BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 780–789
doi:10.1093/bioinformatics/btt030

Bioimage informatics

Advance Access publication February 7, 2013

A new algorithm for context-based biomedical diagram similarity estimation

Songhua Xu[1,*,†], Jianqiang Sheng[2,*,†] and Xiaonan Luo[2]

[1]Information Systems Department, College of Computing Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA and [2]National Engineering Research Center of Digital Life, State-Province Joint Laboratory of Digital Home Interactive Applications, School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, P.R. China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Diagrams embedded in the biomedical literature convey rich contents, which often concisely and intuitively highlight key thesis of a research article. Despite their vital importance and informative clues for biomedical literature navigation and retrieval; currently, we miss an effective computational method for automatically understanding and accessing these valuable resources.

Proposed Method: To address the aforementioned gap, we propose a novel context-based algorithm for estimating the similarity between a pair of biomedical diagrams. The main difference of the proposed algorithm with respect to the existing methods lies in the new algorithm's incorporation of the semantic context associated with diagrams in their source documents into the diagram similarity estimation process. In addition, the new approach also performs a series of advanced image processing and text mining operations to comprehensively extract the semantic content graphically encoded inside diagram images.

Results: The new algorithm can be deployed as a reusable component providing a fundamental function for building many advanced, semantic-aware applications on biomedical diagram processing. As a case study, in our experiments, we demonstrate the advantage of the new algorithm for diagram retrieval. A set of biomedical diagram search and ranking experiments were conducted, where the performance of the new method was compared with that of five peer methods. The comparison results demonstrate the performance superiority of the new algorithm with all peer methods with statistical significance.

Contact: songhua.xu@njit.edu, shengjianqiang@163.com or lnslxn@mail.sysu.edu.cn.

Received on May 27, 2012; revised on January 11, 2013; accepted on January 17, 2013

1 INTRODUCTION

Diagrams are widely used graphical vehicles for illustrating ideas, explaining hypotheses and reporting findings. They provide a powerful communication device for visually sharing key information supplied in a document. These visual elements are well received by readers—people often like to overview key contents of a document through browsing its embedded diagrams, if any. Such a diagram browsing-based practice for document navigation has been popularly adopted in reality by many biomedical researchers to cope with the exploding amount of biomedical literature in existence today. In this article, we propose a new diagram similarity estimation method, which exploits the context information of a diagram latent in its source document for deriving a high-level understanding over the diagram's intended semantic messages. As image similarity is of fundamental importance for many biomedical diagram image processing, understanding and retrieval tasks, our new similarity estimation method can be used for many advanced semantic computing applications relating to diagram images, such as searching, ranking, clustering and categorizing diagrams, to name a few. One important extended application of our algorithm is to apply the algorithm to empower search engines and digital library systems, so that they can more capably return diagrams and the corresponding source documents to meet users' needs and interests in diagram searching and diagram browsing-based visual literature navigation. Because of space limit, we will only report the results of our experiments that demonstrate the advantage of our method for diagram retrieval and ranking.

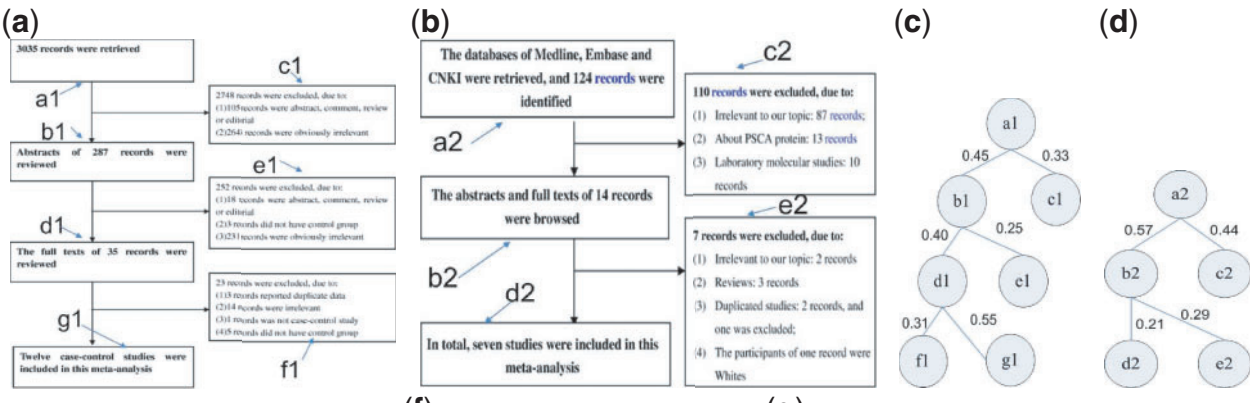
In this article, we introduce a new method for context-based diagram similarity estimation via a probabilistic reasoning approach. Based on the new method, we build an algorithmic framework for deriving context-based diagram similarity, including procedures to detect nodes and edges from an input diagram image using off-the-shelf computer vision tools, a method to represent the extracted nodes and edges from the input diagram image as a graph and the procedure to apply the new method for deriving pairwise diagram similarity through cross-referencing the diagrams' graph representations and their source documents (see Fig. 1 for an example). In the end, we also present extensive experimental results for validating the effectiveness of our new method in the application context of diagram retrieval and ranking.

2 RELATED WORK

We will now briefly look at two aspects of work closely related to our study here, including (i) diagram similarity estimation and (ii) context-based image retrieval.

Diagram similarity estimation. Significant efforts have been dedicated to designing algorithms and methods for estimating pairwise diagram similarity, most of which focus on processing

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



(e)

	a1	b1	c1	d1	e1	f1	g1
a1	0	0.45	0.33	0	0	0	0
b1	0.45	0	0	0.40	0.25	0	0
c1	0.33	0	0	0	0	0	0
d1	0	0.40	0	0	0	0.31	0.45
e1	0	0.25	0	0	0	0	0
f1	0	0	0	0.31	0	0	0
g1	0	0	0	0.55	0	0	0

(f)

	a2	b2	c2	d2	e2
a2	0	0.57	0.44	0	0
b2	0.57	0	0	0.21	0.29
c2	0.44	0	0	0	0
d2	0	0.21	0	0	0
e2	0	0.29	0	0	0

(g)

	a1	b1	c1	d1	e1	f1	g1
a2	0.09	0	0	0	0	0	0
b2	0	0	0	0.44	0	0	0
c2	0	0	0	0	0.11	0	0
d2	0	0	0	0	0	0	0.31
e2	0	0	0	0	0	0.05	0

Fig. 1. (**a** and **b**) are two diagrams commonly related to the query theme of gastric cancer, where (a) is Figure 1 in the article of Liu *et al.* (2010), and (b) is Figure 1 in the article of Qiao and Feng (2012); (**c** and **d**) are the two diagrams' respective attributed graph representations extracted by the method introduced in this article; (**e** and **f**) are the two diagrams' respective weighted adjacency matrices **W**s constructed by our method according to their respective source documents; (**g**) is the optimal matching matrix $\widehat{\mathbf{M}}$ computed for the two diagrams

