(TBD) THESIS TITLE:INTERFACING INFORMATION IN USER STUDIES WITH MIXED METHODS

A Proposal

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Kyeongan Kwon

December 2015

(TBD) THESIS TITLE:INTERFACING INFORMATION IN USER STUDIES WITH MIXED METHODS

Kyeongan Kwon

APPROVED:

Ioannis Pavlidis, Chairman
Department of Computer Science

Zhigang Deng
Department of Computer Science

Guoning Chen
Department of Computer Science

Ricardo Vilalta
Department of Computer Science

Brian Uzzi
Northwestern University

Dean, College of Natural Sciences and Mathematics

(TBD) THESIS TITLE:INTERFACING INFORMATION IN USER STUDIES WITH MIXED METHODS

An Abstract of a Proposal

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Kyeongan Kwon

December 2015

Abstract

Evaluation of scholarly achievements in academia is largely based on the researcher's publication record. This record is communicated in exhaustive detail in the researcher's curriculum vitae (CV) or in summary via her/his h-index. The h-index, although a convenient abstraction, considers neither the time of the publication nor the impact factor (IF) of the journal where it appeared. In this article we present a novel method that visually complements the h-index, revealing at a glance the nature of a researcher's scholastic record. This method (aka Scholar Plot) is particularly appropriate for web interfaces, as it produces information that is compact and simple, yet highly illuminating. The method uses Google Scholar, Impact Factor and NSF/NIH Funding data to create a temporal representation of a researcher's publication/funding record that blends publication prestige with paper popularity and funding information. Scholar Plot aids to obtain an insightful appraisal of academics at one's fingertips.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A curriculum vitae (CV) provides a synopsis of an individual's achievements. The CV content varies by profession. Academic CVs feature prominently a publication section. This section references the researcher's journal papers and other scholarly products.

Search, promotion, and award committees that screen CVs go through lists of publications trying to form opinions about the candidates' records. Does candidate A or B have enough publications? Are they of high quality? Did they have any impact on the research community? In a highly competitive context, these questions do not always have clear answers. Another question that needs to be addressed is whether the candidate has been funded. If so, has the candidate done justice to the amount of funds obtained? This also enables one to decide if the candidate's output is in proportion with the input.

There has been some work on the quantification of academic careers, focused on a quest for a 'number' that sums up an academic's scholarship. The most well-known outcome of this line of research is the h-index, proposed by Hirsch [**?**]. A scholar has an index of h if s/he has published h papers each of which has been cited in other papers at least h times.

The h-index depends on both the number of publications and the number of citations. Hirsch demonstrated that h can predict honors, such as National Academy membership and Nobel prize. He also suggested that it could predict advancement to tenure, although with some uncertainty. Despite its value, the h-index has weaknesses and when used, context should be carefully taken into account; such context includes the academic field and the academic age of the candidate [**?**].

With the advent of Google Scholar, information about a researcher's publication record and her/his h-index has become easily accessible. Then, with the ease of access of the internet, this information has become ubiquitous.

In this article we introduce a data visualization tool that complements the publication information contained in a standard CV and summarized by the h-index. The tool produces a temporal visualization that connects the h-index with the paper citations and the journal impact factors along with the funding data.

There have been other efforts in visualizing patterns of scientific production and impact [**?**, **?**, **?**]. Recently, a mobile app (DBIScholar) has also appeared that interfaces information from Google Scholar [**?**]. A social tool named Scholarometer has been developed to facilitate citation analysis and to evaluate the impact of authors

[?]. This tool helps to visualize author and discipline networks. There is another tool called SciVal to visualize the collaboration and research output of institutions [?]. This tool uses data from Scopus. But these tools do not provide a visual picture of a single scholar's achievements.

Our method and application differ from the prior art. Scholar Plot helps the reviewer determine at a glance from where the researcher's impact (if any) arises from.

