Interactive System for Scientific Publication Visualization and Similarity Measurement based on Citation Network

Hanadi Humoud A Alfraidi

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements for the degree

Master of Computer Science

Ottawa-Carleton Institute for Computer Science

uOttawa

School of Electrical Engineering and Computer Science

University of Ottawa

Abstract

Online scientific publications are becoming more and more popular. The number of publications we can access almost instantaneously is rapidly increasing. This makes it more challenging for researchers to pursue a topic, review literature, track research history or follow research trends. Using online resources such as search engines and digital libraries is helpful to find scientific publications, however most of the time the user ends up with an overwhelming amount of linear results to go through.

This thesis proposes an alternative system, which takes advantage of *citation/reference relations* between publications. This demonstrates better insight of the hierarchy distribution of publications around a given topic. We also utilize *information visualization techniques* to represent the publications as a network. Our system is designed to automatically retrieve publications from Google Scholar and visualize them as a 2-dimensional graph representation using the citation relations. In this, the nodes represent the documents while the links represent the citation/reference relations between them.

Our visualization system provides a better view of publications, making it easier to identify the research flow, connect publications, and assess similarities/differences between them. It is an interactive web based system, which allows the users to get more information about any selected publication and calculate a similarity score between two selected publications.

Traditionally, similar documents are found using Natural Language Processing (NLP), which compares documents based on matching their contents. In the proposed method, similar documents are found using the citation/reference relations which are

represented by the relationship that was originally inputted by the authors. We propose a new path based metric for measuring the similarity scores between any pair of publications. This is based on both the number of paths and the length of each path. More paths and shorter lengths increase the similarity score. We compare our similarity score results with another similarity score from Scurtu's Document Similarity [1] that uses the NLP method. We then use the average of the similarity scores collected from 15 users as a ground truth to validate the efficiency of our method. The results indicate that our Citation Network approach yielded better scores than Scurtu's approach.

Acknowledgements

First and foremost, I would like to thank “Allah” the almighty for providing me with the health, strength, patience, and perseverance to complete this thesis.

I would like to express my sincere thanks to Professor Won-Sook Lee for her extensive support, direction, advice, and patience during my learning journey. This work could have never been completed without her guidance and experience.

Special thanks go to Professor David Sankoff for his support and directions in the statistical analysis part of this thesis.

A deeply hearty thank you to my husband Turki, my angels Faisal and Mohammed, my father, my mother, my grandmother, my sister, my brothers especially Ali, my aunts-especially Nahed, my uncle, and all my family members who have supported me during my study. Their love, prayers, and encouragements have brightened my life.

I would never forget to thank my friends Fatimah, Hanan, Hawazin, Modhawi, Nora and Shahad for their unlimited support and caring for both my children and myself.

Finally, special thanks go to my sponsor: King Abdulaziz University in Jeddah, Saudi Arabia, represented by the Saudi cultural bureau in Canada, for their financial support in the form of a generous academic scholarship.

Dedication

To my parents, grandmother,

husband, children (Faisal and Mohammed),

sister and brothers

who have supported me since the beginning of my studies.

Table of Contents

List of Figures

List of Tables

Glossary

GS	Google Scholar
2D	2- dimensional
IR	Information Retrieval
OS	Operating System
NLP	Natural Language Processing
HCI	Human- Computer Interaction
SDS	Scurtu’s Document Similarity
DFD	Data Flow Diagram
WWW	World Wide Web
CN	Citation Network
JSON	Java Script Object Notation
DOM	Document Object Model

Chapter 1. Introduction

In this chapter we will present the motivation behind this thesis, define the research question, describe the research objectives and contributions, and outline the organization of the rest of this thesis.

Information visualization is a set of technologies which use computer supported, interactive, and visual processing of abstract data to gain cognition [2]. As the volume of available data increases, the use of information visualization to represent data is increased as well. This leads to a large amount of unstructured information requiring organization in a way that amplifies the users understanding. This is done via selecting one of many appropriate graphical representation techniques such as a table, chart, tree, 2-dimensional (2D) graph, or 3-dimensional (3D) landscape model. Various research has been done on the use of information visualization in different fields such as social network visualization, geographic information visualization, financial data analysis, scientific research, medical applications, and scientific literature visualization (our focus in this thesis).

1.1 Motivation

Information visualization becomes a powerful approach to analyze different bibliometric networks, ranging from Citation Network (CN) to co-authorship relation network or co-occurrence relations between keywords network [3]. In spite that extensive research has been conducted in this domain (will be discussed in Chapter 2), the needs of advanced publications visualization and analysis systems is still under research due to literature retrieval limitation and large citation network analysis limitation.

In this thesis, we will retrieve the publications automatically from a scholarly search engine and visualize them as a 2D graph representation using the citation/reference relations, where the nodes represent the documents and the links represent the citation/reference relations between them. Moreover, we calculate the similarity score between each pair of publications based on both the number of paths and the length of each path between the two nodes. More paths and shorter lengths result in scoring higher on the similarity score. Thus, we analyze the connected publications. By applying the Breadth-First Search algorithm on any selected pair of nodes, we explore the paths between them and calculate the similarity scores between based on the number and the length of paths.

We observe that defining the hierarchical relationship between scientific papers, using 2-dimensional graph based citation relations is helpful in tracking research history and following given knowledge in a research field. Citation relations include all of the publications that cite a certain article, as well as the publications that this particular article cites (i.e. references). The publications that cite a specific article demonstrate the research which has followed the very publication, while the references of an article demonstrate the scientific history. Co-citation occurs when multiple papers cite the same article [4], which is difficult to visualize in linear dimension (list) of publications, like the output of a Google Scholar search.

As time passes, the number of publications and resources rapidly increase. Researchers and students need to update their knowledge and keep up with the advancements in their field, so they look for research trends and review literature. Most of the search database systems return lists of papers, and some of them provide the ability to analyze the results and explore the related works, citations, versions and publications. Users may need to read individual papers, understand

the ideas, and extract the related literature. This process may consume a lot of time and effort, especially when undertaking new research in a field.

This problem motivates us to build a system which helps researchers and students to inquire a topic, retrieve the results, and visualize the resulting publications in an interactive way. We demonstrate the relationships between them visually rather than spending time between papers extracting their relations. Visual representation of papers can help users understand the relationship between groups of literature, rather than going through a list of papers one by one.

1.2 Research Question and Thesis Goals

Based on the analysis of existing systems for scholarly publication visualization, here is our research question:

"*Is it feasible to have a system that is dynamic, accurate, readable, valid, and accessible to explore scientific publications that helps researchers to find related articles through interactive visualization, while providing the similarity scores between articles?*"

The research question drives the possibility of building an automated scientific publications retrieval and visualization system. We will focus on following characteristics: (i) dynamics in terms allowing the user to search for a publication in multidisciplinary fields from dynamic database, (ii) accuracy in terms of search results and decisions, (iii) readability for the user in terms of understanding the user interface including the Citation Network, (iv) validity in terms of similarity score calculations, and (v) accessibility on the web for all researchers from anywhere, regardless of the institution to whom they belong.

Consequently, the main goal of this thesis is to answer the research question by:

1. Exploring technical and bibliography fields for designing Citation Network visualization and a similarity measurement system which includes the above characteristics.
2. Demonstrating the results obtained from the system (and to compare them to one of the available methods for document similarity measurement and human estimation scores).

1.3 Research Contributions

Our system makes the following contributions:

- A novel *dynamic* web based system that is *directly* connected to a scholarly search engine, which allows the users to *search* in *multidisciplinary* domains and *retrieve* the scientific publications *automatically*.
- An interactive visualization system which *visualizes* the search results as *interactive* graph representation based on *citation/reference relations* between publications, where the nodes represent the literature while the links represent the citation/reference relations to make the *Citation Network*.
- It provides *visual cues* on the graph, such as node and link colors, sizes, labels, and positions. These are used to indicate the type of citation/reference relation so that users can understand the hierarchy structure of the Citation Network.
- A Novel *similarity measurement method* that calculates the similarity between a pair of selected publications. This is based on both the number and the length of paths between them. We have conducted *a case study* to evaluate the effectiveness of our proposed system compared to other methods for document similarity; (i) our proposed method uses citation/reference network (ii) the similarity measurements judged by the human user

which is assumed to be the truth value for the comparison (iii) Natural Language Processing method to measure the similarity between two documents [1].

1.4 Thesis Organization

In this section, we will explain the organization of the rest of this thesis as follows:

Chapter 2 elaborates a literature review on the concept of information retrieval, information visualization, and document similarity measurements. Furthermore, it presents related works and compares the similarities and differences from our work.

Chapter 3 outlines the framework of this research and the methodology used to create the system. It covers the design of our system, methods used for Search query and collecting data, retrieving publications and their attributes, Citation Network Visualization, and similarity measurements.

Chapter 4 presents the evaluation of the proposed system through a comparison study with another document similarity tool which uses the Natural Processing approach. Additionally, it explains the statistical analysis and depicts the results.

Chapter 5 concludes the thesis by summarizing the results and discussing some open problems suggested for future work.

Chapter 2. Literature Review and Related Work

In this chapter, we shed light on several methods related to the main concept of the thesis, which is bibliographic visualization focusing on the Citation Networks. We describe some of the available information visualization techniques and documents similarity measuring methods. In addition, we present other existing bibliometric network visualization systems and explain their properties compared with our system. This chapter is composed of the following sections: information retrieval, information visualization, Citation Network, and document similarity measurements.

2.1 Information Retrieval

The term information retrieval (IR) was introduced in 1951 by Calvin Moores [5], who defined it as:

"Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information. It is another, more general, name for the production of a demand bibliography. Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation. Information retrieval is crucial to documentation and organization of knowledge".

Nowadays, IR systems are considered a widespread research field for both computer scientists and library and information science researchers in different domains. These areas include education, research, business, entertainment and pleasure [6]. Further, it supports users

who use the technologies for browsing or filtering collections of documents, or processing a group of retrieved documents to develop a cluster of the documents based on the contents. [7].

Looking for information on the internet can be done either using a web search engine (such as Google, Yahoo and Bing) or browsing categorized directories (such as Yahoo and BOTW). Three main examples for information retrieval systems are as follows: (i) *Web Search engines* are the most popular example of IR services which allow users to access the needed information such as the daily news and events, to locate people and organizations, to observe diverse electronic shopping services, to answer any queries, and much more [6]. (ii) *Digital libraries* are another important example of IR services. These help students and academic researchers to discover and obtain the required information regarding journals, scientific articles, conferences, and to connect them to recent research news in their specific domain. Additionally, (iii) *Desktop and file search systems* allow users to search for files, documents, and e-mail on their local disks. We discuss them in further detail in the next subsections. Note that we do not discuss multi-media retrieval, but focus on text-based information retrieval.

2.1.1 Web Search Engines

Searching on the World Wide Web (WWW) is the daily activity done the most for many computer users [8]. Many computer applications including "search" on the internet are widely used either via the search engine directly (such as Google and Yahoo) or implicitly such that using applications for finding health and medical sciences, foods, restaurants, museums, bus routes, airline reservations, places or information everywhere throughout smart phones, tablets and computers. Some of the search engines are general like Google and Bing and some are in a specific domain such as an academic search engine (Google Scholar and Web of Science) or a

health and medical search engine (PubMed, HealthFinder and iMedisearch). Users send the query and expect to get accurate answers. The search process behind this simple interface is divided into three main steps: web crawling and indexing, ranking algorithm, and filtering the results.

Web search engines work by storing the web pages information from the html itself by a web crawler. The web crawler, also called web spider or web bot, is automated software or a script that browses the WWW and discovers the new and updated web pages. It then follows the hyperlinks and parses those pages for the new links recursively [6] [9]. In order to obtain accurate results, the contents of each page will be analyzed to determine how it should be indexed. Web page information (addresses) is stored in an index for later use to organize the information and make the information retrieval process quick [9]. Furthermore, the search engine's ranking algorithm selects the top-ranked pages based on different clues such as the title, the content and structure of the pages, the hyperlinks between pages, the user geographic location, or previous searching behavior [6]. These clues play an important role in selecting which pages are the best matches and in which order the result should be ranked. The effectiveness of the search engines are determined by the accuracy of the results and how relevant they are to the entered query by the user.

In this thesis, we focus on one of the popular academic search engines, which is *Google Scholar* search engine. This search engine was released by Google in November 2004 (Figure 2-1). It is a free search engine used to search for scholarly literature, theses, books, abstracts, and papers in multidisciplinary scientific fields across different sources such as academic publishers, universities, professional societies, other web sites, leading journals, conferences and other scholarly organizations [10]. Google Scholar uses the Googlebots web

crawler to browse and gather documents from the WWW to then be added to the index [11]. New websites, or any update or deletion of a webpage is noted by the crawler and then used to update the index. Google Scholar uses an automated algorithm "*parser*" to identify if the document is scholarly or not. Furthermore, it ranks the results according to relevance based on combined ranking algorithms considering the title, the full text, the author, the publication, and how often it has been cited in other articles (Citation count) [12] [10]. The exact ranking algorithm is unknown. However, Beel et al. in [12] concluded in their reverse engineering of Google Scholar's ranking algorithm that the article's citation count and its title are the most important factors for its ranking in Google Scholar.

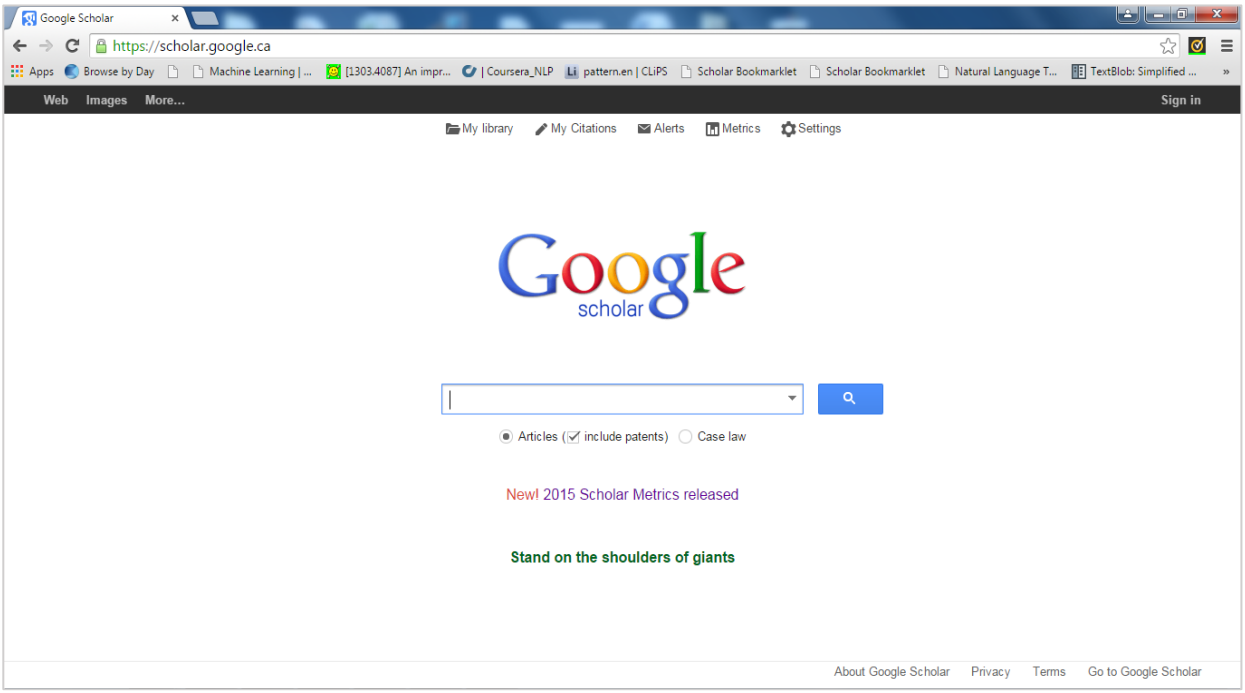

Figure 2-1: Google Scholar Search engine web interface, screen shot from [10].

2.1.2 Digital Libraries

A digital library is a special library with a collection of electronic text documents and digital resources such as multimedia sources including visual, audio and video materials. They

are selected according to certain criteria, stored in electronic format, organized in a specific way to make it available for access through the network, and to archive it for future uses [13], [14].

The purpose of having digital libraries is to assist users in seeking the needed information at anytime and anywhere. Wide varieties of the provided services are different according to the type of the library or the organization who owns the library. These differences are as follows: *(i) Academic digital libraries* involve institution's books, papers, theses, and other electronic publications from the institution which may be available for access to public users with some restrictions. The users belonging to that institution such as Digital Library Systems and Services (DLSS) at Stanford University, which is the information technology production of Stanford University Library (SUL), serves as the research gate for new technologies and methodologies related to library systems [15]. *(ii) Catalogue and archive digital libraries* involve historical information and documents representing a culture such as the American Memory, which is an archive library within the Library of Congress in United States [16], [17]. *(*iii) *Specific domain digital library* has sources for certain knowledge fields such as Health Sciences Digital Library at Michigan State university (MSU). This includes materials and services related to health sciences and medicine at MSU and provides access to medical databases and electronic journals for all faculty staff and students in MSU's medical college or other colleges who are in clinical and biomedical area [18].