a specific type of diagram. For example, within the software engineering community, statecharts are a specialized type of diagrams widely used for illustrating calling logics in a software architecture design. Nejati *et al.* (2007) studied the problem of optimally matching statecharts, where they offered a special-purpose image similarity metric for measuring statechart similarity. Wombacher (2006) validated a number of metrics for measuring the similarity between a pair of workflow diagrams, where each workflow is represented as finite state automata. All the metrics he evaluated can be broadly categorized into language and structure-based approaches. His study suggested that the relatively simple n-gram sets-based approach achieves the best performance among all peer methods. Li *et al.* (2008) proposed a structural approach using high-level change operations for measuring similarity between two process model diagrams. Their idea is to find a minimum set of addition, deletion and moving operations to transform one diagram into another diagram, where the minimum number of transformation steps needed is used as the pairwise diagram distance. Dijkman *et al.* (2011) experimentally compared three classes of similarity measurement methods for business process model diagrams—label similarity, structural similarity and behavioural similarity (element labels and causal relations captured in a process model). Their result shows that structural similarity attains the best performance. Ehrig *et al.* (2007) introduced a method for measuring the similarity between a pair of business process model diagrams. Their method derives diagram similarity from three aspects: syntactic, linguistic and structural measures. Minor *et al.* (2007) studied the problem of workflow diagram similarity estimation and retrieval by introducing a structure-based approach using a weighted graph edit distance. Melnik *et al.* (2002) proposed a versatile graph matching algorithm, called 'similarity flooding', for matching data schema diagrams. Their algorithm first derives a similarity propagation graph to denote pairwise node similarities between two diagrams and then performs a similarity spreading process in the graph for deriving an optimal matching between the two diagrams. Madhusudan *et al.* (2004) introduced a structural method using Artificial Intelligence (AI) planning techniques for comparing diagrams illustrating workflow models. Their method first uses a domain-independent AI planning-based approach to represent diagrams of business workflow models as cases and then adopts a case-based reasoning framework for deriving pairwise diagram similarity.

Context-based information retrieval. A decade ago, Lawrence (2000) pointed out the importance of context in web search. Recently, Belkin (2008) discussed the challenges associated with characterizing context for building information retrieval applications. Within the image retrieval field, people have explored contextual information for image processing, annotation and retrieval, e.g. Sinha and Jain (2008); Luo *et al.* (2009); O'Hare and Smeaton (2009); Lopes (2009); Segev and Toch (2009); Yang *et al.* (2010); Fisher and Hanrahan (2010); Choi *et al.* (2010); Yang *et al.* (2011). For the diagram retrieval problem, textual information carried inside a diagram only intends to shed highlights rather than to provide detailed explanation over the diagram's content; in addition, the caption of a diagram does not always cover every message illustrated in the diagram. Both factors incur computational difficulties for automatic diagram understanding and retrieval.

In the bioinformatics literature, Meekers and Rahaim (2005) observed the importance of socioeconomic context when developing social marketing models for improving reproductive health. Moskovitch *et al.* (2007) designed a context-sensitive search method for retrieving medical text with better accuracy.

Sinha and Jain (2008) described an unsupervised context analysis method for inferring context-specific gene regulatory networks from publicly available gene expression data. Rodriguez-Esteban and Iossifov (2009) introduced a figure mining method that jointly leverages image understanding, text mining and optical character recognition (OCR) techniques to retrieve tables and figures embedded in the biomedical literature that match a certain user-prescribed image type.

Among all the methods surveyed earlier in the text, none has looked into the contextual information of diagrams as a clue for estimating diagram similarities. To develop more comprehensive and accurate understanding over a diagram's semantic contents, we propose to acquire supplementary contextual information of a diagram from its source document for estimating diagram similarity in a semantically meaningful way. Our proposed approach is also applicable for processing other generic types of figures, such as statecharts and workflow diagrams, even though our method works particularly effectively with biomedical contents because of the prevalence of diagrams in biomedical publications. To the best of our knowledge, no published efforts have previously pursued this idea in the biomedical informatics field.

3 REPRESENTING A DIAGRAM WITH ITS DOCUMENT CONTEXT AS A GRAPH

Given a diagram G embedded in its source document D, in our method, we represent it as an attributed undirected graph. Each entity in G is represented as a node in the graph; each visually illustrated relationship between entities in G is represented as an edge in the graph. When no ambiguity arises, we will not differentiate the diagram from its graph representation. We can formally characterize G as $(\mathbf{N}(G), \mathbf{E}(G), \mathbf{W}(G))$, where $\mathbf{N}(G) = \{N_i | N_i \in G\}$ is the set of nodes in G; $\mathbf{E}(G) = \{E_{i,j} | N_i, N_j \in \mathbf{N}(G), E_{i,j} \in G\}$ is the set of edges in G; $\mathbf{W}(G)$ is a weighted adjacency matrix, which describes whether two nodes in the diagram are connected, and if so, how intensively the source document D discusses the semantic relationship represented by the edge, or how saliently the semantic relationship is embodied in D. We will look at how to derive $\mathbf{W}(G)$ later. As from the matrix $\mathbf{W}(G)$, we can readily understand the node connectivity information, we can more compactly characterize a diagram as $(\mathbf{N}(G), \mathbf{W}(G))$, without losing any information. Figure 1 gives an example of two sample diagrams' graph representations constructed by our method.

3.1 Detecting nodes from a diagram

Given an input diagram G in the form of a static image, we detect its nodes and edges through a set of image processing steps as follows. We first apply the Gaussian filter function offered by the OpenCV 2.2.0 package (http://sourceforge.net/projects/opencvlibrary) to remove local image noise from G. We then detect from G a collection of basic shape elements, including quadrilaterals, circles and ellipses. In our current implementation, we adopt the method proposed in the study by Qin *et al.* (2010) to detect these elements.

For each detected shape element, we further attempt to recognize any text that may be carried inside the interior image region of the shape. This text recognition task is accomplished by parsing the image region to the OCR tool provided in the Microsoft Office 2007 Document Imaging package. In our OCR process, currently, we only process English contents. For each recognized word from the OCR process, we match the word against the full text of the diagram's source document D. Only words that occur in D will be retained; the rest of the words will be considered OCR errors and, hence, discarded. For all remaining OCR result words, we further remove stop words. Finally, we perform a stemming process to restore each word to its basic root form.

Each shape element detected previously will be represented as a node N_i. In this way, we establish our node set $\mathbf{N}(G) = \{N_1, N_2, \cdots\}$, where each N_i is a shape element. After this step, we remove from the image G all the detected shape elements, including their interior image regions, to make the downstream image processing steps more reliable.

3.2 Detecting edges from a diagram

To construct the edge set $\mathbf{E}(G)$ for representing the node connectivity information in the diagram, we first detect arrows and line pieces, the latter of which are composed of one or multiple line segments in the image G (see Fig. 1 for an example). To detect line pieces in G, we applied the application programming interface (API) named 'cvHoughLines2' in the OpenCV 2.2.0 package; to detect arrows in G, we adopted the algorithm proposed in the study by Wendling and Tabbone (2003), which is relatively easy to implement and is capable of producing satisfying performance in our experiments. Occasionally, line pieces and arrows in a diagram may be accompanied by annotation text. To capture such text, we first remove from G all recognized line pieces and arrows. For the remaining image region, we then perform a text detection procedure using the method suggested in the study by Xu and Krauthammer (2010). For each recognized text region, we will anchor the text region onto its nearest line piece or arrow according to the Euclidean distance. Finally, we will call another OCR procedure to recognize these annotation text strings.

For each line piece or arrow detected previously, we need to associate both its end points to their respective closest shape elements according to the Euclidean distance. In our work, we define the distance between a line piece or an arrow to a shape element as the minimum distance between one end point of the line piece or arrow and a pixel on the contour of the shape element. As each shape element has been represented as a node N_i, any pair of nodes commonly pointed to by a line piece or an arrow are considered linked, in which case, we will introduce an edge to connect the two nodes in G. Through the aforementioned process, we construct our edge set $\mathbf{E}(G)$ for G. The aforementioned procedure of transforming an input diagram image into its corresponding structural graph representation is implemented as a fully automatic module in our prototype system.

