Influence Visualization of Scientific Paper Through Flow-based Citation Network Summarization

Yue Su
Institute of Software
Chinese Academy of Sciences
Email:suyue@ios.ac.cn

Sibai Sun
Institute of Physics
Chinese Academy of Sciences
Email:sunsibai14@mails.ucas.ac.cn

Lei Shi
Institute of Software
Chinese Academy of Sciences
Email:shil@ios.ac.cn

***Abstract*—This paper presents VEGAS - an online system that can illustrate the influence of one scientific paper on citation networks via the influence graph summarization and visualization. The system is built over an algorithm pipeline that maximizes the rate of influence flows in the final summarization. Both visualization and interaction designs are described with respect to a real usage scenario of the VEGAS system.**

I. Introduction

Citation networks are important in studying the relationship among scientific papers and discovering their topic evolution over time. For example, starting from one source paper in the citation network, we can obtain a graph containing all the papers reachable from the source paper by following the reversed citation links. These reversed citation links indicate the influence relationship between papers, such that if A cites B, then B influences A. The graph spanned from the source paper is called the maximal influence graph (or influence graph for short) that illustrates all the papers that the source paper can influence in a maximal range. While there has been quite a few numeric measures to evaluate the importance/value of one scientific paper (e.g., the total number of citations), the problem of how the scientific paper influences and advances the research field and topic is still an open question. Visualization of the influence graph is one possible solution to this question, as this kind of graph can provide an overview of the influence hierarchy from the source paper and is more consumable from the perspective of end users.

However, displaying the maximal influence graph in its entirety is challenging, mainly because the size of the graph can increase beyond the feasible boundary for a readable and real-time visualization. In our academic data set, the most influential paper can lead to an influence graph of more than 100,000 nodes, where an appropriate summarization is necessary. We define this as the Influence Graph Summarization (IGS) problem. On the IGS problem, traditional graph mining algorithms [1] can be applied, notably graph clustering and compression methods. These methods typically look for coherent regions in the graph by optimizing a pre-defined loss function to group the graph nodes with direct linkages together. We argue that although these works can help the users to some extent, they are not enough in the context of the influence graph summarization, where the goal is to reveal the influence flows between paper clusters, instead of keeping directly connected papers in the same cluster.

On the other hand, social graph simplification problems [2] in the scenario of information diffusion over social networks have been studied before. They extract the most important social paths based on information propagation logs to optimize applications such as the viral marketing. These works, though close to, are quite different from the IGS problem studied here. On citation networks, there is hardly an underlying social network (at least difficult to acquire or predict), over which the influence propagates. For example, a new researcher can cite seminal papers in his field without connecting to the authors in person. In this sense, the problem is more an "unsupervised" summarization problem.

In this demonstration paper, to solve the IGS problem, we present an online system called VEGAS (Visual influEnce GrAph Summarization), that generates flow-based, localized visual influEnce graph summarization over large-scale citation networks. The system is built on an algorithm pipeline which integrates several algorithms to extract the maximal influence graph, solve the IGS problem, and prune the resulting influence graph for the effective visualization. This is a challenging job, because the IGS problem exhibits both nonlinear and combinatorial natures, making it hard to solve in a close form. Over the existing algorithm paper in [5], this paper makes several additional contributions: 1) we describe in more details both the visualization and interaction design in the VEGAS system; 2) we conduct case studies which validate the usefulness of the proposed system; 3) We demonstrate the online VEGAS system which can visualize the IGS result according to user queries.

The rest of the paper is organized as follows: Section II overviews the design of VEGAS, Section III describes the visualization and user interface of the VEGAS system. Finally, Section IV details the summarization algorithms.

II. System Overview

A. Example Usage

We demonstrate how VEGAS system helps to analyze the influence of one scientific paper through the visual summarization. Consider a user interested in the paper *Manifold-ranking based image retrieval* published in ACM Mulitmedia 2004. She started by issuing a search in the VEGAS system. She was then provided with the visual summarization in Figure 1. Focus on Figure 1 (c), the summarization graph shows several paper clusters, connected by influence links. Each influence link is a group of reversed citation links. She discovered that the source paper directly influenced three clusters. As she was more

