Temporal Analysis of Bibliographic Databases
-When are academics most productive?

Yiting Zhao
Supervisors: Dr Qing Wang

Dr Peter Christen

COMP4560: Advanced Computing Project
Australian National University Semester 2, 2014
31 October, 2014
Acknowledgement

I would like to express my greatest gratitude to my supervisors Dr Qing Wang and Dr Peter Christen for their great guidance, assistance, knowledge sharing and patience along the project process. I would also like to thank my family and friends for their help and encouragement.
Abstract

In this project, we present a method for the analysis of productivity of researchers using data from the Google Scholar digital online library. To implement the method, we developed techniques to crawl and parse Google Scholar profile web page data for individual researchers as well as groups of researchers working on certain research topic. Different analysis and plotting techniques were adopted to analyse the productivity of researchers based on their publications and citations counts. To evaluate the method, we firstly present a way for measuring the accuracy of our algorithm towards calculating the citation counts where the result indicated a higher accuracy when calculating the average monthly. Secondly, we performed the analysis on four groups of researchers working on the topics ‘catalysis’, ‘data mining’, ‘genomics’ and ‘pharmacokinetics’ and we compared the results across these groups.
## Contents

Chapter 1 Introduction........................................................................................................... 5

1. 1 Background .................................................................................................................... 5

1.1.1 Publications ............................................................................................................... 5

1.1.2 Citations ...................................................................................................................... 6

1.1.3 Research Topics ........................................................................................................ 6

1.2 Objectives ..................................................................................................................... 7

1.3 Contributions ............................................................................................................... 7

1.3.1 Web Data Crawling and Parsing .............................................................................. 8

1.3.2 Citations and Publications Analysis ........................................................................ 8

1.3.3 Evaluation ................................................................................................................ 8

Chapter 2 Methodology ........................................................................................................... 9

2.1 Web Data Crawling and Parsing .................................................................................. 9

2.1.1 Individual Researcher Citation Crawling and Parsing ....................................... 9

2.1.2 Researchers Citations Crawling and Parsing Based on Research Topics ............ 11

2.2 Citations Analysis ....................................................................................................... 13

2.2.1 Individual Researcher Citations Analysis ............................................................ 13

2.2.2 Group Citations Analysis ....................................................................................... 14

2.3 Publications Analysis ................................................................................................. 15

2.3.1 Individual Researcher Citations Analysis ............................................................ 15

2.3.2 Group Publications Analysis .................................................................................. 16

Chapter 3 Evaluations .......................................................................................................... 17

3.1 Individual Citation Counts Accuracy ........................................................................... 17

3.2 Evaluation of Analysis for Group of Researchers ...................................................... 18

3.2.1 Calendar Year Citation Plot Evaluation ................................................................ 18

3.2.2 Career Year Citation Plot Evaluation ..................................................................... 19

3.2.3 Calendar and Career Year Publication Plot Evaluation ....................................... 21

Chapter 4 Conclusion and Future Work .............................................................................. 23

4.1 Conclusion ................................................................................................................... 23

4.2 Future Work ............................................................................................................... 23

Appendix: Project Contract ................................................................................................. 25
Chapter 1 Introduction

In this chapter, we will give a general introduction and background information of our project. In Section 1.1, we will introduce a public bibliographic database - Google Scholar and 3 major features it provides regarding authors’ publications, publication citations, and research topic groups. Based on the information Google Scholar provides, we will outline the objectives of this project in Section 1.2. The contributions and achievements will then be discussed in Section 1.3.

1.1 Background

Google Scholar [1] is a bibliographic database which collects multi-disciplined research output across the world. It provides a simple way to explore scholarly literature from different sources including books, articles, journals and online publications. It also helps users to search for related works and publications regarding different research topics. In addition to the feature on searching for research output, there are some other features Google Scholar does provide. In the following sections, we will discuss the three major features that we are interested in this project.

1.1.1 Publications

The first feature Google Scholar provides is that it allows users to create an author profile containing basic information about the author such as their names, affiliations, area of interests (research topics) and email contacts. An author will be able to add and update their publications they have published. From this profile, we will know the research topics one author has been working on and browse through their publications from a single place. Figure 1 below shows an example of Dr Peter Christen’s Google Scholar profile.

