A Collaborative Support for Recommending References in Papers

Orlando Fonseca Guilarte [1]
Department of Mathematics PUC-Rio
Rio de Janeiro, Brazil
Email: ofonsek0702@mat.puc-rio.br

Simone Diniz Junqueira Barbosa
Department of Informatics PUC-Rio
Rio de Janeiro, Brazil
Email: simone@inf.puc-rio.br

Sinesio Pesco
Department of Mathematics PUC-Rio
Rio de Janeiro, Brazil
Email: sinesio@puc-rio.br

***Abstract*—Understanding citations to scientific publications is a task of vital importance in the academic world. This task can be supported by appropriate data structures and visualization mechanisms. One challenge is the amount of existing relationships and the difficulty of determining which of the references of a document are considered the most potentially relevant to it. In this paper, we propose a visual approach based on graphs to recommend important references. Also, we propose a procedure to build and update the citation graph in a collaborative way.**

I. Introduction

Interactive data visualizations help to clearly and efficiently interpret the underlying information, so that, through exploration, users can acquire knowledge that will assist them in making decisions. Graphs are used in numerous applications within the field of information visualization, enabling visual representations of the relationships between nodes in network data.

A commonly used visualization model for academic institutions and researchers involves mapping scientific documents onto nodes of a graph, and citations or bibliographic references onto relationships between those nodes [1]–[3]. This model is useful because the set of citations is an indicator of the quality of scientific articles, which can be analyzed in the graph. The authors of scientific documents are usually careful in selecting the works that will be listed. This is crucial because in only a small sample they must provide which articles are potentially relevant to their research to cite them in the development of the document, as seen in [4]. The list of references is an index of the selectivity where authors have filtered the trivial information from the relevant one, but in several occasions the amount of irrelevant references provided by authors is excessive.

Representing and viewing papers and their reference relationships as graphs can be confusing, because there are usually many citations to each node of the graph. Therefore, it can be arduous to find a sequence of closely related works, i.e., works with relevant citation relationships between them. The problem is how to rank and visualize the most relevant articles that reflect the evolution of the different branches of studies.

A classification of references in importance classes would be an alternative solution to the ranking problem, as seen in [5]. Some strategies to classify important references include incidental citation using supervised classification [6], predicting academic influence with machine learning [7], and using a topic model approach [8]. Although there have been proposed algorithms to analyze the classification or ranking of references, we consider that the opinion of the author or of an expert is essential in this decision. This is because the notion of ”relevance” according to the criterion of ”references with more academic influence that the others” is subjective. In addition, getting a picture of the relevant material in a specific topic can be of help to novice researchers who are newcomers in a field or even to experienced researchers.

For the reasons mentioned above, we have developed a collaborative web application that presents a visualization of the relationships between scientific publications. In this way, it is possible to obtain an overview of the field on a topic through a visual recommendation of the main works. This proposal facilitates users searching for potentially relevant articles with respect to a query paper and helps to understand the flow of research topics in the scientific literature.

The remainder of this work is structured as follows. In Section II, we briefly discuss related works. We describe in Section III some notations and the collaborative characteristic of the system. In Section IV we discuss the visual structure. We present our experimental results in Section V. Finally, we conclude and give an outlook to future work in Section VI.

II. Related Work

Several visualizations of scientific documents have been proposed to help researchers explore citation graphs. PyScholarGraph [9] shows a framework for indexing, searching, and viewing references, based on data recovered from the CiteSeerX repository and indexed in a graph-oriented database. Waumans and Bersini [1] adopt an evolutionary perspective, building the graph in the form of a genealogical tree. In PyGraphviz [10], a graph was created to represent the evolutionary process of the deep learning field in the last 25 years. Wei *et al.* [2] propose an interactive visualization method that uses citation paths [11]. Berger *et al.* [12] treat the documents as a collection of special words. Ginde [13] showed how the data can be processed previously and sent to a database with information on scientific journals, authors and research documents [14].