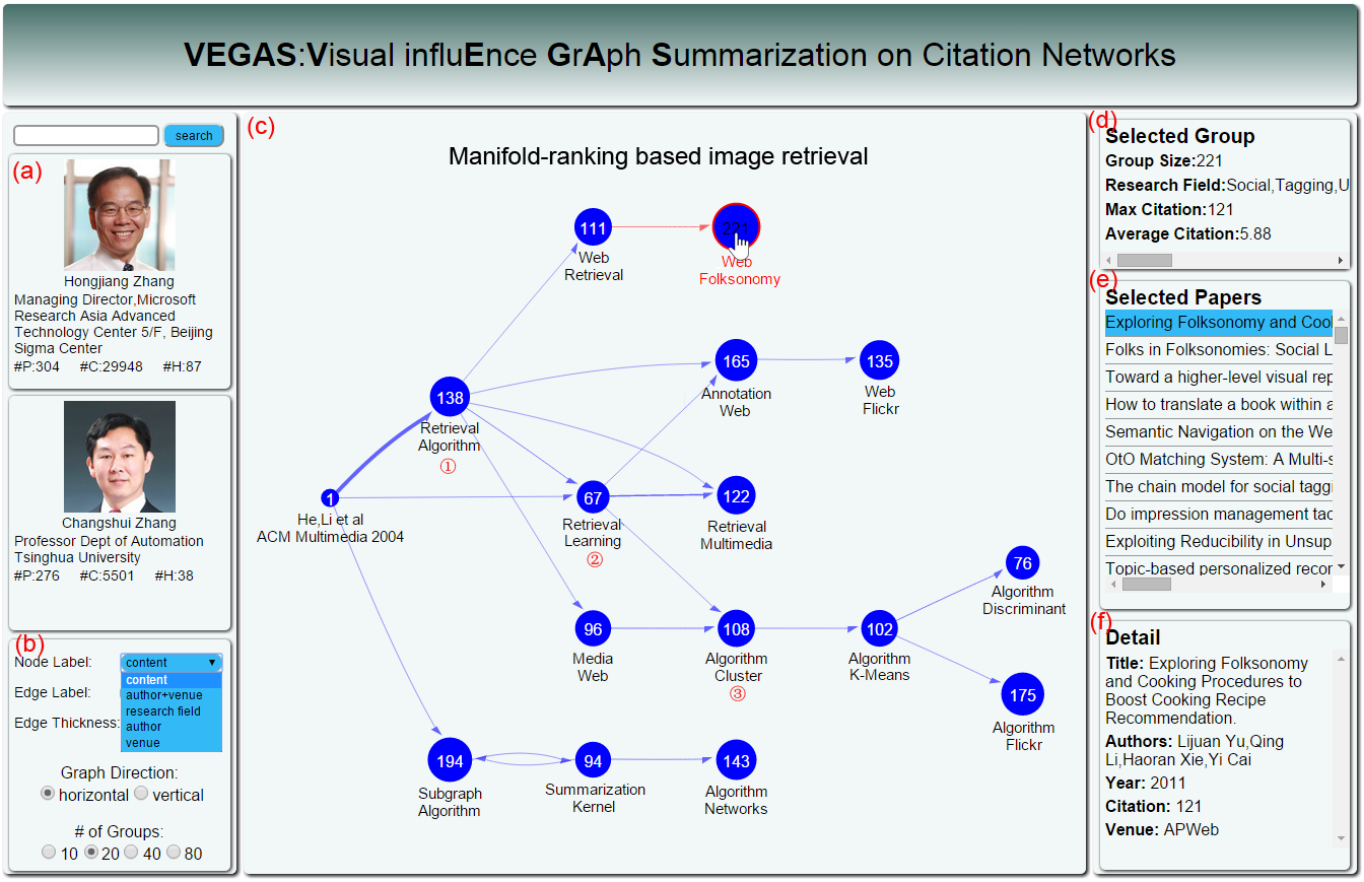

Fig. 1: Influence Graph Summarization of [Manifold-ranking based image retrieval]: (a) the author information, (b) the control panel, (c) the influence graph summarization, (d) the information of the selected cluster, (e) the papers list of the select cluster, (f) the detail information of the selected paper, ①∼③ represent the corresponding clusters. The link to our VEGAS system is http://vis.ios.ac.cn/usr/suyue/Influence%20graph/index.html.