![Figure 1 Dr Peter Christen’s Google Scholar Profile](image-url)

Figure 1 Dr Peter Christen’s Google Scholar Profile (Red rectangles: raw information we want to crawl and parse in Chapter 2. Green Circles: labels of research topic. Orange rectangles: citation metrics that will be used in Chapter 3 for the evaluation purpose.).
1.1.2 Citations
The second feature is that Google Scholar also allows authors to keep track of the citations to each of their publications. Authors will not only be able to know who has cited their work but also a total count of citations for each publications in their profile. In addition, they can view the metrics on how their citations grow every year as shown in Figure 1 (highlighted in orange rectangles showing the citation metrics for all publications) and Figure 2 (bar chart showing the citation metrics for one publication named ‘A comparison of fast blocking methods for record linkage’).

![Figure 2 Citation details for one of Dr Peter Christen's papers — ‘A comparison of fast blocking methods for record linkage’](image)

1.1.3 Research Topics
The last feature of interest in the Google Scholar is that, when users editing their own profiles, authors can add different research labels as shown in a green circle in Figure 1 to their profile, such as ‘data mining’ or ‘database’. These labels are used to indicate the research topics they are interested in or currently are working on. This feature enables us to be able to search for researchers who work on the same research

![Figure 3 list of publications that have cited the paper — ‘A comparison of fast blocking methods for record linkage’](image)
topic through Google Scholar. Figure 4 below shows an example of the result when searching for researchers with the label ‘data mining’.

![Google Scholar search result](image)

Figure 4 Web page example of a list of researchers labelled with 'Data mining' research topic. (Red rectangles: name of the researchers and the link to their profile we aim to crawl and parse in chapter 2)

1.2 Objectives
Google Scholar provides us with a simple way to search for individual researchers and research topics. This gives us the opportunity to perform various analyses on these real world data. Thus the main aim of this project is to investigate, develop, implement and evaluate techniques that analyse the research outputs of researchers from Google Scholar. The specific tasks are:

- To collect a data set with research outputs by individual researchers and their citation information from Google Scholar.
- To develop techniques that support the temporal analysis of research outputs of individual researchers in terms of their research impact and productiveness over time, as measured by the number of publications and attracted citations over time.
- To develop approaches to summarise the impacts of a group of researchers according to some group definition.
- To develop visualisation method that allows a presentation of the analysed and summarised results.

1.3 Contributions
The aim of this project is to use the information Google Scholar provides us on author publications and citations for the analysis of the productivity of researchers. The following sections will describe the works have been done in order to fulfil the tasks listed.
1.3.1 Web Data Crawling and Parsing
There were two programs developed for the purpose of crawling and parsing Google Scholar web page data. The first program ‘Individual_Parser’ was focusing on individual researchers, that is, to crawl and parse the information shown on the Google Scholar profile page for a single researcher.

The second program ‘Group_Parser’ was developed in order to search researchers who working on a certain research topic and collect information on those researchers as a group. More detailed information on these two programs will be introduced in Section 2.1.

1.3.2 Citations and Publications Analysis
In order to analyse the data collected from Google Scholar web pages, various techniques were developed for this purpose. The program ‘Individual_Cites’ was developed in order to calculate the yearly citation counts for an individual researcher which will be introduced in Section 2.2.1. For analysing the performance of groups of researchers, four techniques each designed regarding the researchers’ citation and publication counts respectively. These techniques will be explained with more details in Section 2.2.2.

1.3.3 Evaluation
The last part of work was focusing on evaluating the techniques developed for the data analysis. We have developed an algorithm ‘DifferencesCalc’ in order to measure the accuracy of the calculation done in program ‘Individual_Cites’. Four groups of researchers chosen from topics ‘Catalysis’, ‘Data Mining’, ‘Genomics’, and ‘Pharmacokinetics’ were used for the evaluation of the analyses techniques developed for groups of researchers.
Chapter 2 Methodology

In this chapter, we will introduce the techniques we used for the crawling and parsing of web data from Google Scholar online digital library as well as other techniques used for the analysis of these retrieved data. Figure 5 below provides a high-level overview for the methods used in this project.

2.1 Web Data Crawling and Parsing

In this section, we will introduce the techniques we developed for the crawling and parsing of web data from Google Scholar for an individual researcher and a group of researchers working on a certain research topic.

