Pontifícia Universidade Católica
do Rio de Janeiro

Orlando Fonseca Guilarte

A Graph-based Collaborative Support for Expert Finding and Recommending References in Scientific Publications

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Matemática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Matemática.

Advisor : Prof. Sinesio Pesco
Co-advisor: Prof. Simone Diniz Junqueira Barbosa

Rio de Janeiro
May 2019

Pontifícia Universidade Católica
do Rio de Janeiro

Orlando Fonseca Guilarte

A Graph-based Collaborative Support for Expert Finding and Recommending References in Scientific Publications

Thesis presented to the Programa de Pós–graduação em Matemática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Matemática. Approved by the undersigned Examination Committee.

Prof. Sinesio Pesco
Advisor
Departamento de Matemática – PUC-Rio

Prof. Simone Diniz Junqueira Barbosa
Co-advisor
Departamento de Informática – PUC-Rio

Prof. Alex Laier Bordignon
Instituto de Matemática e Estatística – UFF

Prof. Marcos de Oliveira Lage Ferreira
Instituto de Computação – UFF

Prof. Lis Ingrid Roque Lopes Custódio
Centro de Tecnologia e Ciências – UERJ

Prof. Giseli Rabello Lopes
Departamento de Ciência da Computação – UFRJ

Prof. Helio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Rio de Janeiro, May 14, 2019

Orlando Fonseca Guilarte

The author completed his undergraduate studies in Computer Science, obtained the Bachelor's degree in Computer Science in 2012 from Universidad de La Habana. The author also obtained the degree of Master in Mathematics in 2014 at the same institution. Research interests include information visualization, data science and data structure.

Bibliographic data

Guilarte, Orlando Fonseca

A Graph-based Collaborative Support for Expert Finding and Recommending References in Scientific Publications / Orlando Fonseca Guilarte; advisor: Sinesio Pesco; co-advisor: Simone Diniz Junqueira Barbosa. – Rio de Janeiro: PUC-Rio, Departamento de Matemática, 2019.

v., 92 f: il. color. ; 30 cm

1. Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Matemática.

Inclui bibliografia

1. Matemática – Teses. 2. Visualização de informação. 3. Encontrar especialistas. 4. Classificação de referências. 5. Sistema colaborativo. 6. Interação humano-computador. I. Pesco, Sinesio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Título.

CDD: 510

To my little prince

Acknowledgments

I would like to thank the following people:

My advisor, Prof. Dr. Sinesio Pesco, for his valuable supervision and great help during the development of this work, and my co-advisor, Prof. Dr. Simone Diniz Junqueira Barbosa, for her professional advice and support. This thesis would not be completed without their guidances.

The members of my thesis committee for generously offering their time.

My parents, Adriana y Orlando, and my wife Dania, for their unconditional support and love.

The staff of the Mathematics Secretariat of PUC-Rio, specially for Creuza, Katia and Carlos, for their assistance throughout my study period.

Prof. Ricardo Alonso and Prof. Débora Mondaini, for all the help.

My colleagues and friends, specially for Yunelsy, Viviana, Pablo Vinicius, Rafael, João Marcos, Renan, Tiago, Luis, Tahiz, Thiago, Tamires, for their friendship and help.

A special thanks to my loving family.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Finally, thanks to CNPq for the financial support.

Abstract

Guilarte, Orlando Fonseca; Pesco, Sinesio (Advisor); Barbosa, Junqueira Simone Diniz (Co-Advisor). **A Graph-based Collaborative Support for Expert Finding and Recommending References in Scientific Publications**. Rio de Janeiro, 2019. 92p. Tese de doutorado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

The scientific literature review is a critical account of the main papers in a particular subject area or topic. In this way, the authors surveys the literature and present the relevant articles in an organized way by publication date and evolution of the research topic, which gives an overview of the state of the art in a subject. Through these relevant papers it is also possible to identify the most expert authors in the area or in certain papers, thus providing a solution to the problem of finding potential expert candidates. The main challenge of making a literature review is to identify the most relevant articles that reflect the evolution of the different research topics. In this thesis, we propose a visual collaborative approach that uses graphs to recommend important references. In addition, we introduce the task of searching and ranking authors given a target paper using relevant citation paths. From a ranking of references, the value of the authors' expertise is calculated. A methodology is proposed in order to build and update the citation graph in a collaborative way with the expert's votes.

Keywords

Information Visualization; Expert Finding; Ranking References; Collaborative System; Human-Computer Interaction.

Resumo

Guilarte, Orlando Fonseca; Pesco, Sinesio; Barbosa, Junqueira Simone Diniz. **Suporte colaborativo baseado em grafos para localizar especialistas e recomendar referências em artigos científicos**. Rio de Janeiro, 2019. 92p. Tese de Doutorado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

A revisão da literatura científica é um relato crítico dos principais trabalhos em uma área ou tópico específico. Dessa forma, os autores buscam a literatura e apresentam os artigos relevantes de forma organizada por data de publicação e evolução de um tema de pesquisa. Revisões da literatura fornecem uma fotografia do estado da arte de um tópico de pesquisa. Através da seleção dos trabalhos mais importantes de uma certa área é possível identificar os autores mais especializados na área ou em determinados artigos, proporcionando assim uma solução para o problema de encontrar potenciais candidatos especialistas. Nesta tese estudaremos o problema de selecionar e visualizar os artigos mais relevantes que refletem a evolução de um tópico de pesquisa. Para isso, propomos uma abordagem visual colaborativa baseada em grafos para recomendar referências importantes. Apresentamos também a tarefa de encontrar e classificar os autores dado um artigo científico usando caminhos de citações relevantes. A partir de um ranking de referências, o valor da expertise dos autores é calculado. Uma metodologia é proposta para construir e atualizar o grafo de citações de forma colaborativa com os votos dos especialistas.

Palavras-chave

Visualização de informação; Encontrar especialistas; Classificação de referências; Sistema colaborativo; Interação humano-computador.

Table of contents

List of figures

1
Introduction

The search for experts in a specific field has received increased interest in recent years, both in industry and in academia. The identification of relevant people on a specific academic topic can be of great value in many applications. For example, determining the potential experts who will provide their knowledge on the research topic [1, 2, 3]. Another example is the assignment of papers to referees to be peer reviewers in conferences or scientific journals, as explained in [4]. Finally, ranking researchers is one of the most central issues in scientometrics studies [5].

Expert finding systems help in the search of a suitable person with the appropriate skills and knowledge [6]. Conventionally, expert finding in academia is treated as an information retrieval task and it models the rank of experts, measuring the similarity between a query paper and the publications of experts. The goal of this information retrieval task is to rank authors based on their expertise in the research topic of a given paper submission. We define expertise as enough knowledge to understand and judge the value of the contributions. The expertise is often considered as the main criterion in the problem of finding experts for a given paper, although other factors could also be considered as the researcher recognition in the scientific community and the analysis of the expert's profile [4]. Publications and citations are essential to solve this problem. Publications provide an effective way to evaluate one expert in a specific field [7], either by the content or by the number of papers published in this field. Also, the publications of the authors could be a good indicator that reveals the research interest trend of these authors and his/her diversity [8], that is, when these authors have diverse research interests and background. Citations provide a measure of the quality of published work [9]; they evidence the impact or popularity of an author's publications. The traditional approach for measuring the expertise is to estimate a similarity value between the candidate expert's publications and a query paper. For example, it can be measured by the topic model-based similarity, see [4]. This approach has one important problem: it requires a huge amount of calculation for the similarity of research topics, as discussed in [7].

On the other hand, to model the problem of ranking researchers, a

graph-based approach can be considered. For example, Computer Science papers form a huge directed graph named citation graph, whose nodes are articles and edges are links to the articles cited in a paper, see [10]. Analysis in this type of graph is used to suggest potential experts for a paper on the premise that the subject of a paper is characterized by its references and corresponding authors [11]. In fact, one important aspect of the content of scientific papers is their relation to previous work through its references. This is because they provide a way to measure the quality of the work, to detect emerging research topics, and to follow evolving ones [9]. If the paper has a very strong relationship with some of its references, then basically the authors of those references would be potential experts.

1.1 Problem Statement

In this work, we address the problem of **ranking authors** by the values of expertise. A Citation Graph is used to calculate the expertise of the authors given a query paper. To provide a better understanding, our approach is illustrated by a little example.

Suppose, for example, that three different papers or documents p_1, p_2 and p_3 are referenced by the query paper q, and that all these documents have the same relevance for q, represented by the weight value of 1 on each corresponding edge. These documents are related by the same research domain. Also consider that the document p_1 was published in 2010, p_2 was published in 2011 and p_3 in 2018. And these documents were authored by a_1, a_2 and a_3, respectively. An authorship relationship is built for each document and its author, as shown in Figure 1.1. In this case, the authors a_1, a_2, and a_3 have the same experience or expertise to deal with a paper q. When relationships between the documents are considered, the expertise of the authors can change. Consider the case in which these relationships are given by the citations between them; more precisely, document p_1 was cited by p_2 and p_3. The most cited document, that is, p_1, would be more important that p_2 and p_3. Consequently, we can expect that a_1, the author of p_1, has a higher probability of being more expert than a_2 and a_3 given the query q, as shown in Figure 1.2. However, if we consider a collaborative procedure where p_2 became a strong reference of q, that is, has high weight value, intuitively a_2 would be more expert than a_1 and a_3, as shown in Figure 1.3. Note that, although a_1 is the most cited author and a_3 is the most active person, that is, author with publications in a more recent year, the author a_2 would probably be the most indicate to, for, example, review the paper q.

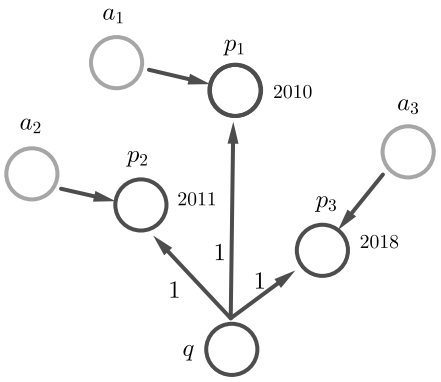


Figure 1.1: Documents p_1,p_2 and p_3 with the same relevance for q

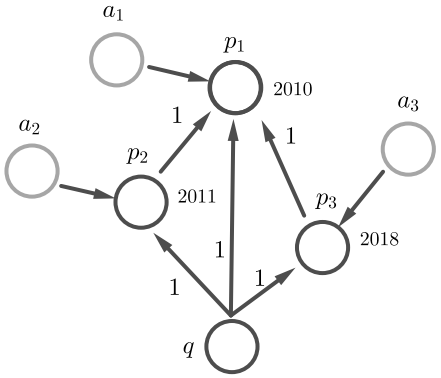


Figure 1.2: Documents with citation relationships

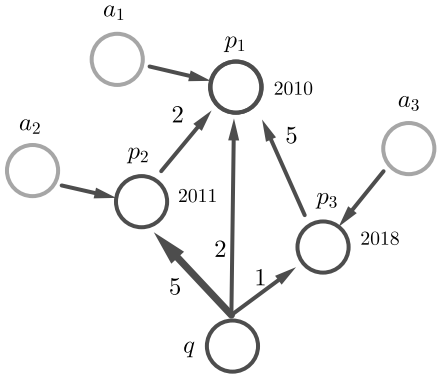


Figure 1.3: Citation networks with edge weights resulting from a collaborative procedure

For the problem of discovering experts based on a query paper, each author has a value of expertise that will be deduced by paths of relevant references and other metrics that consider, for instance, quantity

of publications and citations. Our approach is also based primarily on the assumption that two documents connected by a path of the strongest references share similar research issues. Therefore, the smaller the path between two articles, the more similar the research issues found in them. In this way, an author with several publications in paths containing the query paper q will have more probability to be selected as a potential expert for q. We also consider the proximity of the publication of the authors to the paper q.

The proposal for ranking authors via citation paths is viable if a weight function is defined for each reference, so that the strongest references have greater weight. Therefore, the procedure to rank authors requires solutions to the problem of ranking of references.

In this work, we also address the problem of **ranking references** for each paper. We propose a collaborative approach to establish this ranking. More explanation of this choice is given below.

The classification of references in importance classes would be an alternative solution to the problem of ranking references. In this, each class has a specific weight. For example, the references can be grouped into five classes: Agree(the citing paper agrees with the cited paper), PRecycle(the citing paper uses an algorithm, data, method or tool from the cited paper), Negative(the paper is cited negatively/contrastively), Neutral(the paper is cited neutrally) and Undef(it is impossible to determine) [12]. The value for the weight would be 2(greater value) for classes Agree, PRecycle and Negative, 1 for class Neutral, and 0 for class Undef. The problem is that performing this classification automatically is a hard task [12]. Other variants separate important references, that is, extended in the new publication, from incidental citation using supervised classification [9], predicting academic influence with machine learning [13] and using a topic model approach [14].

We are consider that the task of ranking references using a weight function, which quantifies the important references of a paper that directly impacted the new effort, can be more precise with the collaboration of experts. This is because the term “important” according to the criterion of “direct impact in the new effort” or “references with more academic influence that the others” is subjective. Only authors and experts have the appropriate knowledge to affirm the reasons why the references were included in the work, and identify references with more probability of having a very strong relationship with the paper. For this reason, we establish a collaborative classification in which users with their votes play a principal role. First, we perform an automatic classification considering the date of publication of the papers. Finally the classification will be improved as soon as users give their votes. In this way, a

human effort is necessary to obtain more precise and reliable results.

To deal with these ideas, we need to consider three fundamental questions:

1) What conditions are necessary for a user's vote or rating to be considered?
2) How to establish a general ranking of references considering all votes of users, who may have different experience?
3) How to visualize the ranking of references in a clear and comprehensive way?

In response to the questions, we consider that, unlike a conventional recommendation system, a selection of users must be made. In that sense, every researcher or author with at least one paper in the research area of the target paper can contribute with his/her votes in the rating. This vote will be more or less important depending on how expert in the subject the researcher is. To establish a general rating of a cited paper, we propose a weighted arithmetic mean with all authors' votes for each reference. This measure is calculated dynamically: as the authors vote or modify the vote, the value of the metric will be recalculated. A general ranking of the references of a specific paper is possible when ordering these references by the value of the general rating.

We also developed an interactive and collaborative web tool for voting and viewing references. It highlights one of the most important or most influential references for each paper, that is, the top one in the ranking. This reference will be fundamental in our proposal to have a simplified outline of the state of the art of a specific topic and to calculate the authors' expertise with more precision, once those references have the number of votes necessary for this purpose.

1.2 Thesis Contribution and Applications

The novelty of this work lies in the following points:

1) It proposes a new collaborative model for ranking references, in which the evaluation of the experts and their expertise is decisive.
2) It establishes a new function for calculating author's expertise via citation paths. This function includes the concept of proximity between the publications of the candidate experts and a specific paper.

3) It develops web support for visualizing the most important relationships among the scientific publications in the area of Computer Science, enabling users to analyze the influence patterns of papers. In this way, it is possible to obtain an overview of the field or a state of the art on a topic through a visual recommendation of the main works, result of the collaborative evaluation of experts. We consider that expert evaluation in the collaborative model assigns credibility to the visual result.

Our visual collaborative system can have various applications. Some of these uses can be:

- Looking up the most cited paper.
- Looking up the most important references for a given paper.
- Looking up the path of relevant or influential papers from a given paper.
- Visually comparing two or more papers by the number of references and by the differences in publication dates of these references.
- Comparing two or more authors by productivity and quality of publications.
- Comparing two or more authors by expertise in a general issue.
- Comparing two or more authors by expertise in a particular paper.

This proposal facilitates users searching for potentially relevant articles with respect to a query paper; helps to understand the flow of research topics in the scientific literature; and helps making decisions about which paper should be studied by them. The usefulness of this work is to support not only the authors of scientific articles, but also reviewers, conference program committee chairs, and journal editors. The reviewers will be able to verify whether the submitted works cite relevant contributions, both seminal and recent. The editors will be able to find reviewers or validate the expertise of reviewers.

1.3 Thesis Outline

The remainder of this work is structured as follows. In Chapter 2, we briefly discuss related works. We describe in Chapter 3 some important notations and definitions, the graph-based model, the collaborative characteristic of the system and the new proposal of the calculation of expertise of the authors via citation paths. In Chapter 4 we discuss the visual structure and the exploration and interaction with the system. Some applications of the system are shown in Chapter 5. We present our experimental results in Chapter 6. Finally, we conclude and give an outlook to future work in Chapter 7.

2
Related Work

2.1
Visualizing Scientific Documents and Ranking References

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. Visualizations help users to perceive the relationships between items and identify the most representative ones. Graphs are used in numerous applications within the field of information visualization, enabling visual representations of the relationships between nodes in network data, as shown in [15]. They assist in the perception process and increase the level of understanding [16].

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 [17, 18, 19, 20]. Several visualizations of scientific documents have been proposed to help researchers explore graphs of references. PyScholarGraph [21] 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 [18] adopt an evolutionary perspective, building the graph in the form of a genealogical tree, showing the rate of growth of the number of citations for each article. In PyGraphviz [17], a graph was created to represent the evolutionary process of the deep learning field in the last 25 years. Wei et al. [19] propose an interactive visualization method that uses citation paths, as a result of a variation in the TF-IDF scheme [22]. Berger et al. [23] treat the documents as a collection of special words. As shown in [24], the data can be processed previously and sent to a database with the information of scientific journals, authors and research documents, from where they can be consulted using the Cypher language [25].

Several papers have presented a literature review and proposed algorithms to analyze the classification or ranking of references. Sibaroni et al. [26] proposed to build relationships between documents using basic approaches like analysis of co-citations, reference lists and bibliographic

coupling. Singh et al. [27] pointed out that the quality of the citations is very important and propose a modified version of the PageRank algorithm [28] to rank the research papers. Nallapati et al. [14] presented two novel topic models to address the problem of joint modeling of link and text. Zhu et al. [13] focused on automatically identifying the subset of key references that have a great academic influence on the citing paper, and Valenzuela [9] described a classification approach for identifying important and incidental citations. In [29], Zhou et al. implemented a visualization framework for visual ranking of academic influence of papers.

However, in [24, 29] a hierarchical structure to facilitate user understanding is not presented. 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. In [18], although such a structure is considered, it is not easy to perform an analysis per period of time. In [17, 21] it is difficult to visually obtain a sequence of closely related works due to the number of relationships that it presents. In [19, 27, 14, 13, 9], the opinion or change of opinion of the specialist is not considered in the classification of key references. Finally in the network view in [29], the user cannot easily filter the citation links of a certain paper sorted by relevance, in order to reduce visual clutter and find relevant or influential citation paths.

Until now, we have not found in the literature a collaborative visualization of scientific publications that shows data and recommends references in a simple and comprehensive way for users. In our representation, a hierarchical structure is defined by year and the general ranking of references is visualized, taking into consideration the evaluations of specialists.

2.2 Expert Finding and Ranking Authors

Expert finding systems help find “someone who is an expert on X”. Essentially, expert finding is considered a ranking problem [30]. Given an input or query, which can be a topic or a paper, the system tries to rank authors assigning a score to each author according to the relevance with the query. For this task, in this work we consider that a query refers to a paper in general and not to a specific topic of the paper. Typically, a user profile is used for each author, who determines their relevance with respect to the query, that is, the value of expertise they have in relation to the query. These profiles can be generated manually by the user or automatically by the system [1], which is characterized mainly by the authors’ publications. For this purpose, several

metrics were developed, for example, h-index [31] and g-index [32] to measure the productivity and impact of publications. Also, some classical information retrieval methods have been used to identify authors by the text associated with their publications, such as topic modeling [33, 34] and the vector space model [35].

A practical application of expert finding systems is to assign paper submitted to the reviewers of conferences or scientific journals [4, 36, 37]. The sets of papers written by the candidate reviewers have been used as evidence of content-based expertise for these reviewers [38].

In the aforementioned approaches, the entire set of author's papers is essential for the calculation of the value of expertise for ranking authors. This value depends mainly on the number of publication of the authors, their number of citations, and a term weighting scheme in all documents. In addition, the topology of the citation network can be used to determine a better expert profile. In this way, we consider that for the calculation of expertise other factors should be take in account, such as the influence of the author in the query paper via citation paths. Therefore, we propose an approach that considers relevant citation paths and the proximity of the papers of an author with the query paper.

3
Terminology and Overview of the Proposed Approach

In this chapter, we describe our proposal. We deal with a graph, our fundamental tool to solve the problem of visually recommending references and finding experts. First we define a graph to represent the scientific papers, the authors, and the relationships between them. We also define a *Citation Graph* for visualizing and ranking references in our system, the *Max-Generator Graph* and the *Relevant References Paths* (RRP), very useful to determine the expertise of the users and to find experts. Then, we discuss an important novelty of our work, which consists of taking into account the votes and expertise of the users. To do this, we are going to explain in detail the collaborative characteristics of the system. From this collaborative approach we introduce a ranking of authors for each paper in particular, via citation paths. Finally, a methodology is proposed in order to build and update the node set in the *Citation Graph*. The example graphs presented in this chapter aim to exemplify the concepts discussed here, so they do not present the characteristics of the visual structure defined for the system, explained in the following chapter.

3.1
Graph–based Model

We propose a collaborative system following a graph-based model. This model is formulated from a directed, acyclic and weighted graph represented by a triplex $G = \langle V, E, \omega \rangle$, where V is the node set that indicate authors and documents, E is the edge set that indicate relations of authorships and citations and ω is a function that assigns to each edge a positive weight. This weight is determined by the collaboration of the users and will be defined later.

We are motivated to develop a directed graph because the citation has a source node and destination node, which are not interchangeable and determine a direction. We avoid cycles by guaranteeing that a document can only refer to another document that was previously published.

We now explain the structure of this graph.