3.3 Identify counterpart text for diagram nodes

For each node N_i detected from G, we need to identify text fragments in the source document D that embody the semantics represented by N_i's corresponding visual symbol on the diagram. To locate a text position in a document, we use a sentence's sequence number in the document as the location index. For text appearing in the main body of an article, we separate it into sentences

according to the presence of punctuation marks in the text; in particular, for text occurring in the title of an article, an article section or sub-section, as long as the source document allows automatic detection of these title regions, we treat all text displayed in one title region as a single sentence. Let the set of sentences semantically related to the node N_i be $\mathbf{S}_i = \{s_{i,1}, s_{i,2}, \cdots\}$, where each $s_{i,j}$ is a sentence in D that explains or discusses the meanings of N_i. Each $s_{i,j}$ is associated with a significance score $\rho_{i,j} \in [0, 1]$ that indicates the semantic relatedness between $s_{i,j}$ and N_i. For an arbitrary sentence $s_x \in D$, to measure its semantic relatedness with N_i, we compare the alignment of the semantics represented by s_x and N_i, respectively, according to their text. Recall that the text of the node N_i has been previously recovered through the OCR process. To estimate the aforementioned semantic alignment, we use the algorithm proposed in the study by Li *et al*. (2006), which is specifically designed for measuring semantic similarity between two pieces of short text. To determine the sentence set $\mathbf{S}_i$ for a given node N_i, we start with an empty set and scan all sentences in D. We respectively derive each sentence's semantic relatedness with N_i following the aforementioned procedure. If the detected semantic relatedness exceeds an empirically chosen threshold (0.05 in all our experiments), we consider the sentence noticeably related to the node and collect the sentence into the set $\mathbf{S}_i$. In this document sentence scanning procedure, we consider all sentences in the full text of the document, including those in the document's title, abstract, footnotes and figure captions.

3.4 Identify counterpart text for diagram edges

Once we have identified the counterpart text in D for every node in G, we can further identify the corresponding text in D that reflects the semantic meanings denoted by each edge in the diagram. Our edge counterpart text detection procedure is based on the node counterpart text detection result. Recall that $E(i, j)$ is the edge that connects the nodes N_i and N_j in G; N_i, and N_j's counterpart text in D is organized as two sentence sets $\mathbf{S}_i$ and $\mathbf{S}_j$, respectively. To locate counterpart text for the edge $E(i, j)$, we essentially pair sentences from $\mathbf{S}_i$ and $\mathbf{S}_j$, one from each set. Let $|\mathbf{S}_x|$ be the number of sentences in the sentence set $\mathbf{S}_x$. Our aforementioned edge counterpart text identification procedure leads to $|\mathbf{S}_i| \times |\mathbf{S}_j|$ instances of the counterpart text for $E(i, j)$. For the sentence pair $s_{i,u} \in \mathbf{S}_i$ and $s_{j,v} \in \mathbf{S}_j$, we estimate its significance in representing the semantic meanings of $E(i, j)$ as $\rho_{i,u}\rho_{j,v}\theta(s_{i,u}, s_{j,v})$. Recall that $\rho_{i,u}$ and $\rho_{j,v}$ are the significance of the sentences $s_{i,u}$ and $s_{j,v}$ in embodying meanings of the nodes N_i and N_j, respectively; $\theta(s_{i,u}, s_{j,v})$ is a newly introduced measure that quantifies how likely the two sentences $s_{i,u}$ and $s_{j,v}$ embody the semantic meanings intended by the edge $E(i, j)$. We assume the farther apart the two sentences are, the less likely the pair of sentence describes the relationship represented by the edge. Note that $s_{i,u}$ and $s_{j,v}$ may both refer to the same sentence, in which case, it is most likely the meanings of the edge $E(i, j)$ are reflected by the sentence. In our current design, $\theta(s_{i,u}, s_{j,v})$ is estimated as follows: $\theta(s_{i,u}, s_{j,v}) = exp(-\frac{dis(s_{i,u}, s_{j,v})}{ave\_dis(D)})$, where $dis(s_{i,u}, s_{j,v})$ is the number of non-stop words separating the two sentences $s_{i,u}$ and $s_{j,v}$; $ave\_dis(D)$ is the average number of non-stop words in a sentence in the document D.

We further introduce a function $\varrho(s_{i,u}, s_{j,v}, t_{i,j}) \in [0, 1]$ to measure the degree of relevance (the larger, the more relevant) between text of the two sentences $s_{i,u}$ and $s_{j,v}$ and the annotation text of the edge $t_{i,j}$. Recall that we mentioned earlier that occasionally a line piece or an arrow in a diagram may be accompanied by some annotation text. We empirically define $\varrho(s_{i,u}, s_{j,v}, t_{i,j})$ as follows:

$$\varrho(s_{i,u}, s_{j,v}, t_{i,j}) = \max\{\xi(w_a|_{w_a \in t_{i,j}}, w_b|_{w_b \in s_{i,u} \cup s_{j,v}})\}, \quad (1)$$

where $\xi(w_a, w_b) \in [0, 1]$ computes the semantic relatedness between a pair of words, w_a and w_b, according to WordNet::Similarity (http://search.cpan.org/dist/WordNet-Similarity/doc/intro.pod). If the edge does not have any anchoring text, i.e. $t_{i,j} = \emptyset$, $\varrho(s_{i,u}, s_{j,v}, t_{i,j}) = 1$. The reason why we made such a value assignment is due to the following logic: if an edge does not carry any anchoring text, no particular semantic relationship is specified to govern the edge's two end nodes. Hence, we give the benefit of doubt by assuming any text can be relevant in some way to the relationship represented by the edge. If there is indeed some anchoring text associated with the edge, then only those counterpart text instances that embody the same semantic relationship specified by the anchoring text shall be considered well matched and relevant to the edge.

Based on the significance of each sentence pair, we can further estimate the significance of the edge $E(i, j)$ embodied in the entire document D by aggregating the significance of all its counterpart sentence pairs across the document as follows:

$$w_{i,j} = \frac{1}{z} \sum_{s_{i,u} \in \mathbf{S}_i, s_{j,v} \in \mathbf{S}_j} \rho_{i,u}\rho_{j,v}\theta(s_{i,u}, s_{j,v})\varrho(s_{i,u}, s_{j,v}, t_{i,j}), \quad (2)$$

where z is a scaling factor to ensure the significance value for the most significant edge in G is 1. Based on the estimated significance for each edge in G, we can construct a weighted adjacency matrix $\mathbf{W}(G)$, for characterizing whether two nodes in the graph are connected, and if so, how saliently this connection is embodied in the source document D. Let $\mathbf{W}_{i,j}(G)$ be the element on the i-th row and j-th column of the matrix $\mathbf{W}(G)$; we define $\mathbf{W}(G)$ as follows: $\mathbf{W}_{i,j}(G) = w_{i,j}$ if $E_{i,j} \in \mathbf{E}(G)$ and 0 otherwise.

4 MEASURING DIAGRAM SIMILARITY BY LEVERAGING DOCUMENT CONTEXT

Given two diagrams G_1 and G_2, to measure their similarity, we need to compute an optimal matching between these two diagrams. Let $|\mathbf{N}(G_1)|$ and $|\mathbf{N}(G_2)|$, respectively, be the number of nodes in the two diagrams. We can then represent any matching relationship between G_1 and G_2 using a $|\mathbf{N}(G_1)| \times |\mathbf{N}(G_2)|$ dimensional matching matrix $\mathbf{M}(G_1, G_2)$. The element on the i-th row and j-th column of $\mathbf{M}(G_1, G_2)$, denoted as $M_{i,j}(G_1, G_2) \in [0, 1]$, represents the matching degree between the i-th node in G_1 and the j-th node in G_2. Note that $M_{i,j}(G_1, G_2)$ can be any number between 0 and 1, which implies one node from a diagram can match to one, multiple or no node in the other diagram. Such a fuzzy matching mechanism allows our method to deal with a wide range of diagram matching scenarios without enforcing a strict one-to-one matching between two diagrams' nodes, the situation of which does not always exist in reality. For simplicity,

we will abbreviate $\mathbf{M}(G_1, G_2)$ for $\mathbf{M}$ from now on when no ambiguity arises. Also, we will abbreviate $\mathbf{W}(G_1)$ and $\mathbf{W}(G_2)$ as $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively, for easy notation. In the following, we will introduce a method for deriving an optimal matching relationship matrix $\widehat{\mathbf{M}}$ for an arbitrary pair of diagrams G_1 and G_2. The optimality in the matching relationship matrix refers to the estimation of the optimal matching degree between two matrices using our heuristic-based estimation framework.