Chapter 2

Background

Chapter 3

Methods

3.1 Visualization and User Interface

Scholar Plot obtains the Impact Factor (IF) for a particular journal from our database. The data of Impact Factor is acquired from The Thomson Reuters Impact Factor - Web of Science. Based on all this information it constructs the plots as per the design outlined in the Visualization and User Interface section, using nvd3 library [**?**].

The NSF/NIH funding datasets are available at the respective US government websites in various file formats such as XML, CSV and so on [**?**, **?**]. We implemented a script to parse this massive XML dataset into our data structure that consists of AwardID, AwardAmount, First name, Last name, Investigator by RoleCode (Principal Investigator, Co-Principal Investigator and Former Principal Investigator), using

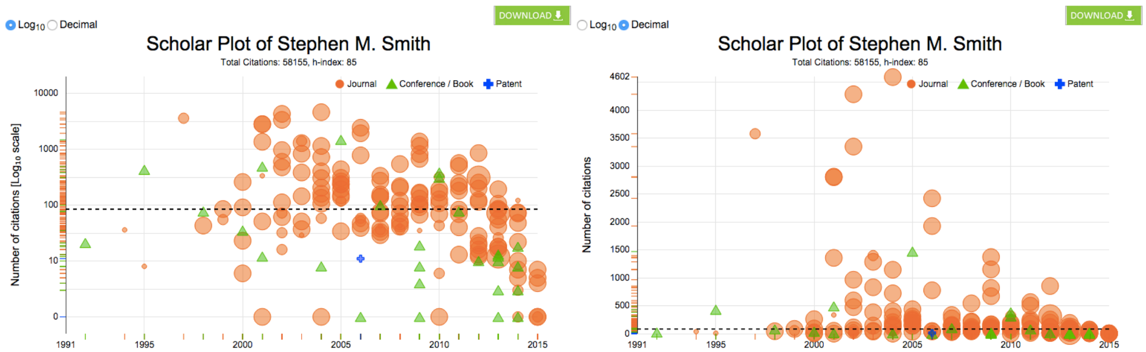


Figure 3.1: The log_{10} view and *decimal* view: The radio button allows to switch between different scale views without reloading the entire page.

XMLStarlet [**?**]. We imported this data to our database using Toad DBMS tool.

Scholar Plot depicts the publications of an individual as a scatter plot and the NSF/NIH funding as a multiline plot. The publications are represented in a 2D diagram (number of citations vs. year of publication) with the h-index line (Figure 3.2). The horizontal axis is time, starting with the year of the researcher's first publication ending with the current year. The vertical axis is the number of citations. The default plot is in log_{10} scale. The user can also view the plot in the decimal scale by a toggle option using a radio button at the top left corner (Figure 3.1). The log scale provides a standardized scale which helps to compare the plots of multiple scholars.

3.1.1 Visualizing Publication Data

Each publication i is represented with a symbol. The center of the symbol has coordinates (i_{PY}, i_C), where PY stands for Publication Year and C for Number of