2.1.1 Individual Researcher Citation Crawling and Parsing

We firstly developed a program ‘Individual_Parser’ that crawls and parses individual researcher’s profile (the web page shown in Figure 1) and save it locally for further analysis. That is, given a web link to a researcher’s Google Scholar profile, the program will send a HTML request and parse the information from the returned webpage data (lines 1-2, Algorithm 1). There are two different parts of information that we focus on.

The first part is the table (shown in Figure 6) that contains the citation indices showing the citation summary of the researcher for the past few years. The parsed summary information will then be saved locally into a user defined file. This is done by looking for parameters and HTML tags that associate with the start of the table. Since the table for all researchers fall into the same format, it is quite straightforward to locate the end of the indices (lines 5-8, Algorithm 1). Table 1 below shows an example of a sample result of saved summary information.

<table>
<thead>
<tr>
<th>Citation indices</th>
<th>All</th>
<th>Since 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>2317</td>
<td>1831</td>
</tr>
<tr>
<td>h-index</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>i10-index</td>
<td>49</td>
<td>43</td>
</tr>
</tbody>
</table>

Figure 6 An example of citation summary table we aim to crawl and parse.
Table 1 parsed summary information for citation summary table in Figure 6. (RID: Google Scholar ID of a researcher. RName: name of a researcher.)

<table>
<thead>
<tr>
<th>RID</th>
<th>RName</th>
<th>All</th>
<th>All 2009</th>
<th>H All</th>
<th>H 2009</th>
<th>i10 All</th>
<th>i10 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>2317</td>
<td>1831</td>
<td>23</td>
<td>22</td>
<td>49</td>
<td>43</td>
</tr>
</tbody>
</table>

To facilitate a more comprehensive analysis for future studies, we have to get more detailed information of each publication. Thus the second part we focused on is the citation information for all articles that a researcher had published which involves parsing the title, citation count and the publish year. The parsed citation information, similar to the summary table, will be saved into a different user defined CSV file. Since one researcher could have more than one page of publication lists, the program has to keep sending HTML request until a blank list returned (lines 10-15, Algorithm 1). For more efficient parsing, the program has set the parameter for publications per page to 100 which is the maximum number Google Scholar could support. Table 2 below shows a sample output for this part.

Table 2 Example output for publications information of Dr Peter Christen. (RID: Google Scholar ID of a researcher. RName: name of a researcher.)

<table>
<thead>
<tr>
<th>RID</th>
<th>RName</th>
<th>Title</th>
<th>Count</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>a comparison of fast blocking methods for record linkage</td>
<td>253</td>
<td>2003</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>a comparison of personal name matching: techniques and practical issues</td>
<td>186</td>
<td>2006</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>a survey of indexing techniques for scalable record linkage and deduplication</td>
<td>117</td>
<td>2012</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>quality and complexity measures for data linkage and deduplication</td>
<td>113</td>
<td>2007</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>peter christen</td>
<td>febrl- a parallel open source data linkage system</td>
<td>112</td>
<td>2004</td>
</tr>
</tbody>
</table>

Algorithm 1 below describes the overall procedure for the program 'Individual_Parser' discussed in this section above.

**Input:** url to an individual researcher’s Google Scholar profile, name of output citation file, name of output summary file

**Output:** updated citation and summary file

**Individual_Parser** (ResearcherURL, CitationFile,SummaryFile):

1. Send HTML request with ResearcherURL

2. htmldata ← html returned data

3. tableParsed ← False

4. for data in htmldata:

5. if tableParsed is False:

6. Look for start tag of summary table
As mentioned earlier in Section 1.1.3, Google Scholar allows users to search for researchers that work in the same research topic i.e. having the same research label. To be able to perform analysis on the productivity of a group of researchers, a new program ‘ResearchTopic_Parser’ was developed to parse and collect the citation information for a group of researchers working on certain research topic from the web data. The program will firstly send a HTML request (lines 2-3, Algorithm 2) with keywords as input setting to the designated research topic such as ‘data mining’. Upon receiving the request, the Google Scholar web server will return a HTML page showing a list of researchers that are working on the same topic (shown in Figure 4). The next step of the program was to parse each listed researcher with their names, Google Scholar author IDs and the web page addresses that link to their profile. The program will keep sending requests for more researchers (on the next web page) until an empty list or the designated researcher count for the group of researchers is reached (lines 8-16, Algorithm 2). With the utilization of the program ‘Individual_Parser’ introduced in Section 3.1.1 the program will then parse and collect the information for all researchers on that topic. There are three output files for this program:

- A summary file, similar to the individual summary file, containing the citation summaries of each researcher in the group.
- A citation file, similar to the individual citation file as well, but containing the publication information for all researchers in that group.
• A query result file, recording the returned list of researchers in the research group (Table 3 shows a sample output of this result file).

Table 3 Example of query result CSV file for 'Data mining'

<table>
<thead>
<tr>
<th>RID</th>
<th>RName</th>
<th>url</th>
</tr>
</thead>
<tbody>
<tr>
<td>wUJ2bXgAAAAAJ</td>
<td>Jeffrey Ullman</td>
<td><a href="http://scholar.google.com/citations?user=wUJ2bXgAAAAAJ&amp;hl=en">http://scholar.google.com/citations?user=wUJ2bXgAAAAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>tQVe-fAAAAAJ</td>
<td>Trevor Hastie</td>
<td><a href="http://scholar.google.com/citations?user=tQVe-fAAAAAJ&amp;hl=en">http://scholar.google.com/citations?user=tQVe-fAAAAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>Kv9AbjMAAAAJ</td>
<td>Jiawei Han</td>
<td><a href="http://scholar.google.com/citations?user=Kv9AbjMAAAAJ&amp;hl=en">http://scholar.google.com/citations?user=Kv9AbjMAAAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>XPdhXUUAIAAJ</td>
<td>Rakesh Agrawal</td>
<td><a href="http://scholar.google.com/citations?user=XPdhXUUAIAAJ&amp;hl=en">http://scholar.google.com/citations?user=XPdhXUUAIAAJ&amp;hl=en</a></td>
</tr>
<tr>
<td>CCOAVjAAAAAJ</td>
<td>Jeongkyu Lee</td>
<td><a href="http://scholar.google.com/citations?user=CCOAVjAAAAAJ&amp;hl=en">http://scholar.google.com/citations?user=CCOAVjAAAAAJ&amp;hl=en</a></td>
</tr>
</tbody>
</table>

Algorithm 2 below illustrates the overall procedure for the program 'ResearchTopic_Parser' described in this section above.

**Input:** keywords of the research topic, name of output citation file, name of output summary file, number of researchers aim to retrieve

**Output:** updated query results, citation and summary file

**ResearchTopic_Parser**(Keywords, CitationFile, SummaryFile, ResearcherNumber):

1. `ResultFile ← new spreadsheet file named Keywords`
2. `Send html request with label set to Keywords`
3. `htmlData ← html returned data`
4. `count ← 0`
5. `if ResearcherNumber not specified:
   6. `ResearcherNumber ← 100`
7. `nextPageNotEmpty ← True`
8. `while count < ResearcherNumber and nextPageNotEmpty:
   9. `Send html request for next page`
   10. `for data in htmlData:
        11. `Look for start tag of an researcher`
        12. `rname, rid, url ← Parse the researcher information`
        13. `Write rname, rid, url to ResultFile`
        14. `Algorithm1(url, CitationFile, SummaryFile)`
        15. `count ← count + 1`
   16. `nextPageNotEmpty ← Check if there's next page`

**return** (CitationFile, SummaryFile, ResultFile)

Algorithm 2 Algorithm for program ‘ResearchTopic_Parser’.
2.2 Citations Analysis
In the following sections, we will introduce the techniques we developed for the analysis performed on the publication citations information retrieved in Section 2.1. We will start with describing the technique we used to analyse individual researchers and utilize it for the analysis of groups of researchers based on their publications and citations.