To conform the first term of the triplex in G that defines the graph, we consider authors and papers (or documents), as was previously mentioned. 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. In this way, the node set can be defined as $V = A \cup P$. The second term of the triplex, the edge set E, is determined by the union of the set of authorship edges E_A and the set of reference edges E_R, in the following way: $E = E_A \cup E_R$, in which,

$$E_A = \{(a_j, p_i) \ / \ a_j \in A, p_i \in P, a_j \sim p_i\} ; i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, m\},$$

where $\sim$ defines a relation between an author and a paper in the following sense: We say that $a_j \sim p_i$ if a_j is one of the authors of the paper p_i.

The set E_R is defined by,

$$E_R = \{(p_i, p_k) \ / \ p_i, p_k \in P, p_i \simeq p_k, i \neq k\} ; k \in \{1, 2, \ldots, n\}, \tag{3.1}$$

where $\simeq$ defines a relation between two papers. We said that $p_i \simeq p_k$ if the paper p_i has the paper p_k in its reference list.

For example, if we consider the case represented in Figure 3.1, then the set $V = \{a_1, a_2, a_3\} \cup \{p_1, p_2, p_3\}$, and the set E can be obtained by the union of subsets E_A and E_R, where

$$E_A = \{(a_1, p_1), (a_2, p_2), (a_3, p_3)\},$$

and,

$$E_R = \{(p_2, p_1), (p_3, p_1)\}.$$

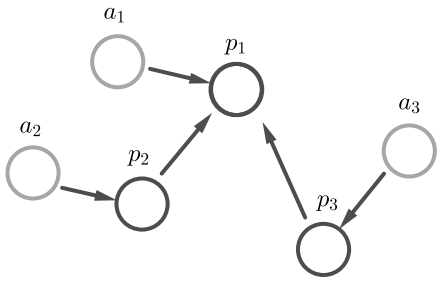


Figure 3.1: Documents and Authors

Note that, in the set E_R, if $p_i \simeq p_k$ then paper p_i is more recent than the paper p_k. Also, in case of the set E_A, if we have a paper with more than one author, the order of authorship is not determined in the relationship. In this way there is no distinction between the first author and the other authors of the paper.

The last term in the triplex of G, the weight function, is defined below.

Definition 3.1 (Edge Weight) *Let $e \in E$. We define the weight map ω as a non-negative function that satisfies $\omega : E \rightarrow [0, 5]$, where*

$$\omega(e) = \begin{cases} \omega(p_i, p_k), & e = (p_i, p_k) \\ 0, & e = (a_j, p_i). \end{cases}$$

We associate for each edge in E_R a weight from 0 to 5. This definition is relevant in our work for ranking references. In fact, the sets P and E_R are very important in our model. These sets represents a subgraph of the initial graph G and are defined as follows.

Definition 3.2 (Citation Graph) *A Citation Graph, denoted by C, is a directed acyclic graph represented by the sets P and E_R, where each edge has a weight defined by the function ω.*

We consider this particular graph in the visualization of important references, which will be discussed in Chapter 4.

For each paper $p_i \in P, i \in \{1, 2, \ldots, n\}$, we denote by $|In(p_i)|$ the number of citations, where

$$In(p_i) = \{p_k \ / \ p_k \in P, (p_k, p_i) \in E_R\}; k \in \{1, 2, \ldots, n\}.$$

Similarly, the number of references of a paper p_i is denoted by $|Out(p_i)|$, where

$$Out(p_i) = \{p_k \ / \ p_k \in P, (p_i, p_k) \in E_R\}.$$

Also, for each author a_j, $j = \{1, 2, \ldots, m\}$, we denote by n_j the number of publications, $P^j = \{p_1^j, p_2^j, \ldots, p_{n_j}^j\}$ by the set of papers authored by a_j, where $P^j \subseteq P$, and by c_j the number of citations received by the author, which is defined as follows.

Definition 3.3 (Number of Citations to an Author) *The number of citations received by the author a_j, denoted by c_j, is the sum of the citations received for each paper of his/her authorship, that is, $\sum_{k=1}^{n_j} |In(p_k^j)|$, where $p_k^j \in P^j$, $k \in \{1, 2, \ldots, n_j\}$.*

The next elements to be defined are the *Max-Generator Graph* (MG) and the *Relevant References Path* (RRP) between two nodes in this graph.

Definition 3.4 (Max-Generator Graph) *Let $C = \langle P, E_R, \omega \rangle$ be a Citation Graph, where P is the node set, E_R is the edge set and ω is a weight function. We define a Max-Generator Graph, denoted by $MG = \langle P, \widehat{E}, \omega \rangle$, the subgraph of C, such that if the edge $(p_i, p_k) \in \widehat{E}$, then $\omega(p_i, p_k) = max\{\omega(p_i, p_j)\}$, where $p_i, p_k \in P$ and $p_j \in Out(p_i)$ in the graph C.*

Finally, we define the *Relevant References Paths* (RRP) between two nodes in the *Max-Generator Graph*, the length of the RRP, and the sum of the weights of the edges in this path.

Definition 3.5 (Relevant References Path) *The Relevant References Path between two nodes p_i and p_k, $p_i \neq p_k$, in the Max-Generator Graph, denoted by $RRP(p_i, p_k)$, is the list of nodes (papers) starting with p_i and ending with p_k, such that from each node x in the list exists a directed edge (reference) from x to the successor node.*

Definition 3.6 (Length of the RRP) *The length of the Relevant References Path between two nodes p_i and p_k, that is, the distance between these nodes, is denoted by $L(p_i, p_k)$ and defined by,*

$$L(p_i, p_k) = \begin{cases} |RRP(p_i, p_k)| - 1, & RRP(p_i, p_k) \neq \emptyset \\ 0, & otherwise. \end{cases}$$

Definition 3.7 (Weight of the RRP) *Let $RRP(p_1, p_n) = \{p_1, \ldots, p_n\}$ a Relevant References Path and the length of this path is $L(p_1, p_n) = n - 1$. The weight of this path, denoted by $W(p_1, p_n)$, is defined by,*

$$W(p_1, p_n) = \sum_{i=1}^{n-1} \omega(p_i, p_{i+1}),$$

where $p_i, p_{i+1} \in RRP(p_1, p_n)$ and ω is a weight function.

For example, consider the *Citation Graph C* in Figure 3.2 below.

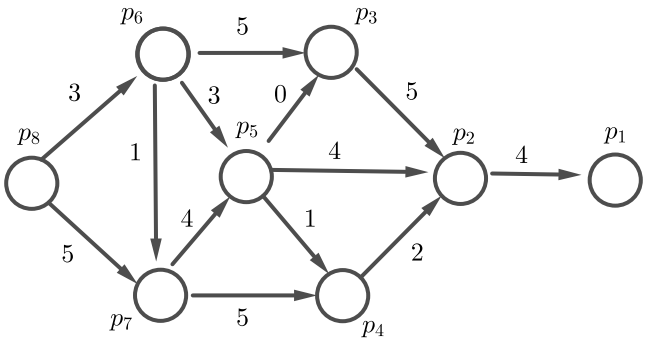


Figure 3.2: Citation Graph

The *Max-Generator Graph* obtained from the graph C is shown in Figure 3.3.

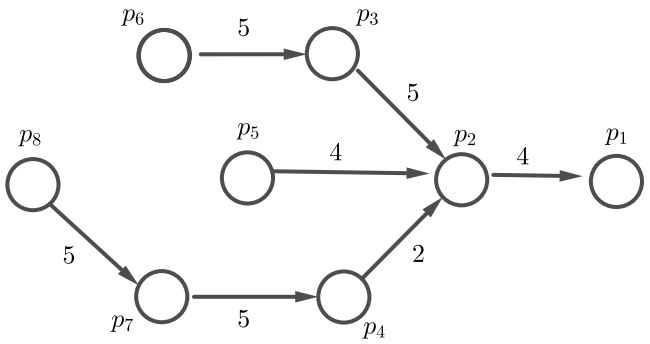


Figure 3.3: Max-Generator Graph of the Citation Graph

Thus, the RRP between the nodes p_8 and p_2 is $RRP(p_8, p_2) = \{p_8, p_7, p_4, p_2\}$. This is because in this particular case the weight function between the papers satisfies, $5 = \omega(p_8, p_7) > \omega(p_8, p_6) = 3$ and $5 = \omega(p_7, p_4) > \omega(p_7, p_5) = 4$. The weight and length of the $RRP(p_8, p_2)$ is 12 and 3, respectively.

3.2 Collaborative Approach

Next, we discuss how weights are assigned to each edge in the citation graph through a collaborative process.

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 of these references are the most representative for the paper q according to our proposal. Following this ranking, the users obtain a recommendation of, for example, which reference is more convenient to study from paper q.

Thus, our approach has a starting point: web recommendation systems. Any system of this type, such as the global superstore Amazon, see [39], has a user history or profile and a user rating, in which the recommendation algorithm finds users with common characteristics or also focuses on finding similar items, as shown in [40]. Amazon allows users to submit ratings or votes for each item in the system, then automatically provides one ranking in the final recommendation. The users must see the general rating on a scale from one to five stars and the number of votes. In Figure 3.4, we can observe that the book "Introducing Communication Theory" obtained 61 votes, which provided the value of 3.8 in the general rating.

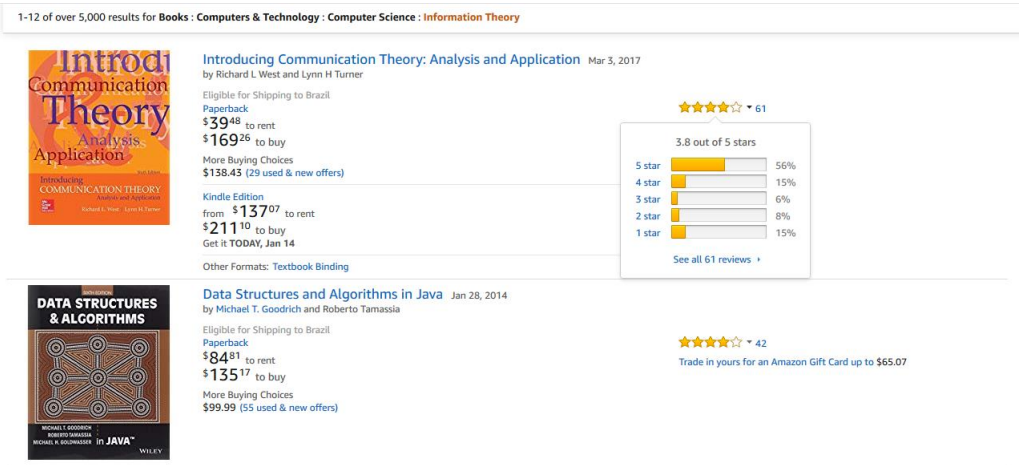


Figure 3.4: Book's ratings in Amazon.

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. The most logical thing is that people with robust knowledge of the subject are those whose votes are most valued. In this sense, we propose a collaborative system between previously selected users, where each user's vote depends on their expertise. The users can vote at any edge in the system and the votes and expertise of these users determine the weight of each edge in the *Citation Graph*.

To define the values of expertise, different factors will be considered, such as the structure of the *Citation Graph* and the impact of the publications of the authors on the subject. Although it is important to consider the expertise of an author locally, that is, in a specific paper in the graph, we consider this calculation globally in the initial stage of the system. In Section 3.4 we propose a new method for calculating the authors' expertise given a specific paper. Thus, the expertise of the author in the initial stage is defined in the same way for all the papers in the graph, and its value depends, in general, on a count of citations.

Next, we define the global expertise of an author in the graph.

Definition 3.8 (Global Expertise of the Author) *The global expertise of the author a_j is a non-negative function defined as follows,*

$$Exp(j) = \begin{cases} \dfrac{c_j + 1}{c_l + 1}, & n_j \neq 0 \\ 0, & n_j = 0, \end{cases} \tag{3.2}$$

where c_j is the number of citations received by the author a_j (by other authors), 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 of author* a_j.

This function reflects the impact of the author's publications in comparison with the author that most impacted the growth of research in the area. This function describes whether the author has authority in the subject, that is, if he/she has a good recognition in the scientific community specialized in this subject.

Inspired by Amazon's voting system explained above, we use a weighted arithmetic mean to calculate the general rating. Formally, given a graph of authors and citations, we consider that, in the initial state of the system, the reference edges still have no associated voting value. That is, for all $e \in E_R$; $\omega(e) = 0$.

Then, the vote will be considered. Each author may give a vote. We denote $v_j(q, p_k)$ the vote of author a_j for reference p_k of a target paper q, where $k \in \{1, 2, \ldots, |Out(q)|\}$, that is, the vote of the author in the edge $(q, p_k) \in E_R$.

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 or furthest issue with respect to q.

The first vote of the author $a_j \in A$ in a specific edge modifies the initial zero weight value of this edge, as follows:

$$\omega(q, p_k) = \frac{v_j(q, p_k) * Exp(j)}{Exp(j)},$$

where $v_j(q, p_k)$ is the author's vote for the edge (q, p_k), and the expertise $Exp(j)$ of the author 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.

For the value of an author's expertise, we consider two fundamental conditions:

- A user expert in a subject must have at least one paper published on the subject.
- The expertise of the author in a subject increases when the author's papers, in this subject, gain more impact, that is, the author's papers receive more citations.

This means that users without publications in the graph will not have their votes considered and, therefore, the weight of the edges will not change.

This justifies the fact that our system is selective to guarantee the credibility of the process, but it is not excessively selective, in that it does consider the votes of an author who has not yet received citations. It is important to point out that authors with at least one publication in the graph and zero citations can vote in our model, but their vote will be affected by the value of their expertise, which will be minimal. This is the main reason why we do not consider a metric like h-index[1] in this collaborative system.

Now, suppose that the graph has evolved with respect to the edge (q, p_k), that is, the edge has already been voted one or more times. For example, m experts have already voted on this edge. Then the weight value associated with the edge (q, p_k) after the vote of author $m + 1$ is:

$$\omega(q, p_k) = \frac{Y + v_{m+1}(q, p_k) * Exp(m + 1)}{X + Exp(m + 1)} \tag{3.3}$$

where

$$Y = v_1(q, p_k) * Exp(1) + \ldots + v_m(q, p_k) * Exp(m),$$

and $X = Exp(1) + \ldots + Exp(m)$.

It is necessary to point out that the values of Y and X in the previous expression do not need to be recalculated at the moment of incorporating the vote of the author $m + 1$ into the system.

Next, we illustrate with an example how weights are attributed to edges in our collaborative system.

Consider a part of the *Citation Graph*, in which the target paper q has three papers, p_1, p_2, p_3, in their references list, that is, $Out(q) = \{p_1, p_2, p_3\}$. These references have not been valued by votes by any user. Then, $w(q, p_1)$, $w(q, p_2)$ and $w(q, p_3)$ have zero value by default, as shown in Figure 3.5.

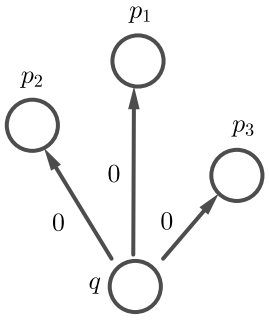


Figure 3.5: Edges without user evaluation

Suppose, also, that the number of citations of all authors in the graph

[1]An author $a_j \in A$ has index h in the *Citation Graph* if h of his n_j papers have at least h citations each, and the remaining papers have no more than h citations each.

are between 0 and 99, and the author a_1 has zero citation in all the graph, the a_2 has 9 citations and a_3 has 99 citations, the maximum value. Then, using the function 3.2 in Definition 3.8, it is easy to verify that:

$$Exp(1) = \frac{1}{100} = 0.01,\ Exp(2) = \frac{10}{100} = 0.1 \text{ and } Exp(3) = 1.$$

According to the values of expertise for these authors, a_3 is the most expert in the subject, and a_1 is the least expert.

Now, consider that a_1 and a_2 are users of our system and the votes of these users are included. For example, if the author a_1 submits the following votes: $v_1(q, p_2) = 4$, $v_1(q, p_1) = 5$, and $v_1(q, p_3) = 1$, as shown in Figure 3.6, then $\omega(q, p_2) = 4$, $\omega(q, p_1) = 5$, and $\omega(q, p_3) = 1$, establishing his/her rank of references for q, represented by the set $\{p_1, p_2, p_3\}$. At this time the *Citation Graph* has evolved with respect to the three references of q.

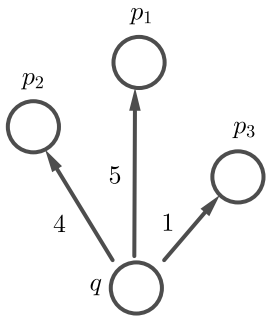


Figure 3.6: Weight of the edges after the votes of the first user

If the author a_2 submits the following votes, $v_2(q, p_2) = 5$, $v_2(q, p_1) = 4$ and $v_2(q, p_3) = 1$, that is, his/her rank of references for q is represented by $\{p_2, p_1, p_3\}$, then the weight for these edges, see Figure 3.7 below, is calculated according to the measure in function 3.3, resulting that:

$$\omega(q, p_2) = \frac{4 * 0.01 + 5 * 0.1}{0.01 + 0.1} = 4.90,$$

$$\omega(q, p_1) = \frac{5 * 0.01 + 4 * 0.1}{0.01 + 0.1} = 4.09,$$

$$\omega(q, p_3) = \frac{1 * 0.01 + 1 * 0.1}{0.01 + 0.1} = 1.$$

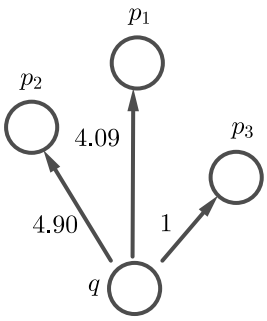


Figure 3.7: Weight of the edges after the votes of the second user

Note that although the edges (q, p_1) and (q, p_2) received in general the same values of votes, 5 and 4, their final weights are not equal due to the value of expertise of the users who voted on them. In this way, the edge (q, p_2) stands out as the most relevant for q in the general ranking set for this paper, represented by $\{p_2, p_1, p_3\}$.

3.3 Edit User's Votes

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. A user can select a relevant reference cited in a paper, and once it is selected, the user can change his/her mind to consider alternative references. For this reason, we consider a possible change of opinion by the users, which is represented by the votes and introduce a modification in the weight functions defined in function 3.3. Suppose that the author a_j has already evaluated the edge (q, p_k) and now he/she gives a second vote (different from the first vote) in this edge. Then the modification is made in the following way,

$$\omega(q, p_k) = \frac{Y + \overline{v_j}(q, p_k) * Exp(j) + Z}{Exp(1) + \ldots + Exp(j-1) + Exp(j) + Exp(j+1) + \ldots + Exp(m)},$$

where q is a target paper in P, p_k is one of the references of a target paper q for $k \in \{1, 2, \ldots, |Out(q)|\}$, m is the number of authors who voted in the edge, Exp is the expertise function as in Definition 3.8, $\overline{v_j}(q, p_k)$ is the new user's vote and

$$Y = v_1(q, p_k) * Exp(1) + \ldots + v_{j-1}(q, p_k) * Exp(j-1),$$

$$Z = v_{j+1}(q, p_k) * Exp(j+1) + \ldots + v_m(q, p_k) * Exp(m).$$

This modification means that if an author votes two or more times in the same edge, only the last vote is considered in the weight of this edge. If an edge received n votes means that exactly n users evaluated this edge, regardless of how many times the values of the votes of these users have changed. In that sense, in the weight function 3.3, the factor $v_j(q, p_k) * Exp(j)$ is substituted by a new product $\overline{v_j}(q, p_k) * Exp(j)$.

In this process, it is also natural that the expertise can change. When new papers by the authors existing in the system are added to the citation graph and their citation relationships are included, the expertise of these authors may change. This possibility is taken into account. Thus, denoting by $\overline{Exp(j)}$ the new value of expertise of author a_j and considering that this author gave at least a second vote in the edge (q, p_k), the weight function in function 3.3 is modified as follows,

$$\omega(q, p_k) = \frac{Y + \overline{v_j}(q, p_k) * \overline{Exp(j)} + Z}{Exp(1) + \ldots + Exp(j-1) + \overline{Exp(j)} + Exp(j+1) + \ldots + Exp(m)},$$

where Y and Z are as in the previous weight expression.

3.4 Characterization of Expertise by Citation Paths

As seen in function 3.2, the value of the global expertise for each author is calculated in a general way in the initial stage of the system. This function depends on the citations of scientific publications obtained throughout the author's scientific career. The number of papers by the author, the date of publication of their works and the number of citations characterize the author's career and its influence in the graph. In our system, this trajectory is obtained by the construction of the citation graph.

An underlying information that can be found in the *Citations Graph* is the Relevant References Path (RRP), which helps to understand the development of a line of research from a specific paper. For each paper, its most relevant or influential reference is what characterizes it and positions it with higher probability in a certain area of study. Thus, if the author a_j of a paper p_i references a paper p_k, and this reference is considered relevant or influential, then it is said that a_j is a good candidate to comment on any issue with respect to p_k, because it is very likely that he/she studied it in depth. This means that a value of expertise of the author a_j can be defined in relation to p_k. In the same way, the author of p_k would be a good candidate to review the paper p_i. If, in addition, other publications of the author a_j references the paper p_k indirectly through paths of relevant publications, or also from

the paper p_k other author's publications are found in one path of relevant references, then this author will have a greater value of expertise. In this way, the author's scientific career is analyzed through citation paths.

Based on the aforementioned ideas, in our system we propose a new local characterization of the expertise. We made a first approximation to the calculation of the value of the expertise of an author given a target paper through relevant citation paths. To do this, we consider two fundamental conditions for the definition of this function:

1) An expert in the subject of a target paper q should have publications close to q. This proximity refers to the number of citations (edges) in the RRP between each publication of the author and the paper q.
2) The expertise of an author given a target paper q increases when:
   - the author has more papers connected by relevant citations with q.
   - the distance between his publications and the paper q decreases.

