.

3.4 Open-data Platform

We provide access to our labeled image corpus and the derived viziometrics data. We also plan on expanding our dataset to include other large scholarly databases, including JSTOR, SSRN, arXiv, DBLP, etc. The image corpus that we used to train our classifiers and our raw data on from the PMC archive. The image corpus contains 782 photos, 436 tables, 394 equations, 890 visualizations, and 769 diagrams. We plan to enlarge this corpus to 20,000 images in each category. Second, the raw data was used to discover visual patterns (see Figure 3) [8]. It includes paper and figure meta data, machine-labels, and ALEF scores. We also plan to build a gateway for Tableau users so that they can connect to our data directly. For researchers with a background in statistics and programming, we will open our database via REST APIs.

4. FUTURE WORK

We have many future directions. We plan to expand our data to include other archives (arXiv, JSTOR, SSRN, etc). We will continue to grow the corpus of labelled images and will continue to improve our classifiers using cutting-edge deep learning models [5, 7]. Moreover, we plan to utilize natural language processing (NLP) to analyze figure captions that describe the images we are classifying. Extracting keywords in a caption can improve the searching quality and help us determine the most important figure in an article. We also aim to extract axes labels and quantitative data from plots using methods from one of our colleagues [12]. For the search engine, we plan to import scholarly article recommenders to improve the related figure results, and we plan to build a user-centric interface for browsing these relations. For crowdsourced labelling, we would like to introduce tag functionality and utilize these results for improving document retrieval, support NLP models and derive other applications. The Open-data platform is also under construction. We also plan on improving the faceted search.

5. CONCLUSIONS

We present VizioMetrix, a platform for mining millions of figures from the biomedical sciences. Our hope is that the platform will catalyze future research for improving scholarly search and facilitating large-scale analysis of these figures and new figure-centric applications. VizioMetrix provides a figure-oriented search service for general academic users and an open data resource for researches interested in mining scholarly figures. We provide a data processing pipeline that extracts the metadata from scholarly papers, extracts the figures and classifies them into different types. We also develop a crowdsourced labelling application for further labeling and subsequent improvement of our machine learning methods and methods of others. This platform is needed since mechanical turkers do not usually have the domain knowledge for labeling figure types[2]. We plan to assign a Document Object Identifier (DOI) to our dataset and making it freely available to the general public. We hope this platform reduces the activation energy needed to analyze scholarly figures at scale and provokes new and exciting questions.

6. ACKNOWLEDGMENTS

The authors wish to thank the authors of the papers from which we drew the examples in this paper. This work is sponsored in part by the National Science Foundation through S2I2 award 1216879 and IIS award III-1064505, a subcontract from the Pacific Northwest National Lab, the University of Washington eScience Institute, and an award from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

7. REFERENCES

[1] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM, 2009.

[2] S. Bhatia, P. Mitra, and C. L. Giles. Finding algorithms in scientific articles. In *Proceedings of the 19th international conference on World wide web*, pages 1061–1062. ACM, 2010.

[3] Z. Chen, M. Cafarella, and E. Adar. Diagramflyer: A search engine for data-driven diagrams. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15 Companion, pages 183–186, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[4] T. W. Fawcett and A. D. Higginson. Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences*, 109(29):11735–11739, 2012.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[7] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE, 2010.

[8] P.-S. Lee, J. D. , West, and B. Howe. Viziometrics: Analyzing visual information in the scientific literature. *In Prep.*, 2016.

[9] P.-s. Lee and B. Howe. Dismantling composite visualizations in the scientific literature. In *International Conference on Pattern Recognition Applications and Methods, ICPRAM, Lisbon, Portugal*, 2015.

[10] S. Ray Choudhury and C. L. Giles. An architecture for information extraction from figures in digital libraries. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 667–672. International World Wide Web Conferences Steering Committee, 2015.

[11] M. Rosvall and C. T. Bergstrom. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One*, 6:e18209, 2011.

[12] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. ReVision: Automated Classification, Analysis and Redesign of Chart Images. In *UIST '11*, pages 393–402, 2011.

[13] I. Wesley-Smith, R. Dandrea, and J. West. An experimental platform for scholarly article recommendation. In *Proc. of the 2nd Workshop on Bibliometric-enhanced Information Retrieval (BIR2015)*, pages 30–39, 2015.

[14] J. West, M. , D. Vilhena, and C. Bergstrom. Ranking and mapping article-level citation networks. *In Prep.*, 2016.

[2] Mechanical turkers could label the four different types of figures in this paper, but we plan on extending beyond this set and including more complicated figure types (e.g., different protein gels, phylogenetic trees,etc).