We first introduce a heuristic function $H(G_1, G_2, \mathbf{M})$ to represent the semantic relatedness between two diagrams G_1 and G_2, assuming that nodes in the two diagrams are matched according to the correspondence relationship $\mathbf{M}$. Intuitively, the more semantically related the two diagrams are, the larger the value of $H(G_1, G_2, \mathbf{M})$ becomes. This modelling perspective is inspired by the work of Li and Hsu (2008), who studied a related problem of content-based natural image retrieval with relevance feedback using a graph-theoretic region correspondence estimation method. Given the aforementioned function $H(G_1, G_2, \mathbf{M})$, we can search for the optimal matching relationship matrix $\widehat{\mathbf{M}}$ as follows:

$$\widehat{\mathbf{M}} = \arg\max_{\mathbf{M}} H(G_1, G_2, \mathbf{M}). \tag{3}$$

As mentioned earlier, a diagram G can be characterized using its node set $\mathbf{N}(G)$ and its weighted adjacency matrix $\mathbf{W}(G)$. Working with this premise, we empirically assume that:

$$H(G_1, G_2, \mathbf{M}) \propto \prod_{i=1}^{|\mathbf{N}(G_1)|} \prod_{j=1}^{|\mathbf{N}(G_2)|} P^{M_{i,j}}(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M}), \tag{4}$$

where $H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M})$ estimates the pairwise node matching score between $N_i(G_1)$ and $N_j(G_2)$. Recall that $N_i(G_x)$ is the i-th node in the node set $\mathbf{N}(G_x)$, and $M_{i,j}$ is the element on the i-th row and j-th column of the matching matrix $\mathbf{M}$.

It shall be noted that the above property is heuristically assumed. Intuitively, the pairwise graph similarity can be estimated according to the pairwise similarities between corresponding nodes in the two graphs. The higher such node similarities collectively are, the more semantically related the two graphs may be perceived as assumed by our heuristic. Note that if $M_{i,j} \neq 0$, it means that the node $N_i(G_1)$ from $\mathbf{N}(G_1)$ matches the node $N_j(G_2)$ from $\mathbf{N}(G_2)$ according to the matching relationship $\mathbf{M}$. The reason why we raise the pairwise node matching score function $H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M})$ to the power of $M_{i,j}$ is because the matching degree $M_{i,j}$ can be any number between 0 and 1 to emulate the fuzzy nature of such non-binary matching decision. Again, this power factor is empirically introduced, whose effectiveness will be proved through our experimental results to be presented later in this article.

Next, to estimate $H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M})$, we take into account two clues: (i) *self similarity*, i.e. how closely the two nodes' carrying text are; (ii) *context similarity*, i.e. how similar the two nodes' surrounding nodes are in terms of their text similarity. To measure the self similarity $\phi(N_i(G_1), N_j(G_2))$ between a pair of nodes $N_i(G_1)$ and $N_j(G_2)$, we first implement a method for estimating content similarity between a pair of sentences. Let $\psi(s_i, s_j) \in [0, 1]$ be the semantic similarity for an arbitrary pair of sentences s_i and s_j. $\psi(s_i, s_j) = 1$ indicates the two sentences deliver the same semantics, whereas $\psi(s_i, s_j) = 0$ shows the two sentences share no semantic overlap. To derive the value of $\psi(s_i, s_j)$, in our current implementation, we adopt the sentence similarity estimation algorithm proposed in the study by Li *et al.* (2006) because of the algorithm's leading performance among the peer methods. Based on the function of $\psi(s_i, s_j)$, we can now estimate $\phi(N_i(G_1), N_j(G_2))$. Assume $N_i(G_1)$ and $N_j(G_2)$'s counterpart sentence sets detected from diagrams G_1 and G_2's source documents D_1 and D_2, are $\mathbf{S}_i(G_1)$ and $\mathbf{S}_j(G_2)$, respectively. Also recall that the significance for a sentence $s_{i,u}$ ($s_{j,v}$) in $\mathbf{S}_i(G_1)$ ($\mathbf{S}_j(G_2)$) to embody the meanings intended by the node $N_i(G_1)$ ($N_j(G_2)$) is $\rho_{i,s}$ ($\rho_{j,v}$). We can now estimate $\phi(N_i(G_1), N_j(G_2))$ as follows:

$$\phi(N_i(G_1), N_j(G_2)) = \frac{1}{Z(G_1, G_2)} \sum_{s_{i,u} \in \mathbf{S}_i(G_1)} \sum_{s_{j,v} \in \mathbf{S}_j(G_2)} \rho_{i,u} \rho_{j,v} \psi(s_{i,u}, s_{j,v}) \tag{5}$$

where $Z(G_1, G_2)$ is a normalization term to ensure the maximum pairwise node similarity across the two graphs G_1 and G_2 is 1.

To estimate the context similarity of the pair of nodes $N_i(G_1)$ and $N_j(G_2)$, we first consider the similarity of the two nodes' immediately adjacent neighbours, which is denoted as $\vartheta_{1,1}(N_i(G_1), N_j(G_2))$, as follows:

$$\vartheta_{1,1}(N_i(G_1), N_j(G_2)) = \sum_{N_u(G_1) \in \mathbf{N}(G_1)} \sum_{N_v(G_2) \in \mathbf{N}(G_2)} \mathbf{W}_{1,i,u} \mathbf{W}_{2,j,v} \phi(N_u(G_1), N_v(G_2)), \tag{6}$$

where $\mathbf{W}_{1,i,u}$ and $\mathbf{W}_{2,j,v}$ are the short notations for $\mathbf{W}_{i,u}(G_1)$ and $\mathbf{W}_{j,v}(G_2)$, respectively.

Similarly, we can estimate the similarity between one node's immediate neighbour node and the other node's second-level neighbour node. In analogy to the definition of $\vartheta_{1,1}(N_i(G_1), N_j(G_2))$, we can further define $\vartheta_{1,2}(N_i(G_1), N_j(G_2))$ and $\vartheta_{2,1}(N_i(G_1), N_j(G_2))$ as follows:

$$\vartheta_{1,2}(N_i(G_1), N_j(G_2)) = \sum_{N_u(G_1) \in \mathbf{N}(G_1)} \sum_{N_v(G_2) \in \mathbf{N}(G_2)} \sum_{N_x(G_2) \in \mathbf{N}(G_2)} (\mathbf{W}_{1,i,u} \mathbf{W}_{2,j,v} \mathbf{W}_{2,v,x} \phi(N_u(G_1), N_x(G_2))), \tag{7}$$

$$\vartheta_{2,1}(N_i(G_1), N_j(G_2)) = \sum_{N_u(G_1) \in \mathbf{N}(G_1)} \sum_{N_x(G_1) \in \mathbf{N}(G_1)} \sum_{N_v(G_2) \in \mathbf{N}(G_2)} (\mathbf{W}_{1,i,u} \mathbf{W}_{1,u,x} \mathbf{W}_{2,j,v} \phi(N_x(G_1), N_v(G_2))). \tag{8}$$

Similarly, we can further define $\vartheta_{2,2}(N_i(G_1), N_j(G_2))$, whose explicit form is omitted because of space limit. By aggregating all the aforementioned sub-estimates, we can derive the context similarity for the pair of nodes $N_i(G_1)$ and $N_j(G_2)$, denoted as $\vartheta(N_i(G_1), N_j(G_2))$, as follows:

$$\vartheta(N_i(G_1), N_j(G_2)) = \sum_{u=1,2} \sum_{v=1,2} \vartheta_{u,v}(N_i(G_1), N_j(G_2)). \tag{9}$$

In our method, we do not calculate $\vartheta_{u,v}(N_i(G_1), N_j(G_2))$ for $u > 2$ or $v > 2$ because their values are almost always 0.