Thus, to establish this expertise, is necessary to previously establish who the most relevant or influential references of each paper are through a ranking of references, for example, with a collaborative approach explained in Section 3.2, then form the MG graph, see definition 3.4 and finally establish the RRP, see definition 3.5. Each edge of this path will have a weight defined by a function ω obtained in a collaborative way. Therefore, it is necessary that the system is not in the initial stage, that is, some edges of the graph have already been voted.

We are now able to define the function of expertise of an author for a target paper.

Definition 3.9 (Expertise of the Author via Citation Paths)
Let an MG graph and $P^j = \{p_1^j, \ldots, p_n^j\}$ the set of papers of the author a_j. The expertise of the author a_j for a target paper q is a non-negative function defined as follows,

$$Exp(j,q) = \sum_{i=1}^{n_j} \frac{\delta(p_i^j, q) * W(p_i^j, q)}{L(p_i^j, q)^2} + \sum_{i=1}^{n_j} \frac{\delta(q, p_i^j) * W(q, p_i^j)}{L(q, p_i^j)^2}, \tag{3.4}$$

where the functions L and W, as enunciated in definitions 3.6 and 3.7, respectively, are calculated in the MG graph, and the function δ is defined by,

$$\delta(p_i, p_k) = \begin{cases} 1, & L(p_i, p_k) \neq 0 \\ 0, & otherwise, \end{cases}$$

with $p_i, p_k \in P$.

Consider the following example. Suppose that, in the graph of Figure 3.2, the author a_1 is one of the authors of the papers p_6, p_5, and p_1, that is, $p_1^1 = p_6$, $p_2^1 = p_5$, and $p_3^1 = p_1$. According to the function 3.4, the value of expertise of a_1 for the paper p_2 in this graph is calculated as follows.

$$\begin{aligned} Exp(1, p_2) &= \frac{W(p_1^1, p_2)}{L(p_1^1, p_2)^2} + \frac{W(p_2^1, p_2)}{L(p_2^1, p_2)^2} + \frac{W(p_2, p_3^1)}{L(p_2, p_3^1)^2}, \\ &= \frac{W(p_6, p_2)}{L(p_6, p_2)^2} + \frac{W(p_5, p_2)}{L(p_5, p_2)^2} + \frac{W(p_2, p_1)}{L(p_2, p_1)^2} \\ &= 2.5 + 4 + 4 \\ &= 10.5. \end{aligned}$$

Note that the first sum in 3.4 is applied to the author's papers that are more recent than the paper q, and the second sum otherwise. The two terms of the sum in this function are necessary because the RRPs are directed. Also, observe that when $p_i^j = q$, the function δ is equal to zero.

Next, we show how to take into account this new value of expertise in our collaborative system, which introduces a modification in the expression 3.3, explained in Section 3.2.

First, we define the function $\alpha(t)$ that depends on time (number of votes in the system) by the following expression,

$$\alpha(t) = \exp\Big(\frac{-t * \ln 2}{|E_R|}\Big),\ 0 \leq \alpha(t) \leq 1, \tag{3.5}$$

where E_R is the references edge set.

The function $\alpha(t)$ above takes values depending on the time t of the system evolution or, in other words, the total number of votes in the system, and the number of references edges in the graph. For example, when $t \ll |E_R|$, that is, the system is in its initial stage, then $\alpha(t)$ tends to 1. Also, when $t = |E_R|$, the system is more evolved and $\alpha(t) = 0.5$. After a long time of evolution, $\alpha(t)$ tends to 0.

For more generality, we can redefine the function of expertise as:

$$Exp(j, q, t) = \alpha(t) * Exp(j) + (1 - \alpha(t)) * Exp(j, q), \tag{3.6}$$

where $Exp(j)$ is as defined in 3.2, $Exp(j, q)$ is as defined in 3.4 and $\alpha(t)$ is as defined in 3.5.

Now, the expression in 3.3 in our collaborative proposal is modified as follows:

$$\omega(q, p_k) = \frac{Y + v_{m+1}(q, p_k) * Exp(m+1, q, t)}{X + Exp(m+1, q, t)} \tag{3.7}$$

where

$$Y = v_1(q, p_k) * Exp(1, q, t) + \ldots + v_m(q, p_k) * Exp(m, q, t)$$

and

$$X = Exp(1, q, t) + \ldots + Exp(m, q, t).$$

In this way, our collaborative model contemplates two types of expertise, global and local for a specific paper, so that the evolution of the system implies a change in the values of the authors' expertise via citation path. The system users, as part of the scientific community, indirectly help to improve the estimates of the expertise of the authors in the graph, through the ranking of references being established by them in the collaborative process.

3.5 Constructing and Updating the Citation Graph

In Section 3.1 we define a model that considers a set of nodes P and edges. To determine the papers of this set P it is possible to perform a search on 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. The edges of the graph would be determined from the relationship of citations between the papers. Building a graph with all this information is a complex task given the large volume of information to be considered. Also, it is not known in an easy way if really good papers were chosen to build the graph, in the sense of contributing new knowledge to the subject. In our approach we chose to initiate the system using a set of papers mentioned in a survey of a specific subject. This set of papers were selected as important by the survey's authors. We are going to shown this procedure below.

Graph Construction in 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, $i \in \{1, \ldots, n\}$ is such that p_i addresses the subject T and S has the paper p_i in its reference list, that is, $S \simeq p_i$.
2) Consider E_R as defined in expression 3.1.
3) Also, establish $\omega(e) = 0$ for each e in E_R.

4) Finally, build the *Citation Graph* with the triplex $C = \langle P, E_R, 0\rangle$.

Note that in step 4 the last value of the triplex with zero value corresponds to the value of the weight function ω.

With the previous methodology 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. For example, in Section 3.2 we explain how edge weights are modified with the collaboration of the users. Next, we propose a procedure to update the *Citation Graph* when new nodes and their corresponding edges are added to the graph.

Update the Node Set in the Graph

Let $C = \langle P, E_R, \omega\rangle$ a *Citation Graph* and q a published paper. If $q \notin P$ is considered one of the relevant references of any paper in P or any paper in P is considered one of the relevant references of q, then consider:

1) $P = P \cup \{q\}$.
2) $E_1 = \{(q, p_i) \ / \ p_i \in P, q \simeq p_i\}$, $E_2 = \{(p_i, q) \ / \ p_i \in P, p_i \simeq q\}$, $E_q = E_1 \cup E_2$.
3) $E_R = E_R \cup E_q$.
4) $\omega(e) = 0$ for each e in E_q.
5) Update the *Citation Graph* with the new sets P and E_R and the weight function ω.

After this process, the user who considered including the published paper q in the graph can establish a new weight to the outgoing edges of q, as explained in Section 3.2. For example, suppose that in Figure 3.2 an expert considers that p_4 does not have its most influential references represented in the graph. This is checked by the low value of weight in the edge (p_4, p_2), which is 2, given by other users. In this case, the expert who wants to contribute with the extension of the graph must first add the new paper to the system and then give the vote in the new edge created. In Figure 3.8 below, the new paper p_9(reference of p_4) was added to the graph and a value of 5 was established for the new edge (p_4, p_9). This process is repeated in case the user wants to include more than one paper in the graph.

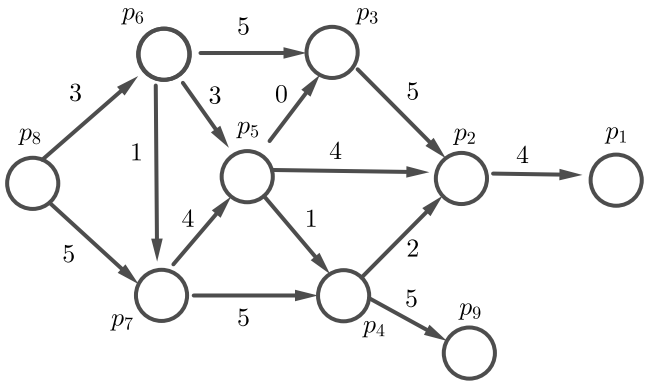


Figure 3.8: Add a New Node and Set a Weight Value in the Edge

Modifications or updates in the graph can be made in our system with a visual interface. This interface allows visualizing the *Citation Graph*, so that the user can extract knowledge in an easier way and it is possible to interact with the system, for example, to modify the value of weight in relation to an edge. The visualization of the information contained in the graph, as well as the interaction and exploration of the system will be discussed in the next chapter.

4
Information Visualization

Our system supports the visualization of citation threads on scientific papers. For this purpose, it has a visual structure, in the form of a directed acyclic graph, which helps users to track publications of interest and understand the growth of a line of research, showing the most relevant and influential papers ordered by their publication date. This visual structure is used to map the data set, which contains the papers and citation relationships, and their location in the display area, as seen in [41]. In this chapter we present the visual mapping of the data and the interaction and exploration mechanisms in the system. We consider that all the papers represented in the graph belong to the same area of study.

4.1
Visual Mapping

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.

In this configuration, the nodes of the graph, which represent papers, are initially associated with blue circles when the out-degree or the number of outgoing edges of the corresponding node is greater than zero, as shown in Figure 4.1.

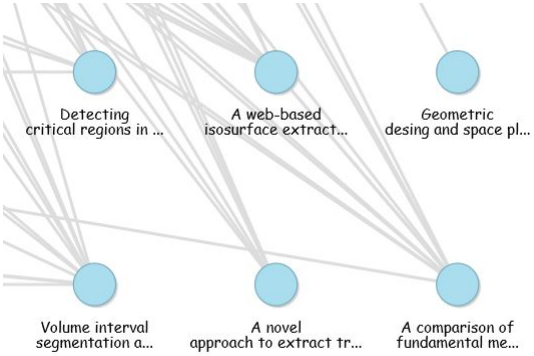


Figure 4.1: Circle Nodes

The nodes with out-degree equal to zero are initially associated with a blue triangle down, as seen in Figure 4.2. The reason for this change of shape is that in our system we consider it important to distinguish the papers that represent either the beginning of a research line or papers disconnected from the graph because they do not reference any other paper contained in the graph. For example, a highly cited paper that does not reference other papers included in the graph can be considered as the initiator of a branch of study. In the same way, the papers that are not cited by other papers in the graph and do not have their references included in the graph are considered to be from a different area of study to the papers in the data set.

Figure 4.2: Triangle-Down Node

The nodes can have different sizes. This indicates the in-degree or the number of incident edges of the corresponding node. The in-degree of a node (paper) is the number of times that this paper is cited. In this way, the papers with the most impact on a branch of study are visually identified with larger shapes. In the example of Figure 4.3, the middle node is larger because it has a higher in-degree.

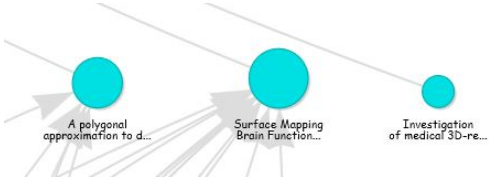


Figure 4.3: Size of nodes

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 gray line, from a citing paper to a cited paper, as shown in Figure 4.4.

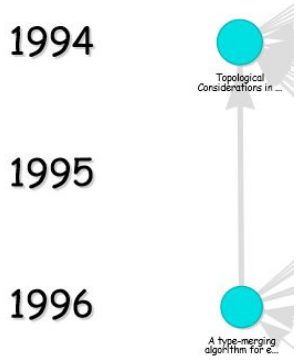


Figure 4.4: Reference edge

Thus, the visualization presented in this work is based on four characteristics:

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

One Node for each Paper. Each node appears only once in the representation. A paper is not represented by multiple nodes in the graph because it would increase clutter and make visualization difficult.

Hierarchical Structure by Publication Year. The graph is organized in levels, defined according to the publication year of the papers. In this visualization, a natural way to observe the progress of different lines of research is to have a hierarchical structure, as observed in [18]. 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. We are starting from the premise that a paper cannot reference a more recent paper. For that reason, in the visualization the vertical position of nodes reflects the publication year, so that those published in a specific year will be on the same level, as seen in Figure 4.5. We chose the position of the nodes in this way to obtain a representation with a much finer granularity and remove the edges as much as possible between nodes on the same level, that is, side-way connections.

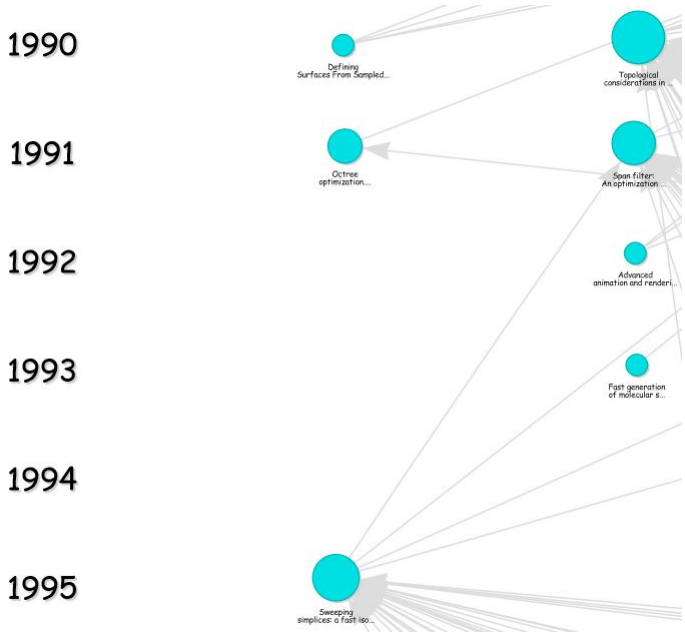


Figure 4.5: Levels by publication year

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, the k most influential references are represented in the graph.

Our system allows visualizing the *Citation Graph* with the most relevant references by node through edge filtering. Figure 4.6 illustrates a part of this graph where five of the most important references for each paper are displayed. In Figure 4.7, a reduction to two relevant references is observed as a result of applying a filter in the edges.

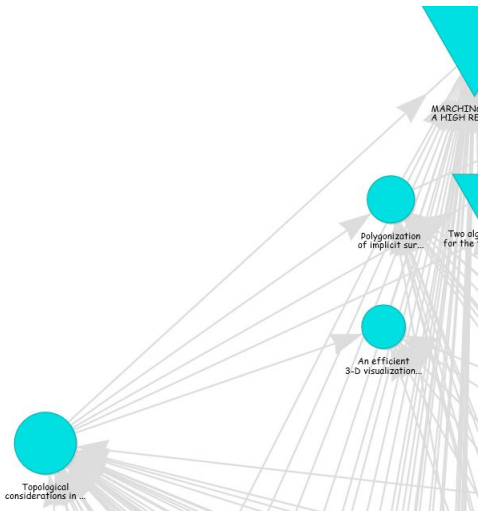


Figure 4.6: Top five most relevant references

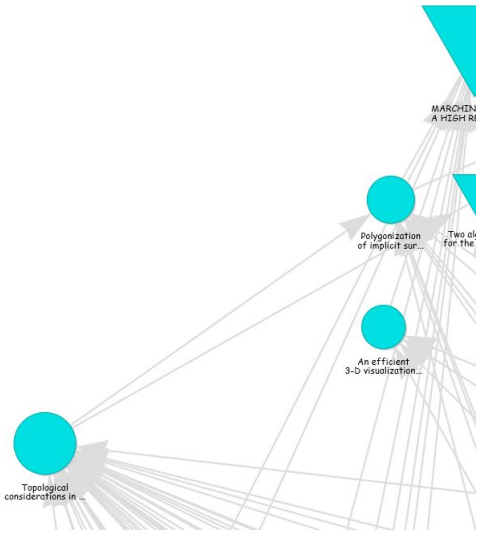


Figure 4.7: Top two most relevant references

4.2 Exploration and Interactive Visualization

The graphic representation proposed in this work allows user interaction and exploration. It is possible to generate other views by zooming, as the example in Figure 4.8, and dragging the nodes, as the example in Figure 4.9.

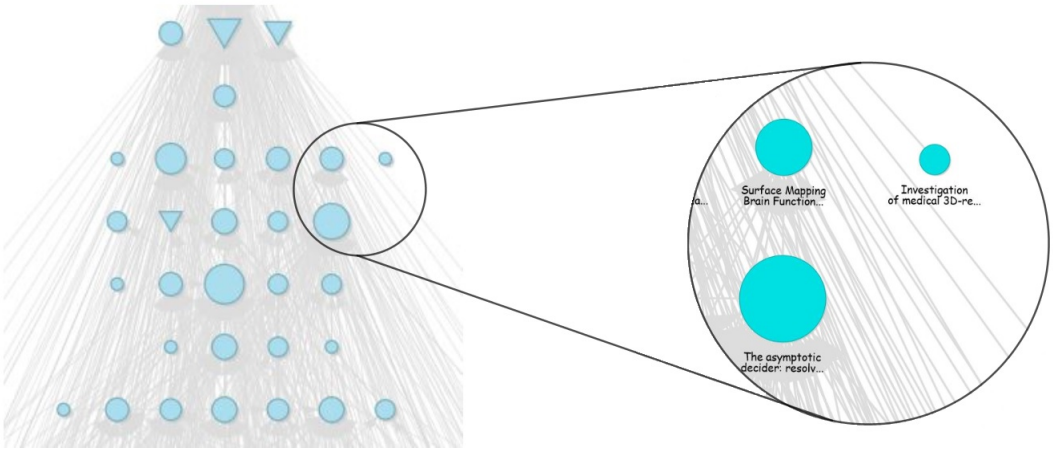


Figure 4.8: Zoom

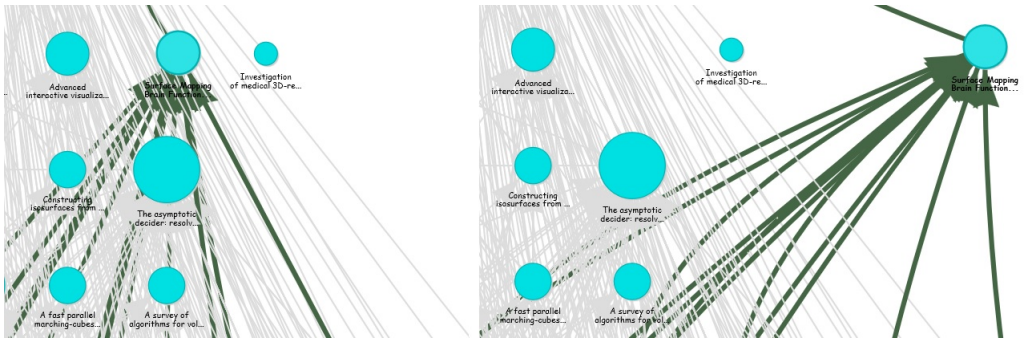
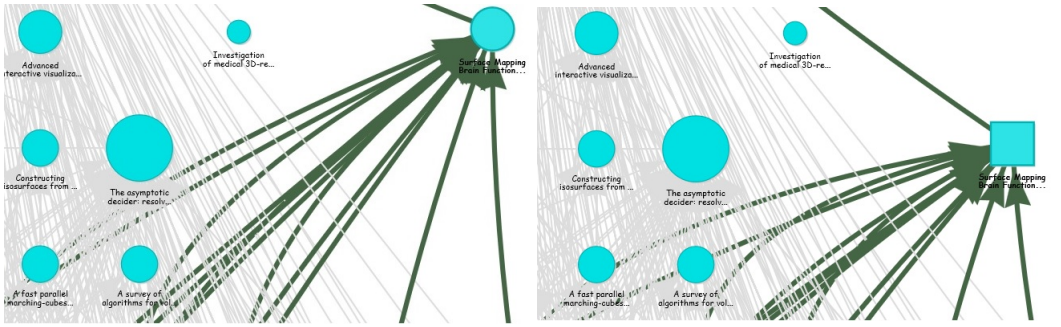


Figure 4.9: Drag

The nodes that are at their corresponding level remain with the shape initially defined – a circle –, whereas those that are dropped outside their level – through user interaction, such as dragging – will change their shape to a square. For example, in Figure 4.12, the paper represented by a square is at a level (year) below that of its original position. When a user double clicks on the square node, the node returns to its original level in the hierarchy.



Nodes at their original levels A node (square shape) at a different level

Figure 4.12: Node positioned at a different level from the original, calculated level

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 (paper of interest) is highlighted in orange and with a thicker border. For more clarity, the papers cited in the selected publication (*aka* Reference nodes) are highlighted in dark green and the papers that cite this publication (*aka* Citation nodes) are highlighted in light green, as shown in Figure 4.16, which facilitates the exploration in the graph. Figure 4.17 shows these graphic elements in the system.

Figure 4.16: Selection of colors for the Selected node, Reference nodes, and Citation nodes

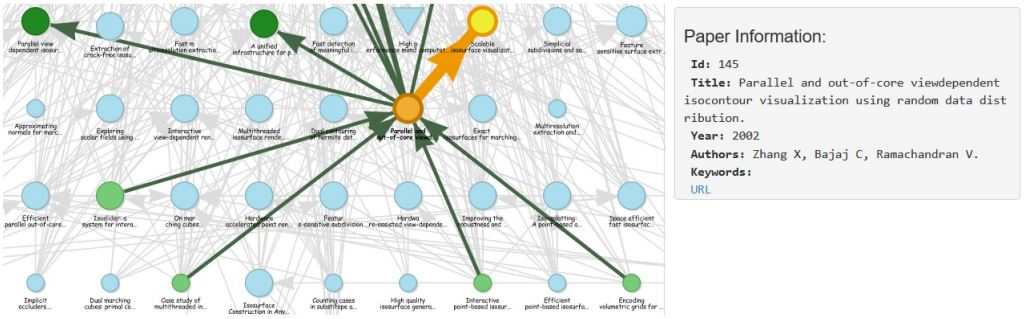


Figure 4.17: Additional information about the selected Node

We also distinguish the edges associated with the selected node, as seen in Figure 4.17 above. More precisely, the edge associated with the selected node is highlighted in green. The edge that represents the most relevant reference of the selected node is highlighted in orange. Also, the node that represents the most relevant reference is highlighted in yellow with a thicker border and orange color border.