[1]This paper is associated to the PhD thesis of the author

Several papers have presented a literature review and proposed algorithms to analyze the classification or ranking of references. Sibaroni *et al.* [15] proposed to build relationships between documents through analysis of co-citations, reference lists, and bibliographic coupling. Singh *et al.* [16] propose a modified version of the PageRank algorithm [17] to rank research papers. Nallapati *et al.* [8] presented two novel topic models to address the problem of joint modeling of link and text. Zhu *et al.* [7] focused on automatically identifying the subset of key references that have a great academic influence on the citing paper, and Valenzuela [6] described a classification approach for identifying important and incidental citations. In [18], Zhou *et al.* implemented a visualization framework for visually ranking the academic influence of papers.

A hierarchical structure can be useful when the number of nodes and relations in the graph increases, allowing the resulting visualization to be more readable when each node is positioned at its corresponding level, thus facilitating user understanding. However, Ginde [13] did not present a hierarchical structure. In [1], although such a structure is considered, it is not easy to perform an analysis per period of time. In [9], [10] it is difficult to visually obtain a sequence of closely related works due to the number of relationships presented. In [2], [6]–[8], [16], the opinion or change of opinion of the specialist is not considered in the classification of key references. Finally, in [18], the user cannot easily filter the citation links of a certain paper sorted by relevance, to reduce visual clutter and find relevant or influential citation paths.

III. Terminology and Overview of the Proposed Approach

A. Graph–based Model

Our model is formulated from a directed, acyclic and weighted graph called Citation Graph and represented by a triplex $G = \langle V, E, \omega \rangle$, where V is the node set that indicates documents, E is the edge set that indicate relations of citations and ω is a function that assigns to each edge a positive weight. We consider this particular graph in the visualization of important references.

We denote the set of authors by $A = \{a_1, \ldots, a_m\}$, where m is the number of authors in the system, and the set of papers by $P = \{p_1, \ldots, p_n\}$, where n is the number of papers.

Initially, the references of each paper are classified in classes of relevance by an automatic method. Thus, a preliminary overview of the state of the art in the subject can be obtained. When the specialist in the area wants to improve this classification, the system gives the option of inserting, in the form of votes, a particular ranking of references of this specialist. Then, we associate to each edge a weight from 0 to 5. The overview of the state of the art will be improved as soon as specialists give their votes. In this way, human effort contributes to obtain more adequate and reliable results.

B. Collaborative Approach

Given a target paper $q \in P$, and its respective reference list, we show in the next line how we define the function ω, which corresponds to determining which reference is the most representative for paper q, according to our proposal.

Our approach was inspired by web recommendation systems, for example, Amazon (see [19]). Amazon allows users to submit ratings or votes for each item in the system, then automatically provides a ranking in the final recommendation. All these systems have a fundamental characteristic that they are open to the user, that is, any user can give his/her opinion or vote. In an academic environment, this option is no longer so convenient. In our proposal, people with more knowledge of the subject are those whose votes are most valued. Thus, we propose a collaborative system where the users can vote at any edge in the system and their votes and expertise determine the weight of each edge in the graph. To define the values of expertise, we consider the impact of the publications of the authors on the subject as follows.

The global expertise of author a_j is a non-negative function defined as follows,

$$Exp(j) = \begin{cases} \dfrac{c_j + 1}{c_l + 1}, & n_j \neq 0 \\ 0, & n_j = 0, \end{cases} \tag{1}$$

where c_j is the number of citations to author a_j's papers, c_l is the number of citations received by the most cited author in the graph, denoted by a_l, $l \in \{1, 2, \ldots, m\}$ and n_j is the number of publications by a_j. This function describes to what extent the author has authority in the subject.

We use a weighted arithmetic mean to calculate the general rating. We consider that, in the initial state of the system, the reference edges still have no associated voting value, that is, for all $e \in E$; $\omega(e) = 0$. The votes will then be considered. Each author may give a vote. We denote by $v_j(q, p_k)$ the vote of the author a_j in the edge (q, p_k). We also assume that the possible values for the user votes will be in a five-point scale, where 5 would indicate the most important or influential reference for q, and 1 would indicate the least important with respect to q.