2.2.1 Individual Researcher Citations Analysis
The first analysis performed was based on the output citation file generated by program 'Individual_Parser' introduced in Section 2.1.1. With the help of the citation file, we developed a new algorithm called ‘Individual_Cites’ that will be used for calculating the citation counts for each publication with respect to their published calendar year. In order to do this, for each publication recorded in the citation file, we firstly calculate monthly averaged citation counts for that publication based on its year of publication whereas the averaged citation counts is calculated through dividing the total counts by the number of calendar months since the papers was published (line 6, Algorithm 3). Using this averaged citation counts, we will then be able to calculate the yearly citation counts for that publication thus be able to output an overall yearly citation counts for that researcher (lines 7-11, Algorithm 3). A more detailed procedure for this method is shown below in Algorithm 3.

**Input:** name of the citation file

**Output:** Sum vector of citation counts for the researcher with respect to calendar year

**Individual_Cites**(CitationFile):

1. `currentYear ← GetCurrentYear()`
2. `currentMonth ← GetCurrentMonth()`
3. `sumCitationCounts ← []`
4. `for each record in CitationFile:
   5.     `year, count ← GetYearAndCount(record)`
   6.     `ave ← count \frac{(currentYear - year) \times 12 + currentMonth}{count}`
   7.     `for i in range (year, currentYear):
         8.         `if i != currentYear:
             9.             `sumCitationCounts[i] ← sumCitationCounts[i] + ave \times 12`
         10.        `else:
             11.            `sumCitationCounts[i] ← sumCitationCounts[i] + ave \times currentMonth`
   return (sumCitationCounts)`

Algorithm 3 Algorithm for ‘Individual_Cites’.
In order to measure the accuracy of calculation in program ‘Individual_Cites’, the calculated yearly citation counts will be compared against the real counts shown on the profile page (highlighted in orange rectangle in Figure 1). This calculation of a differences score follows the Algorithm 5 illustrated below where the lower the differences score the higher accuracy the program has.

**Input:** Citations count shown on the author profile, Citations count calculated by ‘Individual_Cites’

**Output:** a differences score

**DifferencesCalc**(realCitation, calcCitation):

1. \( \text{diff} \leftarrow 0 \)
2. \( \text{for each year in relCitation:} \)
3. \( \text{realCount} \leftarrow \text{realCitation}[\text{year}] \)
4. \( \text{calcCount} \leftarrow \text{calcCitation}[\text{year}] \)
5. \( \text{diff} \leftarrow \text{diff} + \frac{|\text{realCount} - \text{calcCount}|}{\text{realCount}} \)

\( \text{return (diff)} \)

**Algorithm 5 Algorithm for ‘DifferencesCalc’**.

### 2.2.2 Group Citations Analysis

This section introduces techniques adopted to measure the productivity on a group of researchers. Since Algorithm 3 enables us to calculate the yearly citation counts for an individual researcher, this gives us the opportunity to calculate the sum of citation counts for a group of researchers. The following sections will explain in detail how we analyse the productivity of the researchers in each technique.

#### Calendar Year Citation Counts

For a group of researchers, the first analysis is to calculate their average citation counts for each calendar year after 1960. The calculated average counts will then be plotted together with each individual researcher’s citation plots against each calendar year. This plot will show clearly how much a researcher performed better or worse than other researchers in the same group each year.

#### Career Year Citation Counts

We noticed that within a group of researchers some have already started their careers for years where some were still in their early career. Hence, the second analysis instead of plotting the citation counts for each calendar year, are generating a career year plots. Using the results calculated by Algorithm 3, the earliest active year for each researcher will change to career year 1, where the following years are all adjusted to a career year relative to the earliest year. An averaged count will be calculated using the summation of citation counts for each career year for all the researchers in the group and the resulting graph are represented by plotting this average count together with each researcher’s career year citation plots.
Career Year Citation : Average Counts Ratio

The first two analyses were all focusing on the actual counts of citations which sometimes may not be a good measurement. A third analysis technique was developed then to show the productivity of the group of researchers in a more relative way. This analysis made use of the career year citation counts calculated in the second group analysis. A percentage ratio will be calculated for each career year citation counts of an individual researcher against the average career year citation counts. That is, to calculate how well the researcher performed relative to the average counts where a negative or positive ratio indicating a worse or better performance.