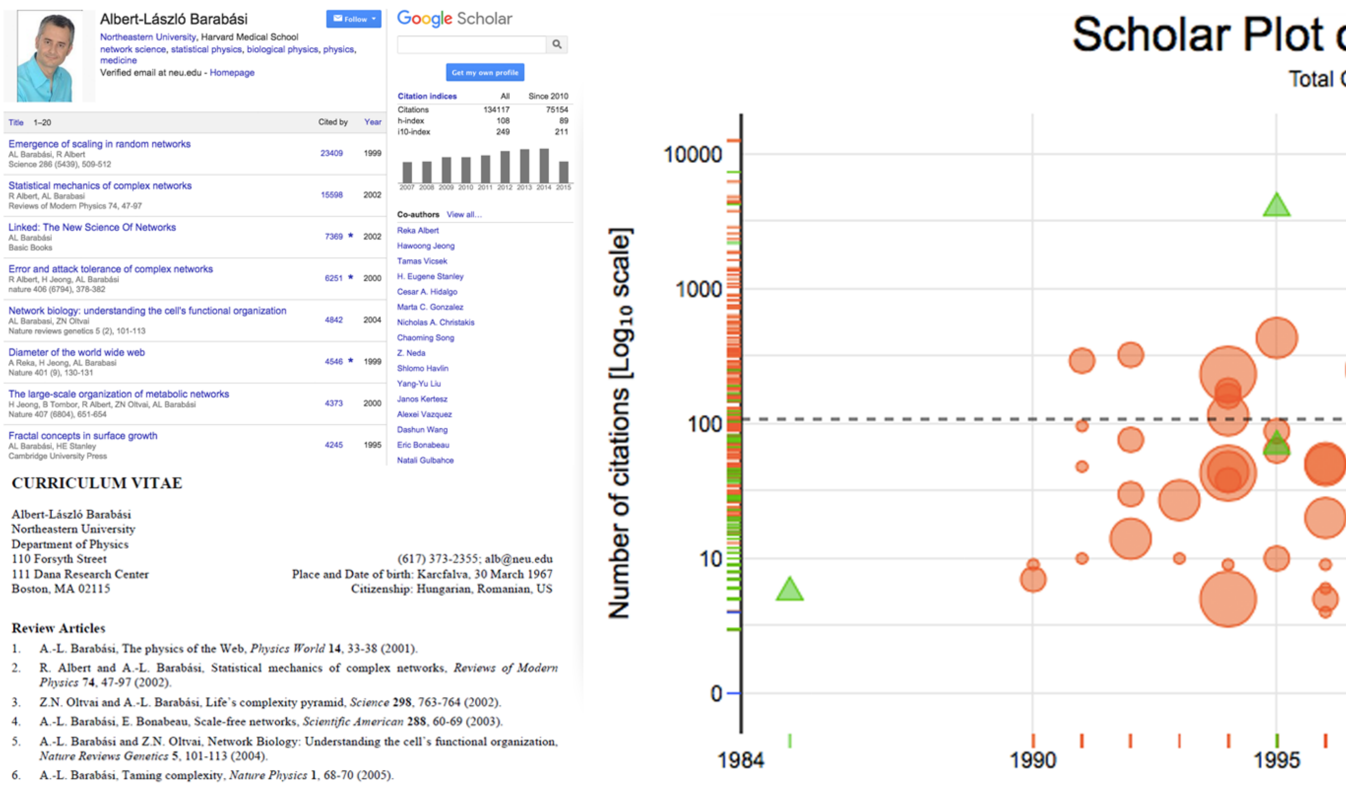


Figure 3.2: An example of Scholar Plot - Visualizing Publication Data

citations obtained by the publication till date. The journals are represented as circles (orange) with area analogous to the impact factor the journal. The conferences / books are represented as triangles (green) and the patents as crosses (blue). By clicking at a symbol you can obtain the publication title, the year, the number of citations, the venue where published and its impact factor (if it is a journal), as well as a breakdown in the authorship, complete with the level of collaboration between the co-authors and the selected scholar (Figure 3.3). The publication title also enables the user to navigate to the Google Scholar page for the selected paper. This helps to quickly verify and obtain further details of the selected publication. It makes user reach out to the PDF file directly if available. To enhance user experience, we customized the tooltip to give detailed information without overlapping the plots.

A dotted horizontal line on the plot denote the h-index of the scholar. We also denote those publications which earn greater than 10,000 citations with diamonds as they represent the great success in publications (Figure 3.2). The title of the plot contains the name of the scholar and her/his total number of citations along with the h-index. At the top right corner of the plot, a legend shows the three different types of publications we distinctly display (Figure 3.4).

You can bring the journals, patents, and conferences / books in and out of the view by clicking at the respective legend. If there is an overlap between journals, conferences and patents, this feature can help the user to selectively view them. The user can also zoom into the plot for closer picture. Also note that the symbols are not completely opaque. So if there are multiple symbols which overlap, the user can see and interact with them by hovering the mouse over them appropriately.