2.1.3 Desktop and File System Search

Computer's users nowadays have many electronic files (documents, pictures, audio, video, web browser history, e-mail archives, etc.) stored in the hard disk in different locations. Finding a specific file is a challenge within a wide hierarchy folder system. Desktop search

engines allow search and browsing for data files stored on a hard disk on a user's computer or on a disk connected over a local network [6]. Some of the desktop search service comes built-in the operating system such as (Windows Desktop Search for Windows OS [19] and Spotlight for Apple's OS X and iOS operating systems [20]); others are downloadable applications such as X1 [21] and Lookeen [22].

Desktop search tools generally used at least one of the following information from the user for the search: the name of the file or the folder, the file format, the file content or the creation date. It benefits from the entered information and indexes all of the data on the desktop to achieve better search efficiency of desktop search results [23]. However, the desktop search applications are still less efficient than web search engines due to their lack of a ranking mechanism --like PageRank algorithm- in the file retrieval system i.e. in the search algorithm [23].

2.1.4 Other IR Applications

While search is the main task in the area of IR, this field covers a wide variety of document and text applications including document routing, filtering, text clustering, text comparison, document summarization, Information extraction, question/answering system, and multimedia information retrieval system. Most of the IR researches and applications overlap with natural language processing and machine learning [6] fields in information science and computer science.

2.2 Information Visualization

After discussing the popular types of information retrieval systems and their uses in our everyday life, we will display the need for visualizing the retrieved information in order to make

it more understandable and usable. This will be done by explaining the different visualization techniques in this section. Nowadays, information plays an essential role in our life, and the amount of it is rapidly increasing. This may lead to losing the value of this information. For that, the need to represent the massive amount of data in a good visual representation demonstrating the characteristics, similarities, differences, and the relationship among this information is of utmost importance. In 1998 Card et al. in [24] defined information visualization as "the use of computer-supported, interactive visual representations of data to amplify cognition".

Some approaches to information visualization are defined by Purchase et al. in [25] as follows:

- Explanation of a visualization based on the exterior physical form (syntactic, semantic and stylistic structures) made by a user.
- Investigation of the external representation model through interactive facilities provided by a visualization tool.
- Investigation and manipulation the interrelationship of the internal data model to achieve appropriate representation.

The information visualization field is connected to Human-Computer Interaction (HCI) in such a way that the human uses the computer-based visualization to enrich their understanding of the needed information. It is a handshake between the human and the machine (see Figure 2-2). So, the main objective of information visualization tools is to facilitate the machines ability including: graphics, data mining, statistical analysis and semantic-based approaches to help users to enhance cognition, to recognize patterns, and to make good decisions that could be used for building an appropriate visualized model [25].

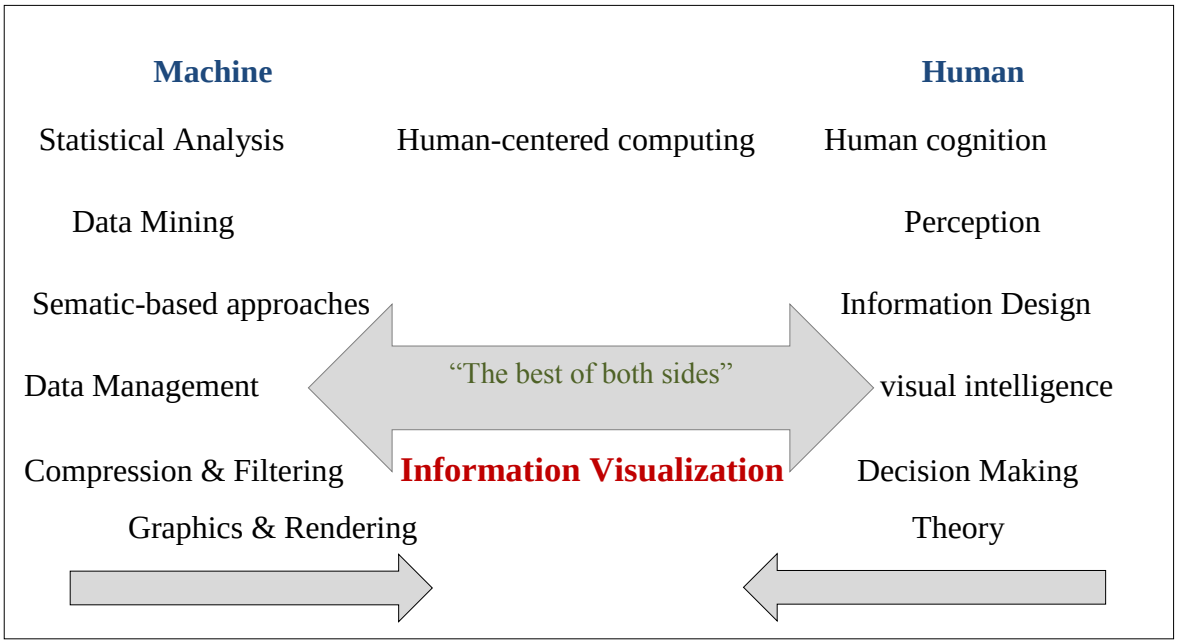

Figure 2-2: Visual analytics combine scientific disciplines to enhance distributing the work between human and machine [26].

Numerous visualization techniques have been proposed for visualizing information in order to transform the textual data set into meaningful visual representations. The most important task is how to choose the appropriate visualization technique to represent the data in order to achieve the users understanding and cognition goal.

2.2.1 Visualization Techniques

Many visualization techniques have been developed during the last decade to represent different types of abstract data and analyze it. A number of visualization toolkits or libraries have also been developed to make the visualization process much easier to the researchers. In this section we will explain some of the common data visualization techniques:

- **Parallel Coordinates :** is a technique used to plot singular data crosswise over many dimensions in such a way that every dimension is linked to a vertical axis, and every data object is shown as a series of connected points along the axes [27]. Stephen Few

discusses in his article [28] the features of this technique to analyze multivariate data by explaining an example of parallel coordinate visualization that measures several aspects of United States (U.S.) counties. This example illustrates seven variables to measure an aspect of U.S. countries (home value, farm acres, average income, population, senior population, college graduates and expectancy) among 3,138 countries, each line in the graph is associated with a series of values that measures the aspects and represents a county in the U.S. The highlighted line represents Alameda (see Figure 2-3).

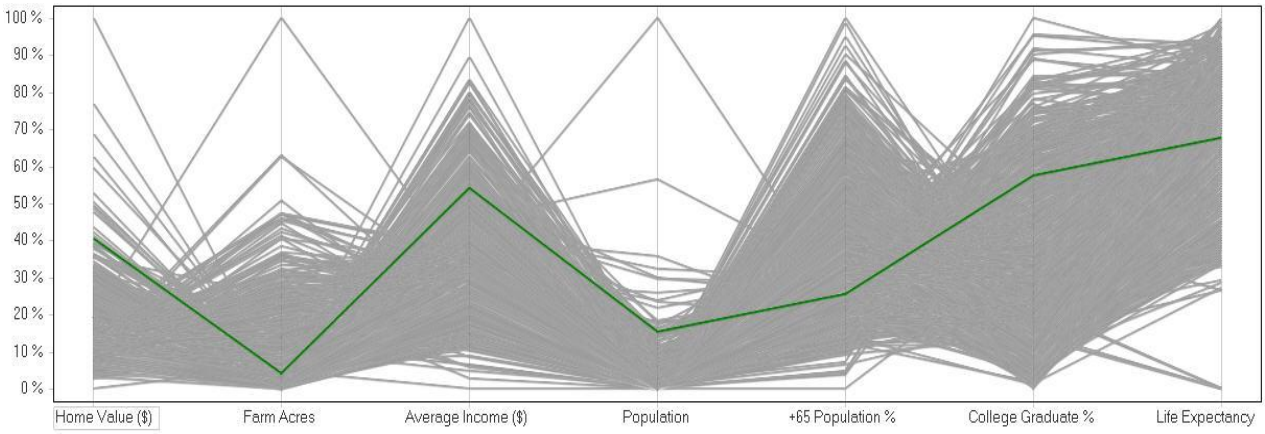

Figure 2-3: Example of Multivariate Analysis Using Parallel Coordinates by Stephen Few (Reprinted from [28]).

- **Tree Map:** This method used to represent hierarchical information. It has been proposed by Johnson et al. in [29] to present hierarchical information structures efficiently on 2-D display surfaces in which the rectangular screen space is divided into smaller rectangles, and then each rectangle is further divided into smaller sizes for each level in the hierarchy. Such that, each data object from the dataset is represented by a rectangle in different size and color [30]. The basic goal of tree-maps was to visualize the contents of hard disks with tens of thousands of files in many levels of directories. It is good to know which directory or file is taking up most of the disk space. As shown in (Figure 2-4) the content

of Shneiderman's hard drive, one picture for the entire hard disk and the screen is subdivided into rectangles representing the files in the hard disk in which the area indicates the file size and the color shows the file type [31].

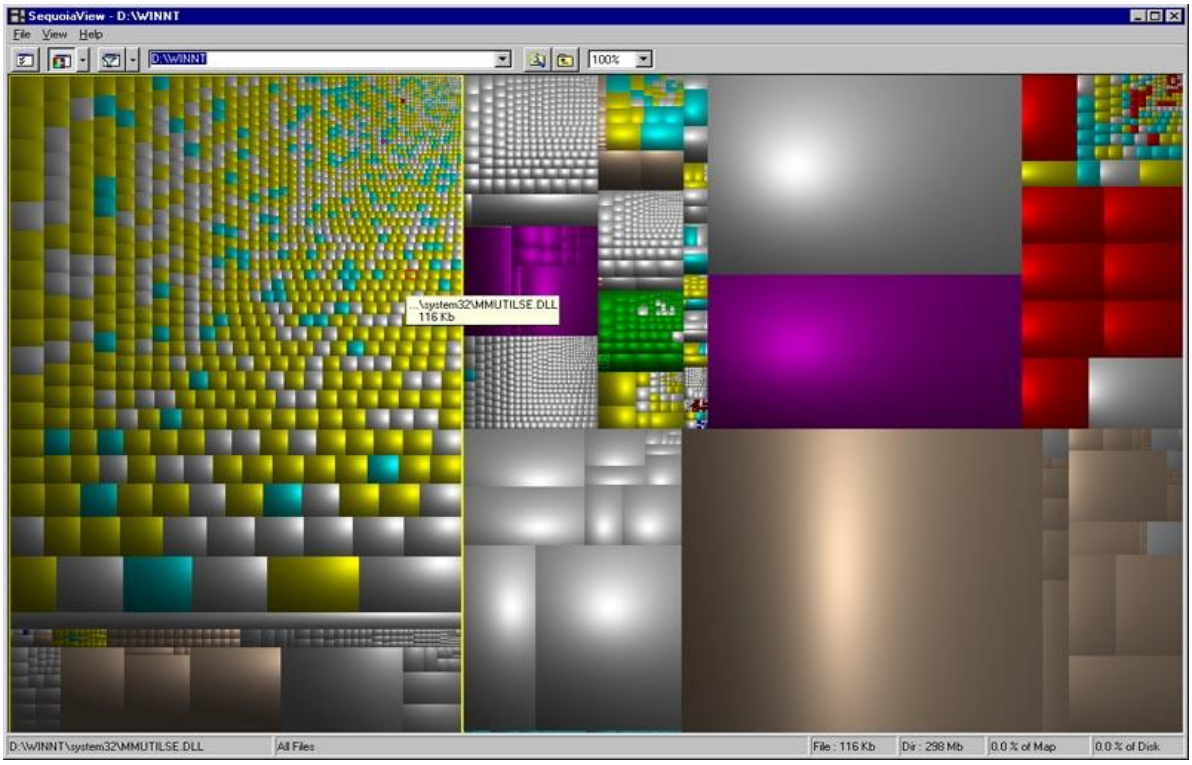

Figure 2-4: Example of tree map technique (SequoiaView tool [32], Reprinted from [31]).

- **Time Line:** is a graphical representation of a chronological sequence of events along a drawn line (horizontally or vertically) to support the understanding of the relationship between different events simply [33]. The time line can be drawn as simple as a line representing the time and text to indicate the events, or it may include more information and images depending on the objective of the visualization (see Figure 2-5) [34].

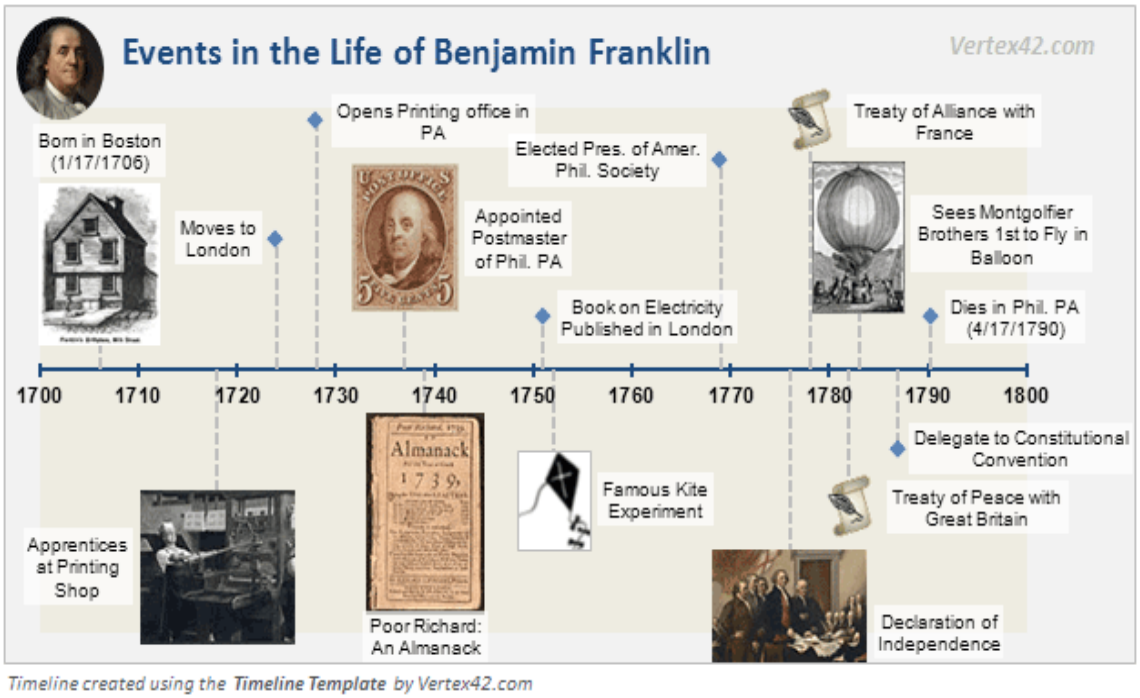

Figure 2-5: Example of time line. It shows time line visualization for events in Benjamin Franklin life (Reprinted from [34]).

- **Data Flow Diagram (DFD):** It is used for structured analysis and software design. It is a visual representation to describe logic models and expresses how the data is transformed in the system (the flow of the data) [35]. Some of the DFD characteristics are: (1) supports the analysis and requirement phase of system design; (2) represents as a diagramming technique; (3) illustrates a network of functions/processes of the system; (4) demonstrates the stepwise improvement through the hierarchical decomposition of processes; (5) explains what type of data input the system requires, where the data came from and where it will be stored, and what type of output the system will produce [35]. Figure 2-6 shows an example of DFD for the flows of information for the registration process [36].