Rank the Researchers

The last analysis was to rank the researchers within the group based on their productivity. Since the average citation counts was already calculated in ‘Calendar Year Citation Counts’, we can calculate the score (illustrated in Formula 1) for each researcher where the higher the score indicates the higher the rank a researcher will have.

\[
Score_i = \sum_{y \in y} - \frac{|\text{CiteCount}_y - \text{Ave}_y|}{\text{Ave}_y} 
\]

\(y: a \text{ calendar year}\)

\(\text{CiteCount}_y: \text{Citation Counts calculated for calendar year } y \text{ by 'Individual Cites'}\)

\(\text{Ave}_y: \text{the average citation counts at calendar year } y\)

2.3 Publications Analysis

Similar to the analysis done with citations for researchers, this part of analysis techniques will be focusing on the number of publications and use it as a marking rubric for measuring the productivity of researchers.

2.3.1 Individual Researcher Citations Analysis

In order to analyse the productivity of each individual researcher based on their publication counts, rather than calculating their citation counts for every calendar year, a new program named ‘Individual_Pubs’ (Algorithm 4) was developed to calculate the total number of publications they published every year since they started their career. Similar to program ‘Individual_Cites’ introduced in Section 2.2.1, we will make use of the output citation file generated by program ‘Individual_Parser’. For each of the publication in the record, we will increase the total publication counts by one for the year this publication was published (lines 2–4, Algorithm 4). As a result, we will be able to have a vector for the total publication counts in each year for the researcher.

\textbf{Input:} \textit{name of the citation file}

\textbf{Output:} \textit{vector of total publication counts for each year.}

\texttt{Individual_Pubs(CitationFile):}
16

1 \text{sumPubCounts} \leftarrow []

2 \text{for each record in CitationFile:}

3 \quad \text{year} \leftarrow \text{GetYearAndCount(record)}

4 \quad \text{sumPubCounts[year]} \leftarrow \text{sumPubCounts[year]} + 1

\text{return (sumPubCounts)}

Algorithm 4 Algorithm for the program ‘Individual_Pubs’.

2.3.2 Group Publications Analysis
In order to analyse publications for a group of researchers who are working in the same research topic, we used similar analysis techniques which were adopted in Section 2.2.2:

- **Calendar Year Publication Counts**: calculate the average publication counts each calendar year for the group of researchers and plot the average counts together with publication counts for all members in the group.

- **Career Year Publication Counts**: transform the publication counts for each researcher to career year counts and calculate the average publication counts each career year for the group of researchers and plot them together in one graph.

- **Career Year Citation Average Counts Ratio**: A percentage ratio will be calculated for each career year citation counts of an individual researcher against the average career year citation counts and plot the ratio for all researchers in the group into a graph.

- **Rank the researchers**: based on the publication counts calculated average publication counts each calendar year, calculate the score (Formula 2) for each of the researcher in the group and rank them with regards to their scores (the higher the score, the higher the rank).

\[ \text{Score}_i = \sum_{y \in \text{year}} \left( \frac{\text{PubCount}_y - \text{Ave}_y}{\text{Ave}_y} \right) \] \hfill (2)

\( y: \text{a calendar year} \)

\( \text{PubCount}_y: \text{Publication Counts calculated for calendar year } y \text{ by 'Individual_Cites'} \)

\( \text{Ave}_y: \text{the average publication counts at calendar year } y \)
Chapter 3 Evaluations

In this chapter, we will evaluate the analysis techniques we introduced in Chapter 2. Section 3.1 will perform the evaluation on the calculation of Individual Citation Counts we described in Section 2.2.1 where our random researchers were chosen for measuring the accuracy of this calculation. Section 3.2 will evaluate the analyses regarding groups of researchers where four research topics from different research areas ‘Catalysis’, ‘Data Mining’, ‘Genomics’, and ‘Pharmacokinetics’ were selected to perform the evaluation.

3.1 Individual Citation Counts Accuracy

At the beginning of this project, rather than calculating the Citation counts using a monthly averaged citation counts adopted in ‘Individual_Cites’, we did calculate a yearly averaged citation counts. In order to evaluate the accuracy of these two types of calculation, four random researchers were selected and difference scores were calculated with both calculations using Algorithm 5 introduced earlier in Section 2.2.1. The results were shown in Table 4 and Table 5 below where we can notice that, the difference score of monthly calculation method for all the four researchers were lower than the yearly method thus having a higher accuracy.