interested in information retrieval algorithms, she decided to view the detailed information of cluster ① which was corresponding to her interest. By clicking the cluster ①, she could find the paper list of cluster ① in Figure 1 (e). In that list, she found a paper *Cross-media manifold learning for image retrieval and annotation* published in Multimedia Information Retrieval 2008. Then she decided to know more about learning algorithms in retrieval field. As the cluster ② that influenced by cluster ① was exactly about the learning algorithms, she clicked on the cluster ② to update the information of panel (d), (e), (f). In paper list of cluster ②, she found a paper *Semi-supervised distance metric learning for collaborative image retrieval and clustering* published in TOMCCAP 2010 which was about the application of learning algorithm in clustering. So she decided to view the paper list of cluster ③ which was mainly about the clustering algorithms to figure out how did *Semi-supervised distance metric learning for collaborative image retrieval and clustering* influence others. Finally, she found a paper *Semi-supervised fuzzy clustering with metric learning and entropy regularization* published in Inf. Sci. 2012 which was directly influenced by *Semi-supervised distance metric learning for collaborative image retrieval and clustering*.

In this case, we learn that how does the influence of a paper diffuse into other fields in a time order.

B. Pipeline

We describe the algorithm pipeline in VEGAS to summarize the influence graph as illustrated in Figure 2. Initially, we import the raw data which includes the citation information and papers profile into database by performing some pre-processings. The pre-processings include data cleaning and the generating of the citation network. From the citation network, we define the maximal influence graph as a graph which is computed by a breadth-first or depth-first search starting from the source paper and store the corresponding paper information as paper profile. We then perform a node summarization algorithm to compute the influence graph summarization from the maximal influence graph. After that, we perform some post-processings to accomplish the link-pruning and insert the paper profile into the influence graph summarization. Generally, the influence graph summarization has a hierarchical structure. We employ GraphViz, an open-source package which initiated by AT&T Labs Research for drawing graphs specified in DOT language scripts, to accomplish hierarchical layout.

C. Data and Analysis

1) Data source: AMiner [3] is a researcher centralized academic social network analysis and mining system. It can automatically mine the relationship between cooperators and it mainly provides research social network, social influence analyzing, etc. CiteSeerX [4] is a search engine and digital library for scientific and academic papers and it can actively crawl and harvest academic and scientific documents on the public web. Table I compares the data scale of two data sources. We learn that the count of papers and authors of CiteSeerX are roughly twice over AMiner and the citation

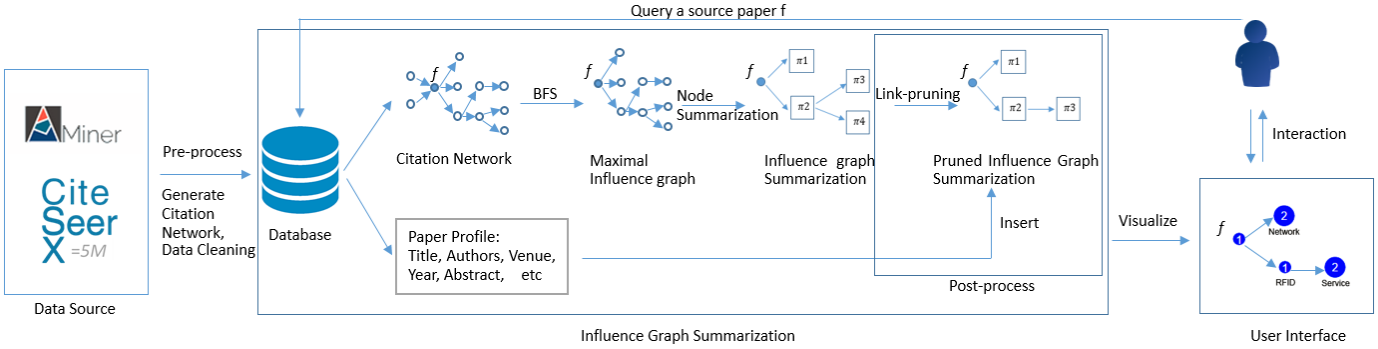

Fig. 2: the Algorithm Pipeline of VEGAS which mainly includes the pre-processing of raw data, the generating of maximal influence graph, the node summarization of maximal influence graph, the post-processing of influence graph summarization, the visualization of influence graph summarization and the interaction between user and VEGAS system

Data source	CiteSeerX	AMiner
#papers	4045694	2146343
#authors	12674545	5834904
#citations	63001253	4191667
#max citation of one paper	8946	4660

TABLE I: CiteSeerX and AMiner

number of CiteSeerX is roughly fifteenfold times than AMiner, which indicates that the CiteSeerX has a larger citation graph.