In addition, a second network is created maintaining a hierarchical structure to better analyze and visualize only the references of a selected node. The node that is lower (bottommost level of the graph) corresponds to the node selected in the original graph. In Figure 4.18, the reference relationships of the node selected in the example of Figure 4.17 are shown.

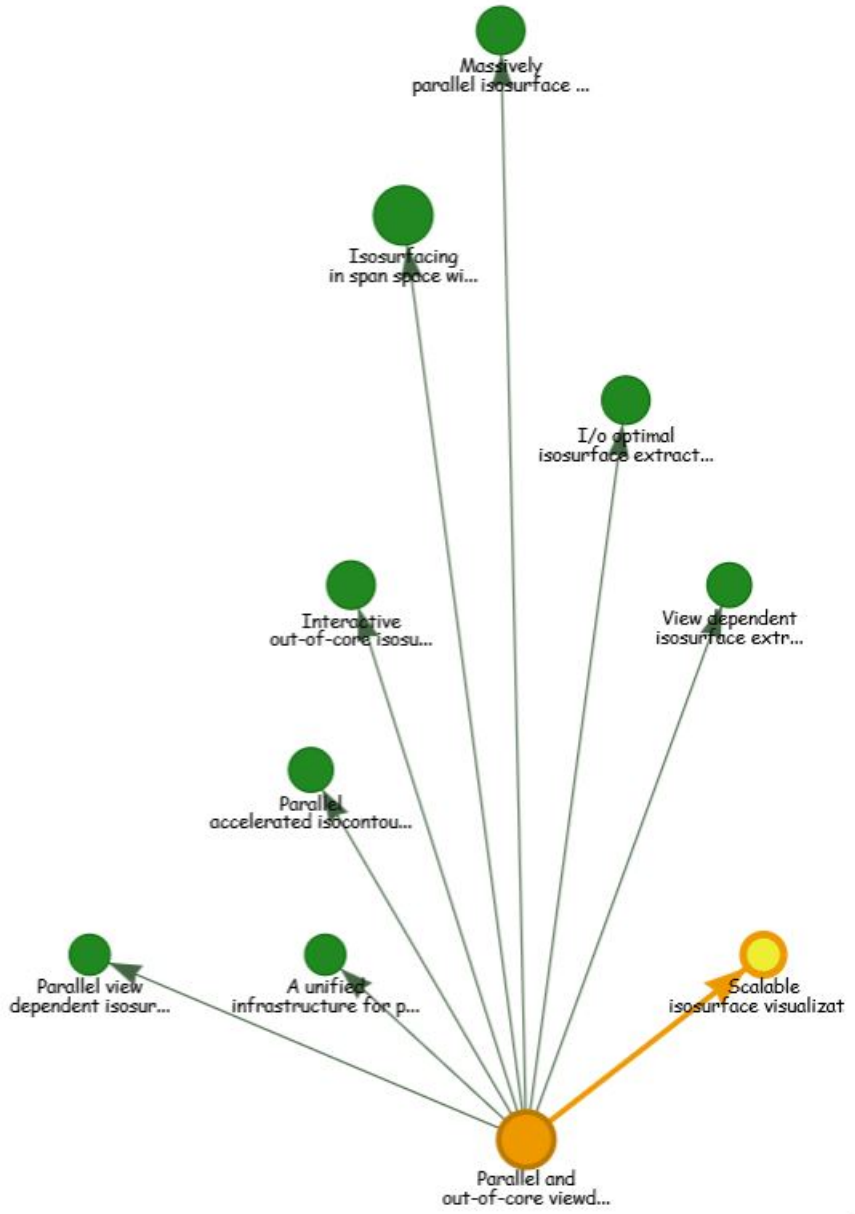


Figure 4.18: Simplified View for the References of a Selected Node

Besides, the users can view the general weight of the edges, which results from the collaborative process explained in Section 3.2. The users can also view, in the second network, the most recent vote given by them to each reference and, by selecting an edge, the number of users who voted on this edge. In Figure 4.19 we can see the two coordinated networks. In the network on the left, a tooltip with the weight or collaborative rating of an edge is shown. In the network on the right, this and additional information of the edge are also shown. The weights of the edges define the ranking of references of the selected node.

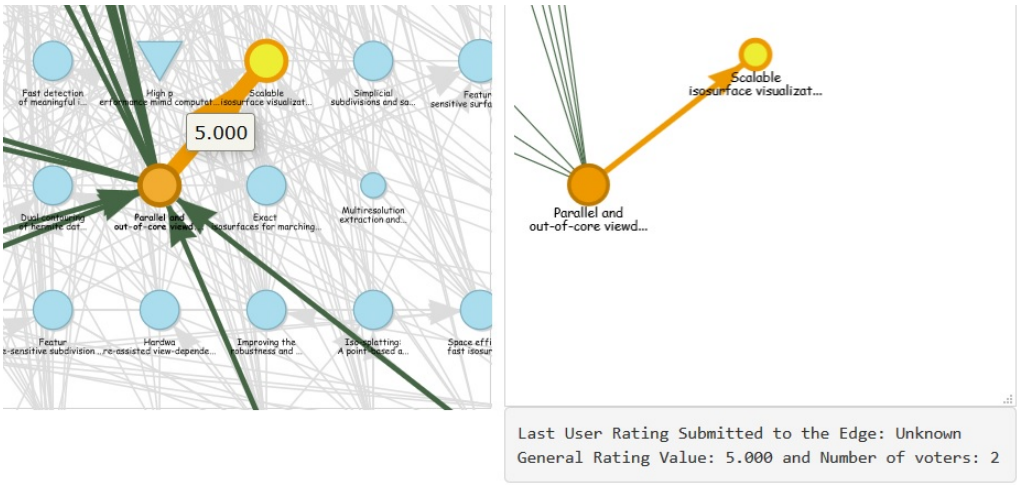


Figure 4.19: Weight or General Rating of the Edge

We consider that highlighting in orange the edge of greater weight for each node in the graph offers simplicity in the visual recommendation and better 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. By ranking and filtering references, the user can easily interpret the resulting visualization.

A registered user who is the author of any publication in the graph can establish, according to their opinion, a ranking of references of each target paper through voting. This can modify the general weight of the edges involved and, consequently, a change in the final ranking obtained collaboratively from the references of this paper. The Figure 4.20 below shows how the user interactively provides his/her rating to a selected edge.

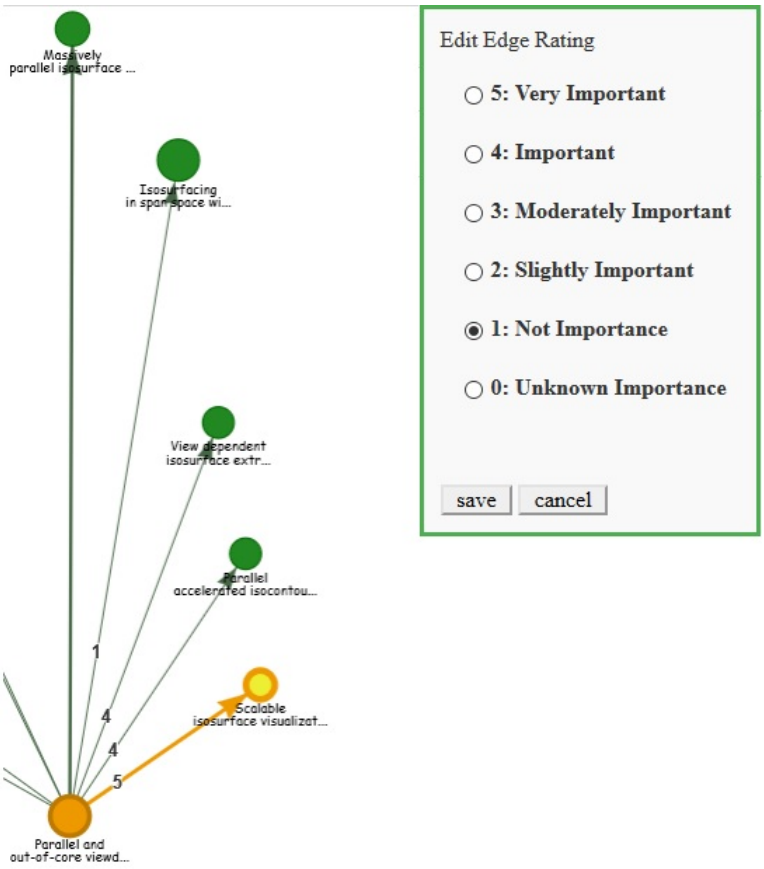


Figure 4.20: Rating an edge

Another functionality of our system is that a document can be found by searching by DOI, author's name, or title, as shown in Figure 4.21, or by visually inspecting the most cited nodes in the area. One can see, for example, that most of an author's papers may be on one or more paths or branches of study.

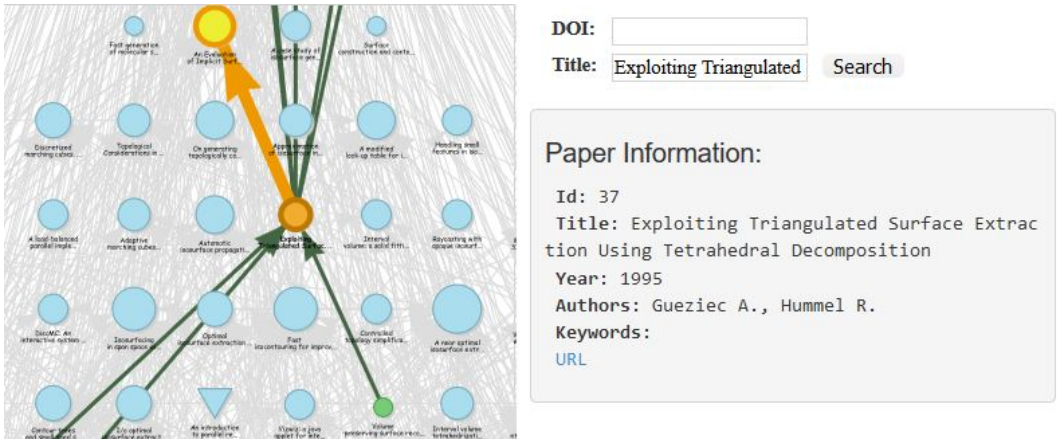


Figure 4.21: Search by title

Finally, we argue that the exploration of the graph to find a specific publication is more convenient in our approach. The user will not only find publications on the same subject, but he/she will have the most relevant or influential papers, in sequential order of citations.

5
Applications of the System

Analyzing citation threads is a relevant challenge. Our visualization can play a important role in facing it, for example, by providing support to the task of ensuring proper coverage of relevant material when conducting literature surveys. The usefulness of this visualization is to support not only the authors of scientific articles, but also their reviewers.

In this chapter we show some examples where we illustrate applications of the system, such as the ranking of references and expert finding.

5.1
Supporting Authors of Scientific Papers

One important application of our system is to determine the references that most impacted the research of a given paper. This procedure for each paper of the graph produces an overview of the state of the art in the subject. For example, in the literature review of the thesis in [42], the author illustrates with a diagram the most influential papers and the different branches that these papers have evolved with their research (see Figure 5.1). Our system allows to visualize the state of the art of a specific topic. For this purpose, we propose to make a ranking of references for each paper in the *Citation Graph*. This ranking will be built by a collaborative process explained in Chapter 3, where specialists define their votes. In the initialization of the system we opted for an automatic method for ranking because even the edges do not have an associated weight value given by the experts. These procedures are explained in more detail below.

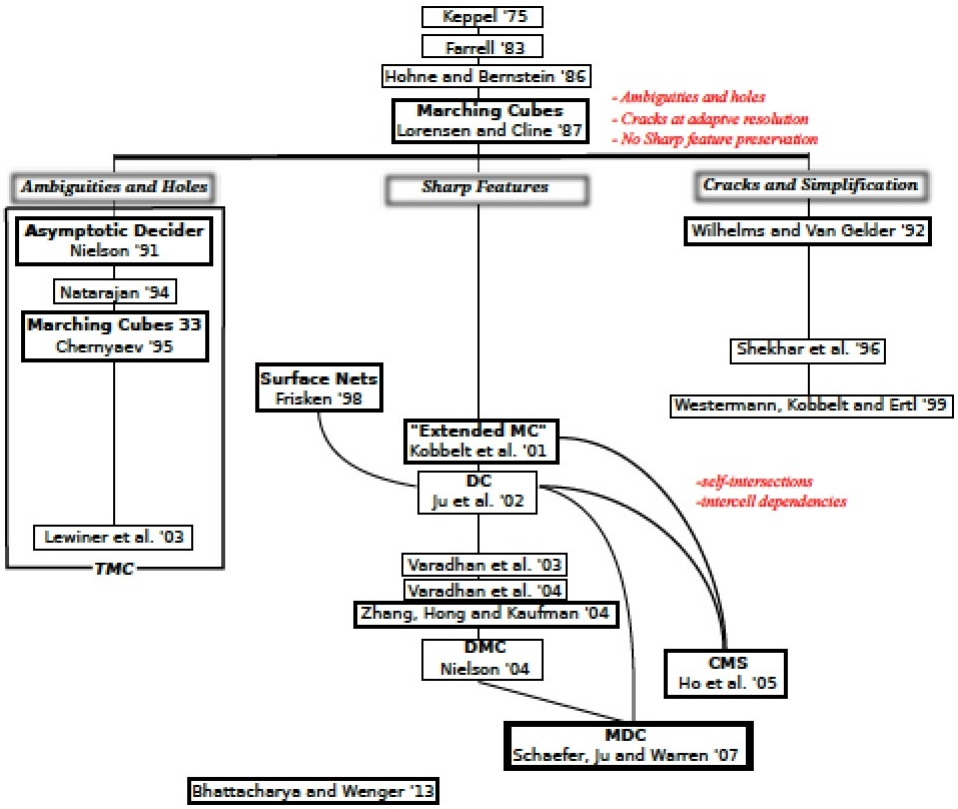


Figure 5.1: Overview of the State of the Art of a Marching Cubes subject. Source: George Rassovsky, 2014

5.1.1 Ranking References without User Collaboration

In the initial stage of our system, each edge of the *Citation Graph* will have zero weight. Therefore, for each paper, all its references have the same relevance in the visualization. Then, to differentiate the references of each paper, we rank them automatically. This differentiation can be done by a topic analysis, for example Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), explained in [33] and [43], respectively, where the reference of the specific paper with more cosine similarity or topic similarity with this paper would be the most relevant. Performing the ranking by this latter approach requires a large number of operations and calculations. For this reason, we propose a simpler process with fewer operations, where this differentiation can be done by the date of publication and the number of citations. For example, these references can be sorted in decreasing order according to the publication date, and in the case of two publications with the same publication date, the references with the highest number of citations will have preference. Even without the user collaboration, this ranking would allow establishing an initial

visual recommendation of papers for authors.

For example, in Figure 5.2 there are eight papers distributed in the seven-year period (2007-2013). Paper p_8 was published in 2013, papers p_7 and p_6 in 2012, and paper p_5 in 2011. Then, the ranking of references of p_8 is the ordered set $\{p_7, p_6\}$, because p_7 is more recent than p_6. Note that p_7 references p_6. Consequently, p_7 is the first visually recommended to a user who is interested in the scientific research of p_8, as p_7 is potentially influential in the development of p_8's research. Paper p_6 is considered potentially the most influential for p_7 and the same is considered for p_5 in relation to p_6.

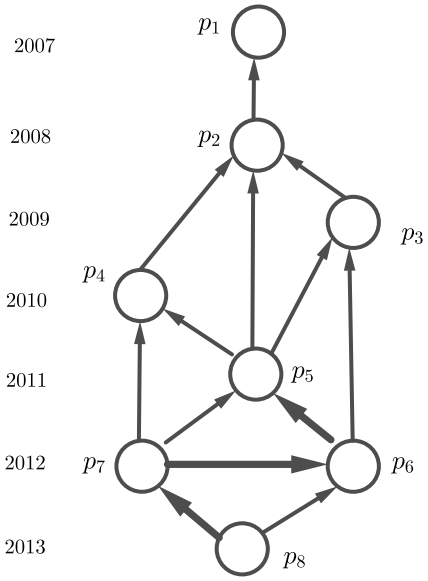


Figure 5.2: Highlight the most influential reference of p_8, p_7 and p_6

This process is a first approximation of finding the most influential papers that define a line of research, forming a portrait of the history of art on a specific subject. Only with the collaborative opinion of experts the system can improve the ranking of references and offer a better recommendation to the user, independent of LDA, LSA, or another automatic method chosen in the initial stage of the system.

5.1.2 Ranking References With the Collaborative Approach

Following the procedure described in Section 3.2, it is possible to obtain a better recommendation of references than the one previously calculated, by allowing the vote of the users who are experts in the subject.

Figure 5.3 shows the evolution of the *Citation Graph* of the previous example after a time t, where only the edge (p_5, p_4) has not yet received votes from users. The weight of the edges is the result of the collaborative process, which assigns more credibility to the ranking of references.

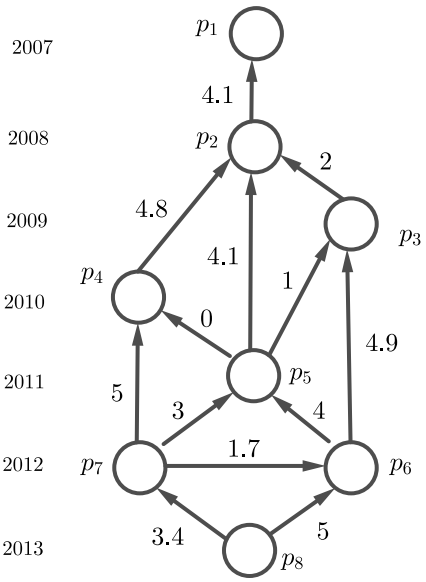


Figure 5.3: Citations Graph after the Collaborative Process

Note that, in the example of Figure 5.2, p_5 was the most relevant reference for p_6; however, now the scientific community has determined that p_3 is really the most relevant reference for this paper.

The ranking of references allows us to establish the RRP, according to Definition 3.5. This path can be used to observe relevant papers in a line of research, ordered by date of publication and following a sequential order of progress in a specific investigation. The user would obtain a visual recommendation of which scientific documents should be studied or considered for inclusion in the list of references in their future papers from this path. These recommended papers directly reflect the evolution of the research in this branch of study.

For example, in Figure 5.3, the reference more influential in the research of p_8 is p_6. Also, p_3, p_2, and p_1 are the papers that will likely help the user to understand the origin of the p_8's research. These publications are obtained through the RRP in the *Max-Generator Graph* shown in Figure 5.4.

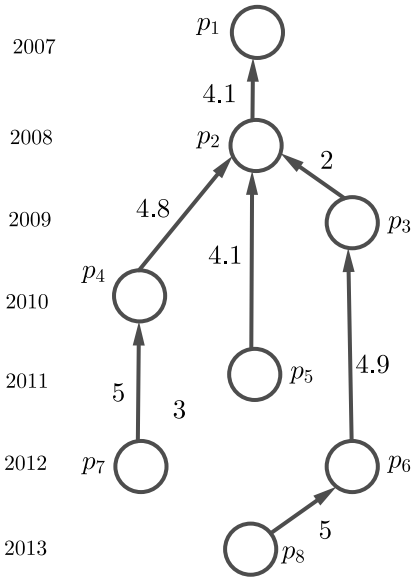


Figure 5.4: Max-Generator Graph

On the other hand, if the author wants, he/she can choose to study the top k most relevant references of a paper, and analyze the papers that are in the RRP from each of these references. It is important to observe that the result is a recommendation for the user, and in no way undervalues other publications. For example, if the user wants another path of relevant citations from p_8 for the references recommendation, then the top second paper, for example, in the ranking of references of p_8 can be chosen as the most valued, which in this case is p_7, and continue with this in-depth search process.

In the process to establish the paths, eventually we can have more than one reference with the same maximum weight. To deal with this, we proposed to perform an automatic ranking of these references with the same weight considering the publication date and number of citations, as explained in Section 5.1.1.

The visual properties of our system are also very useful in this task of recommendation. For example, in Figure 5.3, the paper p_1 is considered the initiator of the branch of study, the paper p_8 has not yet received any citations and the most cited paper is p_2. The position of the paper allows to visually determine whether a paper cites recent and classical references. For example, p_5 cites a recent paper p_4 and a seminal paper p_2. The timeline to the left of the graph is very useful for this purpose. Now, suppose that an author considers contributing with the research of paper p_4. Then p_4 must be included in the reference list of the future paper. The author determines from the visualization which other references are important to study or include in

this list, following the recommendation in the RRP from p_4. In addition, it is possible to verify whether this paper is being cited as a relevant reference in a new contribution. Also, the authors with investigative interests in the subject of the graph in general can identify in our visualization the papers that represent good candidate sources of information, for example p_2, which is the most cited.

5.1.3 Adding New Functionality to Existing Tools

Our proposal improves the current practices of looking for potentially relevant papers with respect to a target paper of a certain topic. Usually there would be several options, as follows:

- Searching in scientific databases the most cited papers that cite the target paper.
- Searching the target paper references for possible important papers and continuing the same process with these references.
- Search for papers with the same topic as in the target paper.

For example, in the web search engine for scholarly literature, Google Scholar[1], when performing a search for relevant papers in the topic "marching cubes"[2], these papers appear in the top 3 as a list, "Marching cubes: A high resolution 3D surface construction algorithm" (published in 1987), with 14,834 citations; "The asymptotic decider: resolving the ambiguity in marching cubes" (published in 1991), with 741 citations, and "Discretized marching cubes" (published in 1994), with 286 citations.[3] Our system can provide this result in a visual way, positioned according to the publication date, where the size of the node that represents the paper "Marching cubes: A high resolution 3D surface construction algorithm" would have the greatest value. In this way the information is better understood by the user (see Figure 6.8 in Chapter 6).