3.1.2 Visualizing Funding Data

Scholar Plot also depicts the NSF/NIH funding of an individual as a multiline (Figure 3.5). Each breakpoint in the multiline corresponds to the individual's total amount in all NSF/NIH awards for the specific year. By pointing at a breakpoint you can obtain the NSF/NIH awards IDs, award amounts, and investigator's role. The total annual funding information per year is also available by clicking the legend.

To place the plots in your personal CV or on your web page we provide a download button at the top right corner of the plot (Figure 3.1). This function enables the user to download plots in a zip file. It includes high resolution vector images in SVG (Scalable Vector Graphics) format of the publication and funding plots.

Scholar Plot also has a projection of the data on the y-axis depicted by small horizontal colored lines. For example, we can clearly see that journals contribute to the *h*-index of scholar in Figure 3.6 (a) and conferences / books contribute to the *h*-index of scholar in Figure 3.6 (b). We can clearly infer the scholar in Figure 3.6 (c)) has many patents. We can also infer the number of publications within a particular range of citations based on the density of the projected lines.

We improve user experience to enable users to quickly find and select from a pre-populated list of scholar names as they type. For each character the user enters, we display similar matching names on the dropdown list. Even entering the space (" "), we display the 10 most recently inserted scholar's names. Scholar Plot follows the approach of responsive web design to provide optimal viewing based on the size of screen.

3.2 System Architecture

Scholar Plot is data visualization tool that uses HTML5, CSS3 and SVG to render a scholar's accomplishment at a glance. We created a MySQL database to store the mapping between the scholar names and their Google scholar IDs. We also designed and created database tables for NSF/NIH funding data. The user can search the name of the scholar in a text field. When the user starts to enter the name of the scholar, the names in our database which are similar to the entered name will be listed as a drop down list. We use jQuery and Ajax (asynchronous JavaScript and XML) method to have this feature, which connects to the database to get the list of names. If there are no matching/similar names, the user can also insert her/his Google Scholar ID to the database by one click event.

The server-side application is implemented in PHP scripting language and MySQL. The HTTP protocol is used for communicating between client-side and server-side to get the basic information via JSON format (JavaScript Object Notation) and JSONP function (Figure 3.7). Scholar Plot also uses htmlSQL library to parse Google scholar's page to extract user basic information [?].

3.3 Name Disambiguation

3.3.1 Within and across profile author name disambiguation

Let i be an index for the Google scholar profile researchers. Within each collaboration profile of i, there are a set of K_0 raw name strings that you have extracted, $Names_k$ indexed by k_i. We will use the fact that these strings are associated with profile i in the process of name disambiguation across Google Scholar profiles. The following provides an outline of this procedure:

A) **Clean last names:** Remove strings at end of all $Names_k$ that are not last names, and which may not consistently be listed for k, e.g. "Jr.", "III" etc. Hence, each name string $Name_k$ consists ideally of a First name string FN_k, a Last name string LN_k, and possibly a Middle name string MN_k.

B) **Clean middle initial strings within each profile i:** Within each i, search for inconsistencies in the use of MN_k. That is, possibly sometimes the author k is listed as *Alexander M Petersen*, sometimes *Alexander Petersen*, and sometimes *Alexander Michael Petersen*. In this example the Last name string $LN_k = Petersen$ and the First name string $FN_k = Alexander$ are clearly consistent. But the Middle name string {␣ , M, Michael} causes some ambiguity if simple string comparison is used, where ␣ is a whitespace.

Then check to see how many different types of *Alexander* $\hat{X}$ *Petersen* occur within each k, where $\hat{X}$ is refers to the middle name. Use the following rules for when there are 2 or more types of $\hat{W}\hat{X}Petersen$.