2) Influence Graph Summarization: In order to make sense of an individual's influence in the context of the citation network and summarize the underlying citation graph to represent this influence, we need to solve the Influence Graph Summarization problem. We perform some post-processings to compute the influence graph summarization over the maximal influence graph G. We can generate several matrices from the graph: the topology similarity matrix, and the generalized similarity matrix. The core of our algorithm is the decomposition of the similarity matrices by assign a constant k to generate k node clusters for the summarization. We carefully design the topology similarity matrix to ensure that the graph summarization can approximate maximal the flow rate. We also perform a link pruning algorithm by assign a constant l to screen out top l flows by the ranking-based filtering algorithm. We will introduce the detail algorithms in Section IV.

D. Interface

Figure 1 shows the main page of VEGAS for one source paper. The panel (a) presents the major author's demographics, including affiliation, #papers, #citations and H-index of the focused paper. The panel (b) is the control panel that can change the label, edge thickness, layout direction and #clusters of the graph in panel (c). The panel (c) presents the influence graph summarization of *Manifold-ranking based image retrieval* and the detail functions and interactions will be introduced in Section III. The panel (d) presents the information of the selected cluster in panel (c) such as group size, research field, max citation and average citation. The panel (e) presents the papers list of the selected cluster. The panel (f) presents the detailed information of the selected paper in the panel (e).

III. Influence Graph Visualization

A. Design

In Figure 1, the panel (c) presents the influence graph summarization of *Manifold-ranking based image retrieval*. The circle in the leftmost represents the *Manifold-ranking* paper we have selected and the labels of that circle are the authors and the venue of *Manifold-ranking*. The right part of the graph represents the influence relationships starting from the *Manifold-ranking* and the number on the cluster means papers number in one cluster as the size of cluster is changing with that number. The default label of these clusters is a text summarization of the main content of papers in one node cluster, including their title and abstract. The detailed information of one node cluster can be found in panel (d), (e), (f). The edge from cluster A to cluster B means papers in A have influenced papers in B and the edge thickness represents the strength of the influence between two clusters. We can change the focus paper by searching some keywords in the search box in panel (a). We provide a demo video to show our system that can be downloaded from ftp://vis.ios.ac.cn/anon_upload/.

We perform several methods to guarantee the clarity of the graph while accomplishing the graph layout. First, we design a plotting scale that can map from an input domain to an output range and the input domain is the set of all the size of clusters and the output range is a set with certain upper bound and lower bound. By using that scale, we can control the size of cluster in that certain output range to prevent the cluster being too huge. Second, we will decrease the font size and amount of the cluster label as the growth of the clusters' count.

B. Interaction

We introduce the interactions in our system. Initially, we can highlight a node cluster and it's correlation by hovering mouse on one node cluster and we can also modify the position of each node by dragging it. We also have some advanced interactions which will be introduced in the below two parts.

1) Control Panel: We provide options for user to change the node label, edge label, edge thickness, layout direction, the number of clusters on the graph by using the option boxes and the ratio buttons in panel (b).

The node label option includes content (the text summarization of one node cluster), author&venue (the author and venue of the most-cited paper in one node cluster), research field(the main research field in one node cluster), author(the most famous author in one node cluster), venue (the most famous venue in one node cluster). By changing the node label, we can know the basic information about one node cluster. The edge label option includes empty (to keep the graph clean), #citation (the number of citations between two clusters), flow rate (the strength of the influence). By changing the edge label, we can know the value of citation's count

or flow rate between two clusters. The edge thickness option includes uniform, #citation, flow rate. By changing the edge thickness, we can also know the citation number or flow rate in an intuitive way. Except for the objects in the graph, we can also change the layout of the graph. The default layout direction is horizontal, and we can change it into vertical by clicking the "vertical" ratio button. The default number of clusters is 10, and we can change it into 20, 40 and 80 by clicking the corresponding ratio button.

2) Information Panel: We can select one node cluster in panel (c) by clicking on it, and the panel (d), (e), (f) will present the detail information of the selected cluster. The panel (d) presents the Group Size(the number of papers in this cluster), Research Field, Max Citation number and the Average Citation number of that cluster. As the number on the cluster represents the papers number in that cluster, the panel (e) presents these papers' titles as a list which is ordered by the citation number from the largest to the smallest. We can drag the scroll bar in the right to view more papers and drag the scroll bar in the bottom to view the full title of one paper. For more detail information of one paper, we can click on the title and view the detail information in panel (f), which presents the title, authors, year, citation, venue of the selected paper.