The first vote of an author $a_j \in A$ in a specific edge modifies the initial zero weight value of this edge, as follows: $\omega(q, p_k) = \dfrac{v_j(q, p_k) \times Exp(j)}{Exp(j)}$, where $v_j(q, p_k)$ is the author's vote for the edge (q, p_k), and the expertise $Exp(j)$ of a_j is considered strictly positive. In the case that $Exp(j) = 0$, regardless of the author's vote, the weight of the edge is not modified.

Now, suppose that the graph has evolved with respect to the edge (q, p_k), that is, the edge has received one or more votes. Then, the weight value associated with the edge (q, p_k) is modified considering the votes and expertise of the m experts who voted on this edge.

To conceive a collaborative model, one of the characteristics that we determine as fundamental is the flexibility of the system in the sense of accepting changes in the data. For

example, users must be able to modify their vote at any time, which causes a change in the weight functions. In this way, if an author votes two or more times in the same edge, only the last vote is considered in the weight of this edge.

In this process, it is also natural that the expertise can change. When new papers by the authors in the graph are added to the citation graph and their citation relationships are included, the expertise of these authors may change.

C. Constructing the Citation Graph

In Section III-A we defined a model with a set of nodes P and edges. To determine the papers of this set P we could search the Internet or in academic databases for papers on a specific topic published in journals and conferences. In this way, it is possible to obtain a large part of the scientific production published over a period of years. Building a graph with all this information is a complex task given the large volume of information to be considered. To prevent irrelevant papers to be included in the graph, we chose to start with a set of papers mentioned in a survey of a specific subject.

Constructing a Graph on the subject T

Let S represent a survey on the subject T.

1) Consider the node set $P = \{p_1, \ldots, p_n\}$, with p_i, $i \in \{1, \ldots, n\}$ such that p_i addresses the subject T and S references paper p_i.
2) Consider the edge set E.
3) Establish $\omega(e) = 0$ for each e in E.
4) Build the *Citation Graph* with the triple $C = \langle P, E, 0\rangle$.

In step 4, the last value of the triple with zero value corresponds to the value of the weight function ω. With this procedure we obtain a *Citation Graph*, which can be updated from several modifications such as adding nodes, edges and establishing a weight value for the edges.

Modifications or updates in the graph can be made in our system through a visual interface. This interface allows visualizing the Citation Graph, so that the user can easily extract knowledge and interact with the system, for example, to modify the value of the weight of an edge.

IV. Proposed Visualization Support

For the information visualization, we establish a geometric configuration in the 2-Dimension space, where the papers and their citation relationships are represented in the form of a graph. A node of the graph, which represents a paper, is initially associated with a cyan circle if its out-degree is greater than zero, and with a cyan triangle down if its out-degree is equal to zero. The reason for this change of shape is that in our system we consider it important to distinguish (with the triangle) the papers that may represent the beginning of a research line. The size of a node indicates the number of its incident edges, which represents the number of times that the corresponding paper is cited. In this way, the papers with the most impact on a branch of study are visually identified with larger shapes. The papers have associated to them categorical attributes, such as title, keywords, abstract, Digital Object Identifier (DOI), authors, and the publication year, which allows knowing how recent the paper is and to establish its location in the coordinate space. Each edge, referring to a citation relationship, is represented as a directed light orange line, from a citing paper to a cited paper. Thus, the visualization presented in this work is based on three characteristics:

Citation Graph Visualization. The data structure to represent the data is a *Citation Graph*.

Hierarchical Structure by Publication Year. The graph is organized in levels, defined according to the publication year of the papers. To reduce clutter in the graph, we define levels by publication year, so that the oldest articles will be located at the top of the graph and the most recent ones at the bottom. The vertical position of the nodes reflects their publication year, see Figure 1, so that those published in a specific year will be on the same level.

Ranking of references for each paper. From each node, a ranking of its references is established. In this way, through a filter of edges with a slider control, only the k most influential references are represented in the graph.