Finally, by combining $\phi(N_i(G_1), N_j(G_2))$ and $\vartheta(N_i(G_1), N_j(G_2))$, we can estimate $H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M})$ as follows:

$$\begin{aligned} &H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M}) \\ &= \kappa exp\{\phi(N_i(G_1), N_j(G_2)) + \vartheta(N_i(G_1), N_j(G_2))\}, \end{aligned} \tag{10}$$

where κ is a fixed constant. Substituting Equation (10) into Equation (4), we further have:

$$\begin{aligned} &\arg\max_{\mathbf{M}} H(\mathbf{N}(G_1), \mathbf{N}(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M}) \\ &= \arg\max_{\mathbf{M}} \sum_{i=1}^{|\mathbf{N}(G_1)||\mathbf{N}(G_2)|} \sum_{j=1} M_{i,j} \log(H(N_i(G_1), N_j(G_2), \mathbf{W}_1, \mathbf{W}_2, \mathbf{M})) \\ &= \arg\max_{\mathbf{M}} \sum_{i=1}^{|\mathbf{N}(G_1)|} \sum_{j=1}^{|\mathbf{N}(G_2)|} M_{i,j} (\log \kappa \\ &+ \phi(N_i(G_1), N_j(G_2)) + \vartheta(N_i(G_1), N_j(G_2))). \end{aligned} \tag{11}$$

To calculate the optimal matching matrix $\widehat{\mathbf{M}}$, we first create a $|\mathbf{N}(G_1)| \times |\mathbf{N}(G_2)|$ dimensional matrix $\aleph$, whose element on the i-th row and j-th column, $\aleph_{i,j}$, takes the value of $\log \kappa + \phi(N_i(G_1), N_j(G_2)) + \vartheta(N_i(G_1), N_j(G_2))$. We thus have:

$$\widehat{\mathbf{M}} = \arg\max_{\mathbf{M}} \sum_{i=1}^{|\mathbf{N}(G_1)|} \sum_{j=1}^{|\mathbf{N}(G_2)|} M_{i,j} \aleph_{i,j}. \tag{12}$$

As $\mathbf{M}$ is a matching matrix, it has the property that $\sum_j M_{i,j} \leq 1$ and $\sum_i M_{i,j} \leq 1$ for all i and j. The inequality in the constraints is introduced for handling the situation that an element from a diagram shall not be matched to any element in the other diagram to yield an optimal matching between the two diagrams. To solve the aforementioned optimization problem, we can use linear programming to find the optimal matching matrix, $\widehat{\mathbf{M}}$, that maximizes $\sum_{i=1}^{|\mathbf{N}(G_1)|} \sum_{j=1}^{|\mathbf{N}(G_2)|} M_{i,j} \aleph_{i,j}$. Once we derive $\widehat{\mathbf{M}}$, we can further derive the value of $H(G_1, G_2, \widehat{\mathbf{M}})$ as the similarity between the two diagrams G_1 and G_2. For information retrieval tasks that only care about rankings where the absolute similarity value is not important, it suffices to use the optimized target function value yielded in the linear programming procedure, i.e. $\sum_{i=1}^{|\mathbf{N}(G_1)|} \sum_{j=1}^{|\mathbf{N}(G_2)|} \hat{M}_{i,j} \aleph_{i,j}$, as the estimated similarity for the two diagrams G_1 and G_2. Figure 1 gives an example of the optimized matching matrix constructed by our method for a pair of sample input diagrams.

5 EXPERIMENTATION

5.1 Experiment set-up

To explore the effectiveness of our new diagram similarity estimation method for diagram retrieval, we conducted a set of evaluation experiments using a PC equipped with a Core i3 2.93 GHz CPU and 4 GB main memory, which ran the Windows XP operating system. To carry out our experiments, we first constructed a diagram image corpus where each diagram is accompanied by its corresponding source document. We acquired these images and their source documents through both downloading from PubMed Central (PMC) and using Google Image Search as follows: (i) we first downloaded all the publicly accessible images from PMC where each image is always accompanied by its source document. We then applied the diagram image recognition algorithm proposed in the study by Qin *et al.* (2010) to identify diagrams from all downloaded images. This procedure lets us acquire 12 500 diagrams. (ii) We then randomly selected 50 diagrams downloaded from PMC in the first step and fed the captions of these images, respectively, as queries into Google Image Search. For each search result image, Google always provides a back link to its source webpage. Following the back link, we can check whether the search result image is associated with a meaningful source document. In this operation, we first removed all the advertisement and navigation content from an image's source webpage using the algorithm proposed in the study by Ntoulas *et al.* (2006). If the filtered webpage contains >500 words, we then consider the webpage as a meaningful document. Otherwise, we discard the search result image. For all the images that passed the preceding test, we ran the algorithm of Qin *et al.* (2010) to detect and select all images of the diagram type and added them into our diagram corpus. Using the second approach, we acquired ~3000 additional diagrams through Google Image Search.

To experimentally explore the performance of a diagram retrieval method, we conducted a collection of diagram image search sessions, which were organized into 16 groups of queries, where each query group consists of multiple query sessions on a common theme. The 16 querying themes are, respectively, as follows, for which we also specify the number of query sessions performed for each theme group using a number in the bracket following the theme's topic phrase—a: breast cancer (9), b: gastric cancer (9), c: non-Hodgkin's lymphoma (9), d: multiple myeloma (9), e: HIV (8), f: detection of chronic kidney (9), g: heart block (9), h: malaria (8), i: thrombosis (8), j: angiogenesis (8), k: tumour angiogenesis (8), l: ochrobactrum (8), m: gene expression (8), n: cardiomyopathy (8), o: respiratory syndrome (9) and p: bone metastase (9). Query theme groups a–c primarily consist of images acquired through Google Image Searches; query theme groups d–m mostly consist of images downloaded from PubMed; the remaining query theme groups, i.e. groups n, o and p, contain images acquired through both means more evenly. For each query session, we randomly selected an image from our diagram corpus whose caption matches the session's theme phrase as the query input image. We then performed diagram retrieval against the whole diagram corpus (excluding the selected query input image) using a retrieval method whose performance is to be evaluated.

5.2 Experimental results

After the aforementioned procedure, we then applied the new diagram similarity estimation method introduced in this article for diagram retrieval and ranking. In each query session, we rank all the retrieved diagram search results according to each result diagram's estimated relatedness to the query diagram. We then recruited five subjects and asked each of them to independently label the relatedness of each search result diagram to the input query diagram according to each subject's personal judgment regarding the two diagrams' semantic similarity. The numeric label ranges from 0 (entirely irrelevant) to 1 (extremely related). We then took an average of the five user labels as the image's

overall user-rated query relevance score. Based on this score, we further calculated the normalized discounted cumulative gain (NDCG) to measure the quality of diagram retrieval and ranking for the query session. To understand the definition of NDCG, we first need to introduce the notation of discounted cumulative gain (DCG), which measures the information retrieval quality of a ranked search result set. DCG takes into consideration the query-relevance of each search result document along with its ranking position in the result list. The DCG score at a particular rank position p can be computed as:

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \tag{13}$$

where rel_i is the average user-rated query relevance of the search result item ranked at the position i in the list. Based on the notation of $\mathrm{DCG_p}$, we can further calculate the NDCG score for a ranked search result set at a particular rank position p as follows:

$$\mathrm{NDCG_p} = \frac{DCG_p}{IDCG_p}, \tag{14}$$

where $IDCG_p$ is the ideal (maximum) DCG score at the rank position p reachable by the current search result set. For more thorough discussions on the NDCG metric, readers are referred to Järvelin and Kekäläinen (2002). In this article, we adopted NDCG as our retrieval performance metric following the predominant practice in the information retrieval research field because such metric gives weighted considerations to search result items at different ranking positions, prioritizing those displayed at the top positions of the list. Plenty of information retrieval research, e.g. Järvelin and Kekäläinen (2002), has pointed out that the NDCG metric can better reflect the search quality perceived by end users than the traditional precision-based evaluation, which ignores rank positions. For all our experiments reported in this article, we use $\mathrm{NDCG_{20}}$ as the metric, as we empirically find it behaves most representatively among all versions of the NDCG scores for our application.