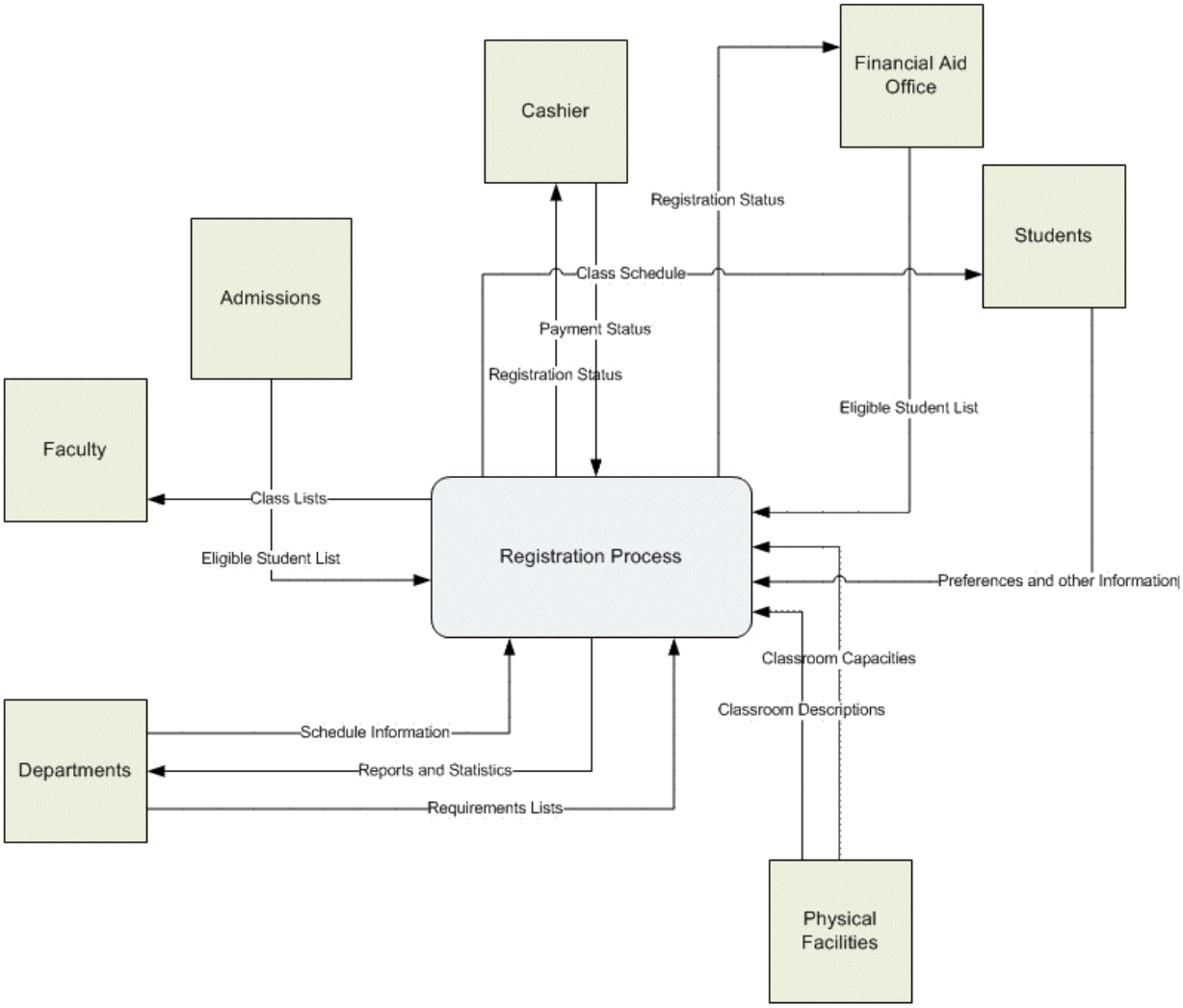

Figure 2-6: Example of DFD. The context diagram of the registration process in a faculty (Reprinted from [36]).

- **Semantic Network:** It is a directed graph consisting of nodes and links. The nodes represent individual entities and the links represent the relationships between those entities. The link is labeled by the name of the relationship it signifies. Each node is represented by a unique value but several links can have the same label [37]. This technique is often used for representing semantic relationships between different concepts. A semantic network is a knowledge representation schema based on human

perception and used to express knowledge about a group of concepts [30]. Many applications in Natural Language Processing are used in semantic network visualization as a form of knowledge representation. Another example that uses the semantic network approach is models based on linguistics such as *WordNet, the lexical database of English* [38]. In this thesis, we will take the benefit of Semantic Network to model our Citation Network such that, the nodes represent the literatures and the arcs represent the citation relations n*ode A cited by node B.* Figure 2-7 shows a simple example of semantic network.

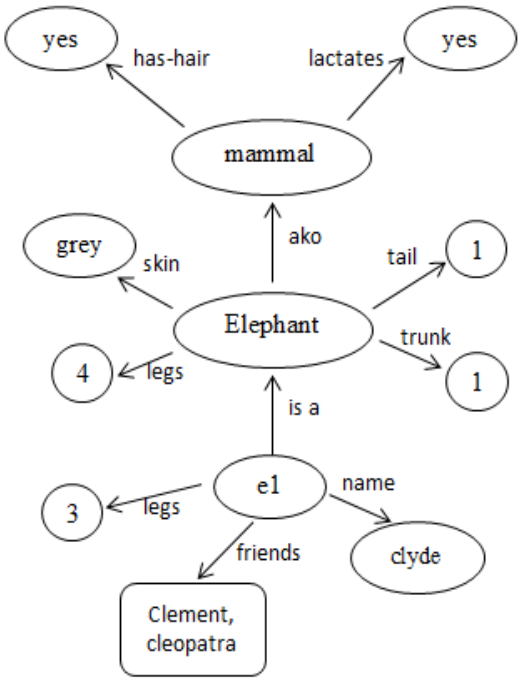

Figure 2-7: Example of semantic network shows the inheritance relationship for the elephant named Clyde (Reprinted from [39]).

There are many other information visualization techniques existing such as the Venn diagram, Entity-Relationship diagram, Flow chart, Cycle diagram etc.

2.3 Bibliometric Networks Visualization

The existence of different visualization techniques has been used as a powerful method to enrich visualizing and analyzing the bibliometric networks. A bibliometric network consists of nodes and links or edges. The nodes can be instances of publications, journals, authors or keywords while the edges indicate the relationship between the nodes which can be citation relations, co-authorship relations or keywords co-occurrence relations. Many studies have been conducted on bibliometric network visualization in order to improve the research knowledge for professional researchers, students, publication agencies and the public.

In this section we will explain the types of bibliometric networks and some popular bibliometric visualization approaches. Then, we will discuss and compare some existing software tools for bibliometric network visualization. We refer to Eck et al. for further reviews in this section [40].

2.3.1 Types of Bibliometric Networks

The most common types of bibliometric networks visualization are Citation Network visualization of publications, co-authorship relations network visualization, or co-occurrence keywords relations network visualization.

2.3.1.1 Citation-based Bibliometric Networks

The citation analysis study has been established by Garfield et al. (1964) [41]. Garfield explains the benefit of the citation in his paper [42] (1979) as:

"The citation is a precise, unambiguous representation of a subject that requires no interpretation and is immune to changes in terminology. In addition, the citation will retain its precision over time. It also can be used in documents written in different languages."

Likewise, Shaw [43] (1979) states "citation establishes a relation among authors which is a measure of the extent to which they communicate indirectly through the literature."

The citation relations can be defined as different classifications. Direct citation relation is defined when there is a direct citation relation between two publications. Co-citation relations proposed by small et al. in [4] is where there is a third publication which cites both publications in a way the more publications which two publications are co-cited means the stronger co-citation relations between them [44]. Bibliographic coupling is the reversed version of the co-citation in which two publications are bibliographically coupled. This occurs when there is a third publication which is cited by both publications, or in other words overlaps in the references list of publications, such that the more common the references in two publications, the stronger relation between them [45].

In this thesis, we combine the direct citation relation with the bibliographic coupling approach to define the relationship between publications. The reason behind choosing this type of bibliometric relations in our thesis is the ability to define the history beyond a publication (predecessors) and/or to follow the trend of a research (successors) by studying the citation relations of that publication. The more successors (citations) or predecessors (references) in common between two publications indicate stronger citation relations.

2.3.1.2 Keywords Co-occurrence Relations Network

Keywords are the most important words that refer to the topic and the scientific field related to the publication. It can be extracted from the title and abstract of a publication or listed by the author in the keyword list. The larger number of co-occurrences of two keywords is the number of publications in which these two keywords occurs together [46].

2.3.1.3 Co-authorship Bibliometric Networks

In this type of network the nodes in the bibliometric network represent the authors or the publishers and the links between them indicate the number of publications that they have authored jointly [40]. This type of bibliometric relations demonstrate the publications linked together between authors, as well as defining which authors are interested in the same research area. However, it does not support our goals to define the similar publications and explore trends of given research.

2.3.2 Bibliometric Networks Visualization

As we explain later in Section 2.2, there are many visualization techniques to represent and visualize the information. The most common bibliometric visualization techniques are the timeline-based visualization approach, distance-based visualization approach, and graph-based visualization approach.

In the timeline-based approach, the nodes are positioned in the network based on a specific time [40]. For example, the publications can be placed in points indicating the date in which they have been published. Usually it is a two dimensional graph; one dimension

representing the time and the other representing the relationship between nodes. Garfield in [47] presents an example of this approach to represent Citation Network of publications.

The distance-based approach network consists of only nodes with no shown edges such that the nodes are placed in a way that the distance between two nodes determines the relatedness between them. The closer the distance indicates higher relatedness. White et al. provides an example of this approach in visualizing the top 100 authors in information science discipline from 1972 to 1979 [48].

Graph-based approach is a two dimensional graph consisting of nodes and links between them to indicate the relationship between the nodes. Chen in [49] and Leydesdroff et al.in [50] are examples of the graph-based approach in their work. We also use this approach in the visualizations phase- the needs of the nodes represent the publications and the links represent the citation relation between the nodes.

2.3.3 Software tools for Bibliometric/Citation Network

Many software tools have been developed for bibliometric networks visualization. We have selected the following tools to describe and discuss their main features and methodology in bibliometric visualization:

- **HistCite** [53] [47]**:** It is a system for mapping and bibliometric analysis of the output of searches using Web of Science. It provides several statistical analysis on the bibliographic data file such that sorting, ranking, and vocabulary analysis which display the frequency of a term. The bibliographic data are represented into tables ordered by author, year, or citation frequency as well as historiographs which take in a small percentile of the most-cited papers and their citation link. HistCite system uses time-based bibliometric network visualization, which

distributes the nodes vertically based on the year of publication. Also, it allows the user to select individual papers from the table to be included in the graph. By default, HistCite includes selecting the top 15 to 25 most cited papers in the visualization of publication Citation Network. These are denoted as historiographs in the system and the algorithm for this visualization called algorithmic historiography.

- **PaperVis: Literature Review Made Easy,** Chou et al. in [51] combined a modified existing version of Radial Space Filling and Bullseye View techniques to place papers as a node-link graph while saving screen space as well. They conducted their experiment on InfoVis 2004 Contest Dataset [52]. *PaperVis* focus on a specific selected paper or keyword according to user defined parameters and places it in the center of the graph. They spatially arranged other papers or keywords relative to the central node while the distance between them corresponding relevant values. It has three modes *Citation-Reference Mode*, *Keyword Mode* and *Mixed Mode*. In The *Citation-Reference Mode,* the relevance value is calculated by measuring the co-reference percentage such that the higher the relevance, the more papers have co-referenced, and the importance of a paper is defined by measuring the percentage of being cited by among the other loaded papers. In the *Keyword Mode,* the relevant value is calculated by counting the common keyword sets among papers such that the papers with more common keywords are placed closer to the central node. As for the *Mixed Mode,* the importance is calculated as in the *Citation-Reference Mode,* and the relevance is calculated as in *Keyword Mode.*

PaperVis provides user interaction in the interface such as *mouse over* or *on click* events to view details about a selected paper and visual clues including node colors and sizes to indicate the citation/reference relationships. Figure 2-8 shows the user interface divided into five regions: A for Modes, levels and color coding parameters, B for review a tree history, C for data filtering,

D for accessing more details about selected paper and E for visualizing the results. On the other hand, it also has some limitations since the system allows the user to have only one focus central node. Furthermore, the lack of semantic analysis in the keyword-clustering algorithm could result in having similar keywords (synonyms) in a different cluster.

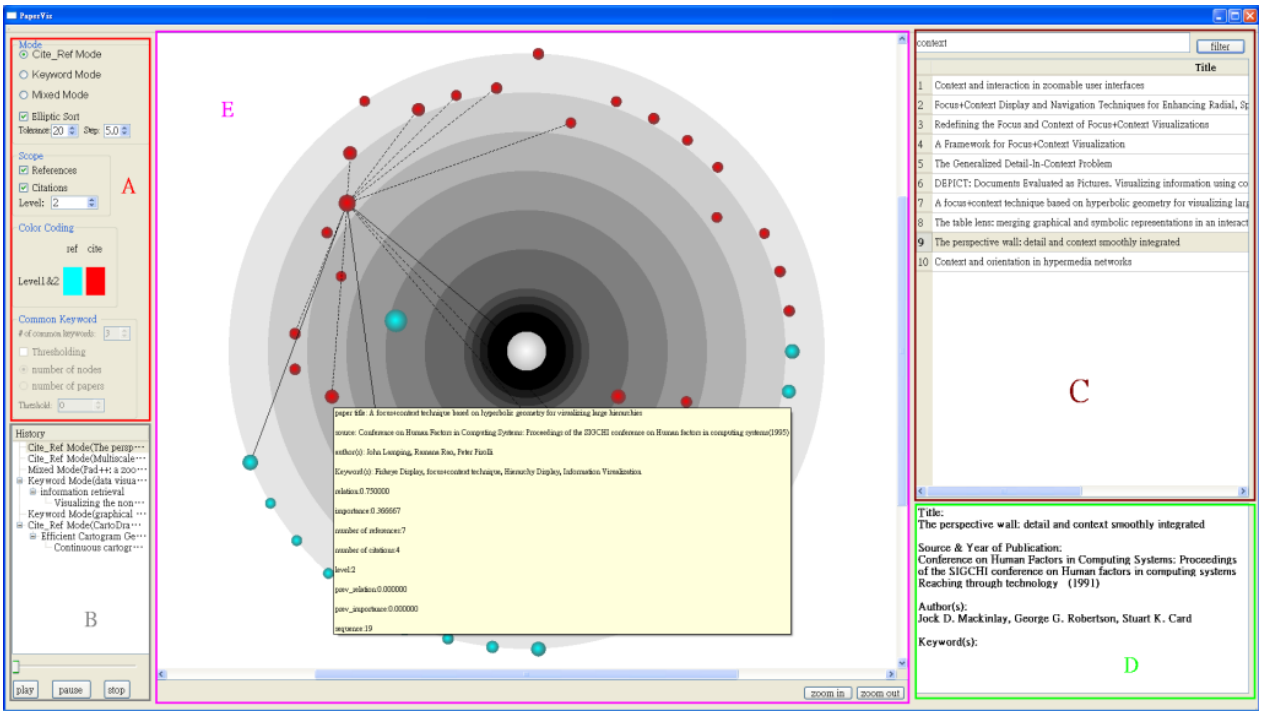

Figure 2-8: The user interface of PaperVis divided into five regions. A, is the system configuration region. B, is the history review tree region. C, is the data filtering and selection region. D, is the detailed information region. Finally, the central region marked in E illustrates the primary visualization results (Reprinted from [51]).

- **Action Science Explorer** [54] is a prototype literature exploration tool for visualizing and understanding the structure of scientific paper collections based on the Citation Network. It uses the JabRef[1] reference manager tool for managing the dataset and SocialAction network analysis[2] tool for the Citation Network visualization and analysis (including ranking, filtering and clustering). These tools are integrated, linked and interacted together to form multiple coordinated views on the screen (see Figure 2-9). Additionally, it uses natural language

[1] http://jabref.sourceforge.net/
[2] http://www.cs.umd.edu/hcil/socialaction/

techniques to analyze the text body of scientific literature, and automatic research summarization techniques of multi-documents of a selected cluster. This integration of the multiple coordinates view in one tool has been developed in order to enrich the rapid understanding of scientific literature. However, it may confuse the user and require a large screen to deal with it.

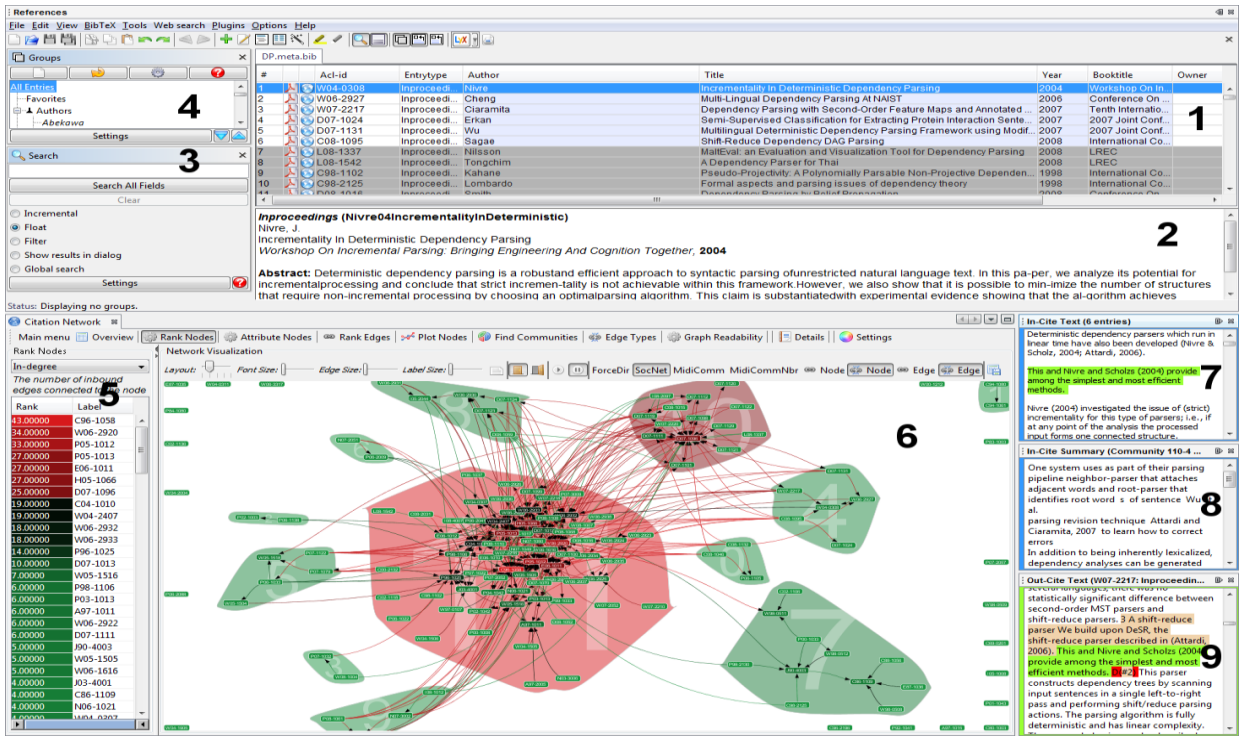

Figure 2-9: The user interface of Action Science Explorer labeled as: (1–4) reference management, (5–6) citation network statistics & visualization, (7) citation context, (8) multi-Document Summaries, and (9) full text with hyperlinked citations (Reprinted from [54]).