Table 4 Actual and Calculated Citation Counts for researchers Brian Anderson, Peter Christen, Lexing Xie, and Arkady Zaslavsky

<table>
<thead>
<tr>
<th>RID</th>
<th>RName</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>6Slzb38AAAAJ</td>
<td>Brian Anderson</td>
<td>1563</td>
<td>1657</td>
<td>1961</td>
<td>2049</td>
<td>2250</td>
<td>2520</td>
<td>2514</td>
<td>2749</td>
<td>1775</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>Peter Christen</td>
<td>1426</td>
<td>1713</td>
<td>2191</td>
<td>2558</td>
<td>2258</td>
<td>2351</td>
<td>2389</td>
<td>2427</td>
<td>2035</td>
</tr>
<tr>
<td>u0xUDSoAAAAJ</td>
<td>Lexing Xie</td>
<td>113</td>
<td>133</td>
<td>168</td>
<td>190</td>
<td>192</td>
<td>201</td>
<td>313</td>
<td>361</td>
<td>400</td>
</tr>
<tr>
<td>sTAoZeUAAAAJ</td>
<td>Arkady Zaslavsky</td>
<td>127</td>
<td>178</td>
<td>196</td>
<td>197</td>
<td>228</td>
<td>245</td>
<td>308</td>
<td>338</td>
<td>283</td>
</tr>
</tbody>
</table>

Table 5 Difference Scores calculated for researchers Brian Anderson, Peter Christen, Lexing Xie, and Arkady Zaslavsky

<table>
<thead>
<tr>
<th>RID</th>
<th>RName</th>
<th>Difference score (year)</th>
<th>Difference score (month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6Slzb38AAAAJ</td>
<td>Brian Anderson</td>
<td>0.0991996</td>
<td>0.067909</td>
</tr>
<tr>
<td>oz-2CXsAAAAJ</td>
<td>Peter Christen</td>
<td>0.212832</td>
<td>0.19961647</td>
</tr>
<tr>
<td>u0xUDSoAAAAJ</td>
<td>Lexing Xie</td>
<td>0.1378674</td>
<td>0.1220656</td>
</tr>
<tr>
<td>sTAoZeUAAAAJ</td>
<td>Arkady Zaslavsky</td>
<td>0.07013629</td>
<td>0.0586942</td>
</tr>
</tbody>
</table>
3.2 Evaluation of Analysis for Group of Researchers

In this section, we will perform and evaluate the analysis techniques described in Section 2.2 and 2.3 on four groups of researchers with research topics ‘Catalysis’, ‘Data Mining’, ‘Genomics’, and ‘Pharmacokinetics’. Table 6 below will show the characteristics of these four groups of researchers.

Table 6 Characteristics of researchers selected on research topics ‘Catalysis’, ‘Data Mining’, ‘Genomics’, and ‘Pharmacokinetics’

<table>
<thead>
<tr>
<th>Researcher</th>
<th>No. of Researchers</th>
<th>Total Citations</th>
<th>Total Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalysis</td>
<td>5</td>
<td>189750</td>
<td>6050</td>
</tr>
<tr>
<td>Data Mining</td>
<td>5</td>
<td>440570</td>
<td>6605</td>
</tr>
<tr>
<td>Genomics</td>
<td>6</td>
<td>713685</td>
<td>4860</td>
</tr>
<tr>
<td>Pharmacokinetics</td>
<td>6</td>
<td>64453</td>
<td>2440</td>
</tr>
</tbody>
</table>

3.2.1 Calendar Year Citation Plot Evaluation

Figure 7 below shows the Calendar Year Citation plot for the four groups of researchers. In the ‘Pharmacokinetics’ group, it can be observed that the researcher Melvin E. Andersen’s has performed the best in the group as his citations are always above the average citation counts. Another observation is the fluctuations in the plot around year 1972, 1980 and 1992. This is because some of the researchers just started their career in these years and don’t have much publications yet.

In group ‘Data Mining’, we can notice that the researcher Jeffrey Ullman performed the best in years 1960-2005 but the citation counts stops to grow after year 2005. It might due to his reduced number of publications. Similar to the reason in group ‘Pharmacokinetics’, the average citation counts drops quite a lot in years 1980-1983 as three of the researchers started their career during those years.