By contrast, searching with Google Scholar papers relevant to a target paper is not an easy task, because we do not have a ranking of references for this purpose. In addition, in general the most cited paper of those citing a target paper is usually considered relevant to the research line, but it is not always so. This is the case, for example, of the paper "Features and Development of Coot" (published in 2010), which is the most cited according to Google Scholar of those who cite the paper "Marching cubes: A high resolution 3D surface

[1] `https://scholar.google.com`
[2] 'Marching cubes' is a computer graphic topic on the extraction of isosurfaces.
[3] Information provided by Google Scholar, 3 February 2019.

construction algorithm". However, this paper is not considered relevant in the topic of "marching cubes". The paper "Features and Development of Coot" presents a model for building and validation of biological macro molecules. It is an application of "marching cubes", but it does not advance the research of isosurfaces extraction. In our approach, the edge between these two papers would probably have a low weight value.

Another possible application of our work would be to introduce these visual features to the ACM Digital Library (DL)[4]. In this way, users would have visual support in the search for relevant references. In the DL, as shown in Figure 5.5 below[5], although there is a visualization in the form of a graph for the references and a filter of papers for several categories, it would be useful to consider our proposal of hierarchical visualization layered by publication year and with an identification of relevant references for each paper, as well as a collaborative building of the state of the art of a specific research topic.

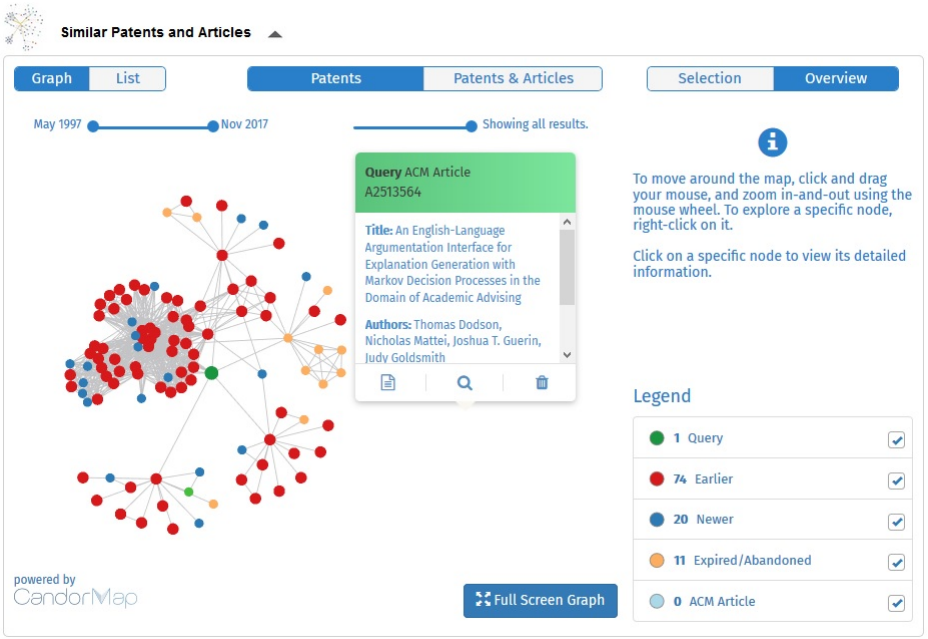


Figure 5.5: ACM Digital Library

5.2 Supporting an Editorial Board

The system proposed in this work can help in the decision making of an editorial board of a scientific journal. The heads of the editorial board can find experts or candidate reviewers for the research of a specific paper. Also,

[4]The ACM Digital Library `https://dl.acm.org` is a comprehensive database of articles and bibliographic literature covering computing and information technology.

[5]`https://dl.acm.org/citation.cfm?id=2513564`

these reviewers can verify how appropriate the selection of the references of the papers submitted to the journal was. This application is explained in more detail next.

Selecting a suitable reviewer is a key step to ensure the quality of the peer review process, but it is laborious to decide which reviewer has enough knowledge of the research areas related to papers [7]. The process of recruiting reviewers is time consuming and it is complex to be done manually [44]. There are tools to reduce the task of editors to find experts such as Conference Management Systems, which are used to invite appropriate reviewers, assigning the papers based on reviewer bidding preferences. However, the preferences of a reviewer may not be essentially consistent with his/her expertise. Thus, the reviewer may get papers for which he/she is not an expert [45].

Also, suppose that a certain paper is not among the preferences of the reviewers, or simply none of the reviewers in the reviewer pool is highly experienced to review that particular paper. The editor-in-chief can then recruit additional external reviewers. One way to do this task would be to ask the known reviewers to suggest some others with a similar research background or to ask some authors of published papers whether they want to become reviewers and give their evaluation regarding that paper, following the same procedure as in the review process at NIPS 2016 [44]. To form a pool with all possible reviewers would yield too many options, and performing information retrieval would not be efficient for the amount of data presented. Even so, the choice of the candidate reviewers can be complicated and there are no guarantees that the new pool will contain relevant reviewers for the paper.

In this sense, the procedure for finding expert for a given paper according to our proposal is the following:

1) The head of the editorial board submits the data of a new scientific paper in to the system, to be (temporarily) included in the graph.
2) The system establishes zero weight value for all the references of the new paper which already included in the graph and ranks them automatically (see Figure 5.6).

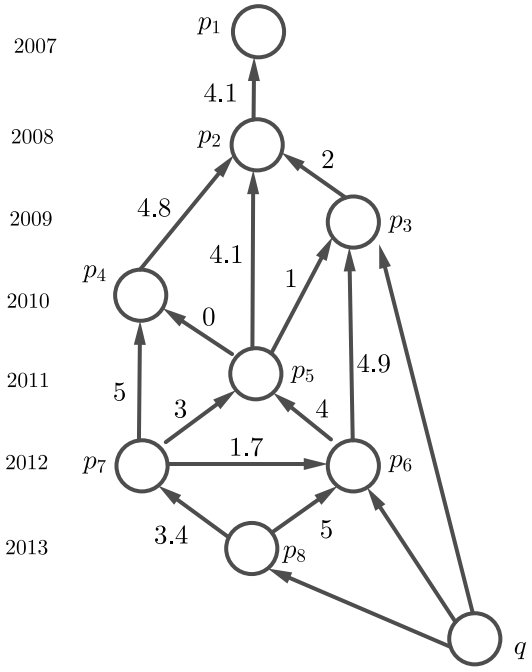


Figure 5.6: New paper q submitted to the scientific journal

3) Considering that several edges in the graph have been voted by the users, the expertise by citation path is calculated for all authors of publications in the RRP between the new paper and any important paper in the graph.

4) The system can provide a list of possible reviewers following the order of their expertise values. The most appropriate reviewers for this paper will have greater expertise.

5) If the paper is accepted by the reviewers for publication, it is included permanently in the graph.

Note that, in step 3, if the new paper does not reference any other paper of the graph (this graph should contain important papers of the subject addresses in the new papers), then the reviewers will not be determined with our proposal. This is consistent with the expected response, since this paper apparently does not relate to the state of the art of the topic. Also, in step 4, authors with conflicts of interest with the new paper should be considered and excluded from the ranking. For example, the author cannot be the reviewer of a publication of its authorship or reviewer of the publications of his/her co-authors.

The editorial board of a journal could verify with this proposal whether the most relevant papers were included in the reference list of the submitted

paper, *i.e.*, verify whether the author failed to include important papers (both seminal and recent) on the topic.

For this application, our system would be managed by the editorial body of a scientific journal specialized in the subject of the papers of the graph. This means that a new paper would be inserted in the graph upon its acceptance for publication. However, the experts could also recommend including some missing papers, or including some references of papers that were not considered in the graph update process.

6
Discussion

This chapter discusses the main results of this thesis. First we begin by explaining the data set that we chose to illustrate these results. We then explain some details of the computational implementation of our system and the resulting visual interface, which is essential in this work. We also show several use cases and the results obtained with the use of our proposal. Finally, we conducted a user study to evaluate the perceived usability, usefulness, and effectiveness of the system.

6.1
Data set

The data collection we chose contains a particular selection of important articles that address the issue of "marching cubes". We called this data set VisMC. All these papers make up the list of references of a survey paper published in 2006, "A survey of the marching cubes algorithm", by Timothy S. Newman and Hong Yi [46]. The data is in Comma-Separated Values format (CSV), where each row identifies a scientific document and the columns describe its metadata: identification number (id), paper title (title), Digital Object Identifier (DOI), paper authors (authors), publication year (year), paper abstract (abstract), keywords, uniform resource locator (url), and the reference by the identifier in the data set (ref-id). With the id, each paper will have a unique key in the data set. The DOI attribute was not selected as a unique key because not all papers in the set contain this information.

6.2
System Implementation

The system was conceived with four modules or main components, which are reusable pieces of software. These modules are specified below.

Data Acquisition, Extraction and Data Cleaning

In this module we developed a semi-automatic process to collect the papers that make up the data set and their corresponding information that will be useful for our system. We chose the Python programming language [47] for

this module because this language has excellent tools for extracting information from PDF documents (Portable Document Format). First we obtain the Survey in PDF format from a specific URL. Then we extract the text with Python using a library such as *pdfminer*, and we identify the references of this paper and assign an identification (id) to each one. Finally, we search in Google for all these references and we identify the relationship of citations between them (ref-id), as shown in Figure 6.1. In this way, for each reference of a paper, a search is made in the data set by the pair reference title - year to obtain the identification of this reference (id).

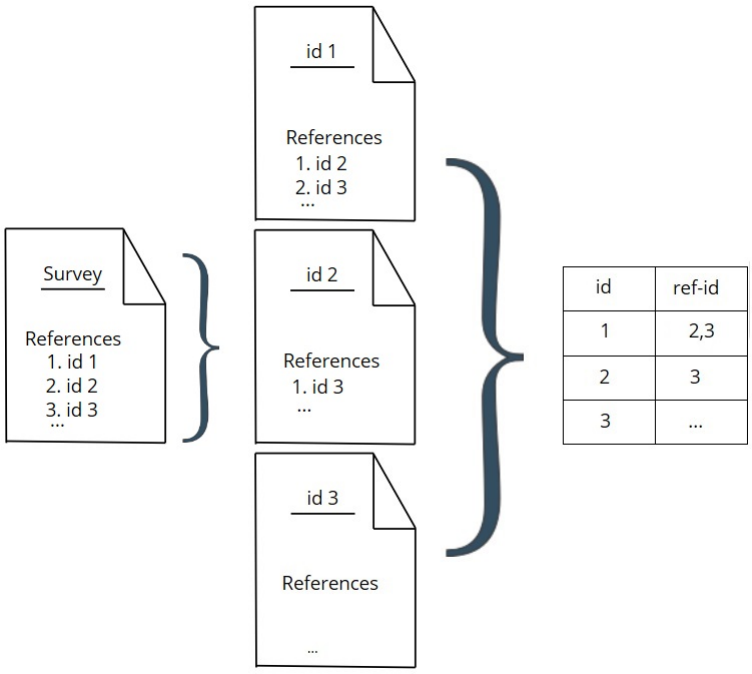


Figure 6.1: Extraction of Information

Although this information can be extracted entirely in an automated way, as explained in [48], was done with human intervention to reduce some errors in the data. For example, the paper "Deformable volumetric model and isosurface: Exploring a new approach for surface boundary construction" has among its references the paper "An evaluation of implicit surface tilers" with two typos, which causes a change in this reference's title in the following way: "An evalulation of implicit surface tillers". These typos make it difficult to automatically identify which papers in the data set are related to one another.

In addition, information was detected and corrected in the data set; for example, some authors with compound names were cited in some cases with their first name and in other cases with their first and second names. To correct

this, an id was established for each author and a list was associated with all variations of names found for this author.

Finally, we disregard cited papers that are not published, identified with "To appear" in the text, and other documents that did not cite or were cited by any paper of the set.

In this way, the VisMC data set was established with 189 papers and 2,334 citations relationship between them.

Modeling by Graph

To build the citation network that defines the graph, following the procedure described in Section 3.5, and to store the information collected, we chose Neo4j [25] as a suitable NoSQL [1] database for this purpose. We chose Neo4j because, although there are other data-oriented databases, such as OrientDB [2], Neo4j has the most information available on the Internet, and it also has Cypher, a declarative query language similar to SQL (Structured Query Language) and which is very easy to learn. The traditional relational databases are not appropriate for modeling graphs because retrieving data row by row results in having expensive joins to query the information. In contrast, graph databases such as Neo4j have adjacency information from the nodes, not requiring complex joins to retrieve connected data. Also, it has a user interface that returns the results using an interactive graph visualization. For these reasons, we implemented a procedure to import the data set to Neo4j using Cypher.

Visualization and Interaction

We decided to implement a dynamic web application in the visualization and interaction module. A web application lends itself well to the process of viewing important papers and the citation relationships between them. Another reason for choosing a web interface is that, due to the collaborative features of our proposal, users at any time and place can interact with our tool.

We chose the Python programming language and its Django framework [49] for developing the web application. The selection is justified because Python has very good libraries to interact with Neo4j, like py2neo. Django was chosen because of its work philosophy and its Model-View-Template (MVT) architecture. It is important to note that it is also possible to interact with the database from the templates and views of Django, with the official Neo4j driver

[1] NoSQL(not only SQL) is a class of database management systems (DBMS).
[2] `https://orientdb.com`

called neo4j-driver. For the interactive visualization, we used the JavaScript library vis.js [50].

The users of our web system are classified as experts in the subject (Registered Users), who are authors of some paper in the data set; non-experts (Non-Registered Users), who are visiting users; and administrator (System Administrator). Registered Users, unlike Non-Registered Users, can vote in a certain relationship. System Administrators can perform other functions, such as adding a new paper or making corrections to the database.

Ranking Algorithms

In this module we implemented the methods to rank references, either by an automatic approach or as a result of expert collaboration. We developed the expertise functions explained in Chapter 3 in order to rank authors. These methods were implemented with Python.

6.3 System Visualization Interface

We show in Figure 6.2 our visualization interface, which is one of the contributions of this work. 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 the user to search for specific papers in the graph.

(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 in the graph and positions them in a hierarchical way.

(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.

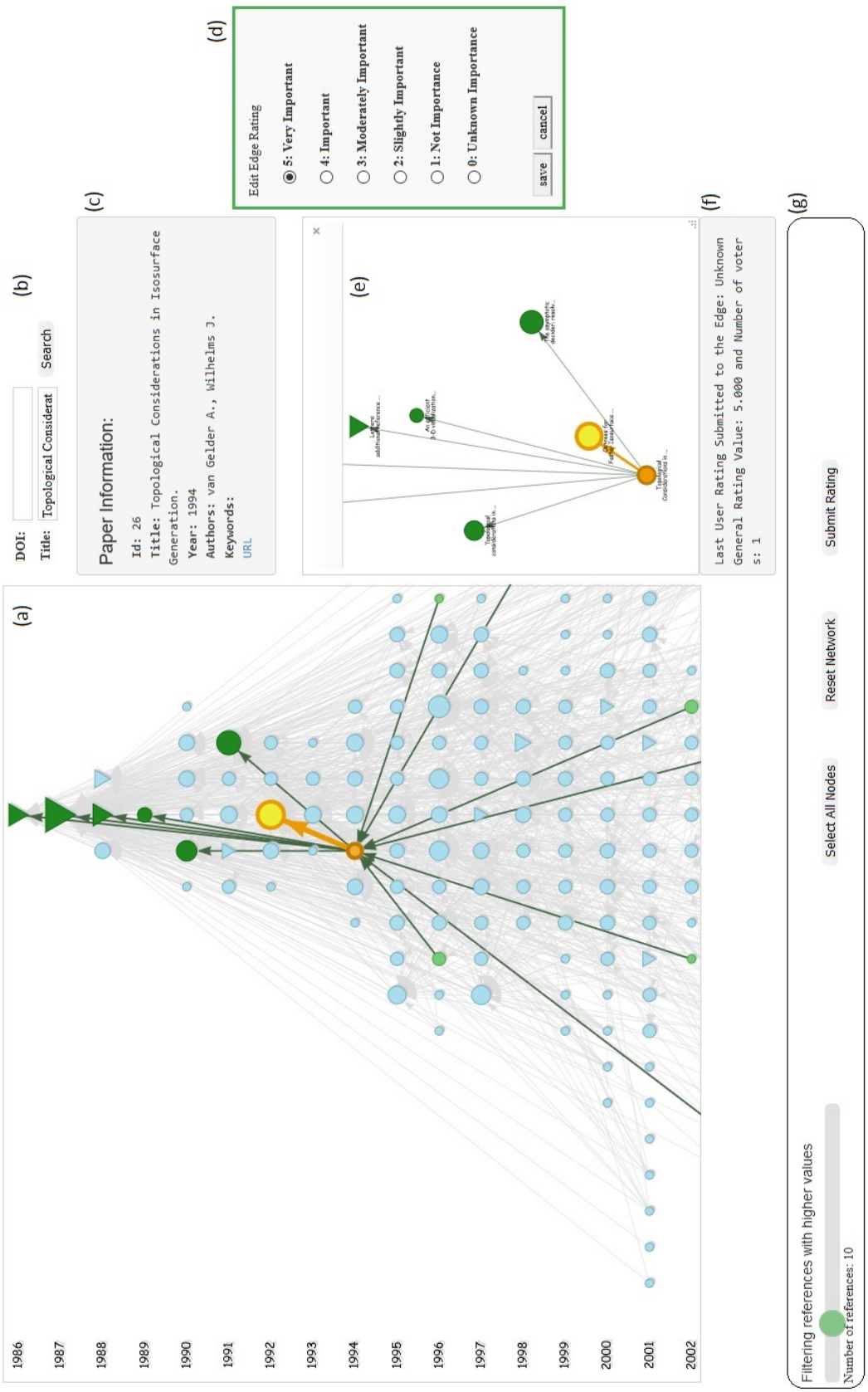


Figure 6.2: The Visualization Interface

6.4 Case Study

In this section we present some case studies that show the main functionalities of our system with the aim of helping to make decisions. First we show some reference visualization tasks and the identification of the root paper of a specific subject, then we capture a picture of the state of the art in the subject and the branches of studies, and finally we show how to identify the authors in an area who are more prolific and with good recognition by the scientific community. We also show how to identify the value of an author's expertise in a particular paper through citation paths.

6.4.1 Visualization of the Research Papers and their References

Our system allows visualizing the papers of a certain set and the relation of citations between them in a graph, as shown in view *a* of Figure 6.2. When a node is selected in this view, only the references related to this node are displayed in the Simplified View, identified with the letter *e* in Figure 6.2. In this way, it is possible to identify the references in time intervals and to visualize their quantity and quality.

For example, the selected paper "Improving the robustness and accuracy of the marching cubes algorithm for isosurfacing", published in 2003 by A. Lopes and K. Brodlie, has 16 references, 12 of which recorded in the data set. These 12 references are visually identified in the graph, as shown in Figure 6.3. They are distributed over nine levels of different years, as shown in Figure 6.4.

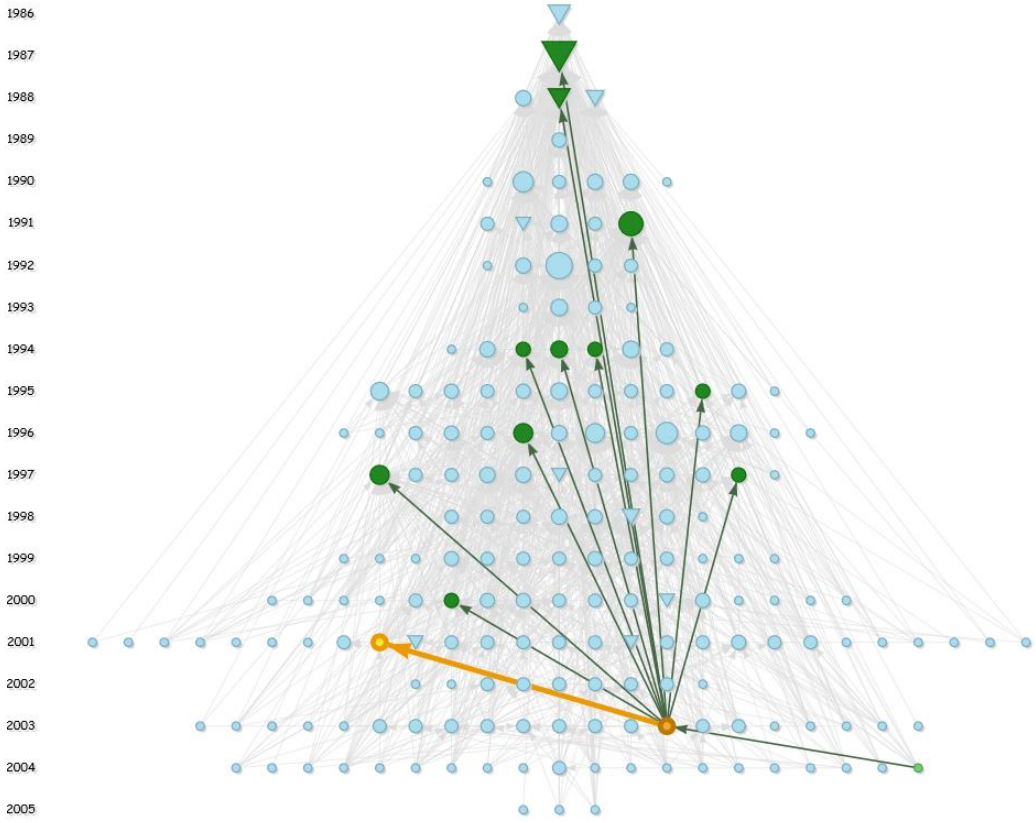


Figure 6.3: Many References Identified in the Graph

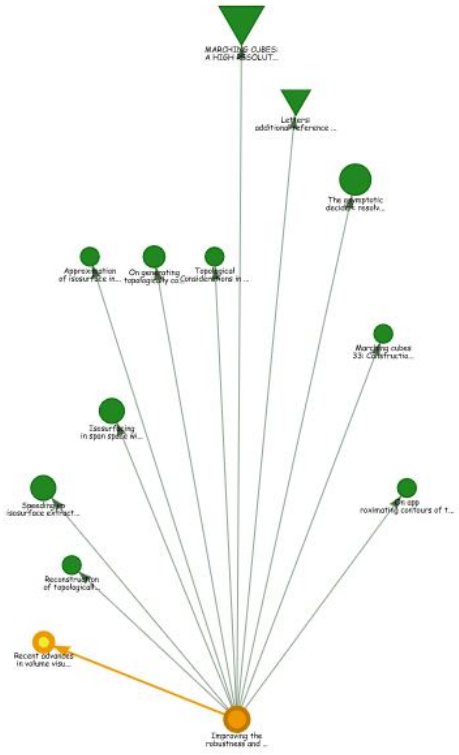


Figure 6.4: References Positioned at Levels Defined by Years

As usual, the paper cites papers from different years, and it is easy to visual identify that it cites two recent papers, whilst the other references are distributed over years prior to 2000. Figure 6.4 also shows that the paper references three publications of the same year. In addition, the selected paper references two of the most cited papers in the graph, identified by larger node sizes. The two most cited papers are "Marching cubes: A high resolution 3D surface construction algorithm" (published in 1987) and "The asymptotic decider: resolving the ambiguity in marching cubes" (published in 1991).