- **Citation Network Explorer (CitNetExplorer)** [55] is a tool for visualizing and analyzing Citation Networks. It is implemented in java so it works on any system supporting java. It is processing data downloaded from Web of Science only. This tool has many features; it has two approaches for providing data on publications. When considering data on citation relation between publications, it can handle millions of publications and citation relations. However, the Citation Network visualization is limited to the 40 most frequently cited publications. Also, CitNetExplorer places publications (nodes) in the graph vertically and horizontally. The vertical

dimension assigns publications to multiple layers according to the year of publication (timeline-based visualization) using a simple heuristic algorithm. The positioning of the publications on the horizontal dimension is based on their closeness in the Citation Network. The more citations between publications, the closer they are horizontally (see Figure 2-10). Furthermore, it provides the user with many options for analyzing the Citation Networks including extracting connected components, clustering, drilling down to small subnetwork and expanding the graph or finding the longest path between two publications.

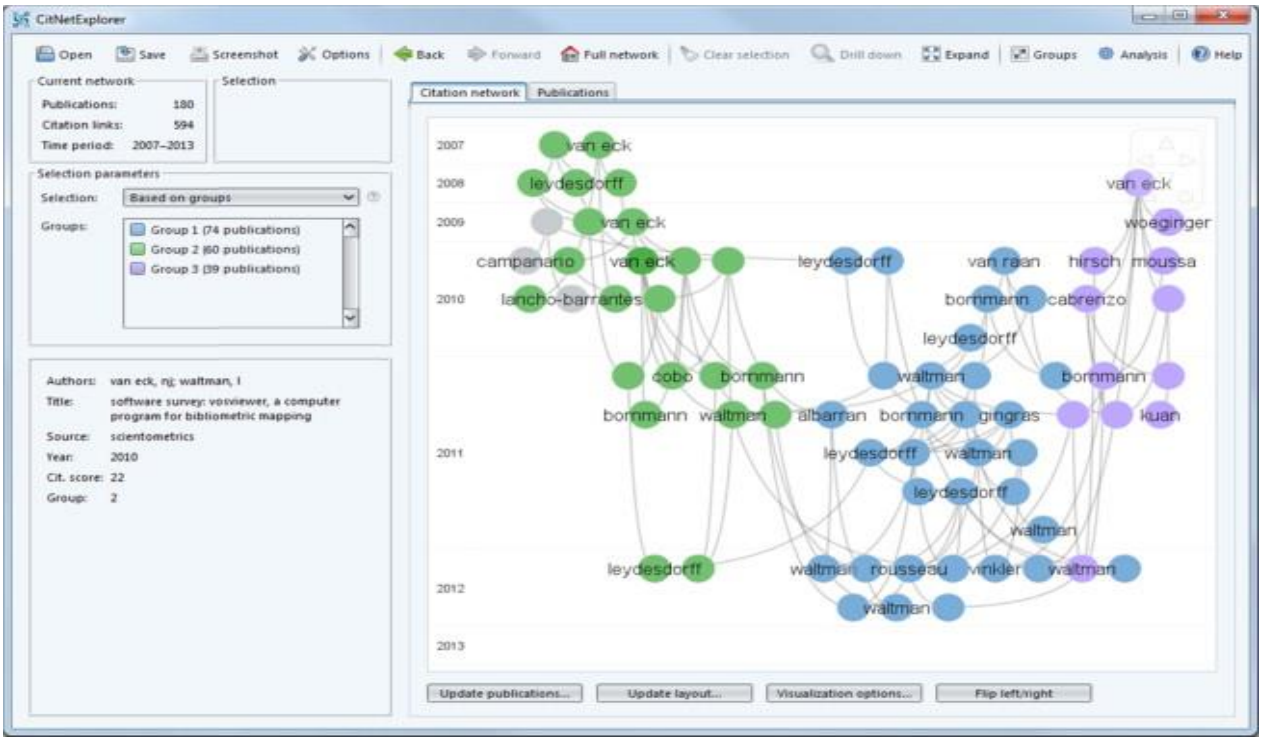

Figure 2-10: The user interface of CitNetExplorer shows the citation network visualization (Reprinted from [55]).

Summary:

After the extensive reviews of the main features of existing software tools for bibliometric visualization *HistCite*, *PaperVis*, *Action Science Explorer*, and *CitNetExplorer*, there are some limitations to be discussed:

- The search and data retrieval process in all of these tools are restricted only within the data stored in the dataset which prevents the user from accessing additional papers or to update the dataset.
- *HistCite* and *CitNetExplore*r accept search results from Web of Science in a way in which users are responsible for providing the publications and citation data as data input. This may lead to errors such as when saving the publication data into a file; only partial information will be available or duplication of the entered data may occur.
- The limitations of the displayed publications or nodes in the Citation Network visualization are such that: in *PaperVis* the visualization focuses on one central node and its connected citation, *HistCite* by default displayed the 15 to 25 most cited papers and in *CitNetExplorer* the visualization is limited to the most 40 frequently cited publications only.
- Another limitation in *CitNetExplorer* is labeling the publications with the last name of the first author and not with the title. This is confusing to the users since there might be two or more publications for the same author but would have same the label.
- Some systems allow the duplicate publication in this Citation Network visualization.
- Finally, some tools have a complicated user interface or multiple view windows like in *Action Science Explorer*, which confuse and distract the users from understanding the layout of the system and the bibliometric network visualization.

Thus, we considered the advantages of the aforementioned research tools in the development of our system and attempted to overcome their limitations. Table 2-1 illustrates a comparison between the main features of the discussed tools and our system based on: the ability to do Search query, the type of the dataset and the data source, the visualization technique they used and the similarity measurements method.

Table 2-1: Summary of Bibliometric Network visualization tools.

Software Tool	Search query	Data set		Data source	Interactive visualization	Similarity Measurement	Similarity Score
		Static	Dynamic				
HistCite	√	-	user needs to upload the dataaset	Web of Science[3] output file	Timeline-based visualization	Citation-based Bibliometric Network	-
PaperVis	√	√	-	InfoVis 2004 Contest Dataset[4]	Distance-based Visualization	Citation-based and Keywords Co-occurrence relations Network	-
Action Science Explorer	√	√	-	The ACL Anthology Network (AAN)[5]	Graph-based Visualization	Citation-based Bibliometric Network	-
CiteNetExplore	√	-	user needs to upload the dataaset	Web of Science output file format	Timeline-based visualization	Citation-based Bibliometric Network	-
Our System	√	-	√	Google Scholar[6]	Graph-based Visualization	Citation-based Bibliometric Network	√

[3] https://isiknowledge.com/

[4] Fekete, J.-D., Grinstein, G., Plaisant, C., IEEE InfoVis 2004 Contest, the history of InfoVis, www.cs.umd.edu/hcil/iv04contest (2004)

[5] http://clair.si.umich.edu/clair/anthology/

[6] https://scholar.google.ca/

2.4 Similarity Measurements

The similarity between documents is determined by the shared features relevant to the nature of comparison. The frequency of observation of such shared features is used to determine the degree of similarity [56]. According to linguistics, it is popular to classify the text features into three classes: syntactic, semantic, and lexical. So, to define if a pair of documents is similar or different is not always done in the same way because it depends on the purpose of the similarity. For example, in this thesis our focus is about searching for scientific publications and measures the similarity between the resulting publications which indicates as "two scientific papers are similar if they share the same thematic topic". This can be *syntactic similarity* in the case of similarity between a query text and corpus documents (text search), or *semantic similarity* in the case of paraphrasing a paragraph or summarizing documents. This produces a new text/document that is semantically similar to the original document [56].

2.4.1 Document Similarity using Natural Language Processing

Natural Language Processing (NLP) techniques play an important role in different types of document processing. This has been explored by Lewis et al. in 1996 [57]. NLP techniques have two main tasks for manipulating with documents: first, identifying important words in a text with syntactic analysis based on its type and on the linguistic context of its appearance in the sentence. Then, clarifying the word sense for document classification. This affects the meaning of the words such that if synonymy words are not recognized to be related, this may lead to classifying two related documents in different classifications. Similarly, if a word has several meanings (polysemy) which are not handled correctly, this may lead to classifying unrelated documents in the same classification [58]. Many similarity measures have been proposed and

applied in previous research work such as Euclidean distance, Jaccard correlation coefficient and Cosine similarity.

To measure the similarity between a set of text documents $D = \{d_1, \dots, d_n\}$, each document *d* is represented as a bag of words, words are counted in the set *D*, each distinct word (term) relates to a dimension in the data space $T = \{t_1, \dots, t_m\}$, where *T* is set of distinct words in *D*. Then, each document represents a vector of non-negative values on m-dimension space $\overrightarrow{t_d}$. Let *tf(d,t)* denotes the frequency of each term $t \in T$ *in* document $d \in D$. Then $\overrightarrow{t_d}$ is the vector representation of a document *d* defines as:

$$\overrightarrow{t_d} = \big(tf(d, t_1), \dots, tf(d, t_m)\big).$$

Generally term frequency signifies the value of a term in the document. The most frequent terms mean the most important or descriptive words in the document. Similarly, when some of the most frequent words appear in a small number of documents in the set, they rarely appear in the other documents within the same set. They tend to be more related and relevant for one group of documents than the other. This is useful when measuring the text similarity between groups of documents. Another important metric to capture and reflect the importance of the term frequencies is to inverse the tf into $tfidf$ (term frequency inverse document frequency), which weights the frequency of term *t* in a document *d* with a factor that measures its importance with its appearance among the whole document collection *D*.

$tfidf(d, t) = tf(d, t) \times \log\left(\frac{|D|}{df(t)}\right)$, where $df(t)$ is the number of documents in which term *t* appears [59].

Before presenting the document as a vector, we have to prepare the document through several steps:

1. Remove the "stop" words which are common but non-important and non-descriptive for the document such as: a, an, the, and are…etc.
2. Stem the words by mapping them into single words such as, removing the suffix and prefix from the words or return the word into the original root.
3. Find the term frequency *tf(d,t).* Also, remove the least frequent words since it does not have an effect on the document.

2.4.1.1 Similarity Metrics

In addition to the previous steps, more preprocessing steps and normalization techniques can be applied depending on the used method and the problem setting. Finally, after preparing the document, we can apply one of the similarity metrics or mathematically "Distance measures". For example, Euclidean distance or Cosine similarity metrics are the most common used methods.

- **Euclidean Distance**: is the basic metric for measuring similarity/dissimilarity. Euclidean distance is the ordinary distance between two points or two vectors. It is defined by calculating the square root of the sum of squared differences between corresponding elements of the two vectors [60]. It is often used for comparing cases and clustering problems. In case of measuring the similarity between two text documents d_a, d_b, first convert the documents to vectors $\overrightarrow{t_a}$, $\overrightarrow{t_b}$ respectively. Then, use *tfidf* .After that, apply the Euclidean formula as:

  $D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{1/2}$, where $T = \{t_1, \dots, t_m\}$ the set of distinct words, and the term weight; $w_{t,a} = tfidf(d_a, t)$, $w_{t,b} = tfidf(d_b, t)$ [59].

- **Cosine Similarity:** is one of the most common similarity metrics for text documents. While documents are represented as vectors, the similarity between two documents on the vector space is calculated by measuring the cosine of the angle between them. It is useful for various information retrieval applications and clustering [59]. The result of the cosine similarity is a non-negative value between [0, 1]. Given $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ are two document vectors, the similarity between them using cosine similarity is calculated as:

  $SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|}$, where $\overrightarrow{t_a}, \overrightarrow{t_b}$ are m-dimensional vectors over the term set *T* [59].

The cosine similarity metric does not depend on the length of the document which allows documents with the same composition and different lengths to be treated identically. This property makes cosine similarity the most common metric for text documents [61]. In this thesis, we compared the performance of our system with the performance obtained from another tool which uses Cosine similarity for Document Similarity. This will be discussed in Chapter 4.

2.4.2 Citation based similarity measures

As we briefly explained in Section 2.3.1 there are several types of bibliometric networks and the citation relations play important roles to analyze and visualize the bibliometric networks. Thus, similarity between publications can be determined by the citation linkages or the co-occurrence similarities.

Citation linkages include direct citation links, co-citation links, bibliographic coupling or a combination of two or all of the above. The citation relation similarities are derived directly from the citation databases. The nature of the citation relations depends on the hierarchy of the provided database [62]. However, determining the strength and the weakness of the relationship

is challenging due to the lack of extensive experiments and standards for comparison on this research issue.

Co-occurrence similarities include co-author, co-citation, or keywords co-occurrence. The common measure of the relatedness or the similarity for this case is to use the NLP similarity metric such that the cosine and Jaccard by counting the frequency of the attribute between the two comparative units [63].

2.4.3 Path Based Similarity Measure

Another semantic similarity measure between objects in ontology is called the path/edge based approach. The shortest path/distance method can be used for measuring similarities between two nodes in a graph which have been used in many applications. Edge counting is another path based similarity method introduced by *Wu et al.* in [64] to find the similarity between two elements in the same ontology based on the distance between each node and the root (N_1, N_2) and the distance which separate the closer common ancestor of them from the root (N) defined as: $Similarity = 2 \times N/(N_1 + N_2)$. *Shenoy.K et al* in [65] built on the work done by Wu at al. and combined the shortest path between the objects and the depth of whole taxonomy together with the distances used in Wu et al. to measure the similarity as follows: $Similarity = (2 \times N \times e^{\frac{\lambda L}{D}})/(N_1 + N_2)$, such that N, N_1 and N_2 as defined by Wu et al., L is the shortest distance between the two objects, D is the depth of the whole ontology tree and λ is 1 for neighbourhood objects and 0 for objects from same hierarchy. *Rada et al.* in [66] also used the shortest path to define the distance as:

"Distance is the average minimum path length over all pairwise combinations of nodes between two subsets of nodes. Distance can be successfully used to assess the conceptual distance between sets of concepts when used on a semantic net of hierarchical relations".

For example: A and B are two objects represented by the nodes a and b in a semantic network, Distance (A, B) = minimum number of edges separating a and b.

All the above studies are different examples of the effectiveness of using the path/edge based approach to calculate the similarities/distances among concepts in a semantic network. This demonstrates the power of hierarchical relations in demonstrating information about the theoretical distance between objects.

In our system, we integrate the citation based similarity measure with the path based similarity measure on the Citation Network of publications in order to define the similarity between two selected publications, and to calculate the similarity score between them.

Chapter 3. Methodology

In this chapter, we explain in detail the techniques used to facilitate our system design and implementation process throughout each phase to achieve our objective. First, our objective in the information retrieval step is to connect the system directly to a scholarly search engine for dynamic access to the Internet database. Then, to import the search results into the system. Following, we focus on visualizing the results as Citation Network graph representation such that the nodes represent the literature while the links represent the citation/reference relations. We utilize the benefit of Citation Network to categorize/cluster the publications based on the citation relation. Furthermore, we came up with a new method to measure the similarity among documents by taking advantage of the graph representation of the Citation Network, and suggest that the paths between the literatures on the graph affect the similarity score between them. In sum, we sketched out our implementation process starting from importing the data collection and creating the Citation Network visualization (CN) to prove our proposed method to measure the similarity by building an interactive web based user interface to connect all the parts of the system.

3.1 System Design

Based on the characteristics and the system requirements discussed in Chapter 1 (including the research problem and from our motivation and objectives), extensive research has been conducted to define the main features and the most significant functionalities of our system. As shown in Figure 3-1, a high-level diagram illustrated our proposed system and its components starting from the Input data including the search query entered by the user and the list of publications retrieved from the Search engine. Then, the main system components are composed

of the parser of the search engine, the publication visualization through an interactive UI, and the proposed method for measuring the similarity between publications. Finally, the output is represented by an interactive CN with the ability to calculate a similarity score between a pair of publications. All these components are connected through an interactive user interface that simplifies the search process for researchers and students and improves the cognitivism of the related publications. In order to fulfill the system requirements, we require the following functionality to be included:

1. Search for publications by allowing the user to enter a search query.
    a. Connect the system to a scholarly search engine directly.
    b. Retrieve the search results including the publications and their scholarly attributes.
2. Visualize the publications as interactive Citation Network graph representation.
    a. Categorize set of related publications by identifying the connected nodes in the CN graph.
    b. Identify the predecessors and the successors of the publications by following the publications cited by link and the list of references.
    c. Display the publication's information on mouse events including title, year of publication, publisher name, URL and number of citation.
3. Define the possible paths between two selected nodes (publications) in the CN.
    a. Find all paths between two selected nodes.
    b. Calculate the similarity score between them.
4. Design an interactive web user interface to connect all the parts together in a simple manner, which allows the user to access the publication easily and interact with the CN visualization from one user interface.