- If there are only two types of $Alexander\hat{X}Petersen$, with $\hat{X} = \_$ or M, then map all of the $Alexander\hat{X}Petersen$ to $AlexanderMPetersen$ for this i
- If there are only three types of $Alexander\hat{X}Petersen$, with $\hat{X} =$ starting with the same initial, $M\_$ or M, then map all of the $A\hat{X}Petersen$ to $AlexanderMichaelPetersen$ for this i
- If there are two or more types of $Alexander\hat{X}Petersen$, say $\hat{X} = O$ and $\hat{X} = P$, then keep these X as they are.

C) **Disambiguate coauthors k across the Google Scholar profiles (connecting i):** Let k and k' be coauthors in profiles i and i', respectively. In this step we would like to identify k and k' that are likely the same person, $k = k'$, allowing us to connect the two profiles i and i' within the coauthor network.

If k and k' have the same initials and same surname, then there is a possibility that they are the same individual. Also, if their full first name strings match, this is clearly very positive evidence of this. Let $A_{k,j}$ be the entire combination of First Name and Middle initial $FM_{k,j}$ with the surname $L_{k,j}$ (e.g. *Adam B Smith*, or *Adam _ Johnson*) of the coauthor j of the coauthor k.

- If the full first name strings and the full last name strings are the same, $FN_{k,j}$=$FN_{k',j}$ and $LN_{k,j} = LN_{k',j}$ (e.g. Adam J. Johnson and Adam Johnson), and they both have at least one coauthors in common, then they are considered the same coauthor.

- If we don't have the added information of their full first names then we must rely more heavily on the information from their coauthors. If the first and last names are the same, $FM_{k,j} = FM_{k',j}$ and $LM_{k,j} = LM_{k',j}$, and there are more than 2 middle names with one of the middle name being empty, we do the following -

  We compute the number of coauthors in common of the empty middle name author with non-empty middle name authors by comparing the sets of coauthors, $\{j\}$.

  We assign the empty middle name to that middle name for which there are more number of co-authors in common.

- If the first name of the author has a hyphen, we check for any other author having the same last name and the first name as the first word of the hyphenated word and middle name starting with the first letter of the second part of the hyphenated word. If any such pair of authors have at least one author in common, we update the first and middle name of the author with the hyphenated middle name to first name and middle name of the matched author.

- If the first name of the author has only two letters, we check for any other author having the same last name and the first name starting with the first letter of the first name and middle name starting with the second letter of the first name. If any such pair of authors have at least one author in common, we update the first and middle names of the author with two letters to first and middle names of the matched author.

Google Scholar data has to be cleaned because it contains many non-english characters. We use regular expression to remove the invalid special characters and translate phonetic characters to english alphabets. We designed and implemented Algorithm **??** to match the author names in Google Scholar with those in NSF/NIH data. This process helps to improve the quality of results.

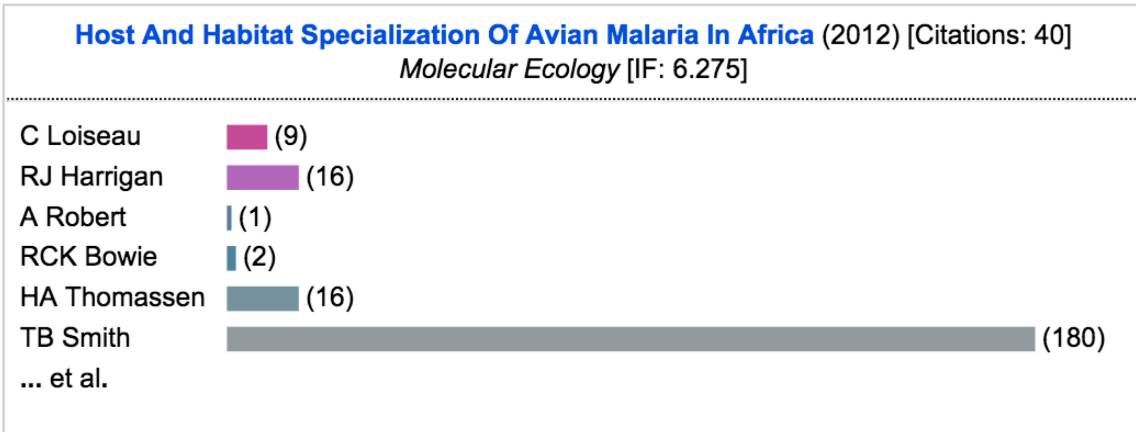


Figure 3.3: An example of the tooltip: The publication title, the year, the number of citations, the venue where published, impact factor, the list of co-authors, the visual horizontal bars with the number of collaboration between the co-authors and the selected scholar.