By contrast, in the example of Figure 6.5, the paper "On marching cubes" by G. M. Nielson, also published in 2003, has only two references in the graph. Through the interface one can see that these papers have very old publication dates and they are the two most cited publications in the graph. The papers are positioned at different levels in the simplified view, as shown in Figure 6.6.

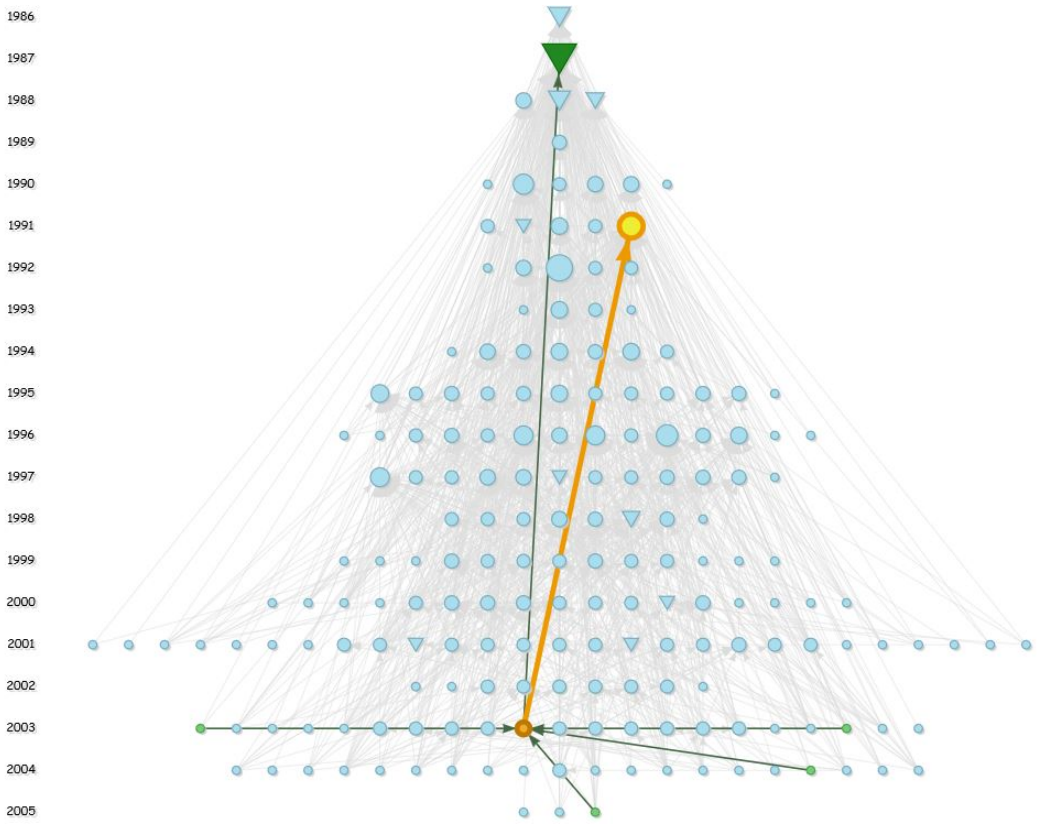


Figure 6.5: Few References Identified in the Graph

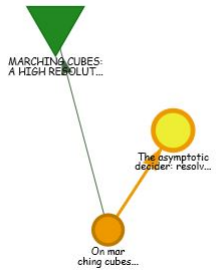


Figure 6.6: References Positioned at Two Levels Defined by Years

In addition, in Figure 6.7, the selected node ("Extracting boundary surface of arbitrary topology from volumetric data sets") has a single reference included in the data set, and the difference between its levels is also large. For this case, some expert users may find it necessary to include some missing papers in the graph. In this way, they would choose from among the nineteen references of the selected article, which ones are considered the most influential to include in the graph. These references, in turn, can be related to other papers already in the graph. Through this procedure, the different research lines are better identified.

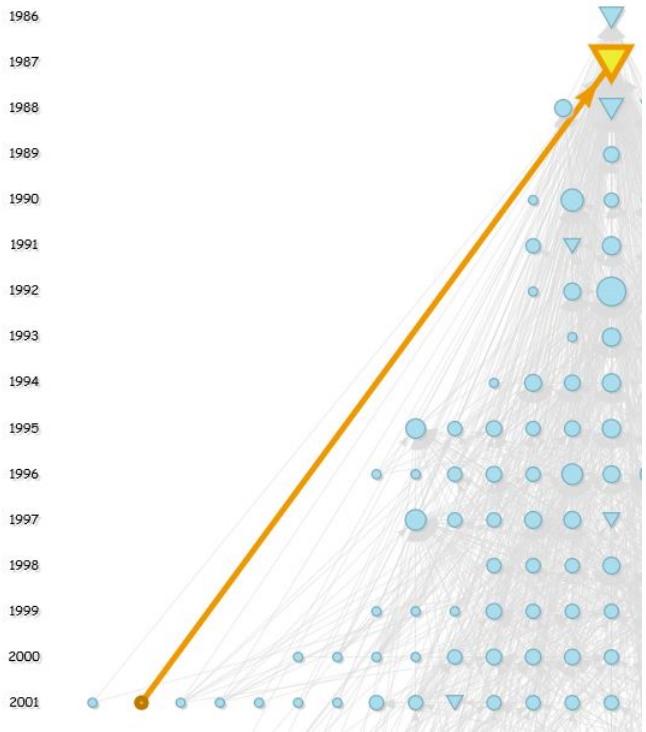


Figure 6.7: Selected Node with a Single Reference represented in the Graph

6.4.2 Identification of the Root Paper of a Scientific Area

We can see that the most cited paper, identified with a large triangle-down shape, can be considered as the initiator of the different branches of studies on the scientific issue addressed in the papers of the graph. Figure 6.8 shows that the vast majority of the publications of the graph reference this root paper.

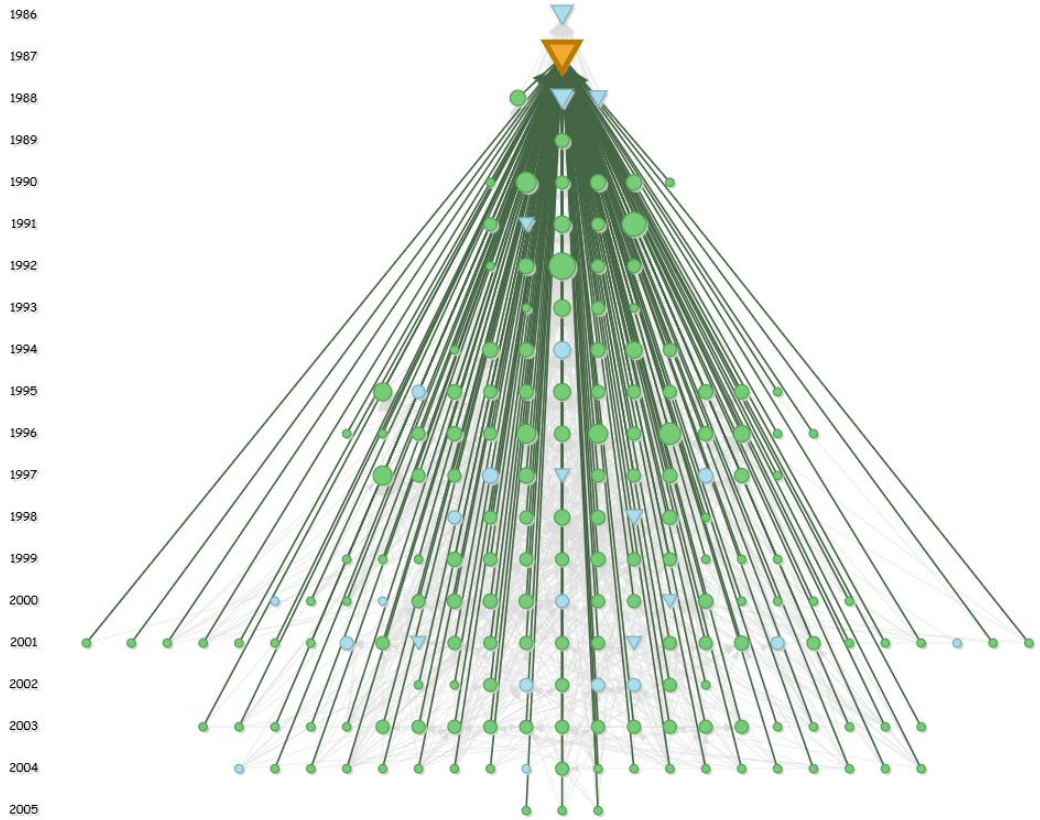


Figure 6.8: Initiator of Different Branches of Studies

In the case that the size of the nodes are similar, it is possible to visually identify the most cited paper by the number of light green nodes that are visualized in the graph when this paper is selected. For example, in Figure 6.9, a node with a large size was selected and we can see that the number of citations represented is smaller than in the example of Figure 6.8.

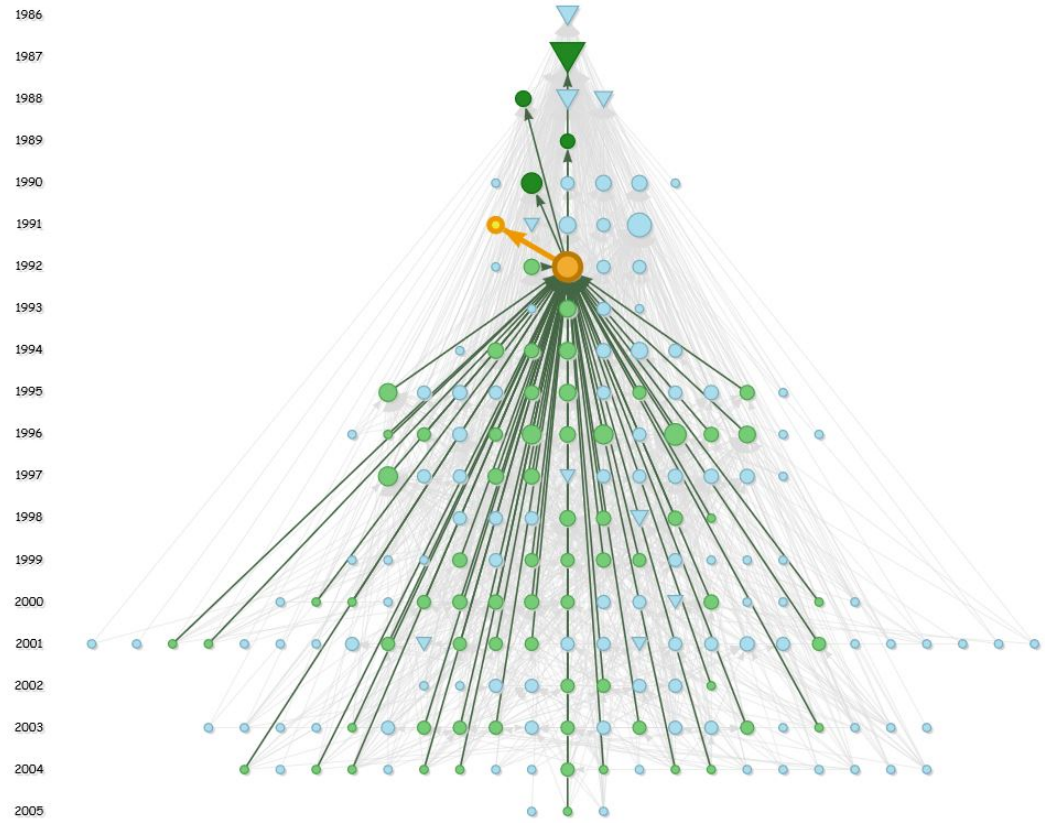


Figure 6.9: Second most cited paper

6.4.3 Look Up a Specific Branch of Study

Our visualization produces a picture of the state of the art of the subject investigated in the papers of the graph. Inspired by Figure 5.1 in Section 5, through edge filtering it is possible to create a simplified visualization for this purpose. Figure 6.10 shows the *Max-Generator Graph*(MG) obtained from the *Citation Graph* of VisMC data set, where the most influential reference of each paper was obtained when performing a filter by the edge of greater weight. Note how the oldest paper in the top of this visualization 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, from the MG graph it is possible to capture the different study lines and the sequence of citations that best reflect the evolution of a certain research.

Figure 6.10: MG Graph in VisMC Data Set

For example, Figure 6.11 shows the branch formed by the papers (in this order): "Efficient implementation of Marching Cubes' cases with topological guarantees" (2003), "Marching Cubes 33: Construction of Topologically Correct Isosurfaces" (1995), "The asymptotic decider: resolving the ambiguity in marching cubes" (1991), "Topological considerations in isosurface generation" (1990), "Polygonization of implicit surfaces" (1988), and "Marching cubes: A high resolution 3D surface construction algorithm" (1987). These papers reflect the evolution of a specific research line, which in this case could be "Ambiguities and Holes", as seen in Figure 5.1. The nodes have a square shape because they have been re-positioned for better visualization of the references path.

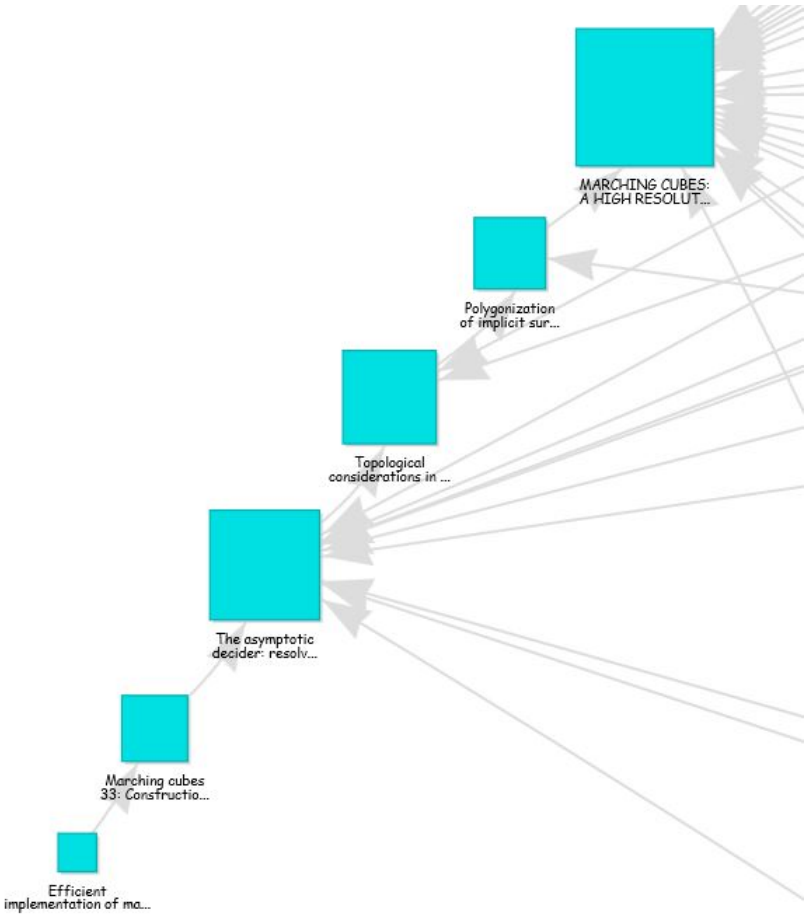


Figure 6.11: Branch of Study "Ambiguities and Holes"

In the example of Figure 6.12, the sequence of papers "Cells octree: a new data structure for volume modeling and visualization" (2001), "Real-time exploration of regular volume data by adaptive reconstruction of isosurfaces" (1999), "Octree-based decimation of marching cubes surfaces" (1996) , "Octrees for Faster Isosurface Generation" (1992), "Octree optimization" (1991), and "Marching cubes: A high resolution 3D surface construction algorithm" (1987) reflect the evolution of the research line "Cracks and Simplification".

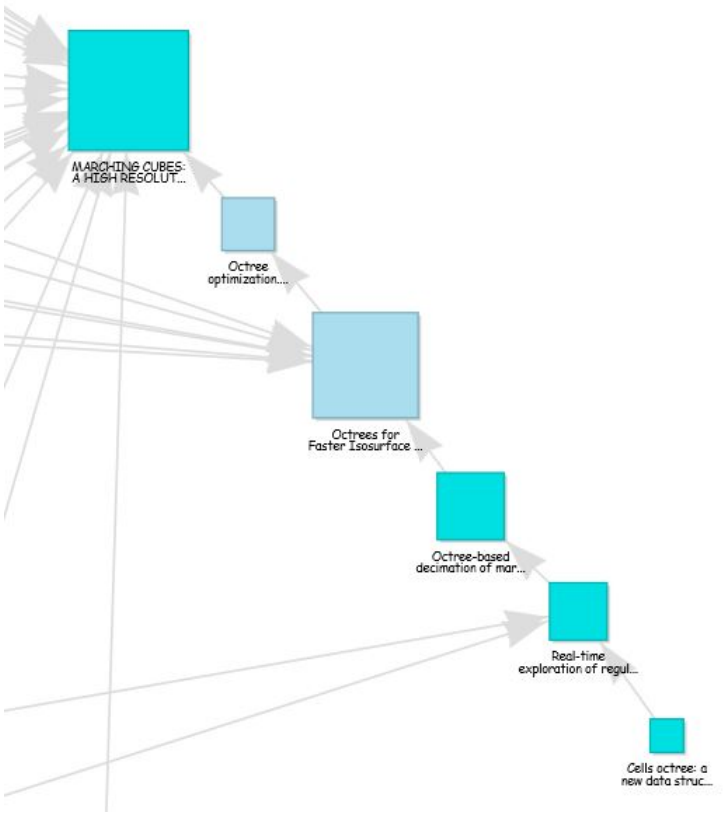


Figure 6.12: Branch of Study "Cracks and Simplification"

The previous examples shown in Figure 6.11 and Figure 6.12 simplify 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 is probably huge. With the filter of relevant references, cluttered designs are avoided.

For the construction of the path in Figure 6.11, 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 6.13, the top one in the ranking of references, the paper "Marching Cubes 33: Construction of Topologically Correct Isosurfaces" (highlighted in orange) is not precisely the most recent. In this case, three specialist users in the subject have collaboratively established with our system a greater weight for the reference highlighted in orange.

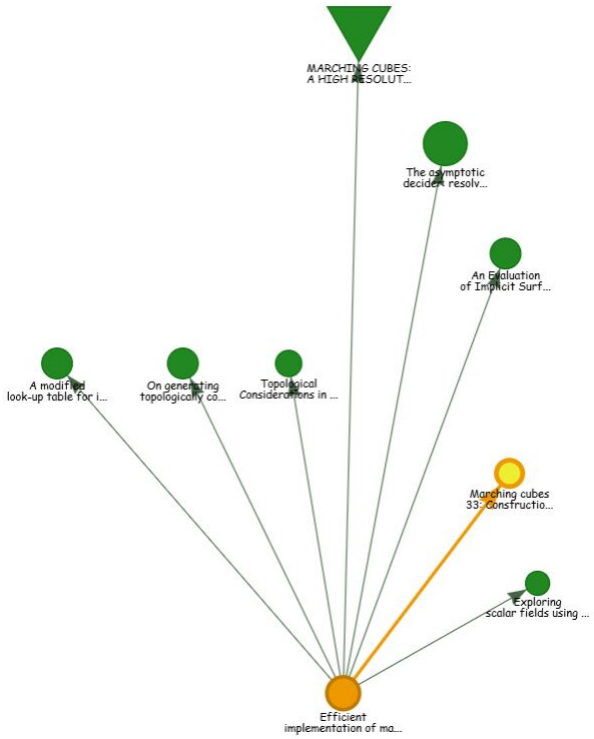


Figure 6.13: Reference top one in the ranking highlighted with orange color

Also, consider the papers in Figure 6.13, listed from top to bottom and from left to right as follows:

1) Marching cubes: A High Resolution 3D Surface construction algorithm.
2) The asymptotic decider: resolving the ambiguity in marching cubes.
3) An evaluation of implicit surface tilers.
4) A modified look-up table for implicit disambiguation of Marching Cubes.
5) On generating topologically consistent isosurfaces from uniform samples.
6) Topological considerations in isosurface generation.
7) Marching Cubes 33 : Construction of Topologically Correct Isosurfaces.
8) Exploring Scalar Fields Using Critical Isovalues.
9) Efficient implementation of Marching Cubes' cases with topological guarantees.