For comparison purposes, we also applied the following peer image search methods to repeat the image search experiments described earlier in the text—Peer Method (PM) 1: each image is represented using its caption text. We then used the text search engine Lucene to retrieve and rank images, whose ranking mechanism is described with details in the study by Hatcher and Gospodnetic (2004). PM 2: the method is to use the text present in title, description and tags of the images for improving the results obtained with a standard content-based search (Barrios *et al*., 2009). PM 3: each image is represented using its caption text and its visually embedded text, as well as the image's anchoring text, i.e. the sentence(s) in the source document that directly quotes the image. This is the image search method proposed in Xu *et al*. (2008). The weighting for mixing the three types of text in the ranking process is also manually tuned to maximize the total NDCG score of the method for all our tested queries. PM 4: the biomedical image metadata manager system proposed by Korenblum *et al*. (2011) that retrieves similar biomedical images using semantic metadata features. PM 5: a state-of-the-art process model diagram search method proposed by Li and Hsu (2008). When conducting all sessions of our comparative experimental studies, we used the same target diagram corpus when executing the five peer methods and our algorithm to ensure fair comparison among all methods.

To explore the diagram retrieval performance of the new method with respect to the five peer methods, we conducted a series of diagram search experiments using the aforementioned 16 theme groups of diagram query sessions. Figure 2 reports NDCG scores of both our method and that of the five peer methods in all these experiments, where the NDCG score of each method for every query is individually reported. We also congregated these individual NDCG scores to derive the distributions of all NDCG scores attained by our method and the five peer methods in all our querying experiments, whose distributions are reported in Figure 3 using boxplots. All the aforementioned experiment results clearly show that our new diagram similarity estimation method performs significantly superior to all peer methods for searching and ranking diagram images.

To further verify our method's performance superiority to the peer methods, we calculated *P*-values for the paired *t*-test following the well-established procedure in statistic hypothesis testing. We tested a series of null hypotheses that the performance of our method and that of a specific peer method is statistically equal. In Table 1, we report the *P*-values as results of two-tailed paired *t*-tests for diagram querying experiments of 16 theme groups. More concretely, for each query theme group, we executed all its constituent query sessions using our method and the five peer retrieval methods, respectively. Without loss of generality, let's focus on the first peer method PM1 initially. For every query session, we paired the NDCG scores for the top 20 diagram retrieval results obtained by our method with those returned by PM1 according to their respective rank positions. That is, every query session will produce 20 pairs of NDCG scores. For each query session in the first query theme group, we repeated the same process and collected all the resultant NDCG score pairs. This gave us 180 pairs of NDCG scores because there are nine query sessions in the first query theme group. Given these NDCG score pairs, we can then derive the *P*-values for the two-tailed paired *t*-test comparing the retrieval quality of our method and that of PM1 for the whole query theme group. The aim is to test the statistical significance of the superiority of our method with respect to PM1. The result is reported in the tabular cell under the column 'Ours-PM1' and on the row for the first query theme group in Table 1. To fill the entire table, we repeated the aforementioned procedure for comparisons against all the peer methods and query theme groups. In Table 2, we further report *P*-values for both one-tailed and two-tailed *t*-tests for all 16 theme groups of query experiments. To calculate the *P*-values, this time we collected all paired NDCG scores comparing our method and one of the peer methods across all query sessions in all query themes. Overall, among all *P*-values reported in Table 1, almost all of them are <0.05, except for a few ones that are marked in bold. In Table 2, all calculated *P*-values both for the one-tailed and two-tailed *t*-tests are <0.05. These small *P*-values consistently indicate a statistically significant superiority of our method with respect to the peer methods in retrieving diagrams semantically relevant to the input query diagram.

To explore the diversity in the search result diagrams, we further calculate the distributions of co-author distances among top ranked diagram retrieval results returned by our method. The purpose is to verify that the new algorithm is capable of

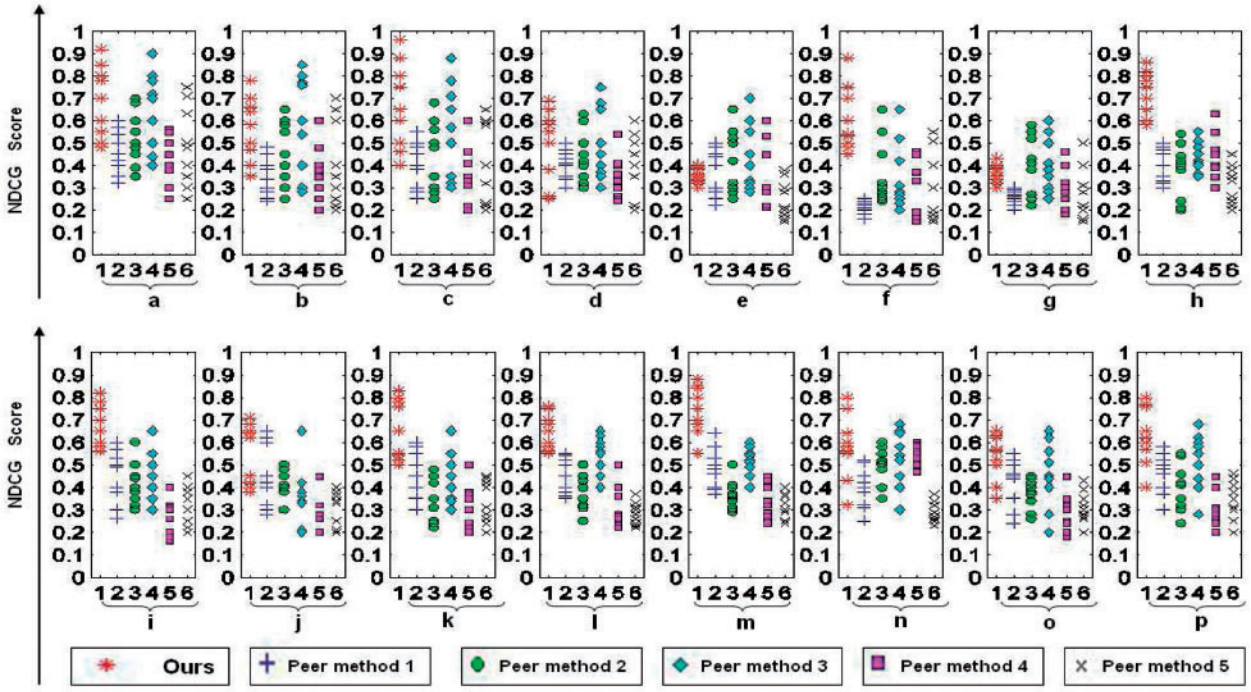


Fig. 2. NDCG scores for all querying experiments of 16 theme groups performed using our method and the peer methods, respectively