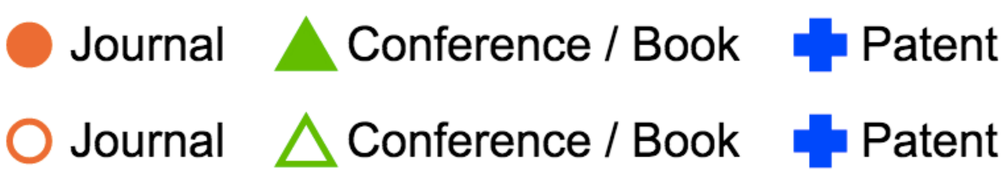


Figure 3.4: The legend allows users to selectively view journals, conferences / books and patents.

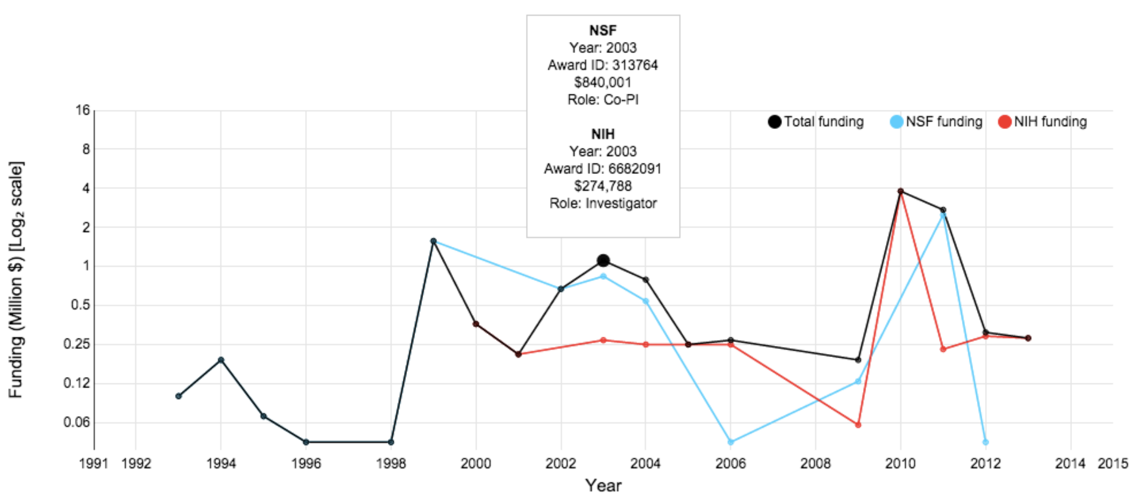
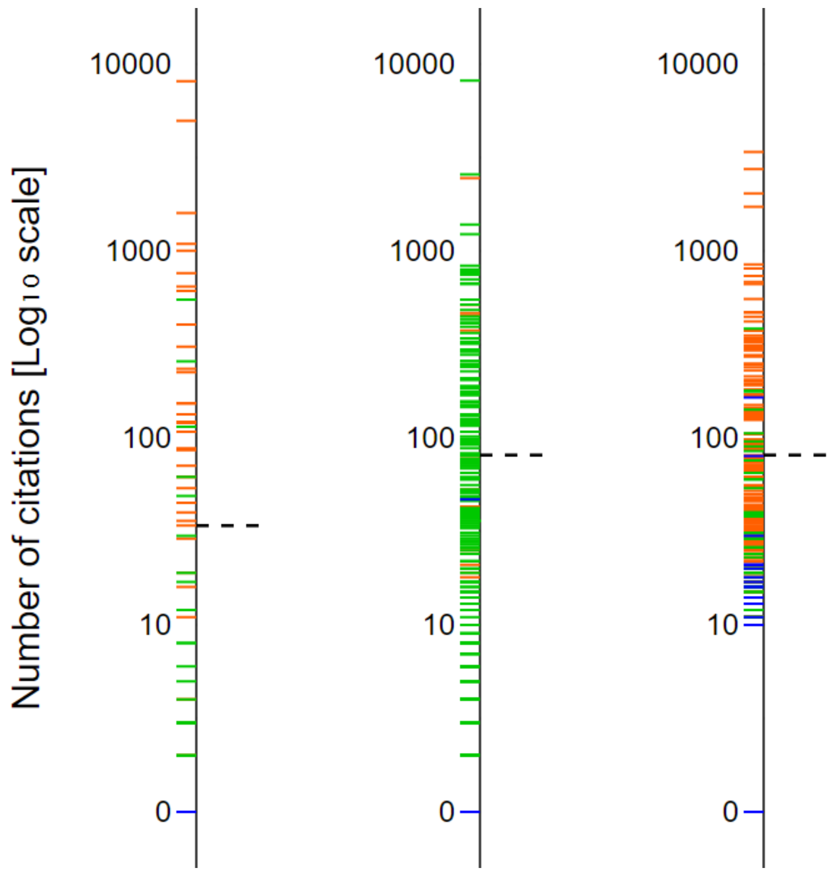