Figure 2 illustrates a part of this graph where five of the most important references for each paper are displayed, whilst Figure 3 presents only the two most relevant references, after applying a filter to the edges.

The graphic representation proposed in this work allows user interaction and exploration. It is possible to generate other views by zooming and dragging the nodes. When a node is selected, that is, when the user clicks on a node, the information pane shows some essential information about the paper. The selected node is highlighted in purple and with a thicker border. For more clarity, the papers cited in the selected publication (*aka* Reference nodes) are highlighted in forest-green and the papers that cite this publication (*aka* Citation nodes) are highlighted in grayish-green. In addition, a second network is created maintaining the hierarchical structure to better analyze and visualize only the references of a selected node. In this structure, the node at the lowest level corresponds to the node selected in the original graph.

The users can also view the general weight of the edges, which results from the collaborative process explained in Section III-B. In the second network, they can view the most recent vote given by them to each reference and, by selecting an edge, the number of users who voted on it.

By highlighting in red the edge of greater weight for each node in the graph we simplify the visual recommendation and exploration of the data. In case the user wants another recommendation, it is enough to observe the ranking of references for each node to obtain the second most important reference, or any that the user considers important according to his/her research interest. References that are relevant to the publication will be identified with high weight values and can be filtered by the user to facilitate the visual recommendation task.