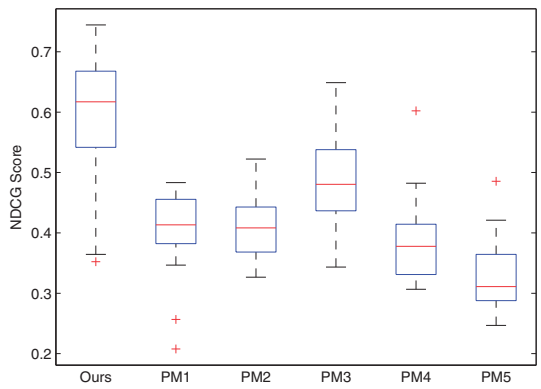


Fig. 3. Distributions of NDCG scores of our method (Ours) and the five peer methods (PM1–PM5) in all our diagram querying experiments

retrieving diagrams composed by people sharing weak or no collaboration relationships. That is, the algorithm will indeed retrieve diagrams according to their semantic similarity rather than common diagram composition styles or practice shared by people who are academically closely related. Let $D_{\text{author}}(A_x, A_y)$ be the co-author distance between a pair of authors A_x and A_y. In this work, we derive the co-author distance between a pair of authors by checking all the publication records in the open access portion of the PubMed corpus. $D_{\text{author}}(A_x, A_y) = 0$ if and only if A_x is the same person as A_y; $D_{\text{author}}(A_x, A_y) = 1$ if A_x and A_y at least co-author one article as captured in the corpus; for the general case, we used the classic Dijkstra's shortest distance algorithm to compute $D_{\text{author}}(A_x, A_y)$. Based on the notation of $D_{\text{author}}(A_x, A_y)$, we

Table 1. *P*-values for two-tailed, paired *t*-tests for search results of 16 query themes using our method and the five peer methods

Theme number	Ours-PM1	Ours-PM2	Ours-PM3	Ours-PM4	Ours-PM5
1	6.72E-6	9.77E-6	0.00418	9.84E-8	4.02E-7
2	4.69E-6	8.54E-8	0.42255	9.62E-8	0.00025
3	2.03E-5	7.47E-7	1.33E-6	1.52E-6	1.96E-7
4	0.02707	**0.0502**	**0.96529**	0.00092	3.14E-5
5	**0.52284**	**0.09693**	**0.06315**	**0.53414**	0.00039
6	7.35E-6	1.53E-9	2.42E-10	4.60E-8	3.03E-9
7	7.26E-10	**0.27657**	**0.21753**	0.0039	0.0074
8	6.69E-10	1.42E-9	1.97E-8	1.88E-8	6.28E-12
9	3.17E-7	2.68E-9	9.58E-8	3.48E-11	1.46E-12
10	0.00074	0.00174	4.15E-5	1.26E-5	6.17E-7
11	1.19E-6	8.51E-9	4.88E-7	2.04E-7	1.42E-8
12	1.46E-11	3.20E-11	4.16E-6	6.03E-8	5.82E-10
13	3.59E-8	3.83E-8	6.36E-7	3.28E-8	1.08E-8
14	2.08E-5	0.00708	0.00651	0.45278	2.98E-5
15	8.76E-8	1.14E-6	0.0017	6.16E-7	5.29E-7
16	5.05E-6	8.63E-8	4.20E-6	4.88E-7	8.42E-8

The first column lists the query theme number.

Table 2. *P*-values of paired *t*-tests with both one-tailed and two-tailed settings

Type	Ours-P1	Our-P2	Ours-P3	Ours-P4	Ours-P5
One-tailed	0.00475	0.00631	0.02423	0.00239	0.00117
Two-tailed	0.00949	0.01263	0.04845	0.00478	0.00234

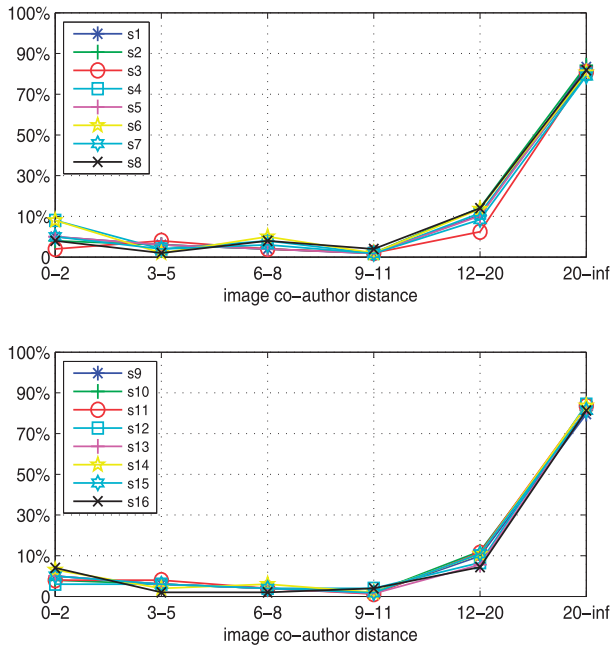


Fig. 4. Distributions of image co-author distances among top diagram search results by our method. We derive the image co-author distances between an input diagram and its top 20 search results returned by our method. For all query sessions in a query theme group, we compute the percentage distribution of these distances. The upper sub-figure shows such distributions for the query theme groups 1–8 (labelled as 's1'–'s8', respectively), and the bottom sub-figure shows results for the query theme groups 9–16 (labelled as 's9'–'s16', respectively). The horizontal axis indicates a specific image co-author distance range, where the last (rightmost) one corresponds to the distance value range of (20, ∞)

can further define the co-author distance between a pair of biomedical images $D_{\text{image}}(I_x, I_y)$. Two images' co-author distance is defined as the minimum co-author distance between one of the authors of one image's source document with one author of the other image's source document. Formally, let the source documents to which the two images I_x and I_y are embedded be Doc_x and Doc_y, respectively. Let the lists of authors of Doc_x and Doc_y be $\mathbf{A}_x = \{A_{x_1}, A_{x_2}, \cdots, A_{x_n}\}$ and $\mathbf{A}_y = \{A_{y_1}, A_{y_2}, \cdots, A_{y_m}\}$, respectively. We then define $D_{\text{image}}(I_x, I_y)$ to be $D_{\text{image}}(I_x, I_y) = \min_{A_x \in \mathbf{A}_x, A_y \in \mathbf{A}_y} D_{\text{author}}(A_x, A_y)$. For each of the 16 theme groups of query sessions conducted in our experiments, we collected all images that were ranked among the top 20 search results by our method in at least one of the query sessions. We then computed the distribution of image co-author distances between the input diagrams and their corresponding search result diagrams for all executed query sessions. In Figure 4, we report the percentage distribution of image co-author distances for each query theme group, respectively. From the reported results, we can clearly see that the new algorithm is able to retrieve semantically related diagrams regardless of whether these diagrams are composed by people that are closely related academically.

6 CONCLUSION AND DISCUSSIONS

We propose a novel context-based method for estimating diagram similarity. The method augments concepts and their relationships illustrated in a diagram by leveraging the contextual information provided by the full text of the diagram's source document. As a diagram usually highlights rather than explains its intended message, expanding the concisely encoded message by cross-referring to the diagram's source document can supply rich supplementary context for more accurately and comprehensively understanding the diagram's intended semantic meanings.

The comparative experiments demonstrate the superiority of our new method for semantically oriented diagram similarity estimation with respect to traditional image similarity metrics, which do not explore such context information. Our enhanced diagram similarity estimation can benefit many information retrieval tasks dealing with diagrams, e.g. improving user experiences with digital library systems for diagram searching and diagram browsing-based visual literature navigation.