Figure 3.5: An example of Scholar Plot - Visualizing Funding Data



(a) (b) (c)

Figure 3.6: Examples of y-axis projection for three different scholars.

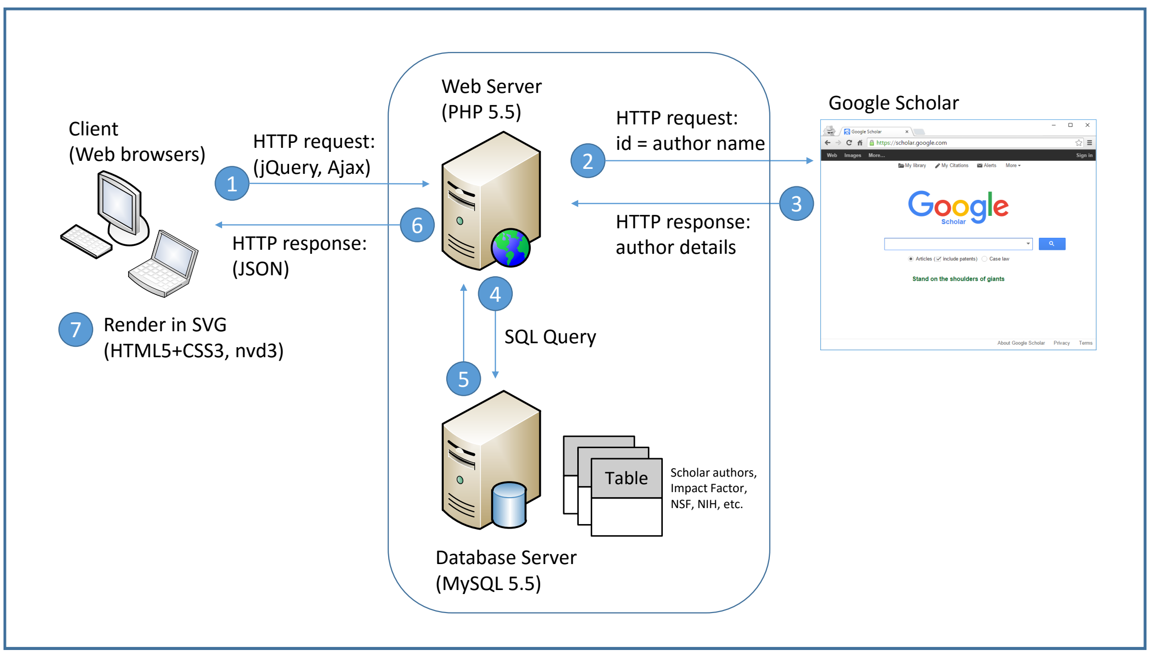


Figure 3.7: System Architecture of Scholar Plot.