The average plot for group ‘Catalysis’ is the smoothest one among these four groups despite a minor drop from year 1977 to year 1994 as four researchers started their careers. The researcher Younan Xia who started his career in 1994 has performed fairly in his early years where he becomes successful after year 1995.

The plot for group ‘Genomics’ has a very different layout when comparing to the plots with other groups where the average citation counts is much steeper. It can be observed that all researchers have started their career after year 1970 which is around 10 years later than the researchers in other groups. This might because that Genomics is a more emerging topic which attracts more concerns in the last 30 years. This also coincides with the graph as the citation counts grows significant since year 1980.
3.2.2 Career Year Citation Plot Evaluation

Figure 8 below presents the four Career Year Citation plots for the groups of researchers. The average citation counts for group ‘Pharmacokinetics’ has grown faster in the early 20 career years of the researchers. An interesting observation is that, although three of the researchers have been working on the topic for more than 30 years, the two other researchers who are still in their early careers performed much better than the three longer career ones.

In group ‘Data Mining’, there is one researcher Jeongkyu Lee with the longest career life has the least citation counts throughout his career which is also much less than the averaged citation counts. The average plot grows significantly between career years 0 to 12 and becomes more gradual afterwards. The two researchers with longer career lives perform worse than other researchers in the group.
Similar to the Calendar Year Citation Plot, the Career Year Citation plot for the group ‘Catalysis’ is the smoothest one among all plots. The majority of the researchers in the group have a similar citation counts plots where one of the researcher Younan Xia becomes more successful than other researchers in the group.

In group ‘Genomics’ the average plots grows significantly between career years 0-8. This is due to one of the researcher named Des Higgins has becomes successful very soon after he started his career where his citation counts plot is always higher than the average plot. The rest of the researchers grow gentler than Des Higgins where they become more successful after around 20 years since they started their career.

Figure 8 Career Year Citation plot for the four groups of researchers
3.2.3 Calendar and Career Year Publication Plot Evaluation
Figure 9 and 10 below present the four Calendar and Career Year Publication plots for the groups of researchers. The average plot for the group ‘Pharmacokinetics’ has the slowest growing trend since year 1970 (Career year 0). This indicates that the productivity of the researchers in this group is the most stable among all other groups.

The plots for ‘Data Mining’ and ‘Catalysis’ are growing slowly before year 1985 (Career year 10) which indicates a low productivity from these two groups of researchers. However, after year 1985 (Career year 10), their productivity has improved gradually as average plots start to increase linearly.

For the group ‘Genomics’, the average plot before year 1985 (Career year 2) indicates a very low productivity of the researchers in this group. The plot grows significantly from year 1985 to 1990 (Career year 2 to 5) which might because that this topic is an emerging topic which starts to attract more concerns in the recent 30 years. The major difference between this group and other groups is that, after year 1990 (Career 5), the average plot stops to grow and remains at a constant level since then.

Figure 9 Calendar Year Publication plots for the four groups of researchers
Figure 10 Career Year Publication plots for the four groups of researchers
4.1 Conclusion
In this project, we present an approach to analyse the productivity of researchers. We developed programs to collect data set from the Google Scholar online digital library by crawl and parse the web page data. We then developed various techniques for the analysis of the productivity of individual and groups of researchers. We then evaluate the accuracy and stability of our analysis which indicates that our analyses are able to support our temporal analysis of the data sets we collected.

4.2 Future Work
In the future, we would like to develop more analysis techniques for supporting a more comprehensive measurement of the productivity of researchers. In addition, clustering of researchers will also be considered depending on their performance.
Reference

# Appendix: Project Contract

**ASSESSMENT**  (as per course's project rules web page, with the differences noted below):

<table>
<thead>
<tr>
<th>Assessed project components:</th>
<th>% of mark</th>
<th>Due date</th>
<th>Evaluated by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report name style: (e.g. research report, software description...)</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Article: name kind: (e.g. software, user interface, robot...)</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presentation:</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MEETING DATES OF KNOWN:**

Will be decided later.

**STUDENT DECLARATION:** I agree to fulfill the above defined contract:

---

**SECTION B (Supervisor):**

I am willing to supervise and support this project. I have checked the student's academic record and believe this student can complete the project.

---

**SECTION C (Course coordinator approval):**

---

Research School of Computer Science

Form updated Jun-12