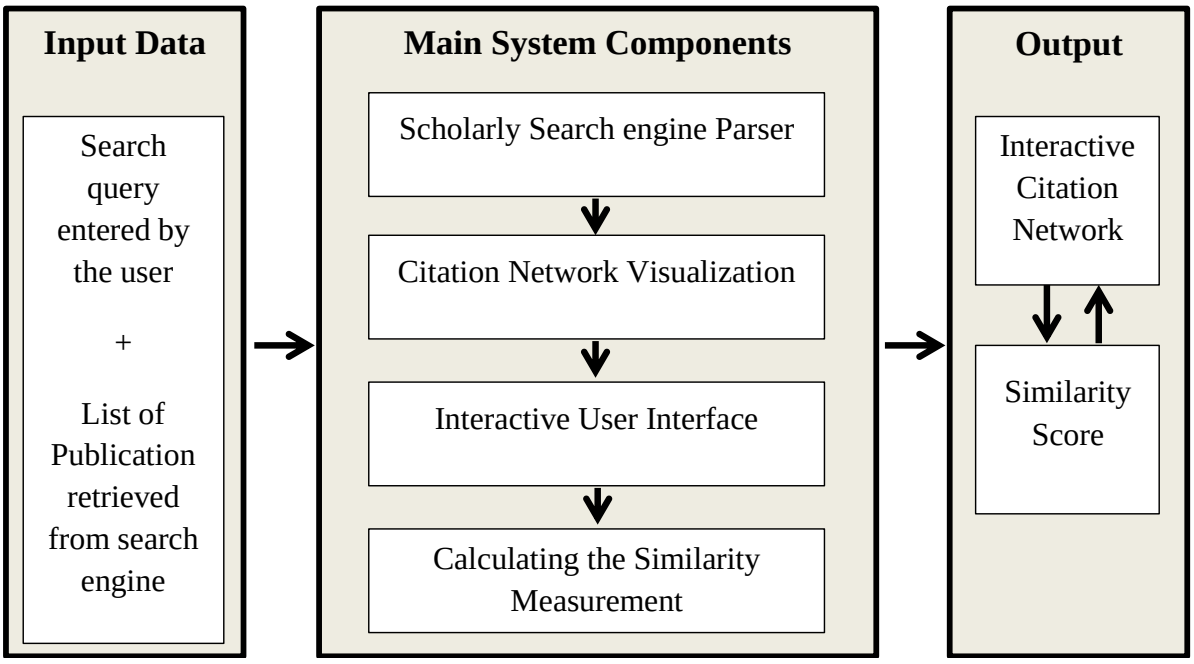

Figure 3-1: A high-level diagram of the proposed system.

3.2 Search Query and Data Import

Searching for publications is the primary task in this research. This process begins with the user entering the query to the system then the system is responsible for making the connection and sending the query to the scholarly database and parsing the search results. Many scholarly search engines exist on the Internet. To select one of them we have defined a list of attributes to be included:

- Free to use for everyone (no need to be register).
- Accessible from everywhere.
- Allowed to search in multidisciplinary scientific fields across many scholarly sources.
- Able to retrieve publications with their scholarly information such as: title, year of publication, URL, number of citations by others, list of citation, and list of references.

Based on the above attributes we utilized Google Scholar search engine as our data source and connected our system directly to it to search and retrieve the requested publications.

In this section, we will explain the techniques we use to connect the system directly to a scholarly search engine for searching and parsing the scientific publications. Then, the format we use for treating the publication list will be used as input data to the visualization process.

3.2.1 Parsing the Publication List

After entering the query by the user, the system will send a request to the connected search engine including the query. While the search process will be executed completely independent on the search engine side, the system will parse the web pages of the search results and save the publication information for the visualization step.

We have chosen Google Scholar as our search engine as it fulfills most of the required attributes listed in Section 3.2. It is a popular search engine used to search scholarly literature in multidisciplinary scientific fields across different sources. It is a port to many scientific literature resources, such as academic publishers, universities, professional societies, other web sites, leading journals, and conferences [10]. Furthermore, they have the ability to find the related publications by tracking the citation (Cited by hyperlink) and linking the user to the original source of the publication through which they can access the full text. However, there is no API for GS, until now, to help programmers make use of it in their applications.

Kreibich implemented an open source python module named *Scholar.py* [67] that includes an inquirer and parser for Google Scholar's first page search results. It parses the search result HTML pages using *BeautifulSoup* python library. It can extract the publication's information such as the title, URL, year of publication, number of citations, number of versions,

and the link to Google Scholar's cluster to save the entries as CSV or text file format. In our system we parse the attributes for each publication as described in (Table 3-1).

Table 3-1: The publication's attributes in details.

Attribute	Description
Title	The title of the publication
Citation list	Link to access the list of other articles citing the publication being viewed.
URL	Link to access the original website where the publication is published. This is good to access the full text and more information about the publication.
Number of Citations	Number of citations cite the current publication.
Version list	Link to access list of all the available version of the publication being viewed.
Year	A year in which a publication has been published.

We used *scholar.py* version 2.4 in our system, with additional features, to make use of it for our purposes. As *Scholar.py* parses only the first page of the search results, we added the ability to iterate the parsing process up to three levels for each publication by making the parsing process a recursive crawler through the *Cited by* hyperlink. In addition, we added a method to parse the search results web pages through Document Object Model (DOM) structure as a Java Script Object Notation (JSON) data format (as described in Algorithm 3-1).

```
1   Create a query object
2   Read the query input from the user interface
3   Generate the query
4   execute the query by send it to the search engine
5   return the results
6   create a list of results as a dictionary
7   for each article in the list
8   {
9     find my children and/or links to my parents/siblings by tracking citation link
10    update the children list
11  }
12  append the results to the article list
```

Algorithm 3-1: Search Query and parsing the search results.

3.2.2 Publications and its Attributes

Retrieving the search results and saving the list of publications and their scholarly information is very important to make a good visualization. As we aspire to visualize the search results as Citation Network visualization, we should take care of the logical structure of the publication list and maintain the relationship between these elements during parsing the HTML search results web pages. We define the relationship as follows: the sibling elements are the

elements located in the same HTML search result page while the children are the elements found by tracking the *cited by* hyperlink for each publication of the search results in order to access the list of citations of a certain publication. So, we chose to parse the search results web pages from HTML format through a document object model (DOM) tree structure known as JSON format. This is done to retrieve the publication list including the scholarly information and the hierarchy tree structure required for identifying the citation relations between publications.

JSON (JavaScript Object Notation) is a lightweight text format that facilitates a structured data interchange format between all programming languages. It is easy to read and write for humans and easy to parse and produce for computer. It is based on a subset of the JavaScript Programming Language. JSON is totally language independent, written in a text format [68].

As we created our JSON data format through a document object model (DOM tree) structure, we state that the *root* to hold the search query and the *children* are the publications with the scholarly information generated by Google Scholar search engine. The results located in the same result web page will be stored in the same level as siblings, and the results parsed from the *cited by* hyperlink for each child will manipulate as successor or sub-children sequentially, and continue the same way if we go from one level to another. In spite of possibly having multiple results of the same publication due to connectivity of the citation relations, it is as simple as two publications being cited by the same publications, or if two publications are citing the same one. To prevent duplicates of the same publication on multiple nodes in the graph visualization, we added an element called *links* to save the redundant publication in rather than in the *children* member. We added a condition before adding any sub-child to a parent in the JSON tree to check if the *title* is already existent in the tree. If so, we add its information under the *links*

member, not under the *children* member. Therefore, in visualization we translate the same concept.

This hierarchy tree structure shows the citation/reference relationship between the publications clearly (Parent cited by children). Figure 3-2 shows the JSON data structure, which illustrates the list of search result information and the relationship between them to produce the Citation Network.

Figure 3-2: The tree structure of our JSON data format.

3.3 Visualization of Publications as Interactive Citation Network

In the visualization, we consider accessibility in terms of the ability to access the visualization directly from our system on the web browser, simplicity in a way the user can understand the content and the structure of the CN, interactivity with the actions that have been taken by the user, and reliability in terms of producing stable and consistent results. Although there are many ways to visualize the publications as discussed in section 2.2.1of the Literature Review and Related Work, we decided to visualize our retrieved data as a Citation Network visualization form. This is done in a way where the nodes are representing the publications and the links between the nodes are representing the citation/reference relations between publications. This is helpful to employ our idea of taking benefit of the citation relations between publications to measure the similarity. In this section, we will explain the algorithm and the technical details used for Citation Network visualization and User Interface implementation of our system.

3.3.1 Citation Network Visualization Algorithm

To make the visualization, we want to convert the list of publications and the scholarly information from the DOM tree structure to a connected directed graph visualization *G*. It consists of the number of nodes, *N* representing the publications and number of edges, *E* representing the citation/reference relations between the publications such that:

$G = (N, E)$; *N*: set of nodes, *E*: set of edges.

$Root = search\ result\ JSON\ data\ labeled\ with\ the\ search\ querey$

$N = \{n_1, n_2, \dots n_k\}$; $n_1, n_2, \dots, n_k$ are the publications.

$E = \{e_{12}, e_{13}, \dots, e_{ij}\}$, set of edges e_{ij} exist if there is a citation/reference relation between publications n_i and n_j and so on which we can extract it directly from the parent/child DOM tree structure JSON data input.

```
1   root= JSON root = search query
2   for each child in children (root)
3       {add node to the node list
4       add edge between the root and the new node}
5   for each sub-child in the children (current)
6       {add node to the node list
7       add edge between the current node and the new node}
8   //if the publication is already exist in the node list
9   for each sub-child in the links (current)
10      {find the existing node in the node list
12      add edge between the current node and the existing node}
13  //adding the references
14  for each node in the references (current)
15      {add node as a reference node
16      add edge between the reference node and the current node}
```

Algorithm 3-2: Algorithm for Drawing the Citation Network Graph

After constructing the foundations of the graph by applying (Algorithm 3-2), the following features have been involved:

- Visual cues to distinguish between the nodes including the color and size properties. The root is colored in yellow and has the largest size, and the publications are colored in blue and are a smaller size. The publications in the references list are colored in green and are the smallest size.
- The direction of the edge expresses the type of relationship between the nodes such that:

$if\ n_j\ in\ the\ children\ list\ (n_i)\ then\ n_i is\ the\ source \rightarrow n_j is\ the\ destination$, i.e. the head of the arrow towards n_j.

$if\ n_j\ in\ the\ references\ list\ (n_i)\ then\ n_j is\ the\ source \rightarrow n_i is\ the\ destination$, i.e. the head of the arrow towards n_i.

Among several available graph visualization tools, we wanted to choose the most efficient tool, which accepts the publication list (JSON data) as an input to produce the Citation Network visualization as an interactive graph on a web-based user interface. Thus, we use a java script library named Data-Driven Documents (D3) [69], a representation-transparent approach for data visualization on the web, for our visualization as this is good for our Citation Network visualization requirements for many reasons:

- The capabilities to web standards such as HTML, SVG and CSS.
- The ability to work on modern web browsers without the need of importing any plugins.
- The ability to support large dataset and dynamic behaviors for interaction and animation.
- The flexibility in transformation of data documents into many visualization forms. It can manipulate any document object model DOM which is the same structure of our JSON data input.

D3 is responsible for mapping the input data into visual elements. The data operator binds input data into nodes in the graph. Since data is represented as an array of arbitrary values, the data is joined to elements by index (the first element to the first datum and so on) [69]. Additionally, D3 takes advantage of new features, web technologies and functionalities that have built in the browsers which simplifies programming issues, especially for mouse interaction. It is easy to manipulate with the graphical objects (SVG for vector graphics objects) and their style

properties such as size, color, position, etc., through a diverse collection of D3's components and plugins.

D3 offers many layouts such as the partition layout, the chord layout, the force layout, the stack layout, and more. In our implementation process, we built our algorithm based on a force-directed algorithm for drawing our graph-based Citation Network since it is the most well-known graph drawing technique [70] [71]. It places nodes randomly into an appropriate layout that fulfils the similarity relations between nodes as well as aesthetics for the visualization (symmetry, minimum edge crossing, avoid overlapping) [63]. Force-directed layout combines physical simulation and interactive constraints. It places forces on the edges and a charge on nodes to position the nodes on the graph as a physical system [72]. We evaluated our algorithm using the force-directed layout because it matches our requirements.

3.3.2 The User Interface

In the user interface we consider the simplicity in the design, easy to learn and use by users, integrated all the parts in one user interface, flexibility in its platform, connectivity to the Internet, dynamic in terms of the ability to search in multidisciplinary fields, interactivity with the user, and reliability in term of producing consistent results. We combined various technologies and programming tools to design and implement our system such as: HTML for page contents, Python programming language for coding the Parser, Cascade Style Sheet CSS for esthetics, Java Script for interaction, Google Scholar for scholarly search engine and so on. The system is running on Google Chrome web browser using a local server Flask[7] web development microframework for Python. In this section we will summarize how to use the

[7] http://flask.pocoo.org/

system for a user starting with a keyword as shown in Figure 3-3. Then, we will illustrate the functionality in detail by expressing the user interface.

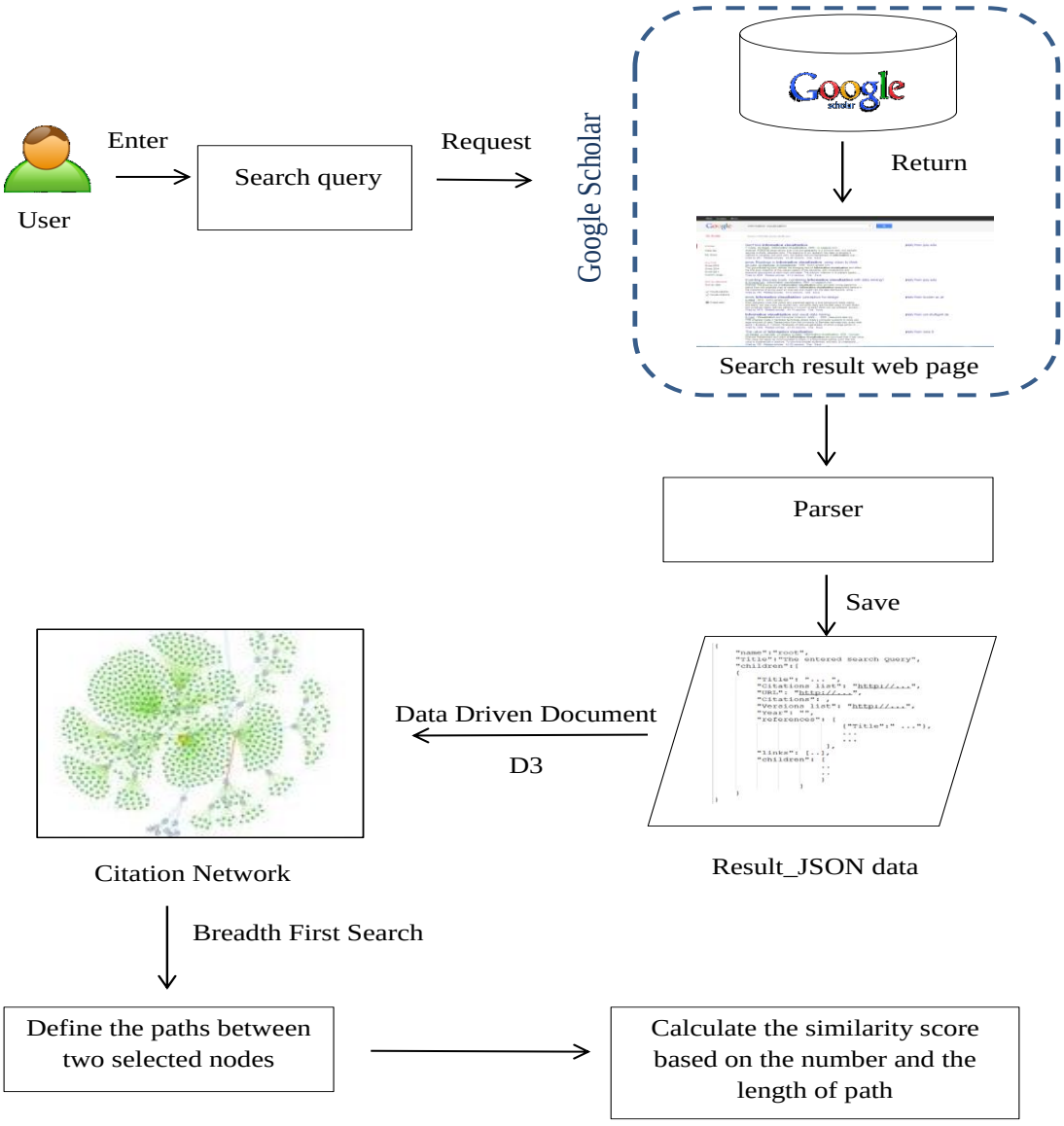

Figure 3-3: How to use the system for a user with keywords

During the execution process of our system, the system goes through four main steps as shown in Figure 3-3:

1. Search for publications by entering a search query in the system and sending the entered query to the Google Scholar search engine to obtain the results.