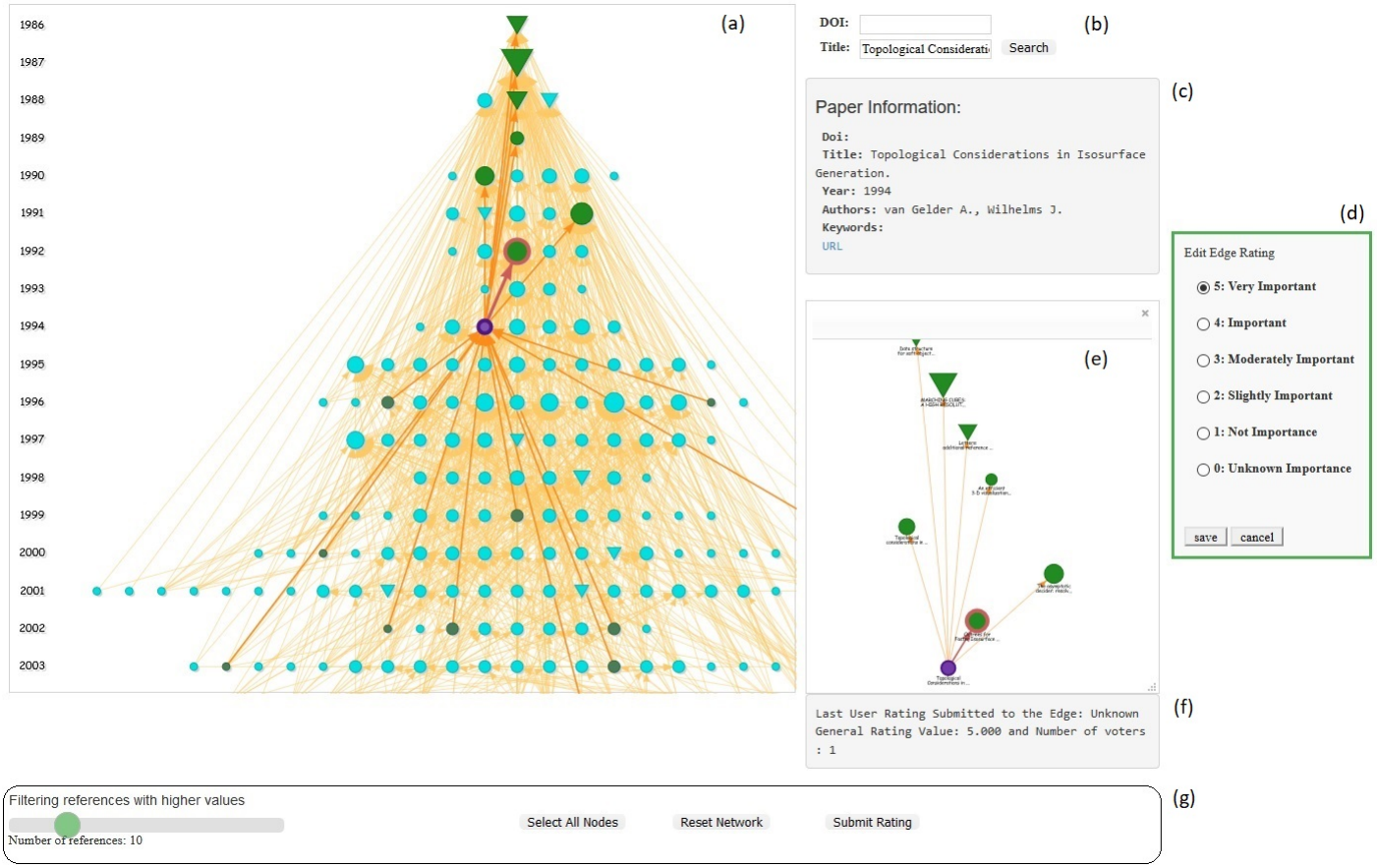

Fig. 1. The Visualization Interface

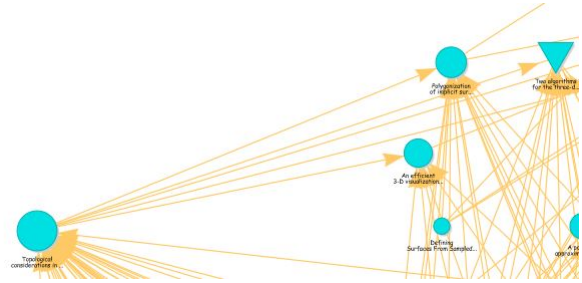

Fig. 2. Top five most relevant references

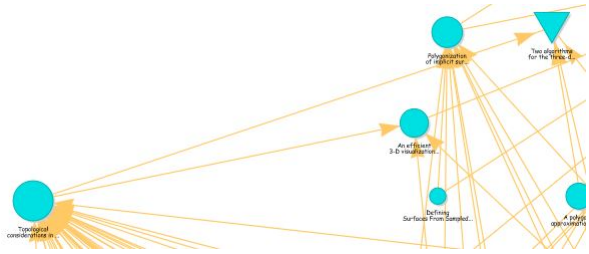

Fig. 3. Top two most relevant references

V. Discussion

A. Data set

The data collection we chose contains a particular selection of important articles that address the issue of "marching cubes", based on a survey paper published in 2006, "A survey of the marching cubes algorithm", by Timothy S. Newman and Hong Yi [20]. The data is in CSV format: each row describes a scientific document and each column its metadata: identifier (id), title, Digital Object Identifier (DOI), authors, publication year (year), abstract, keywords, uniform resource locator (url), and the reference by the identifier in the data set (ref-id).

B. System Implementation

We developed a process to collect the papers that make up the data set and their corresponding information that will be useful for our system. We implemented this module in Python [21] because it has excellent tools for extracting information from PDF documents. First we obtain the survey in PDF format from a specific URL. Then we extract the text using the *pdfminer* library, we identify the references in the survey and assign an identification (id) to each one. Finally, we search in Google for all these references and identify the citation relationship between them (ref-id). In this way, for each paper reference, we search in the data set for the pair (reference title,

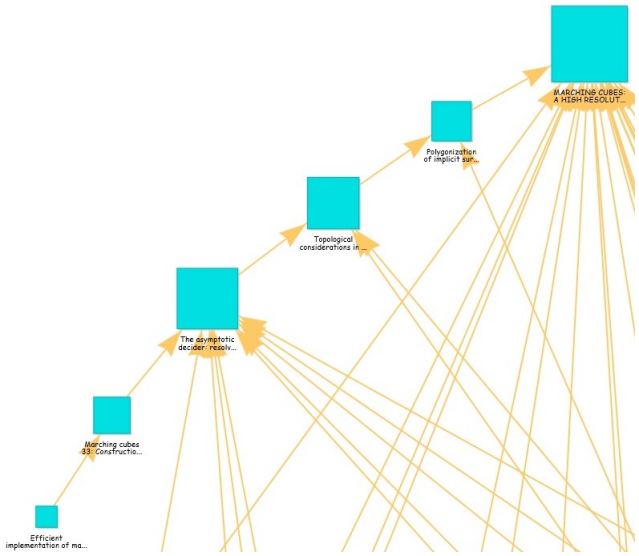

Fig. 4. Branch of Study "Ambiguities and Holes"