IV. Algorithm

A. Influence Graph Summarization Problem

The Influence Graph Summarization (IGS) problem is defined as finding a graph summarization S of the maximal influence graph G, with k clusters and l flows, to maximize a objective function equaling the sum of flow rates [5]:

$$max \sum_{s=1}^{l} r(\xi_s) \tag{1}$$

B. Node Summarization for IGS problem

We apply the node summarization algorithm in [5]. First we compute the topology similarity matrix M^G by the common neighbor heuristic:

$$M^G = \frac{AA^T + A^TA}{2} \tag{2}$$

where A is the adjacency matrix of the maximal influence graph G. In the context of the citation network, the entry in the M^G for the similarity of two nodes indicate their number of commonly cited and commonly citing nodes(i.e., neighboring nodes in the citation graph). Therefore, we name the main algorithm for the node summarization as the bidirectional CommonNeighbor. Meanwhile, two variants of the algorithm are supported, the forward CommonNeighbor algorithm by $M^G = AA^T$ which only considers the outgoing edges of each node, and the backward CommonNeighbor algorithm by $M^G = A^TA$ which only considers the incoming edges of each node. The bidirectional CommonNeighbor is also referred to as the forward+backword CommonNeighbor.

In the second stage, we propose a matrix decomposition based solution to generate k node clusters from the similarity matrix M^G. The decomposition employs a Symmetric version of the Nonnegative Matrix Factorization(SymNMF [6]) which optimizes:

$$\min_{H \geq 0} \|M^G - HH^T\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix. $H = \{h_{ij}\}$ is a n by k matrix indicating the cluster membership assignment of nodes in $G : v_i$ will be clustered into π_c if h_{ic} is the largest entry in the ith row of H.

C. Link-pruning Algorithm

The influence graph summarization computed by SymNMF needs further processing to select top l flows for the final summarization S. In Algorithm 1, we select top l flows by ranking the normalized flow rate. The other flows are filtered out. While, in order to keep the influence graph summarization connected, we need to do link recovery by adding back the most dense flow going to each node cluster.

Algorithm 1 Link Pruning Algorithm

Input: Initial summarization $S_0 = \{V, E\}$,# of flows l, $V = \{\pi_i\}_{i=1}^{k}, E = \{\xi_s\}_{s=1}^{k^2}$, flow rate $r(\xi_s)$
Output: Final Summarization S

```
S ← S_0;
sort E by r(E) in decreasing order;
for s ← l + 1 to k^2 do
    remove E(s) from S;
end for
for i ← tok do
    E_i ← of E having π_i as destination;
    sort E_i by r(E_i) in decreasing order;
    if E_i(0) ∉ S then
        add E_i(0) to S;
    end if
end for
```

V. Conclusion

In this paper , we proposed VEGAS - an online system that can compute the influence graph summarization for a paper on citation networks and display the corresponding visualization via web. Such a system is particularly useful for users who are interested in the influence relationship of citation networks. The major contribution of our work lies in the integration of multiple algorithms together into an algorithm pipeline and the online demonstration of this work. For example, users can learn to which scope one scientific paper has influenced the research field, which is hardly traceable in tradition academic search websites such as AMiner and CiteSeerX.

References

[1] J. Han and M. Kamber, *Data mining*. Amsterdam: Elsevier, 2006.

[2] F. Zhou, S. Mahler and H. Toivonen, *Simplification of Networks by Edge Pruning*, in Bisociative Knowledge Discovery, pp. 179-198, 2012.

[3] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang and Z. Su, *ArnetMiner: Extraction and Mining of Academic Social Networks*. in KDD 08, 2008.

[4] Citeseerx.ist.psu.edu, 'CiteSeerX', 2015. [Online]. http://citeseerx.ist.psu.edu/index. [Accessed: 13- Jul- 2015].

[5] Lei Shi, Hanghang Tong, Jie Tang, Chuang Lin, *VEGAS: Visual influEnce GrAph Summarization on Citation Networks*. in IEEE Transactions on Knowledge and Data Engineering, 2015.

[6] Ding, X. He, and H. D. Simon, *On the equivalence of nonnegative matrix factorization and spectral clustering*. in SDM, 2005.