2. Google Scholar is responsible to do the search process and return the search results in HTML web pages.
3. The system will parse the search result web pages and transform the publication's information into DOM tree structure in Java Script Object Notation (JSON) format.
4. Analyze and visualize the Citation Network using Data Drive Document (D3) java library and force-directed graph layout.
    a. Allow the user to interact with the graph to display the node's information and access the source page of the publication.
    b. Define the paths between any two selected nodes in the network using Breadth First Search Algorithm (BFS).
    c. Assign a similarity score based on the number of paths and the length of each path between them.
    d. Sort the list of publications in descending order based on the number of citations.

First, the system runs on a web browser, which allows the user to access the system from anywhere and enter a search query (including the keyword, the number of publications to be retrieved per level, and the number of citation levels to be parsed) then click on "submit". The system will automatically make the connection to Google Scholar search engine and send the query. Meanwhile, Google Scholar will execute the search process and return the results as HTML web pages. Behind the scene, the system automatically parses the HTML pages and converts the results into JSON data format. This JSON data will be the data source to make the Citation Network visualization. Figure 3-4 illustrates our system user interface. It shows the

search boxes on the top of the screen and the visualization of the resulting Citation Network in the middle. Moreover, detailed information about the selected publication listed on the right side, the similarity measurements and list of the publications sorted by their number of citations also provided on the right side.

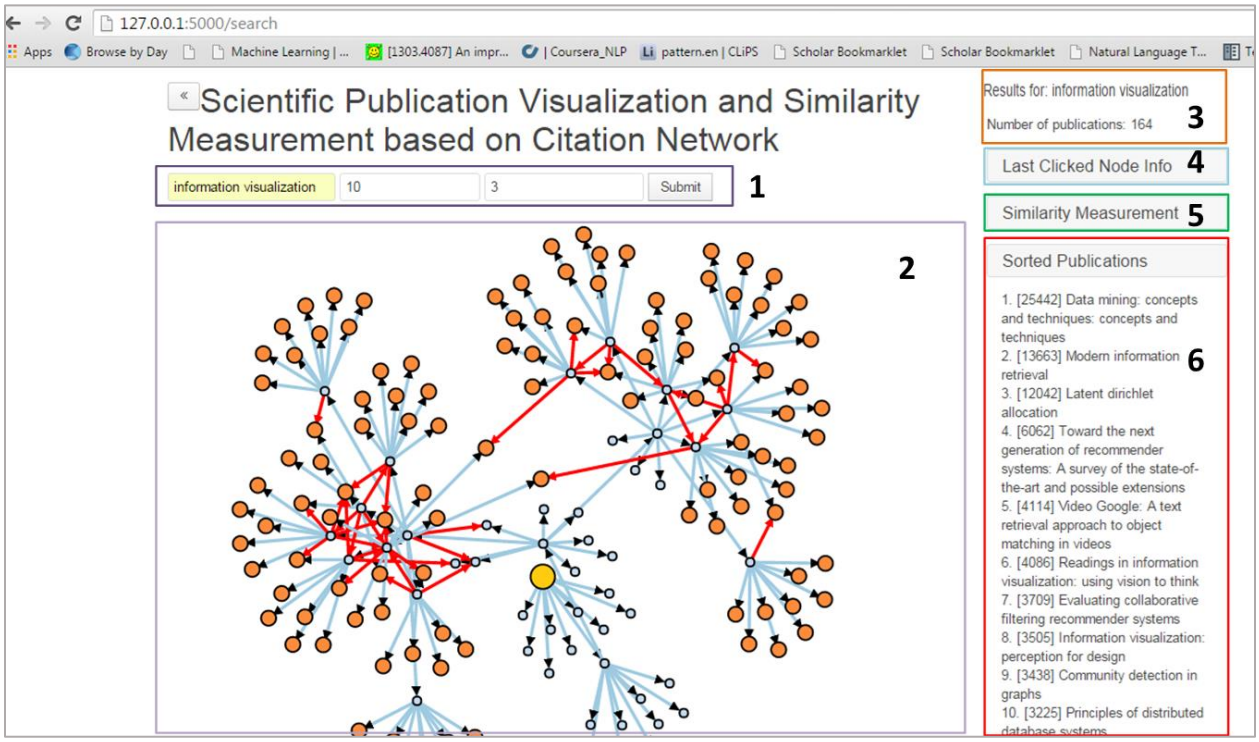

Figure 3-4: The User Interface of our system including (1)the search section to enter the query, (2)the citation network visualization, (3)the search keyword and the number of resulting publications, (4)last clicked node information, (5)similarity measurements and (6)the sorted list of publication.

As our Citation Network visualization is interactive, the system allows the user to access more information about publications in the CN on mouse events (see Table 3-2 and Figure 3-5). In addition, the big collection of nodes in CN demonstrates the cluster of related publications resulting from the same search query and is connected together by the citation relations.

Table 3-2: Actions and Mouse events

Mouse event	Action
On mouse Over	Display the title of the publication
On mouse click	Display the publication information on the right hand side of the screen (Title, Author, Year of publication, URL, Number of Citations)
Double Click	Open the source page in which the clicked publication has been published.

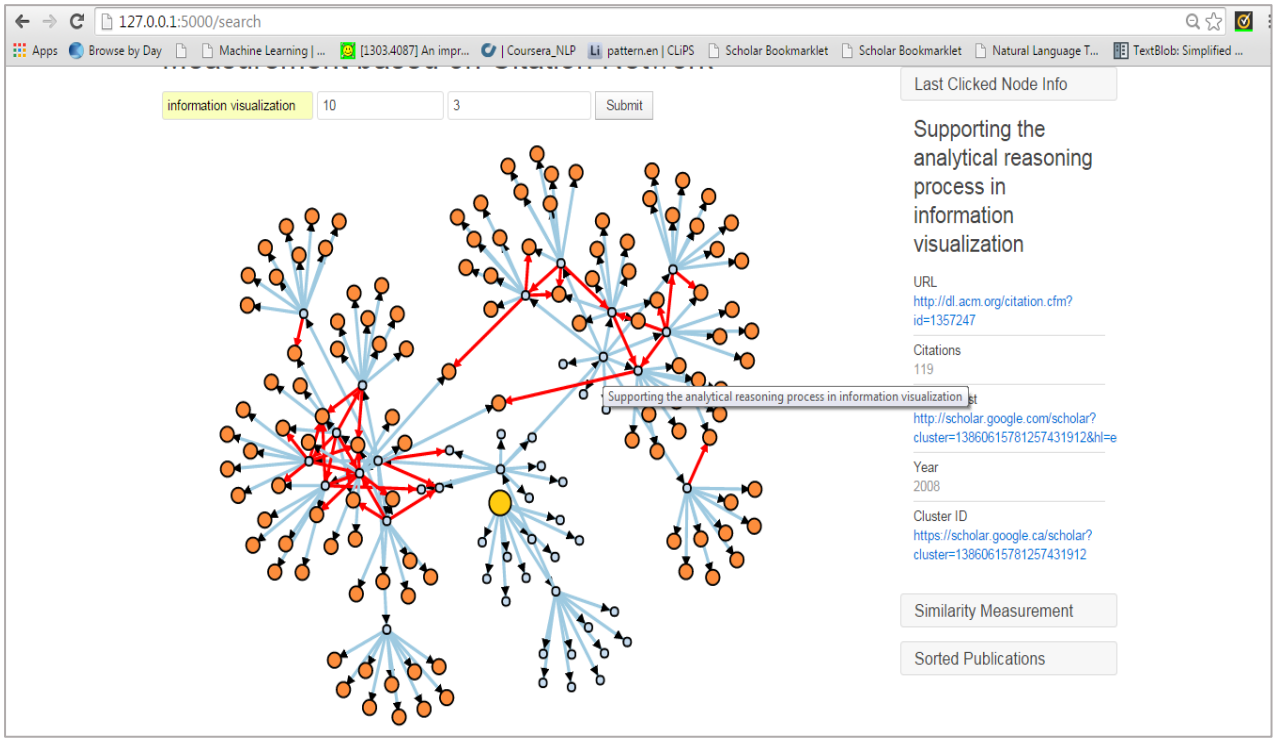

Figure 3-5: Example of the interactivity in our Citation Network graph representation shows the last clicked node's information on the right side.

To find the similarity score between two publications in the CN, the system will ask the user to select the source node and then the destination node to run the similarity measurement algorithm. Furthermore, the system will show all of the possible paths as well as the calculated similarity score on the right hand side of the screen.

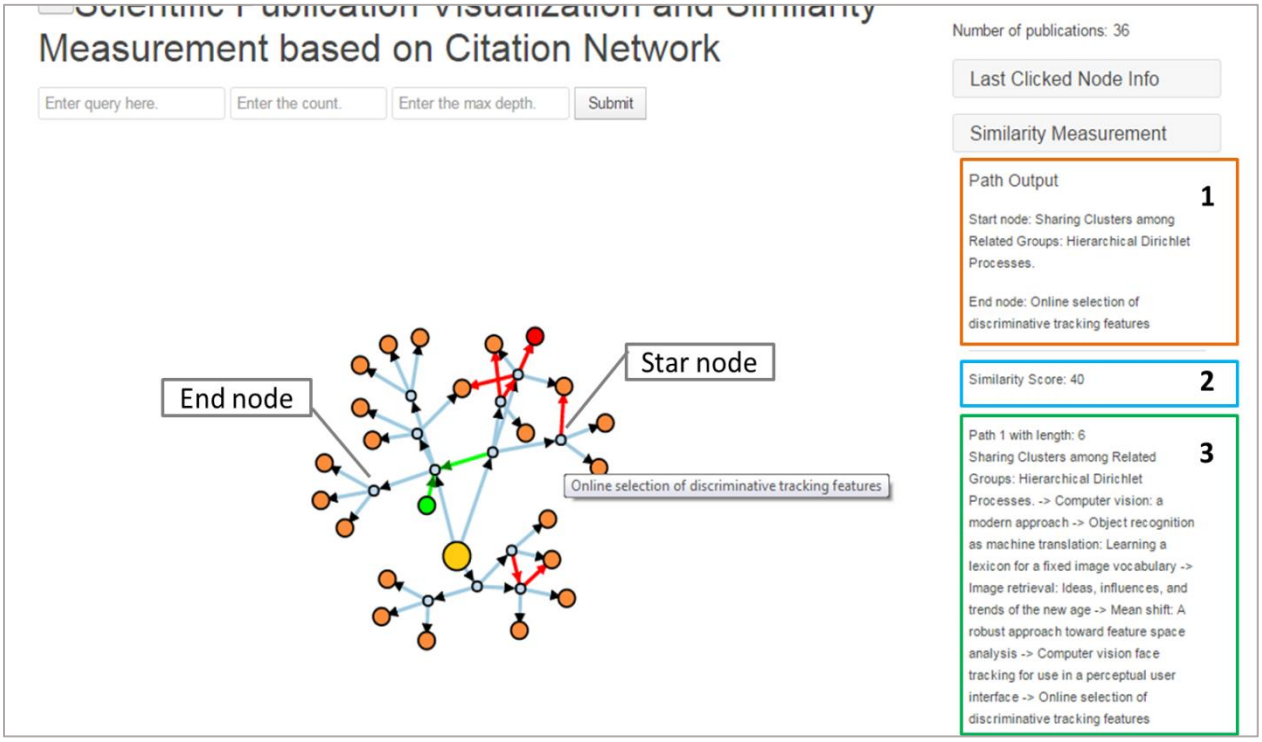

Figure 3-6: The similarity measurement between two selected nodes in our system including (1) the start and end nodes, (2) the calculated similarity score and (3) the paths and the title of the nodes on each path.

3.4 Similarity Measurement Algorithms

As citation/reference information plays an important role in establishing the relationship among scientific publications by researchers [4] [73], we use this advantage to introduce our method in measuring the similarity between any pair of nodes (publications) in our system. We do this by finding the number of paths and the length of each path between them. The more paths with shorter length means a stronger relation, which is represented by a higher similarity score. Our similarity measure is not only between two connected documents, but also from the analysis of multiple levels of the citation related documents. We calculate the similarity score between any pair of scientific publications as follow:

1. Select the source node (n_1), and the destination node (n_2) by the user.

2. Find all paths between the two selected nodes in the Citation Network by running the Breadth First Search algorithm (*BFS*) recursively. BFS technique defines the paths $P(n_1, n_2)$ by exploring the neighbours of each node starting from the source node until reach the destination node. L_p represents the length of each path between the selected publications:

$$P(n_1, n_2) = BFS\ (n_1, n_2)$$

3. Calculate the similarity score between the selected nodes $Sim(n_1, n_2)$, by summation the similarity degree for each path based on the length of the path, in which the shortest is the strongest, as follow:

$Sim\ (n_1, n_2) = \sum_{p=1}^{np} c - 5(L_p - 1)$, $L_p$is the length of the path p and np is the number of paths between n_1 and n_2

This shows that there is an inverse relationship between the length of the path and the similarity score, where c: is a constant value. In our case study, we assigned c to be 45 which means if the length is 1 (direct citation/reference path), then the degree of strengths is 45. This score has been scaled down by 5 if the length has increased as shown in the above equation.

One thing to note is that our similarity measure based on Citation network is being continuously updated over the time as the citation network changes with new citation enters to the network. It is a relative score depends on the depth of the graph and the nature of the citation hierarchy structure. It is very different from a fixed similarity score from natural language processing methods. In our method, two no-citation-related two documents may get a more citation relations after a while and our score reflects the changes. So, our document similarity measurement method reflect how people consider two publications are related over the time.

Chapter 4. Case Study for Similarity Measurement and Evaluation

To measure the similarity between two documents, Natural Language Processing (NLP) method has been a main method, which is a content comparison by text matching. Our Citation Network visualization system (CN) depicts content relation between two documents when they are connected by citation/references directly by one link or through several steps of links, which might be used for document similarity.

We perform a comparison study to check which method between NLP and CN path is closer to human judgement for document similarity.

We collected human judgement based similarity scores to estimate the similarity score between each pair in the dataset as a benchmark (ground truth). We conducted the experiment on the same data set, same evaluation criteria, and procedure. A comparative study applied between the results obtained by our system and the other method. Finally, the system evaluation includes descriptive statistical analysis.

4.1 Data Set

Our dataset was obtained by entering the inquiry "*Information Visualization*" to the search box in our system. We chose this topic because it is related to our work and we can benefit from reading the resulting publications. Then, the system generates the Citation Network visualization of the results automatically. Therefore, we randomly select 10 publications from the entire search result and add one additional non-related article obtained from the GS search engine from the inquiry "*Physics*" on which to prepare our data set for the experiment.

They are as follows:

1. A Novel Approach to Visualizing and Navigating Ontologies.
2. Watson, more than a Semantic Web search engine.
3. Ontology-driven Geographic Information Integration: A survey of current approaches.
4. Inducing Word Senses to Improve Web Search Result Clustering.
5. Graph Visualization and Navigation in Information Visualization: A Survey
6. Active Learning of Expressive Linkage Rules using Genetic Programming.
7. Force-Directed Edge Bundling for Graph Visualization.
8. Winding Roads: Routing edges into bundles.
9. MizBee: A Multiscale Synteny Browser.
10. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines.
11. Bidirectional and efficient Conversion between Microwave and Optical Light.

While our system parses only the publication attributes listed in Table 3-1 including the *title,* year of publication, *URL, citations and versions*, it does not parse the *references* of the publications. This limitation is due to the difficulties in accessing the full text and converting the document to a readable format to extract the references for each publication in the list. Additionally, the publications are in different formats from different sources. Having said that, the system can visualize and process the references correctly. For that, we added the references for each publication in our dataset under the *references* attribute in the JSON data manually to conduct our experiment. This data set will be used for the comparison study.

4.2 Evaluation Setup and Procedure

To assess our system proficiency, we made an experiment that aims to determine whether the system provides useful similarity scores or not. We considered the following evaluation setup including: testing our system, the comparison criteria, the benchmark value for the comparison, another experiment using a different method, and the procedure we follow to conduct the experiment and evaluate the results.

We measured the similarity score between each pair of publications in the dataset using three different methods: the similarity measurement based on Citation Network (our approach) as described in Section 0, Scurtu's Document Similarity API [1] (using Cosine similarity NLP method), and the human sentiment opinions. The goal of this study is to test the effectiveness of our approach in measuring the similarity between publications compared with the similarity scores obtained by Scurtu's Document Similarity API approach in a way that one provides scores closer to the human sentiment opinions.

First, we started the evaluation process with our system by using the data set in Section 4.1 as an input to make the Citation Network visualization automatically (see Figure 4-1). Then, we selected each pair of the nodes representing our data set in the Citation Network to let the system calculate the similarity score between them using our method. After that, we recorded all of the generated scores down in a table for comparison (see Table 4-1).