Chapter 4

Algorithms

Chapter 5

Results

A total of 15 participants from various disciplines including Natural Sciences, Social Sciences, Life Sciences and Computer Science evaluated Scholar Plot. We asked each participant to review the interface and then complete an online survey. Special care was taken to ensure that the participants had correct understanding about the visualization component before they began rating. The participants answered the questions on a Likert scale from 1 to 5 with 1 being strongly disagree and 5 being strongly agree.

Figure 5.1 illustrates the mean evaluation for each visualization component. Accuracy, Usability and understandability of Scholar Plot scored the highest ($\mu = 4.2$) as it is very intuitive and can be used with minimal assistance. Many participants gave us feedback that they mostly liked the visual scheme of Scholar Plot. Another observation is that the participants agree to use Scholar Plot to evaluate themselves

($\mu = 4.1$). They suggested that Scholar Plot can be improved by adding more funding agencies. Overall, this evaluation indicated that Scholar Plot is a user-friendly tool that complements the CV which can be used to review a scholar's accomplishments. The survey has been approved by the University of Houston Institutional Review Board (IRB).

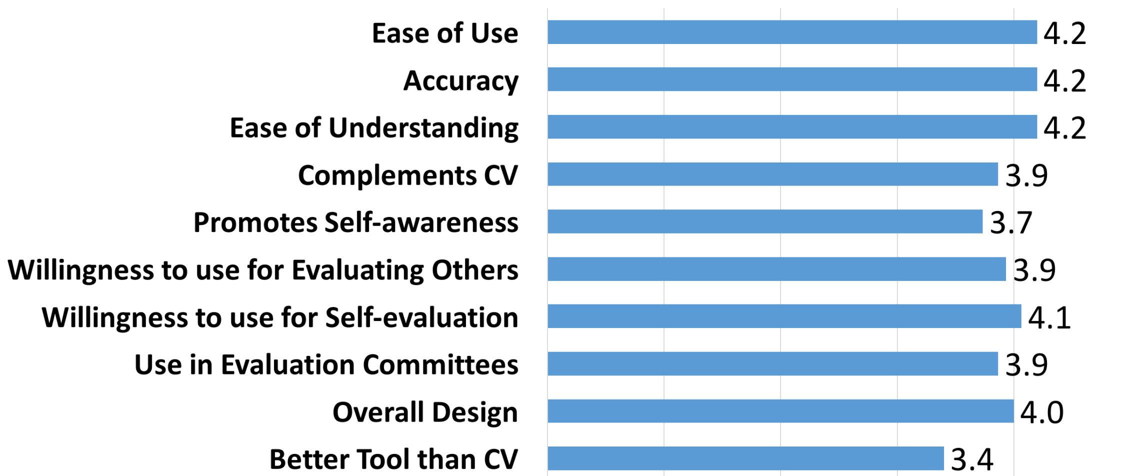


Figure 5.1: Mean evaluation of Scholar Plot. A total of $n = 15$ participants evaluated the survey.

Chapter 6

Conclusion

We have described a visualization method that complements the information contained in a researcher's Google Scholar page and summarized by her/his h-index. One can draw insightful conclusions about the individual's scholastic accomplishments. These conclusions are not supported by the h-index alone and cannot be derived by the CV or the Google Scholar page, unless a significant investigative effort is undertaken. Our user study also supports this.

This approach not only focusses on journal publications, conferences / books and patents but also NSF/NIH funding data. Scholar Plot is a simple, yet valuable visualization scheme. It is likely to have broad appeal not only because it would be useful to evaluation committees, but also because it is available online for free at `http://www.scholarplot.com`.