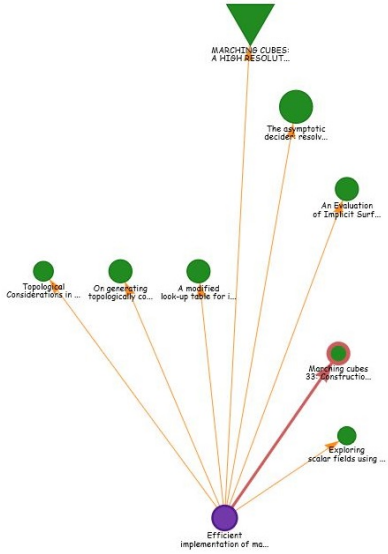

Fig. 5. Reference top one in the ranking highlighted with red color

year) to obtain the identification of this reference (id). To build the citation network and to store the information collected, we chose Neo4j [14] as a suitable NoSQL database, for its graph manipulation and query support.

We chose the Django framework [22] for Python, because Python has very good libraries to interact with Neo4j, like py2neo, and Django follows a Model-View-Template (MVT) architecture. For the interactive visualization, we used the JavaScript library vis.js [23]. Finally, we implemented in Python the proposed methods to rank references, both automatically and as a result of expert collaboration.

C. System Visualization Interface

We showed in Figure 1 our visualization interface. This interface is composed of seven main views. In the following, we explain these views, identified with the letters from *a* to *g*.

(a) Citation Network View. Shows the graph of citations, where the nodes are positioned at levels defined by years.

(b) Search View. Allows searching for specific papers.

(c) Paper Information View. Shows the main information of the selected paper.

(d) Edit View. Allows the user to establish a rating for each edge according to their expertise.

(e) Simplified View of the Selected Node. Shows the selected node and its references, and positions them hierarchically.

(f) Edge Information View. Shows the main information of the selected edge.

(g) Control Panel. Allows the user to control the system, for example, by filtering edges with greater weights.

D. Usage Scenarios

Our visualization produces a picture of the state of the art of the subject investigated in the papers of the graph. The most influential reference of each paper is obtained when filtering by the edge of greater weight. The oldest paper at the top of the original graph loses relevance, in this case by not having any relevant citation to it, so it is probably possible to omit its study without involving a gap in knowledge about "marching cubes". Also, it is possible to capture the different study lines and the sequence of citations that best reflect the evolution of a certain research line.

For example, Figure 4 shows the branch formed by six papers. They reflect the evolution of a specific research line, which in this case could be "Ambiguities and Holes". This example simplifies the display of a citation network by selecting the most relevant references to each paper. This simplification helps because when visualizing the entire network of an area, the resulting graph may be huge. By filtering relevant references, cluttered designs are avoided. For the construction of the path in Figure 4, only the most relevant or influential edge of each node was chosen, that is, the top one in the ranking of references established for each paper. In the case of the paper "Efficient implementation of Marching Cubes' cases with topological guarantees", whose references are shown in Figure 5, the top one in the ranking of references, the paper "Marching Cubes 33: Construction of Topologically Correct Isosurfaces" (highlighted in red) is not the most recent one. In this case, three specialists in the subject have collaboratively established a greater weight for the highlighted reference.