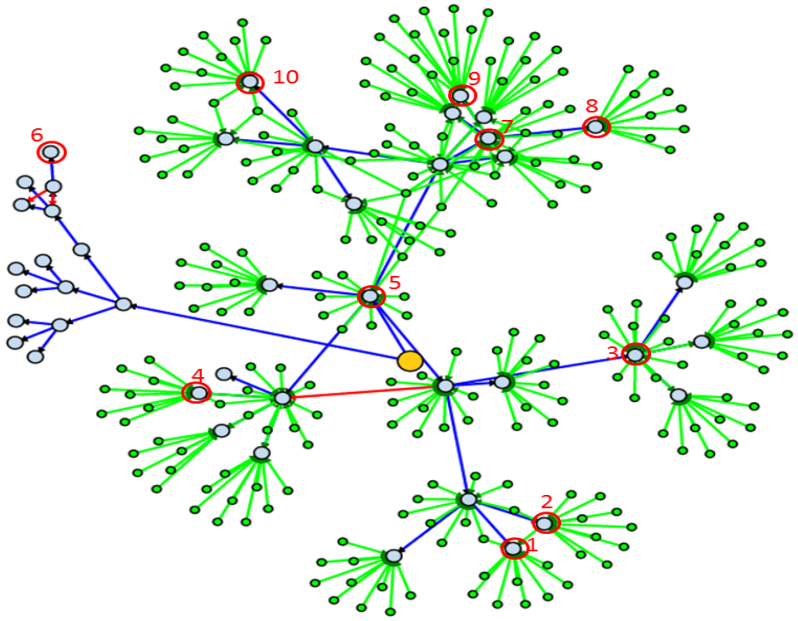

Figure 4-1: The Citation Network visualization for the inquiry "Information Visualization" which represents the distribution of the selected dataset shown in red circles.

Table 4-1: Similarity scores between our Dataset obtained by our system.

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	**Doc7**	**Doc8**	**Doc9**	**Doc10**	**Doc11**
Doc1		45.00	45.00	50.00	45.00	5.00	15.00	10.00	15.00	10.00	0.00
Doc2			45.00	50.00	45.00	5.00	15.00	10.00	15.00	10.00	0.00
Doc3				45.00	30.00	5.00	20.00	15.00	15.00	15.00	0.00
Doc4					55.00	5.00	20.00	15.00	15.00	15.00	0.00
Doc5						10.00	60.00	50.00	50.00	45.00	0.00
Doc6							5.00	5.00	5.00	5.00	0.00
Doc7								75.00	75.00	45.00	0.00
Doc8									55.00	35.00	0.00
Doc9										35.00	0.00
Doc10											0.00
Doc11											

To make a comparison between two methods to decide which method performs better, we need a benchmark value as a truth-value for the comparison. And from our motivation to have a system that measures a similarity score between publications as similar to the human thinking.

Thus, we employ the experiment on measuring the similarity scores between each pair of the data set by collecting human or researcher's sentiment opinions as our benchmark. We asked 15 graduate students in the field of Engineering and Computer Science to estimate the semantic similarity score between each pair of the 11 articles. We defined comparison criteria from a researchers point of view based on the points that any researcher is looking for when searching for related scientific publications as follows:

Publication title: It is the first thing the researcher looks at and checks in any search process because it describes the topic and states the general idea.

Keywords: It contains the most important words or related topic within the whole text

Abstract: It gives a brief summary of any scientific publication including the research goal, the methods and the results.

References: It shows the relationship between this publication and the history behinds this topic. It also defines the predecessors of a node in the Citation Network (CN) graph.

We asked each participant to give a similarity score between each pair of the 11 publications scored from 0 to 100 such that 0 means absolutely dissimilar and 100 means identical. The scoring value is an estimate value from each participant by assessing the similarity between each pair of the set focusing only on the above comparison criteria. Therefore, each participant records the similarity scores in an individual score table. Then, we calculated the average of the similarity scores obtained by all of the participants (Table 4-2). We assumed that the average of the participant's estimation scores to be our benchmark in the comparison study.

Table 4-2: The average of the similarity scores between our Dataset obtained by human judgements.

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	**Doc7**	**Doc8**	**Doc9**	**Doc10**	**Doc11**
Doc1		33.21	43.57	19.29	53.57	14.62	27.14	24.64	19.64	43.93	0.07
Doc2			35.71	56.29	26.79	24.64	19.71	21.07	22.93	21.43	0.07
Doc3				20.00	34.71	23.29	19.64	21.14	20.07	34.29	4.36
Doc4					26.50	17.21	19.64	18.64	18.64	26.07	4.36
Doc5						25.07	57.86	42.86	31.50	48.57	2.93
Doc6							21.79	21.79	18.57	21.79	1.50
Doc7								65.36	27.00	46.57	5.07
Doc8									27.14	41.79	2.93
Doc9										35.71	3.64
Doc10											5.07
Doc11											

Now, we can apply the same experiment on the other method. We utilize the experiment on Scurtu's Document Similarity API [1], which is a free open source library available in three different programming formats (*Python*, *Java* and *PHP*) for measuring the similarity between two text documents. The similarity formula is Cosine similarity (NLP approach mentioned in 2.4.1) for measuring the similarity score between two text documents' vectors. It accepts two texts as inputs; each document is transformed into vectors by splitting it into tokens. It returns a similarity score result between 0 and 1 as output; 0 meaning absolutely different, 1 meaning identical.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

So, we set the comparison criteria (title, keywords, abstract and references) for each pair of our data sets as inputs to Scurtu's Document Similarity (SDS) and recorded the similarity scores in the scores table. Since the resulting scores are between 0 and 1, we scaled these values to be from 0 to 100 for normalization purposes in the comparison (see Table 4-3).

Table 4-3: Similarity scores between our Dataset obtained by Scurtu's Document Similarity.

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	**Doc7**	**Doc8**	**Doc9**	**Doc10**	**Doc11**
Doc1		70.20	57.20	62.1	54.8	68.9	63.5	66	62.3	54.5	47.2
Doc2			51	56.3	47.4	63	49.6	51	50.7	42.4	44.8
Doc3				46.7	43.3	54.3	46.7	49.5	47.4	40.6	36.9
Doc4					53.1	59.8	57.6	59.5	60.8	52	42.2
Doc5						50.3	70.4	65.8	51.6	72.2	48.2
Doc6							61.6	63	61.6	49	52.9
Doc7								84.3	60.8	60.5	47.2
Doc8									60.7	63.7	48.3
Doc9										59.4	47.6
Doc10											49.2
Doc11											

After collecting the similarity scores among the three methods, a comparative study was conducted. We combined all of the results in one similarity scores table in Table 4-4. For each pair of the documents in our dataset there are three similarity scores: the one generated by our system using Citation Network (CN) similarity measure, the average of the human estimation score (AVG) and the score produced by Scurtu's Document Similarity API (SDS). Then, we sorted the scores in the table in ascending order based on the average of the human estimation score column. This to make the comparison more clear and useful for calculating the statistical analysis including Mean, Standard Deviation StD, and Coefficient Correlation r Chi Square test χ^2.

Table 4-4: The similarity scores between each pair of the Dataset among the three methods.

Documents	AVG	CN	SDS
Doc1,Doc11	0.07	0.00	47.2
Doc2,Doc11	0.07	0.00	44.8
Doc6,Doc11	1.50	0.00	52.9
Doc5,Doc11	2.93	0.00	48.2
Doc8,Doc11	2.93	0.00	48.3
Doc9,Doc11	3.64	0.00	47.6
Doc3,Doc11	4.36	0.00	36.9
Doc4,Doc11	4.36	0.00	42.2
Doc7,Doc11	5.07	0.00	47.2
Doc10,Doc11	5.07	0.00	49.2
Doc1,Doc6	14.62	5.00	68.9
Doc4,Doc6	17.21	5.00	59.8
Doc6,Doc9	18.57	5.00	61.6
Doc4,Doc8	18.64	15.00	59.5
Doc4,Doc9	18.64	15.00	60.8
Doc1,Doc4	19.29	50.00	62.1
Doc1,Doc9	19.64	15.00	62.3
Doc3,Doc7	19.64	20.00	46.7
Doc4,Doc7	19.64	20.00	57.6
Doc2,Doc7	19.71	15.00	49.6
Doc3,Doc4	20.00	45.00	46.7
Doc3,Doc9	20.07	15.00	47.4
Doc2,Doc8	21.07	10.00	51
Doc3,Doc8	21.14	15.00	49.5
Doc2,Doc10	21.43	10.00	42.4
Doc6,Doc7	21.79	5.00	61.6
Doc6,Doc8	21.79	5.00	63
Doc6,Doc10	21.79	5.00	49

Documents	AVG	CN	SDS
Doc2,Doc9	22.93	15.00	50.7
Doc3,Doc6	23.29	5.00	54.3
Doc2,Doc6	24.64	5.00	63
Doc1,Doc8	24.64	10.00	66
Doc5,Doc6	25.07	10.00	50.3
Doc4,Doc10	26.07	15.00	52
Doc4,Doc5	26.50	55.00	53.1
Doc2,Doc5	26.79	45.00	47.4
Doc7,Doc9	27.00	75.00	60.8
Doc1,Doc7	27.14	15.00	63.5
Doc8,Doc9	27.14	55.00	60.7
Doc5,Doc9	31.50	50.00	51.6
Doc1,Doc2	33.21	45.00	70.2
Doc3,Doc10	34.29	15.00	40.6
Doc3,Doc5	34.71	30.00	43.3
Doc9,Doc10	35.71	35.00	59.4
Doc2,Doc3	35.71	45.00	51
Doc8,Doc10	41.79	35.00	63.7
Doc5,Doc8	42.86	50.00	65.8
Doc1,Doc3	43.57	45.00	57.2
Doc1,Doc10	43.93	10.00	54.5
Doc7,Doc10	46.57	45.00	60.5
Doc5,Doc10	48.57	45.00	72.2
Doc1,Doc5	53.57	45.00	54.8
Doc2,Doc4	56.29	50.00	56.3
Doc5,Doc7	57.86	60.00	70.4
Doc7,Doc8	65.36	75.00	84.3

We illustrated the scores in Table 4-4 as a line graph representation to investigate the overall appearance of the similarity scores among the three methods as in (Figure 4-2). It demonstrates that the lines representing the AVG and CN are following a kind of similar trend from low to high similarity scores. Despite that, the line representing SDS is always scoring

medium to high. That means that SDS becomes closer or similar to AVG and CN only if the similarity score is medium to high (above 50).

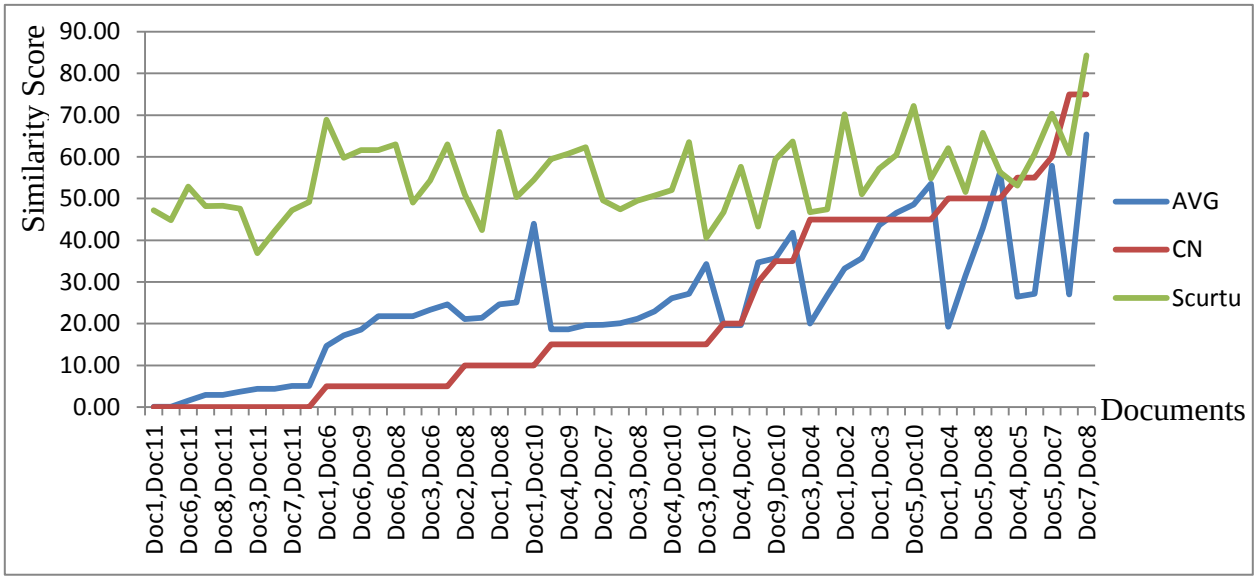

Figure 4-2: Line Graph illustrating the similarity scores between each pair of the dataset among the three methods sorted by AVG scores. The X-axis represents the documents and the Y-axis represents the scores.

4.3 System Evaluation

We conducted our experiment using the three methods to answer the following questions quantitatively:

- Are the similarity score results obtained by our system closer to the results estimated by the human than the similarity scores obtained by SDS?
- How are the similarity scores obtained by our system different from the similarity scores obtained by SDS?
- Is it possible to generalize our method in measuring the similarity between publications?

To answer the above questions, we completed statistical analysis between the similarity scores among the three methods. First, we normalized all the values in the comparison table. Then,

we applied descriptive analysis methods on the values such as the coefficient correlation, the mean value, and the Chi Square test between the averages of the similarity scores obtained by human estimation AVG and the other two methods sequentially.

The findings in Table 4-5 show that the correlation between the average human assessment of the similarity between publications in the data set AVG and the assessment done by our system using Citation Network similarity measurement CN is positive and significant for this sample ($r_1 = 0.746$, $p_1 = 0.007 < 0.01$). Therefore, an increase in the average human assessment by 1 score point implies a 0.746 increase in the score of the Citation Network assessment and vice-versa. The Chi Square test in Table 4-6 shows that the significance of the correlation between the average human assessment AVG and the assessment done by our system CN is systematic, which means that the correlation between these two assessments is valid for other samples and can be generalized. On the other side, Table 4-7 shows that the correlation between the average human assessment AVG and the Scurtu's Document Similarity SDS assessment is significant, medium to high, and positive correlation ($r_2 = 0.548$, $p_2 = 0.118 > 0.01$). Hence, if the average human assessment increases by 1 the SDS increases by 0.548. However, the Chi Square test in Table 4-8 is not significant, which means that the correlation is not systematic and it is only significant in this sample.

Table 4-5: Correlation between AVG and Citation Network (Symmetric Measures).

	Value	**Asymptotic Standardized Error**	**Approximate T**	**Approximate Significance**
Interval by Interval Pearson's R	.746	.068	8.146	.000
Ordinal by Ordinal Spearman Correlation	.764	.070	8.617	.000
N of Valid Cases	55			

Table 4-6: Chi Square Test AVG*CN.

	Value	Df	Asymptotic Significance (2-sided)
Pearson Chi-Square	540.375[a]	462	.007
Likelihood Ratio	231.749	462	1.000
Linear-by-Linear Association	30.022	1	.000
N of Valid Cases	55		

Table 4-7: Correlation between AVG and SDS (Symmetric Measures).

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Interval by Interval	**Pearson's R**	.548	.108	4.764	.000
Ordinal by Ordinal	**Spearman Correlation**	.436	.119	3.528	.001
N of Valid Cases		55			

Table 4-8: Chi Square Test AVG*SDS.

	Value	Df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2048.750[a]	1974	.118
Likelihood Ratio	386.034	1974	1.000
Linear-by-Linear Association	16.191	1	.000
N of Valid Cases	55		

More descriptive statistics have been performed on the results of the three methods in Table 4-9 including the Mean value, Median, the Standard Deviation, the Minimum and Maximum scores for each method. It shows that the Mean, Median, St. Deviation values generated by CN are closer to the values obtained by AVG than the ones from SDS. The boxplot

diagram in Figure 4-3: Boxplot diagram of the mean values of the similarity scores obtained by the three methods. also shows that the trend of the similarity scores obtained from the Citation Network is closer to the human scores trend than SDS.

Table 4-9: Descriptive statistics AVG*Citation Network*Scurtu's.

	AVG	CN	SDS
Valid	55	55	55
Missing	0	0	0
Mean	24.934565434565446	22.91	55.302
Median	21.785714285714285	15.00	54.300
Mode	19.6428571428571420[a]	0[a]	46.7[a]
Std. Deviation	15.400730348336264	21.660	9.2009
Minimum	.0714285714285714	0	36.9
Maximum	65.3571428571428600	75	84.3

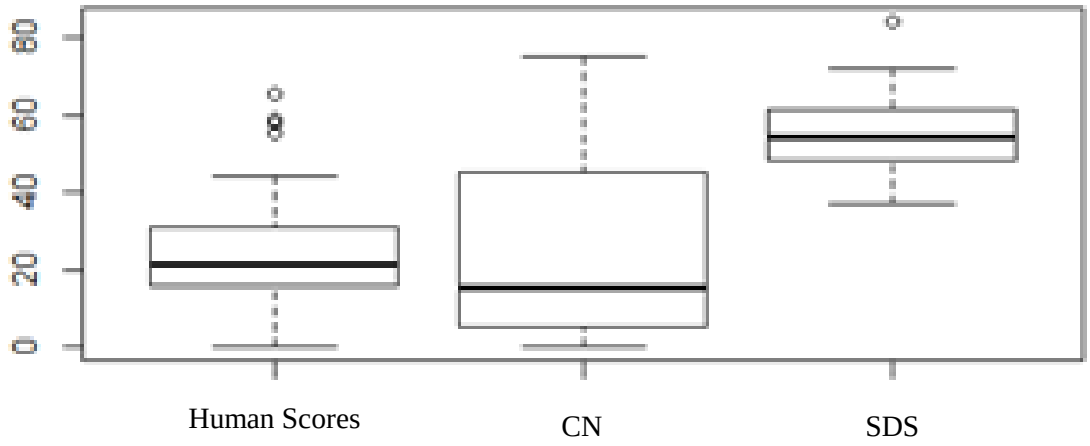

Figure 4-3: Boxplot diagram of the mean values of the similarity scores obtained by the three methods.