The main challenge of estimating the similarity of diagrams embedded in the biomedical literature lies in the following two aspects: (i) unlike diagrams used in the software engineering and many other engineering disciplines that are typically composed of parametric objects using the Unified Modelling Language (UML) or other specialized languages or software packages and, hence, amenable to automatic computer processing, diagrams in the biomedical literature are typically released as bitmap images with no high-level descriptive representation. Therefore, detecting, extracting and automatically understanding entities and their mutual relationships graphically encoded in these bitmap images present a non-trivial technical challenge. To address this issue, we introduce a series of advanced image processing procedures in Section 3.2. For diagrams embedded in the biomedical literature, we witness the novel opportunity of observing and borrowing the context in a diagram's source document to acquire informative semantic clues for enhancing automatic diagram understanding and similarity estimation. Such an opportunity is unique for diagrams carried inside a peer-reviewed research publication because the document usually contains high-quality text that explains its embedded diagrams (such property does not always exist for diagrams from other sources, e.g. those included on casual webpages, as they usually do not have rich and quality explanation text). To leverage the aforementioned opportunity for enhancing automatic biomedical diagram understanding and similarity estimation, we thus introduced a novel and advanced diagram similarity estimation method by incorporating the rich semantic context information supplied in a diagram's source document (see Section 4).

Finally, in terms of the applicability of our method, even though the new method aims to process diagrams embedded in the biomedical literature, it can be applied for dealing with diagrams in the literature of other science or technology fields. Nevertheless, we notice that many science and technology fields do not use diagrams as intensively as by the broad biomedical discipline, which affects the potential of our method in processing diagrams in these fields. One fundamental limitation of our method is that it does not work with stand-alone diagrams that do not have accompanying source documents.

Funding: National Natural Science Foundation of China (NSFC) (60903132); National Key Basic Research and Development Program of China (973) (2013CB329505); NSFC-Guangdong Joint Fund (U1201252, U1135003 and U0935004);

National Natural Science Foundation of China (61232011); National Key Technology R&D Program (2011BAH27B01).

Conflict of Interest: none declared

REFERENCES

Barrios,J. *et al.* (2009) Text-based and content-based image retrieval on flickr: Demo. In: *Proceedings of the 2009 Second International Workshop on Similarity Search and Applications*. IEEE Computer Society, pp. 156–157.

Belkin,N. (2008) Some (what) grand challenges for information retrieval. *ACM SIGIR Forum*, **42**, 47–54.

Choi,J. *et al.* (2010) Automatic face annotation in personal photo collections using context-based unsupervised clustering and face information fusion. *IEEE Trans. Circuits Syst. Video Technol.*, **20**, 1292–1309.

Dijkman,R. *et al.* (2011) Similarity of business process models: metrics and evaluation. *Inf. Syst.*, **36**, 498–516.

Ehrig,M. *et al.* (2007) Measuring similarity between semantic business process models. In: *Proceedings of the Fourth Asia-Pacific Conference on Comceptual Modelling-Volume 67*. Australian Computer Society, Inc, pp. 71–80.

Fisher,M. and Hanrahan,P. (2010) Context-based search for 3d models. *ACM Trans. Graph.*, **29**, 182.

Hatcher,E. and Gospodnetic,O. (2004) *Lucene in Action*. Manning Publications, Greenwich, CT.

Järvelin,K. and Kekäläinen,J. (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, **20**, 422–446.

Korenblum,D. *et al.* (2011) Managing biomedical image metadata for search and retrieval of similar images. *J. Digit. Imaging*, **24**, 739–748.

Lawrence,S. (2000) Context in web search. *IEEE Data Eng. Bull.*, **23**, 25–32.

Li,C. and Hsu,C. (2008) Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation. *IEEE Trans. Multimed.*, **10**, 447–456.

Li,C. *et al.* (2008) On measuring process model similarity based on high-level change operations. In: *Proceedings of the 27th International Conference on Conceptual Modeling (ER '08)*. Springer-Verlag, Berlin, Heidelberg, pp. 248–264.

Li,Y. *et al.* (2006) Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowledge Data Eng.*, **18**, 1138–1150.

Liu,L. *et al.* (2010) Interleukin-8- 251 a/t gene polymorphism and gastric cancer susceptibility: a meta-analysis of epidemiological studies. *Cytokine*, **50**, 328–334.

Lopes,C. (2009) Context-based health information retrieval. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 845–845.

Luo,J. *et al.* (2009) Integration of context and content for multimedia management: an introduction to the special issue. *IEEE Trans. Multimed.*, **11**, 193–195.

Madhusudan,T. *et al.* (2004) A case-based reasoning framework for workflow model management. *Data Knowledge Eng.*, **50**, 87–115.

Meekers,D. and Rahaim,S. (2005) The importance of socio-economic context for social marketing models for improving reproductive health: evidence from 555 years of program experience. *BMC Public Health*, **5**, 10.

Melnik,S. *et al.* (2002) Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In: *Proceedings. 18th International Conference on Data Engineering, 2002*. IEEE, pp. 117–128.

Minor,M. *et al.* (2007) Representation and structure-based similarity assessment for agile workflows. In: *Case-Based Reasoning Research and Development Lecture Notes in Computer Science Volume 4626*. Springer-Verlag, Berlin Heidelberg, pp. 224–238.

Moskovitch,R. *et al.* (2007) A comparative evaluation of full-text, concept-based, and context-sensitive search. *J. Am. Med. Inform. Assoc.*, **14**, 164–174.

Nejati,S. *et al.* (2007) Matching and merging of statecharts specifications. In: *Proceedings of the 29th International Conference on Software Engineering*. IEEE Computer Society, pp. 54–64.

Ntoulas,A. *et al.* (2006) Detecting spam web pages through content analysis. In: *Proceedings of the 15th International Conference on World Wide Web*, ACM, pp. 83–92.

O'Hare,N. and Smeaton,A. (2009) Context-aware person identification in personal photo collections. *IEEE Trans. Multimed*, **11**, 220–228.

Qiao,L. and Feng,Y. (2012) Genetic variations of prostate stem cell antigen (PSCA) contribute to the risk of gastric cancer for eastern Asians: a meta-analysis based on 16792 individuals. *Gene*, **493**, 83–91.

Qin,K. *et al.* (2010) A unified approach based on hough transform for quick detection of circles and rectangles. *J. Image Graph*, **15**, 109–115.

Rodriguez-Esteban,R. and Iossifov,I. (2009) Figure mining for biomedical research. *Bioinformatics*, **25**, 2082–2084.

Segev,A. and Toch,E. (2009) Context-based matching and ranking of web services for composition. *IEEE Trans. Serv. Comput.*, **2**, 210–222.

Sinha,P. and Jain,R. (2008) Semantics in digital photos: a contextual analysis. In: *2008 IEEE International Conference on Semantic Computing*. IEEE, pp. 58–65.

Wendling,L. and Tabbone,S. (2003) Recognition of arrows in line drawings based on the aggregation of geometric criteria using the choquet integral. In: *2003. Proceedings. Seventh International Conference on Document Analysis and Recognition*, IEEE, pp. 299–303.

Wombacher,A. (2006) Evaluation of technical measures for workflow similarity based on a pilot study. In: *Proceedings of the 2006 Confederated International Conference On the Move to Meaningful Internet Systems: CoopIS, DOA, GADA, and ODBASE (ODBASE'06/OTM'06)* - Vol. I. Springer-Verlag, Berlin, Heidelberg, pp. 255–272.

Xu,S. and Krauthammer,M. (2010) A new pivoting and iterative text detection algorithm for biomedical images. *J. Biomed. Inform.*, **43**, 924–931.

Xu,S. *et al.* (2008) Yale image finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics*, **24**, 1968–1970.

Yang,L. *et al.* (2011) Object retrieval using visual query context. *IEEE Trans. Multimed.*, **13**, 1295–1307.

Yang,X. *et al.* (2010) Mobile image search with multimodal context-aware queries. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, pp. 25–32.