For these papers, we apply the LSA method[3]. In this method, the text is first represented as a matrix of terms by documents and subjected

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

to Singular-Value Decomposition (SVD) for representing geometrically the documents in a reduced dimensionality, as shown in Figure 6.14. After this process, the similarity of the paper "Efficient implementation of Marching Cubes' cases with topological guarantees" and its references are calculated by the cosine similarity,

$$sim(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

where A and B are two documents in vector notation. Thus, it is possible to obtain a ranking of references automatically using the LSA method.

In this case, it is automatically determined that the reference identified with the number 3 is the top one in the ranking of references for the paper "Efficient implementation of Marching Cubes' cases with topological guarantees" (identified with the number 9), failing to give greater relevance to the paper identified with the number 7, which is the most influential publication determined by several experts with our system. Hence the importance of our collaborative proposal for the ranking of references.

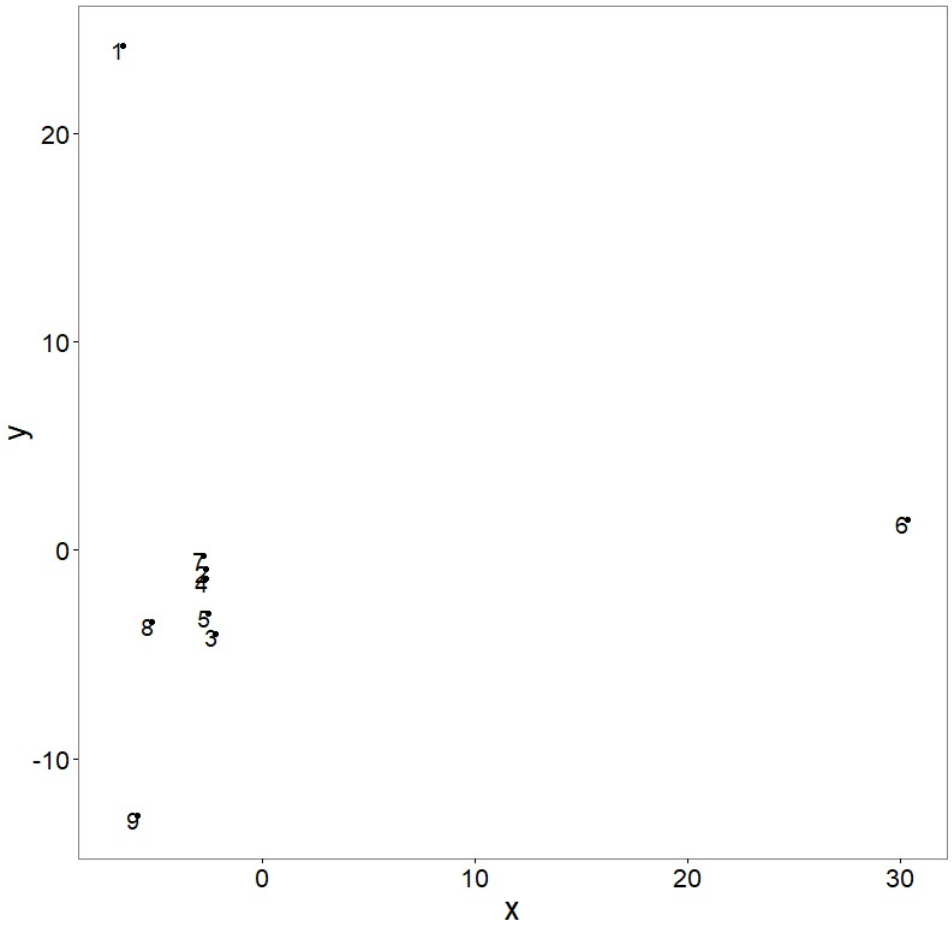


Figure 6.14: Geometrical Representation of References using LSA

6.4.4 Compare Two or More Authors by Expertise via Citation Paths

The search for experts is a task for which our system can provide support. It is possible to obtain a value of expertise from the authors to evaluate a specific paper. Through this value of expertise, which is calculated via citation paths in the graph, we can compare and rank authors.

Table 6.1 show the top five authors in the ranking for the paper "Marching cubes: A High Resolution 3D Surface construction algorithm". This target paper was chosen because it is the most cited paper and also one of the oldest publications in the graph. None of the authors of this paper are in the top five.

Author	Expertise
Gallagher R. S.	8.75
Ertl T.	7.50
Nielson G.	6.16
Hamann B.	5.88
Bloomenthal J.	5.00

Table 6.1: Top five authors with the highest expertise for the paper most cited

By contrast, we show in table 6.2 the top ten authors in the ranking for the paper "Adaptive extraction of time-varying isosurfaces", which was chosen because it has no citations and it is a recent publication (positioned in the lowest level of this graph).

Author	Expertise
Shen H.	6.54
Bordoloi U.	5.00
Newman T.	2.50
Zhang H.	2.50
Gerstner T.	2.25
Pascucci V.	1.66
Joy K.	1.66
Duchaineau M.	1.66
Gregorski B.	1.66
Lindstrom P.	1.66

Table 6.2: Top ten authors with the highest expertise for one of the papers positioned in the lowest level of the graph

Suppose, for example, that the paper "Adaptive extraction of time-varying isosurfaces", whose expert candidates were determined in the

table 6.2, is a new submission to a journal and that the editorial board wants to find candidates to review this article. For this, after obtaining the top k candidates it is necessary to make a filter by the authors who do not have possible conflicts of interest. For example, all authors with publications representing the Lawrence Livermore National Laboratory and the University of California have a conflict of interest with the paper submitted. These are Joy K, Duchaineau M. and Gregorski B. (the authors of the submitted paper) and Pascucci V. In this way, the main reviewer candidate is Shen H, from the University of Alabama, with a 6.54 expertise value.

6.4.5 Compare Two or More Authors by Productivity and Quality of Publications

In our system is possible to obtain visual information about the number of the authors' papers and the number of citations, which allows to establish a comparison between these authors in terms of productivity and quality of their publications in the graph. It is also possible to identify self-citations of the authors.

Through the navigation menu of our application we select the option of a specific view that includes information of authorship relationships. In this way, when selecting in a combobox the name of some author in the graph, his/her publications are shown in the citation graph, highlighted in purple color. Moreover, only the incoming edges for these publications are shown. For example, in Figure 6.15, the author Wilhelms J. has three papers, one of which is the second most cited of the graph, as we see in Figure 6.9. Although this author had few publications, all of them have been highly cited, so the local h-index[4] has a value of 3, that is, the maximum in relation to its number of publications.

[4]The calculation of the h-index is made based on the papers contained in the graph and in no way accounts for this author's other publications.

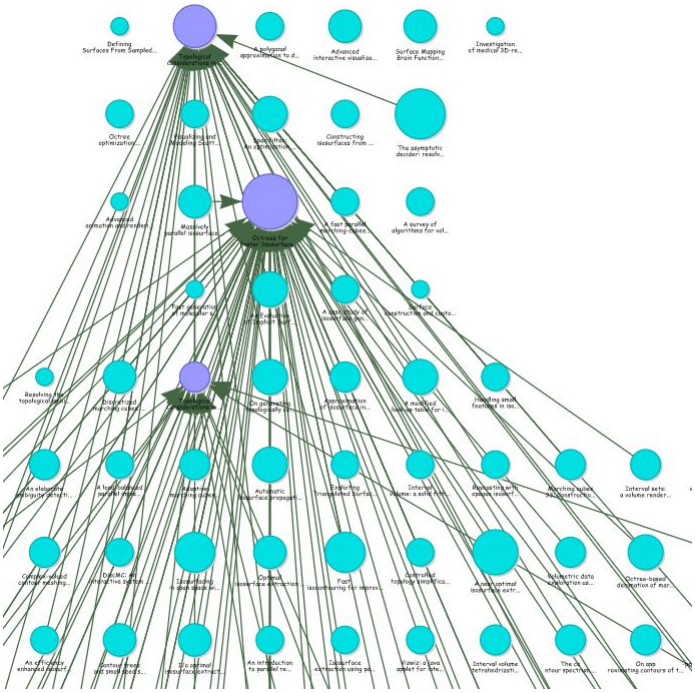


Figure 6.15: Publications of the author Wilhelms J.

By contrast, Banks D. has more publications than Wilhelms J. (four papers), but his articles have been less cited, obtaining a lower h-index (2) (see Figure 6.16). Note that this author has 4 self-citations in this graph.

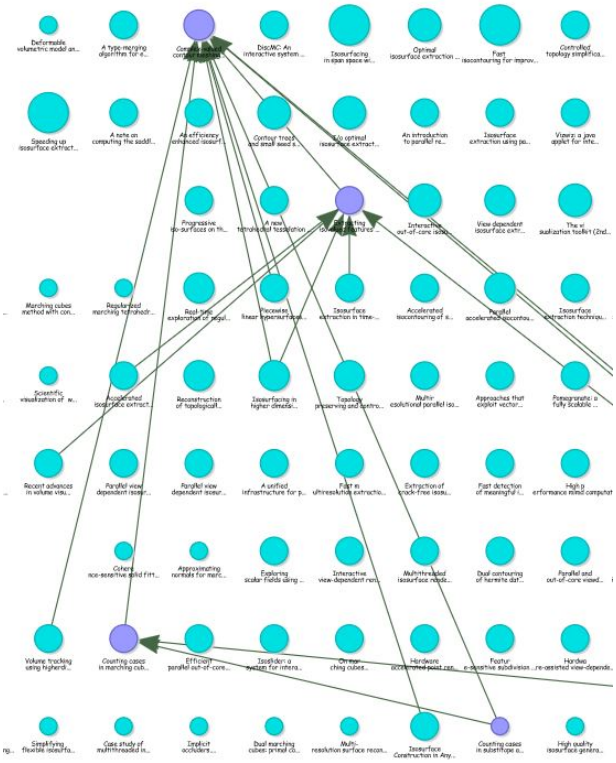


Figure 6.16: Publications by Banks D.

One of the authors with greater productivity (ten publications) and quality in their publications with high h-index value is Shen H., as shown in Figure 6.17. This author obtained h-index 5. In the set there are few authors with h-index 5, as seen in Figure 6.18, whilst 46.5% of the total number of authors in the system have h-index 0 (zero).

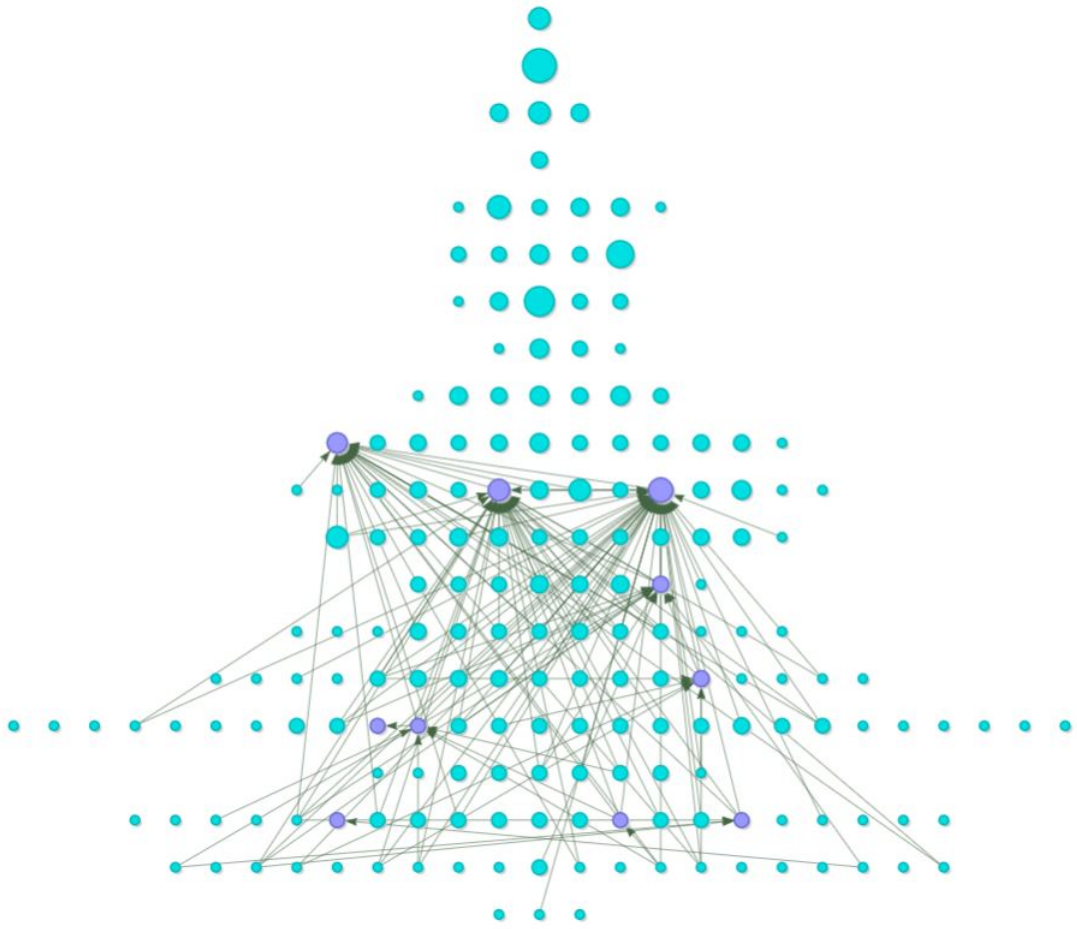


Figure 6.17: Publications of the author Shen H.

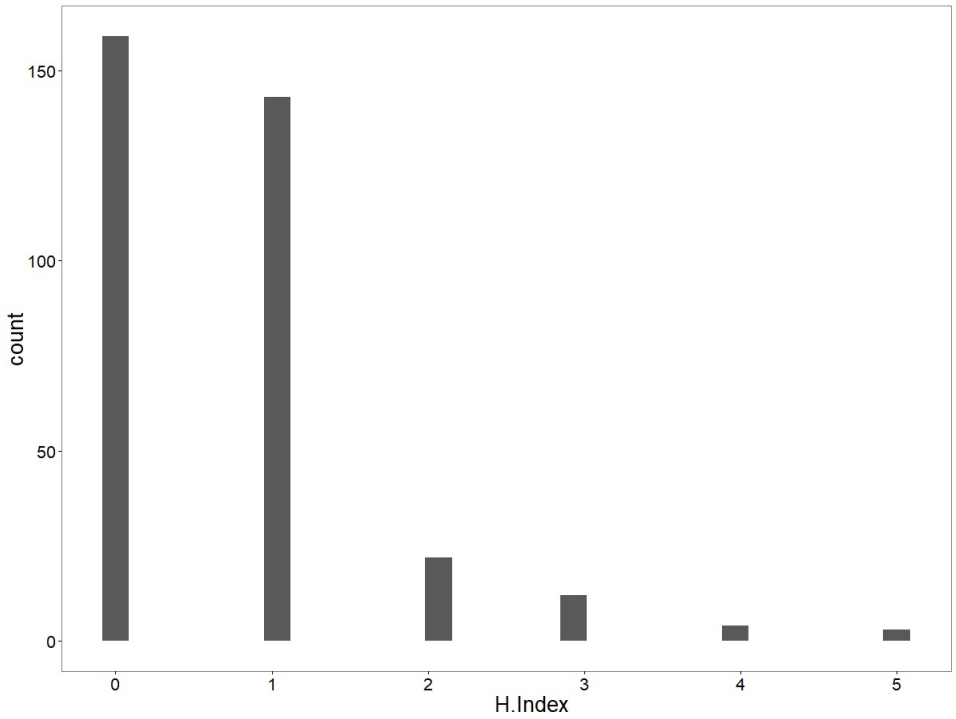


Figure 6.18: Histogram for h-index

6.4.6 Compare Two or More Authors by Global Expertise

In this subsection, we present the main results considering the Global Expertise function in the VisMC data set, according to the Definition 3.8 in Chapter 3. With this objective, table 6.3 shows the top five authors in the ranking according to this function.

Author	Global Expertise
Cline H. E.	1.00
Lorensen W. E.	1.00
Shen H.	0.73
Wilhelms J.	0.66
Van Gelder A.	0.66

Table 6.3: Top five authors with the highest Global Expertise values

We previously showed, in Figure 6.17 and the results of the calculation of the different measurements that the system offers, for example h-index, that Shen H. is one of the authors with the highest value of h-index (value 5). However, he is not in the group of authors with the maximum value of Global Expertise (value 1). For example, table 6.4 shows the top five authors with the highest values for this function and their respective h-index values

for comparative purposes. Note that Cline H. E. and Lorensen W. E. are the authors with the greatest impact and popularity because they are the authors of the most cited paper in the graph. Shen H., despite being an author with many publications and a high h-index, held the third position in the ranking. Montani C. and Scopigno R, authors with h-index 5, are not also within this group.

Author	Global Expertise	h-index
Cline H. E.	1.00	2
Lorensen W. E.	1.00	2
Shen H.	0.73	5
Wilhelms J.	0.66	3
Van Gelder A.	0.66	3

Table 6.4: Top five authors with the highest Global Expertise values and their respective h-index

Figure 6.19 shows that most authors have low Global Expertise values. By definition, authors with h-index 0 obtain a minimum value of Global Expertise that is different from zero. For example, the author Andujar C. has only one paper in the lowest level of the graph (it has no citations), so the value of the h-index is 0 and the value of Global Expertise is 0.005. In this way, we measure the impact and popularity of the authors, ensuring that authors without citations can interact with our system.

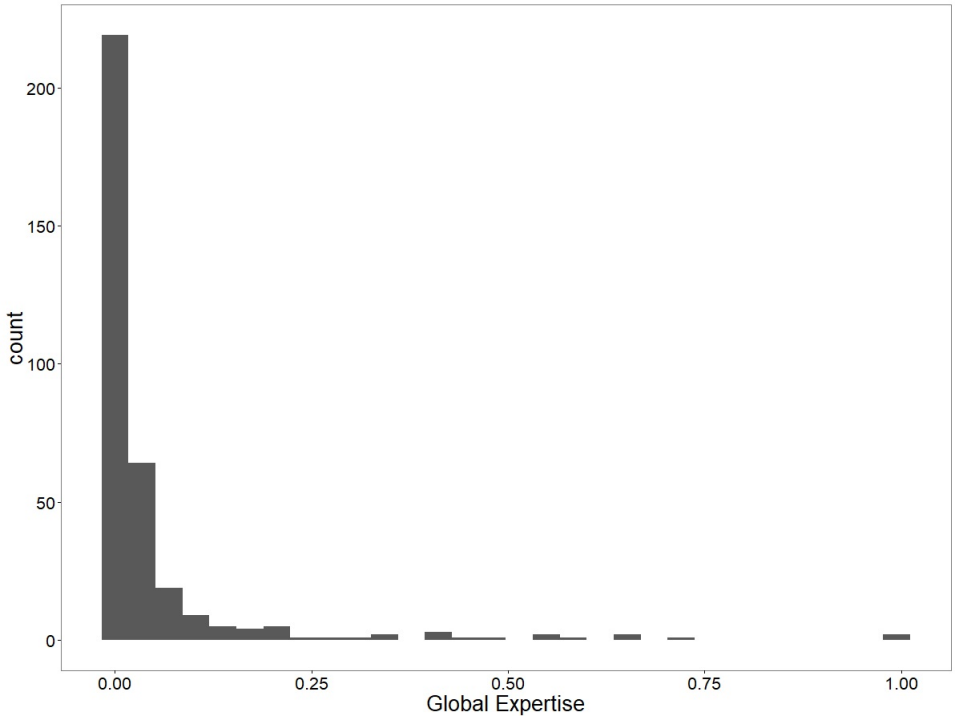


Figure 6.19: Histogram for Global Expertise

6.4.7 Combining Global Expertise and Expertise via Citation Paths

We show in this subsection the value of the function $Exp(j, q, t)$ in the graph, as defined in equation 3.6 in Chapter 3. Table 6.5 shows the expertise value of the authors of table 6.2, considering the Global Expertise, the Expertise by Citation Paths, and $t = |E_R|$, that is, $\alpha(t) = 0.5$. The author Shen H. is considered by our system as the best candidate to review, comment or discuss about the issue of the target paper "Adaptive extraction of time-varying isosurfaces".

Author	Expertise
Shen H.	3.63
Bordoloi U.	2.51
Newman T.	1.27
Zhang H.	1.26
Gerstner T.	1.16
Pascucci V.	1.00
Joy K.	0.85
Duchaineau M.	0.84
Gregorski B.	0.84
Lindstrom P.	0.84

Table 6.5: Expertise values depending on time and given a specific paper

Note that through this function it is possible to establish a differentiation between the expertise values of Newman T. and Zhang H., and also between Joy K. and Duchaineau M. (two authors of the target paper), which was not possible considering only the values in table 6.2.

6.5 User Study

To evaluate the proposed visualization and the methodology for ranking expert and references, we conducted an empirical study in an academic environment according to the following procedure. 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, capturing paths of relevant or influential references, and comparing the expertise between authors. 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). At the end of the questionnaire, the participants can comment, according to their criteria, on the main changes or missing elements in the tool to make it more useful for users. The participation was voluntary and each participant signed an informed consent form. This kind of study is widely used for measuring the performance of visualizations, as explained in [51].

The study was divided in two rounds, with 15 and 12 participants respectively, a total of 27 people evaluating our system. The first round,

denoted by R1, was conducted in order to assess the visual interface and interactions. The 15 researchers who participated in R1 are from the field of pure and applied mathematics, from different areas of knowledge of our university, comprising 3 D.Sc., 4 M.Sc., and 8 B.Sc. None of them had prior knowledge of the marching cubes problem. The form for R1 comprised analysis tasks on the data, defined by the letter T, and statements for the Likert scale, defined 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 if the second paper most cited in the graph cite others 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 correspond to the most influential or relevant citation.