Chapter 5. Conclusion and Future Work

In this chapter, we summarize and conclude our work and discuss some gained insights within this thesis. In addition, we outline the key contributions and shed light on potential solution as our future work.

5.1 Conclusion

We have presented a possible solution for researchers to find related publications within a dynamic, visual and interactive system. Our system is connected directly to Google Scholar, which allows for dynamic access to multidisciplinary fields for the user, and to search about any publications through the system user interface. Then, the system parses the search results' web pages and visualizes the list of the resulting publications as a Citation Network graph representation, where the nodes represent the publications and the links represent the citation/reference relations between them. Our system shows the connected publications in multi-level citation relations in one screen, which allows the user to realize the clusters of connected papers on a topic, to detect which one is a leading paper in the area, to understand the hierarchical structures of the preceding and the following publications. Furthermore, we added interactive features to make our system interactive with the user such as visual cues and several actions on mouse events to enrich the knowledge about detailed information for each publication. This interaction function allows users to navigate the academic paper ocean in their own way, depending on their search purposes. This ranges from exploring a certain publication's information, detecting related publications, or finding the similarity score between two selected papers.

We propose a novel similarity score measurement between any two selected nodes in the Citation Network based on the number of paths connecting them and length of each path. More paths and shorter lengths correspond to a stronger relation between the publications. Through user study, we have proven our new method of comparing documents using Citation Network superior to the pre-existing NLP. The comparative study was conducted on the same dataset in order to evaluate our system in measuring the similarity between publications, compared with another method for measuring document similarity using the NLP cosine similarity formula. This comparison was based on how close the obtained results are from each method to the human judgments similarity scores applied to the same dataset. Our system is implemented using Java Script and python libraries combined with HTML which allows it to work on any platform and access the web for all researchers from anywhere, regardless of the institution to which they belong.

5.2 Contributions

According to the implementation and evaluation of our system, our contributions can be listed as follows:

Interactive System and Similarity Measurement based on Citation Network

- A dynamic web based scientific publication visualization system that is connected directly to a scholarly search engine in multidisciplinary fields. This feature allows researchers to retrieve the scientific publications directly and keeps them up-to-date with any new publication without the need to update the system itself. Other existing systems, which dataset are static, need to make changes in the code of the system every time when they need to update the dataset.

- An interactive Citation Network graph representation based on citation/reference relations between publications. In this, the nodes indicate the publications and the links indicate the citation/reference relations, which enhance the understanding of the hierarchy structure layout easily and allow the users to navigate through the resulting publications.
- Visual cues on the graph, such as node and link colors, sizes, labels and positions which have been used to indicate and distinguish the type of citation/reference relations.
- A new similarity measurement method, which uses a path based similarity method to calculate the similarity between a pair of selected publications. This is based on both the number and the length of paths between them. This has been evaluated through a comparative study to determine the effectiveness of our system compared to an existing NLP method for document similarity: (i) our proposed method using Citation Network (ii) the similarity measurements by the human user judgments which is assumed to be the threshold in the comparison (iii) Natural Language Processing method to measure the similarity between two documents.

5.3 Future Work

Even though this dissertation provides superior results, there remain challenging issues to address, and several research directions can be pursued to enhance the obtained results and the performance of the proposed system. The following points should be taken into consideration when we decide to further develop the work accomplished in this thesis.

- Additional analysis and statistical metrics could be applied to the Citation Network to provide more useful information and capabilities to the system such as filtering, sorting, drilling down, expanding and so on.
- Only three levels of citation relations of the resulting publications were used. This poses a limitation of being banned by Google Scholar from accessing more results and the visualization limitation of handling too many papers in one screen. More efficient visualization techniques to go further in depth regarding publication links is recommended.
- References have been added manually to the data set for the system evaluation and comparison for this time implementation. However, the system can successfully process the references and include them in the Citation Network visualization. Further work is recommended for extracting the references automatically with the other publication's bibliometric information.
- Reduce the complexity in our algorithm for both the publication's information retrieval process and the similarity measurement process as much as possible to increase the performance of the system.
- Add more features to the user interface and conduct a usability study of the system among researchers and public users.

References

[1] V. Scurtu, "Document Similarity," [Online]. Available: http://www.scurtu.it/documentSimilarity.html.

[2] S. K. Card and J. Mackinlay, "The structure of the information visualization design space," *IEEE Symposium on Information Visualization,* pp. 92-99, 1997.

[3] N. J. Van Eck and L. Waltman, "Visualizing bibliometric networks," in *Measuring scholarly impact: Methods and practice,* Springer, 2014, pp. 285-320.

[4] H. Small, "Co-citation in the Scientific literature: A New Measure of the Relationship Between Two Documents," *Journal of the American Society for information science,* pp. 265-269, 1973.

[5] C. N. Mooers, "Making information retrieval pay," Zator Co, Boston, 1951.

[6] S. Buettcher, C. L. A. Clarke and G. V. Cormack, Information Retrieval: Implementing and Evaluating Search Engines, The MIT Press, 2010.

[7] C. D. Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2008.

[8] W. B. CROF, D. METZLER and T. STROHMAN, Search Engines Information Retrieval in Practice, Reading: Addison-Wesley, 2010.

[9] JAWADEKAR and W. S. JAWADEKAR, Knowledge Management: Text & Cases, Tata McGraw-Hill Education, 2011.

[10] "Google Scholar," [Online]. Available: https://scholar.google.ca/intl/en/scholar/about.html.

[11] "Googlebots," [Online]. Available: http://www.google.com/bot.html.

[12] J. Beel and B. Gipp, "Google Scholar's Ranking Algorithm: An Introductory Overview," in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, Rio de Janeiro (Brazil), 2009.

[13] W. Arms, Digital Libraries, the M.I.T. Press in January 2000, 2000.

[14] B. . M. Leiner, "The Scope of the Digital Library," *the DLib Working Group on Digital Library Metrics*, 16 January 1998.

[15] "Digital Library Systems and Services (DLSS)," Stanford University Libraries, 2004. [Online]. Available: https://library.stanford.edu/department/digital-library-systems-and-services-dlss.

[16] "American Memory from the Library of Congress," The Library of Congress, [Online]. Available: http://memory.loc.gov/ammem/index.html.

[17] "Different Types of Digital Libraries," 123HelpMe.com, [Online]. Available: http://www.123helpme.com/different-types-of-digital-libraries-view.asp?id=153209. [Accessed 29 June 2015].

[18] "Health Sciences Digital Library," Michigan State University, [Online]. Available: https://www.lib.msu.edu/health/index/.

[19] "Windows Search," Microsoft, [Online]. Available: http://windows.microsoft.com/en-CA/windows7/products/features/windows-search.

[20] "OS X Yosemite: Search with Spotlight," Apple Inc., [Online]. Available: https://support.apple.com/kb/PH18828?locale=en_GB.

[21] "X1 Search- Premium Windows Desktop Search Software & Tool," X1 Discovery, Inc., [Online]. Available: http://www.x1.com/products/x1_search/.

[22] "Desktop Search Lookeen," Axonic, [Online]. Available: http://lookeen.com/.

[23] P. A. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl and R. Paiu, "Activity Based Metadata for Semantic Desktop Search," in *The Semantic Web: Research and Applications*, vol. 3532, Springer Berlin Heidelberg, 2005, pp. 439-454.

[24] S. K. Card, J. D. Mackinlay and B. Shneiderman, Readings in information visualization: using vision to think, Morgan Kaufmann, 1999.

[25] H. C. Purchase, N. Andrienko, T. Jankun-Kelly and M. Ward, "Theoretical Foundations of Information Visualization," in *Information Visualization Human-Centered Issues and Perspectives*, Springer Berlin Heidelberg, 2008, pp. 46-64.

[26] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon and Jörn, "Visual Analytics: Definition, Process, and Challenges," in *Information Visualization: Lecture Notes in Computer Science*, vol. 4950, Springer Berlin Heidelberg, 2008, pp. 154-175.

[27] Z. Gemignani , "Better Know a Visualization: Parallel Coordinates," 27 04 2010. [Online]. Available: http://www.juiceanalytics.com/writing/writing/parallel-coordinates.

[28] S. Few, "Multivariate Analysis Using Parallel Coordinates," 12 September 2006.

[29] B. Johnson and B. Shneiderman, "Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures," in *Visualization, 1991. Visualization '91, Proceedings., IEEE Conference on*, San Diego, CA , 1991.

[30] M. Khan and S. S. Khan, "Data and Information Visualization Methods, and Interactive Mechanisms: A Survey," *International Journal of Computer Applications*, vol. 34, November 2011.

[31] B. Shneiderman, "Discovering Business Intelligence Using Treemap Visualizations," Business Intelligence Network, 2006.

[32] "the Technische Universiteit Eindhoven-SequoiaView 1.3," November 2002. [Online]. Available: http://w3.win.tue.nl/nl/onderzoek/onderzoek_informatica/visualization/sequoiaview//.

[33] "What is timeline (Internet timeline, history of the Internet)," 2005. [Online]. Available: http://whatis.techtarget.com/definition/timeline-Internet-timeline-history-of-the-Internet.

[34] J. Wittwer, "How to Create a Timeline in Excel," Vertex42.com, 2 September 2005. [Online]. Available: http://www.vertex42.com/ExcelArticles/create-a-timeline.html.

[35] Q. Li and Y.-L. Chen , "Data Flow Diagram," in *Modeling and Analysis of Enterprise and Information Systems*, Springer Berlin Heidelberg, 2009, pp. pp 85-97 .

[36] V. Sauter, "System Analysis Current Page- Data Flow Diagrams Examples," University of Missouri - St. Louis, [Online]. Available: http://www.umsl.edu/~sauterv/analysis/dfd/dfd.htm.

[37] A. Deliyanni and R. A. Kowalski , "Logic and semantic networks," *Communications of the ACM*, vol. 22, no. 3, pp. 184-192, March 1979.

[38] C. Fellbaum , "A Semantic Network of English: The Mother of All WordNets," *Computers and the Humanities*, vol. 32, no. 2-3, pp. 209-220, 1998.

[39] "Semantic Networks and Frames," The University of New South Wales, [Online]. Available: http://www.cse.unsw.edu.au/~billw/cs9414/notes/kr/frames/frames.html.

[40] N. J. v. Eck and L. Waltman, "Visualizing bibliometric networks," in *easuring scholarly impact: Methods and practice*, Springer, 2014, p. 285–320.

[41] E. Garfield, I. H. Sher and R. J. Torpie, THE USE OF CITATION DATA IN WRITING THE HISTORY OF SCIENCE, Philadelphia, USA: Institute for Scientific Information Inc., 1964.

[42] E. Garfield, "Citation Indexing: Its Theory and Application in Science, Technology, and Humanities," *NewYork: John Wiley,* 1979.

[43] W. Shaw , "Entropy, information and communication," in *Proceedings of the American Society for Information Science 16*, 1979.

[44] B. C. Griffith, H. G. Small, J. A. Stonehill and S. Dey, "The Structure of Scientific Literatures II: Toward a Macro- and Microstructure for Science," in *Science Studies*, vol. 4, Sage Publications, Ltd, 1974.

[45] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, no. 61(12):2389–2404, 2010.

[46] M. Callon, J.-P. Courtial, W. Turner and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," in *Social Science Information*, 1983.

[47] E. Garfield, "Historiographic Mapping of Knowledge Domains Literature," *Journal of Information Science,* vol. 30, no. 2, pp. 119-145, 2004.

[48] H. D. White and K. W. McCain, "Visualizing a Discipline: An Author Co-Citation Analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, p. 327–355, 1998.

[49] C. CHEN, "VISUALISING SEMANTIC SPACES AND AUTHOR CO-CITATION," in

Information Processing and Management 35, 1999.

[50] . L. Leydesdorff and I. Rafols, "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology,* p. 348–362, 2009.

[51] E. Garfield and A. I. Pudovkin , "The HistCite System for Mapping and Bibliometric Analysis of the Output of Searches Using the ISI Web of Knowledge," in *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology*, 2004.

[52] J.-K. Chou and C.-K. Yang, "PaperVis: Literature Review Made Easy," *Computer Graphics Forum,* vol. 30, no. 3, pp. 721-730, 2011.

[53] "InfoVis 2004 Contest," [Online]. Available: http://www.cs.umd.edu/hcil/iv04contest/info.html.

[54] C. Dunne, B. Shneiderman, R. Gove, J. Klavans and B. Dorr, "Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization," *Journal of the American Society for Information Science and Technology,* vol. 63, no. 12, p. 2351–2369, 2012.

[55] N. J. v. Eck and L. Waltman , "Systematic retrieval of scientific literature based on citation relations: Introducing the CitNetExplorer tool," *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval*, pp. 13-20, 2014.

[56] E. Grefenstette and S. Pulman, Analysing Document Similarity Measures, University of Oxford, 2009.

[57] D. D. Lewis and K. S. Jones, "Natural language processing for information retrieval," in

Communications of the ACM, 1996.

[58] P. Merlo, J. Henderson, G. Schneider and E. Wehrli, "Learning Document Similarity Using Natural Language Processing," Linguistik Online, Geneva, 2003.

[59] A. Huang, "Similarity Measures for Text Document Clustering," in *the New Zealand Computer Science Research Student Conference 2008*, Christchurch, 2008.

[60] S. Borgatti, "Distance and Correlation," Boston College, Boston.

[61] A. Strehl, J. Ghosh and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000.

[62] B. Gipp, N. Meuschke and M. Lipinski, "CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central," in *iConference 2015 Proceedings*, 2015.

[63] K. Börner, C. Chen and K. W. Boyack, "Visualizing Knowledge Domains," in *Annual Review of Information Science & Technology*, 2003.

[64] Z. Wu and M. Palmer , "VERB SEMANTICS AND LEXICAL SELECTION," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994.

[65] M. Shenoy.K, D. Shet and D. U. Acharya, "A NEW SIMILARITY MEASURE FOR TAXONOMY BASED ON EDGE COUNTING," *International Journal of Web & Semantic Technology (IJWesT)*, vol. 3, 2012.

[66] R. RADA, H. MILI, E. BICKNELL and M. BLETTNER , "Development and application of a metric on semantic nets," in *IEEE Transactions on Systems, Man and Cybernetics*, 1989.

[67] C. Kreibich, "A Parser for Google Scholar written in Python," November 2013. [Online].

Available: http://www.icir.org/christian/scholar.html.

[68] "Introducing JSON," JSON, 2001. [Online]. Available: http://www.json.org/.

[69] M. Bostock, V. Ogievetsky and J. Heer, "D³ Data-Driven Documents," *Visualization and Computer Graphics, IEEE Transactions* , vol. 17, no. 12, pp. 2301 - 2309, 2011.

[70] M. J. Bannister, D. Eppstein, M. T. Goodrich and L. Trott, "Force-Directed Graph Drawing Using Social Gravity and Scaling," in *Graph Drawing*, Springer Berlin Heidelberg, 2013, pp. 414-425.

[71] J. Hua, M. L. Huang and Q. V. Nguyen, "Drawing Large Weighted Graphs Using Clustered Force-Directed Algorithm," in *18th International Conference on Information Visualisation (IV)*, Paris, 2014.

[72] D. Holten and J. J. van Wijk, "Force-Directed Edge Bundling for Graph Visualization," *The Eurographics Association and Blackwell Publishing Ltd*, vol. 28, 2009.

[73] L. C. SMITH , "Citation Analysis," *Library trends 30*, pp. 83-106, 1981.

[74] M. Bastian, S. Heymann and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," *ICWSM*, pp. 361-362, 2009.

[75] C. Kreibich, "A parser for Google Scholar, written in Python," November 2013. [Online]. Available: http://www.icir.org/christian/scholar.html.

[76] T. Dwyer, K. Marriott and M. Wybrow, "Integrating edge routing into force-directed layout," in *Graph Drawing*, vol. 4372, Springer Berlin Heidelberg, 2007, pp. 8-19.

[77] R. Chernobelskiy, K. I. Cunningham, M. T. Goodrich, S. G. Kobourov and L. Trott, "Force-directed Lombardi-style graph drawing," in *Graph Drawing*, Springer Berlin Heidelberg, 2012, pp. 320-331.

[78] Z. Shen, M. Ogawa, S. T. Teoh and K.-L. Ma, "BiblioViz: a system for visualizing bibliography information," in *APVis '06 Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation*, Darlinghurst, 2006.

[79] A. Bellamy, "Force Layout," GitHub, 2015. [Online]. Available: https://github.com/mbostock/d3/wiki/Force-Layout.

[80] S. G. Kobourov, "Spring embedders and force directed graph drawing algorithms," *arXiv preprint arXiv:1201.3011*, 2012.

Publications by the Author Related to the Thesis

1. "Literature Visualization and Similarity Measurement based on Citation Relations", published in the 19th International Conference Information Visualisation, 21- 24 July 2015, The University of Barcelona, Barcelona, Spain.
2. “A survey of Bibliometric/Citation Network Visualization Systems”, (under preparation).