For the papers in Figure 5, we first apply the LSA method[1]. In this method, the text is first represented as a matrix of terms by documents and subjected to Singular-Value Decomposition (SVD) for geometrically representing the documents with reduced dimensionality. Next, the similarity of the paper "Efficient implementation of Marching Cubes' cases with topological guarantees" (2003) and its references are calculated by the cosine similarity. Thus, it is possible to obtain a

[1]Considering abstract, title, and keyword as the corpus for each document

ranking of references automatically using the LSA method. In this case, the reference with id 3 is the top one in the ranking of references for the paper "Efficient implementation of Marching Cubes' cases with topological guarantees" (id 9), instead of the paper with id 7 (edge highlighted in red), which is the most influential publication determined by several experts with our system. Hence the importance of our collaborative proposal for ranking references.

E. User Study

To evaluate the proposed visualization we conducted an empirical study in an academic environment. Each participant was given a general description of the system and its characteristics. The interviewer then asked the user to interact with the system to perform a series of tasks related to identifying important papers and capturing paths of relevant or influential references. The interviewer recorded the main problems users faced when performing these tasks. After performing the tasks, each participant answered a questionnaire. In this questionnaire we used a five-point Likert scale, from 1 (Completely Disagree) to 5 (Completely Agree). The 15 researchers who participated in the study are from the field of pure and applied mathematics at the same university, comprising 3 D.Sc., 4 M.Sc., and 8 B.Sc. The questionnaire for R1 comprised analysis tasks, represented by the letter T, and statements for the Likert scale, represented by the letter S.

R1-T1. Determine the most important papers in the graph published in the year 2003.

R1-S1. I can easily determine visually any important paper published in a specific year.

R1-T2. Identify the references of the second paper most cited in the graph.

R1-S2. It would be easy to identify the references of the second paper most cited in the graph.

R1-T3. Determine whether the second most cited paper in the graph cites other papers with high number of citations published in the two previous years.

R1-S3. It would be easy to identify the most cited references of a specific paper and their respective years of publication.

R1-T4. Determine visually the top one in the ranking of references for any paper in the data and repeat the process with this reference.

R1-S4. I can easily determine visually a sequence of publications in which each paper corresponds to the most influential or relevant citation.

Figure 6 shows the results of the questionnaire. The average number of participants who answered Completely Agree (dark blue bars) for all the statements was 13, a positive evaluation of our system as a tool to visualize and recommend scientific papers. R1-T3 was a little more complex, which led one participant to rate the corresponding statement, R1-S3, as Undecided (gray bar), and another one as Slightly Agree (light blue bar). They had some difficulties in visually comparing the size of the nodes involved in the analysis. However, the large majority (86.6%) of the participants selected the option Completely Agree for R1-S3. The results evidence that

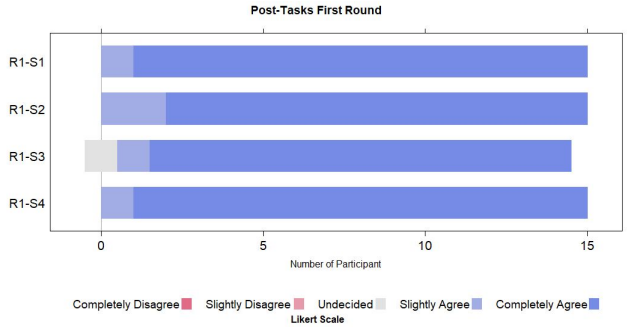

Fig. 6. Post-Tasks First Round

the participants, who are prospective system users, liked the proposed visual interface. Our system made it possible to easily identify important papers in the area, as well as the most relevant references as evaluated collaboratively by the experts. They believe that the proposed system is a great tool for finding important references and visually determine the sequence of publications that belong to a specific branch of study.

VI. Conclusions and Future Work

Assessing the relevance of cited references is far from being straightforward. This research has devised a strategy to visualize and recommend citations within a corpus of scientific papers. The overall idea of the proposed visualization is to create a directed acyclic citation graph arranged in a hierarchical layout, following a chronological order of influential publications.

With our approach it is possible to obtain an overview of the field on a specific subject (in our example case, within the area of Computer Science), visualizing the most influential articles that reflect the evolution of the different branches of study through the collaboration of experts. The strength of the system lies in considering the different opinions and expertise of experts in ranking references, a variant not explored so far in this type of system.