The second round, R2, aimed to assess the usability in the sense of verifying if the participants managed to perform the proposed tasks in a specific interface and not to verify how easy was to navigate the system. Also, R2 aimed to verify the effectiveness of our approach. The participants in this second round were 4 D.Sc, 7 M.Sc. and 1 B.Sc. All these researchers have a vast knowledge and experience on the "marching cubes" subject and applications in Computer Graphics. The tasks and statements were designed to assess not only to what extent the visual application and the ranking of references was useful, but what underlying information in the data was useful for the ranking of authors. The analysis tasks supported by the tool and the statements in this second round are the following.

R2-T1. Select the Visualization tab. Using the visual interface, identify

the paper "Cells octree: a new data structure for volume modeling and visualization". Filter the edges of the graph to get only the most influential edge of each node. Determine the publications that follow an order of influential citation from this paper, that is, position this paper in the most appropriate branch of study.

R2-S1. It is possible to visually determine the sequence of publications that belong to a specific branch of study.

Consider that Global Expertise is a measure of expertise attributed to an author based on citations and popularity of this author.

R2-T2. Select the Expertise tab and select the option Visual Analysis. Identify in the graph the publications of the authors Zhang H. and Gregorski B.

R2-S2. Zhang H. has a higher value of Global Expertise than the author Gregorski B.

Consider that the Expertise by Citation Paths is a measure of expertise attributed to an author based in his/her publications that are close to a specific paper.

R2-T3 Identify in the graph influential or relevant citation paths of maximum length equal to 2 from and to the paper "Space efficient fast isosurface extraction for large data sets". Consider this publication as the target paper in the calculation of the Expertise by Citation Paths. Figure 6.20 shows these paths and the target paper, identified with the number 3, and the names of the authors positioned near the publications of their authors.

R2-S3.1. The author Zhang H.(author of the paper identified with the number 4) has a lower value of Expertise by Citation Paths than the author Gregorski B. (author of the papers identified with the number 1 and 5).

R2-S3.2. The author Senecal J.(author of the paper identified with the number 1) has a high value of Expertise by Citation Paths than the author Lindstrom P. (author of the paper identified with the number 5).

R2-S3.3. The Expertise by Citation Path of Gregorski B. is greater than the Expertise by Citation Path of Pascucci V. (author of the paper identified with the number 5).

R2-S3.4. The authors Gregorski B. and Joy K.(authors of the papers identified with the number 1 and 5) have the same Expertise by Citation Paths.

R2-T4 Consider the Global Expertise and Expertise by Citation Path

given a target paper “Space efficient fast isosurface extraction for large data sets” of Zhang H., Gregorski B. and Joy K. Also consider a general value of Expertise combining these two measures. Determine the best expert candidate to comment the target paper “Space efficient fast isosurface extraction for large data sets”.

R2-S4. The best expert candidate is Joy K, occupying the first position in the ranking.

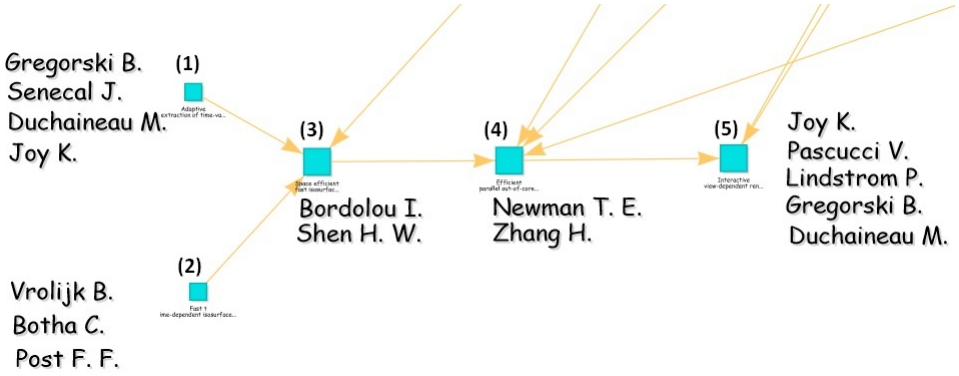


Figure 6.20: Author’s papers in a path of influential references

We now present the results of the user study. The participants in the first round of our study, R1, performed the proposed tasks with our system and evaluated each statement. The results show that the number of participants who answered Completely Agree (dark blue bars) for all the statements was on average 13, as shown in Figure 6.21. This allows us to positively evaluate our system as a tool to visualize and recommend scientific papers. The most complicated task to perform by the users was R1-T3. This led to one participant select the option Undecided (gray bar in Figure 6.21) for the corresponding statement, R1-S3, of the questionnaire. Also, one user selected the option Slightly Agree (light blue bar) for the same statement. This was mainly due to the fact that they had some difficulties to visually comparing the size of the nodes involved in the analysis. However, the 86.6% of the participants selected the option Completely Agree for R1-S3. These results provide evidence that the participants, who are prospective system users, liked the proposed visual interface for its usefulness. Our system made it possible to easily identify important papers in the area, as well as the most relevant references evaluated collaboratively by the experts.

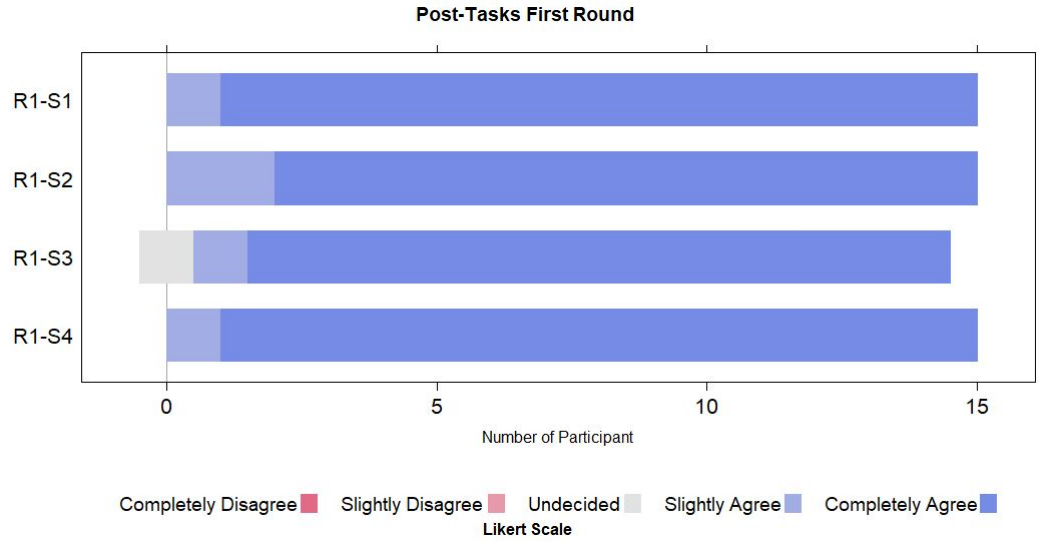


Figure 6.21: Post-Tasks First Round

In the second round, R2, more than half of the participants answered Completely Agree for all the statements (see Figure 6.22). The tasks with the most positive responses were R2-S3.2 and R2-S3.3, where 83.3% of the participants answered Completely Agree and the remaining 16.7% answered Slightly Agree. By contrast, some participants were confused in the task R2-S3.4, to determine which of the two authors involved in the analysis would have more Expertise by Citation Path. This led to 3 out of 12 participants answering Slightly Disagree (light red bar) and 1 participant answering Undecided (gray bar). After the questionnaire was completed, the reason for these answers was identified. The main point is that these users considered the order of the authors in their publication, and also prioritized the first author of a publication that cites the target paper. This result does not have a negative impact on our work because the response of the participants was based on considerations different from ours. Quite the contrary, this and other considerations can be a valid extension of our work. Therefore, it is worth advancing to the next phase of documentation and registration of the software and finally publication of the web system for academic purposes.

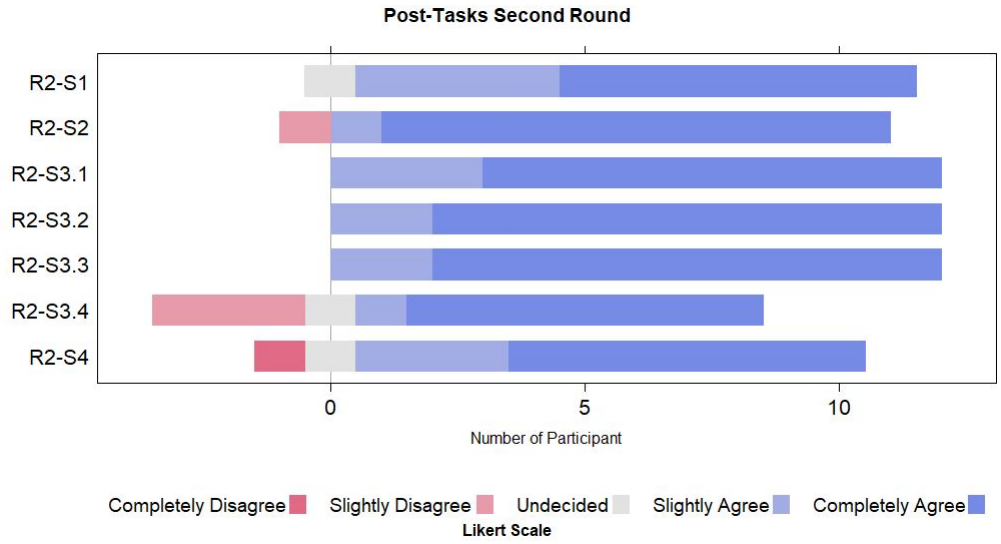


Figure 6.22: Post-Tasks Second Round

As we mentioned before, at the end of the questionnaire, all suggestions of the expert participants were collected to help improve our system. We now summarize some of these suggestions. One participant mentioned that "having a new tab to compare two or more authors by expertise on the same chart could help a lot" and another user suggested that "a legend in the visual interface can help make the system more intuitive". Three participants suggested a common improvement: to consider the authors' self-citations as a factor that may negatively affect their popularity. The expert participants also appreciated the visual interface of our system. One of them said that "At first, the number of arrows in the visualization tool scares a bit, but once you click on a node paper, the arrows connected to it are shown in the second view and the most influential reference is highlighted. It helps very much". In general, participants believed that the proposed system is a great tool for finding important references and expert candidates.

7
Conclusions and Future Work

Assessing the relevance of the references cited is far from being straightforward. This research has devised a strategy to visualize and recommend citations within a corpus of scientific papers. For this, we proposed a new collaborative approach, making use of expert analysis to deal with the problem of selecting and ranking the most influential or relevant references. 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. We proposed a new method to find experts based on citation paths. Also, an empirical study was conducted to evaluate the methodology presented in this work.

In 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 studies through the collaboration of experts. More precisely, the strength of the system lies in considering the different opinions and expertise of the specialists in the ranking of references, a variant not explored so far in this type of system. This supports the user in the task of ensuring proper coverage of relevant papers when conducting literature surveys. Moreover, the system can support the board of scientific journals when analyzing or comparing the expertise of candidate reviewers through relevant citation paths in the network. It also helps the reviewers to verify whether the author failed to include (study) important papers in their documents. The Expertise by Citation Path is a new contribution of our work in the search for a measure of expertise that considers the topology of the citation network and the proximity of the papers of the authors to a specific paper.

As future work, we propose the following considerations.

- Evaluate the procedures with other data sets, for example, vispubdata, see [52]. In this way, we would be able to visualize, among other things, several graphs of different topics where, in general, very few nodes of a graph (or none) are related to the nodes of the other graphs.

- Comparing our in-depth proposal for ranking experts with other

automatic approaches. For example, an automatic method for expertise could form a candidate expert pool and then apply a standard topic model to the target paper and the published paper of the candidate expert, as seen in [4]. Thus, the best expert candidates would be obtained first by our method and then by the other method. This would make it possible to draw conclusions on which of these methods would be more suitable to find reviewers or whether one method complements the other.

- Implement some modifications in the calculation of Expertise by Citation Paths given a target paper, for example, taking into account that other authors whose publications cite any paper in the paths of influential citations from the target paper. With this modification, we would obtain other candidate experts. Also, we could consider the second and the third most influential references of each paper. Another modification that might interesting would be to consider that the Expertise by Citation Paths of an author increases when the author cites the target paper as influential. These modifications would extend the calculation of expertise in this work and provide new tools to the paper-expert assignment systems.

- Establish a co-authoring network. In this way it is possible to identify the frequency of co-authorship and avoid conflicts of interest in the search for expert reviewers.

- Incorporate other measures for the importance and influence of each paper in the network. For example, in addition to the number of citations of a paper (as represented by its node size), we could have another attribute to reflect the number of publications that cite this paper as the most influential to their work.

- 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 value and the references that still have zero weight. In this way, users would be able to 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.

- Collaborate with design specialists to offer users a visual interface that can be introduced in an academic environment following best practices of web design.

Bibliography

[1] Craig Macdonald e Iadh Ounis, *Voting for candidates: adapting data fusion techniques for an expert search task*, Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006, pp. 387–396.

[2] Hongbo Deng, Irwin King, e Michael R Lyu, *Formal models for expert finding on dblp bibliography data*, Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE, 2008, pp. 163–172.

[3] G Alan Wang, Jian Jiao, Alan S Abrahams, Weiguo Fan, e Zhongju Zhang, *Expertrank: A topic-aware expert finding algorithm for online knowledge communities*, Decision Support Systems **54** (2013), no. 3, 1442–1451.

[4] Xiang Liu, Torsten Suel, e Nasir Memon, *A robust model for paper reviewer assignment*, Proceedings of the 8th ACM Conference on Recommender systems, ACM, 2014, pp. 25–32.

[5] Xiaorui Jiang, Xiaoping Sun, e Hai Zhuge, *Graph-based algorithms for ranking researchers: not all swans are white!*, Scientometrics **96** (2013), no. 3, 743–759.

[6] Krisztian Balog, Maarten De Rijke, et al., *Determining expert profiles (with an application to expert finding).*, IJCAI, vol. 7, 2007, pp. 2657–2662.

[7] Xinlian Li e Toyohide Watanabe, *Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers*, Procedia Computer Science **22** (2013), 633–642.

[8] Jian Jin, Qian Geng, Qian Zhao, e Lixue Zhang, *Integrating the trend of research interest for reviewer assignment*, Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 1233–1241.

[9] Marco Valenzuela, Vu Ha, e Oren Etzioni, *Identifying meaningful citations.*, AAAI Workshop: Scholarly Big Data, 2015.

[10] Yuan An, Jeannette Janssen, e Evangelos E Milios, *Characterizing and mining the citation graph of the computer science literature*, Knowledge and Information Systems **6** (2004), no. 6, 664–678.

[11] Simon Price e Peter A Flach, *Computational support for academic peer review: A perspective from artificial intelligence*, Communications of the ACM **60** (2017), no. 3, 70–79.

[12] Ulrich Schäfer e Uwe Kasterka, *Scientific authoring support: A tool to navigate in typed citation graphs*, 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.

[13] Xiaodan Zhu, Peter Turney, Daniel Lemire, e André Vellino, *Measuring academic influence: Not all citations are equal*, Journal of the Association for Information Science and Technology **66** (2015), no. 2, 408–427.

[14] Ramesh M Nallapati, Amr Ahmed, Eric P Xing, e William W Cohen, *Joint latent topic models for text and citations*, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 542–550.

[15] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, e Ioannis G Tollis, *Graph drawing: algorithms for the visualization of graphs*, Prentice Hall PTR, 1998.

[16] Raga'ad M Tarawaneh, Patric Keller, e Achim Ebert, *A general introduction to graph visualization techniques*, OASIcs-OpenAccess Series in Informatics, vol. 27, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[17] 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.

[18] Michaël Charles Waumans e Hugues Bersini, *Genealogical trees of scientific papers*, PloS one **11** (2016), no. 3, e0150588.

[19] Hui Wei, Youbing Zhao, Shaopeng Wu, Zhikun Deng, Farzad Parvinzamir, Feng Dong, Enjie Liu, e Gordon Clapworthy, *Management of scientific documents and visualization of citation relationships using weighted key scientific terms.*, DATA, 2016, pp. 135–143.

[20] Steven A Greenberg, *How citation distortions create unfounded authority: analysis of a citation network*, Bmj **339** (2009), b2680.

[21] Nicholas Crouch e MW Powers David, *Pyscholargraph: A graph-based framework for indexing, searching and visualising relationships between academic papers*, The ANU Undergraduate Research Journal **161** (2015).

[22] Gerard Salton e Chung-Shu Yang, *On the specification of term values in automatic indexing*, Journal of documentation **29** (1973), no. 4, 351–372.

[23] Matthew Berger, Katherine McDonough, e Lee Seversky, *cite2vec: Citation-driven document exploration via word embeddings*, IEEE Transactions on Visualization & Computer Graphics (2017), no. 1, 1–1.

[24] Gouri Ginde, *Visualisation of massive data from scholarly article and journal database a novel scheme*, arXiv preprint arXiv:1611.01152 (2016).

[25] Aleksa Vukotic, Nicki Watt, Tareq Abedrabbo, Dominic Fox, e Jonas Partner, *Neo4j in action*, Manning Publications Co., 2014.

[26] Yuliant Sibaroni, Dwi Hendratmo Widyantoro, e Masayu Leylia Khodra, *Survey on research paper's relations*, Information Technology Systems and Innovation (ICITSI), 2015 International Conference on, IEEE, 2015, pp. 1–6.

[27] Aditya Pratap Singh, Kumar Shubhankar, e Vikram Pudi, *An efficient algorithm for ranking research papers based on citation network*, Data Mining and Optimization (DMO), 2011 3rd Conference on, IEEE, 2011, pp. 88–95.

[28] Lawrence Page, Sergey Brin, Rajeev Motwani, e Terry Winograd, *The pagerank citation ranking: Bringing order to the web.*, Tech. report, Stanford InfoLab, 1999.

[29] Zhiguang Zhou, Chen Shi, Miaoxin Hu, e Yuhua Liu, *Visual ranking of academic influence via paper citation*, Journal of Visual Languages & Computing **48** (2018), 134–143.

[30] Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, Luo Si, et al., *Expertise retrieval*, Foundations and Trends® in Information Retrieval **6** (2012), no. 2–3, 127–256.

[31] Jorge E Hirsch, *An index to quantify an individual's scientific research output*, Proceedings of the National academy of Sciences **102** (2005), no. 46, 16569–16572.

[32] Leo Egghe, *Theory and practise of the g-index*, Scientometrics **69** (2006), no. 1, 131–152.

[33] David M Blei, Andrew Y Ng, e Michael I Jordan, *Latent dirichlet allocation*, Journal of machine Learning research **3** (2003), no. Jan, 993–1022.

[34] David Mimno e Andrew McCallum, *Expertise modeling for matching papers with reviewers*, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007, pp. 500–509.

[35] Gianluca Demartini, Julien Gaugaz, e Wolfgang Nejdl, *A vector space model for ranking entities and its application to expert search*, European Conference on Information Retrieval, Springer, 2009, pp. 189–201.

[36] Humayun Kabir Biswas e Md Maruf Hasan, *Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment*, Information and Communication Technology, 2007. ICICT'07. International Conference on, IEEE, 2007, pp. 82–86.

[37] Muhammad Tanvir Afzal e Hermann A Maurer, *Expertise recommender system for scientific community.*, J. UCS **17** (2011), no. 11, 1529–1549.

[38] Toine Bogers, Klaas Kox, e Antal van den Bosch, *Using citation analysis for finding experts in workgroups*, Proc. DIR, Citeseer, 2008, pp. 21–28.

[39] George Packer, *Cheap words*, The New Yorker **17** (2014).

[40] Greg Linden, Brent Smith, e Jeremy York, *Amazon. com recommendations: Item-to-item collaborative filtering*, IEEE Internet computing (2003), no. 1, 76–80.

[41] Riccardo Mazza, *Introduction to information visualization*, Springer Science & Business Media, 2009.

[42] George Rassovsky, *Cubical marching square implementation*, 2014.

[43] Susan T Dumais, *Latent semantic analysis*, Annual review of information science and technology **38** (2004), no. 1, 188–230.

[44] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, e Ulrike Von Luxburg, *Design and analysis of the nips 2016 review process*, The Journal of Machine Learning Research **19** (2018), no. 1, 1913–1946.

[45] Ngai Meng Kou, Nikos Mamoulis, Yuhong Li, Ye Li, Zhiguo Gong, et al., *A topic-based reviewer assignment system*, Proceedings of the VLDB Endowment **8** (2015), no. 12, 1852–1855.

[46] Timothy S Newman e Hong Yi, *A survey of the marching cubes algorithm*, Computers & Graphics **30** (2006), no. 5, 854–879.

[47] Guido Van Rossum e Fred L Drake, *The python language reference manual*, Network Theory Ltd., 2011.

[48] Tin Huynh e Kiem Hoang, *Gate framework based metadata extraction from scientific papers*, 2010 International Conference on Education and Management Technology, IEEE, 2010, pp. 188–191.

[49] D Framework, *Django the web framework for perfectionists with deadlines*, `https://docs.djangoproject.com/en/2.0/` **1** (2016).

[50] BV Almende, *vis. js–a dynamic, browser based visualization library*, `http://visjs.org/` **1** (2016).

[51] Shixia Liu, Weiwei Cui, Yingcai Wu, e Mengchen Liu, *A survey on information visualization: recent advances and challenges*, The Visual Computer **30** (2014), no. 12, 1373–1393.

[52] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, e John Stasko, *vispubdata. org: A metadata collection about ieee visualization (vis) publications*, IEEE transactions on visualization and computer graphics **23** (2017), no. 9, 2199–2206.