As future work, we propose to implement some modifications in the calculation of expertise of the author, for example, taking into account the average number of citations. To help users identify more easily whether a paper has several references identified with the same maximum value of influence, as well as the references that have not yet received a vote from the experts, we can establish a visual mapping of the k topmost influential references, as well as an attribute that reflects which of these references have the highest weight and the references that still have zero weight.

References

[1] M. C. Waumans and H. Bersini, "Genealogical trees of scientific papers," *PloS one*, vol. 11, no. 3, p. e0150588, 2016.

[2] H. Wei, Y. Zhao, S. Wu, Z. Deng, F. Parvinzamir, F. Dong, E. Liu, and G. Clapworthy, "Management of scientific documents and visualization of citation relationships using weighted key scientific terms." in *DATA*, 2016, pp. 135–143.

[3] S. A. Greenberg, "How citation distortions create unfounded authority: analysis of a citation network," *Bmj*, vol. 339, p. b2680, 2009.

[4] D. Gough, S. Oliver, and J. Thomas, *An introduction to systematic reviews*. Sage, 2017.

[5] U. Schäfer and U. Kasterka, "Scientific authoring support: A tool to navigate in typed citation graphs," in *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: Writing processes and authoring aids*. Association for Computational Linguistics, 2010, pp. 7–14.

[6] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations." in *AAAI Workshop: Scholarly Big Data*, 2015.

[7] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.

[8] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550.

[9] N. Crouch and M. P. David, "Pyscholargraph: A graph-based framework for indexing, searching and visualising relationships between academic papers," *The ANU Undergraduate Research Journal*, vol. 161, 2015.

[10] D. E. Ciriello, "Five hundred deep learning papers, graphviz and python." http://dnlcrl.github.io/projects/2015/10/10/500-deep-learning-papers-graphviz-python, 2015. [Online]. Available: http://dnlcrl.github.io/projects/2015/10/10/500-deep-learning-papers-graphviz-python

[11] G. Salton and C.-S. Yang, "On the specification of term values in automatic indexing," *Journal of documentation*, vol. 29, no. 4, pp. 351–372, 1973.

[12] M. Berger, K. McDonough, and L. Seversky, "cite2vec: Citation-driven document exploration via word embeddings," *IEEE Transactions on Visualization & Computer Graphics*, no. 1, pp. 1–1, 2017.

[13] G. Ginde, "Visualisation of massive data from scholarly article and journal database a novel scheme," *arXiv preprint arXiv:1611.01152*, 2016.

[14] A. Vukotic, N. Watt, T. Abedrabbo, D. Fox, and J. Partner, *Neo4j in action*. Manning Publications Co., 2014.

[15] Y. Sibaroni, D. H. Widyantoro, and M. L. Khodra, "Survey on research paper's relations," in *Information Technology Systems and Innovation (ICITSI), 2015 International Conference on*. IEEE, 2015, pp. 1–6.

[16] A. P. Singh, K. Shubhankar, and V. Pudi, "An efficient algorithm for ranking research papers based on citation network," in *Data Mining and Optimization (DMO), 2011 3rd Conference on*. IEEE, 2011, pp. 88–95.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[18] Z. Zhou, C. Shi, M. Hu, and Y. Liu, "Visual ranking of academic influence via paper citation," *Journal of Visual Languages & Computing*, vol. 48, pp. 134–143, 2018.

[19] G. Packer, "Cheap words," *The New Yorker*, vol. 17, 2014.

[20] T. S. Newman and H. Yi, "A survey of the marching cubes algorithm," *Computers & Graphics*, vol. 30, no. 5, pp. 854–879, 2006.

[21] G. Van Rossum and F. L. Drake, *The python language reference manual*. Network Theory Ltd., 2011.

[22] D. Framework, "Django the web framework for perfectionists with deadlines," *https://docs.djangoproject.com/en/2.0/*, vol. 1, 2016.

[23] B. Almende, "vis. js–a dynamic, browser based visualization library," *http://visjs.org/*, vol. 1, 2016.