REPUBLIC OF TURKEY

AKDENIZ UNIVERSITY

PERFORMANCE EVALUATION SYSTEM FOR ACADEMIC PUBLICATIONS USING DATA MINING TECHNIQUES

Taha Yiğit ALKAN

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

JUNE 2019

ANTALYA

REPUBLIC OF TURKEY

AKDENIZ UNIVERSITY

PERFORMANCE EVALUATION SYSTEM FOR ACADEMIC PUBLICATIONS USING DATA MINING TECHNIQUES

Taha Yiğit ALKAN

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

JUNE 2019

ANTALYA

REPUBLIC OF TURKEY
AKDENIZ UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PERFORMANCE EVALUATION SYSTEM FOR ACADEMIC PUBLICATIONS USING DATA MINING TECHNIQUES

Taha Yiğit ALKAN

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

This thesis unanimously accepted by the jury on 26/06/2019.

Prof. Dr. Melih GÜNAY (Supervisor)
Assoc. Prof Dr. Bekir Taner SAN
Asst. Prof. Dr. Asım Sinan YÜKSEL

ÖZET

VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK AKADEMİK YAYINLAR İÇİN PERFORMANS DEĞERLENDİRME SİSTEMİ

Taha Yiğit ALKAN

Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Melih GÜNAY

Haziran 2019; 43 sayfa

30.000 ve üstü öğrenci sayısına sahip büyük üniversitelerdeki araştırma alanları genellikle sosyal bilimler, fen bilimleri, uygulamalı bilimler, sağlık bilimleri, güzel sanatlar ve atletizm gibi geniş bir yelpazedeki disiplinleri içerir. Bu nedenle, bireysel araştırmacıların ve üniversite içindeki bölümlerin araştırma performansını değerlendirmek ve karşılaştırmak zordur.

Bu çalışmada, Akdeniz Üniversitesi'nde performansın değerlendirilmesi, araştırmaların desteklenmesi ve işbirliğinin geliştirilmesi için bir yazılım geliştirilmiştir. Bu çalışma için veriler İnsan Kaynakları, Web of Sciences (Wos) ve InCites veri tabanından elde edilmiştir. Veriler veri madenciliği teknikleri ile analiz edilmiş ve araştırma alanlarına göre değerlendirilmiştir.

Bu çalışma sonucunda yükseköğretim kurumlarında akademik yayın performansının değerlendirilmesi için özgün bir yazılım uygulanmıştır. Bir akademik birim veya personel için araştırma performansı gerçek zamanlı olarak erişilebilir hale gelmiştir. Yayın performansı, araştırma kalitesinin ve etkinin iyi bir göstergesi olduğundan, akademisyenlerin ve üniversitelerin etkinliğini belirlemek için kullanılabilir.

ANAHTAR KELİMELER: Akademik Performans, Araştırma Göstergeleri, Veri Madenciliği, Web of Science

JÜRİ: Prof. Dr. Melih GÜNAY
Doç. Dr. Bekir Taner SAN
Dr. Öğr. Üyesi Asım Sinan YÜKSEL

ABSTRACT

PERFORMANCE EVALUATION SYSTEM FOR ACADEMIC PUBLICATIONS USING DATA MINING TECHNIQUES

Taha Yiğit ALKAN

MSc Thesis in Computer Engineering
Supervisor: Prof. Dr. Melih GÜNAY
June 2019; 43 pages

Research areas at large universities with a student body of 30K+ often include a wide range of disciplines from Social Sciences, Natural and Applied Sciences, Health Sciences, Fine Arts to Athletics. Therefore, it is challenge to evaluate and compare the research performance of individual researchers and departments within the university.

In this study, a software has been developed to evaluate performance, support research and collaboration at Akdeniz University. The data for this study is obtained from the database of HR, Web of Science (Wos) and InCites. The data has been analyzed by data mining techniques and evaluated according to research areas.

In this study, an original software for the evaluation of academic publication performance was implemented for the Higher Education Institutes. Research performance for an academic unit or staff may be accessible in real time. As publication performance is a good indicator of quality of research and impact, it can be used to determine the effectiveness of academicians and universities.

KEYWORDS: Academic Performance, Data Mining, Research Metric, Web of Science

COMMITTEE: Prof. Dr. Melih GÜNAY
Assoc. Prof. Dr. Bekir Taner SAN
Asst. Prof. Dr. Asım Sinan YÜKSEL

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Dr. Melih GÜNAY who helped, guided and motivated me during this study.

I would like to thank Assoc. Prof. Dr Bekir Taner SAN, Assoc. Prof. Dr. Olgun KİTAPÇI and Fatih ÖZBEK for their supports and advices during the development of Academic Performance Evaluation System.

I would like to thank my all professors, especially Asst. Prof. Dr. Asım Sinan YÜKSEL, for their teaching and effort throughout my college education.

Special thanks are for Gamze KIRAN for their endless patience, love, support.

LIST OF CONTENTS

TEXT OF OATH

I declare that this study "Performance Evaluation System for Academic Publications Using Data Mining Techniques", which I present as master thesis, is in accordance with the academic rules and ethical conduct. I also declare that i cited and referenced all material and results that are not original to this work.

26/06/2019

Taha Yiğit ALKAN

ABBREVIATIONS

ESI	: Essential Science Indicators
HR	: Human Resources
IDF	: Inverse Data Frequency
JIF	: Journal Impact Factor
KDD	: Knowledge Discovery in Databases
TF	: Term Frequency
WoS	: Web of Science

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Nowadays, many indicators are used to evaluate the academic performance of an academic organization, department or academician. Among indicators perhaps publication performance is the most commonly used of all. Today, with the increase of online publications and ease of publishing through Internet, the number of publications exploded. However, with this quantitative increase, the average quality of a publication dropped and consequently it became increasingly difficult the judge the contribution of the academic work hence the performance of the researcher.

In order to evaluate publication performance of a researcher, many metrics such as H-index, M-index and G-index were developed. The most known of these indicators is H-index. H-index measures individual research performance in 2 dimensions namely quantity and quality. Quantity stands for number of publications and quality stands for number of citations. On the other hand, H-index doesn't show information such as high referenced publications, seniority and current publication activity. Moreover, in different research areas the habits of publications and citations differ. For this reason, researchers and publications should be evaluated within their research area.

The subject of this study is to examine the characteristics of the publications of researchers and establish research profiles of academic units at Akdeniz University and quantify their impact within the respected research area. Akdeniz University is established at 1982 with 6 faculties including medicine, engineering, agriculture, science and literature, fine arts, economics and administrative sciences. As of today, the university has 23 different faculties and 5 institutes. The total of undergraduate students in the 2018-2019 academic year was over 70.000, while there were about 3000 postgraduate students. All academic staffs, including researcher, assistant, associate and professors was over 2000 at the end of 2018. Akdeniz University has wide range of research areas including social sciences, health sciences, and engineering.

Within the scope of this study, the publications of Akdeniz University researchers were evaluated with data mining methods. The concept of data mining can be explained as obtaining valuable data from large datasets (Özkan 2016). Today, data mining has been used in a wide variety of industries including marketing, earth science, computer science, finance, health services, and social media. With the data extracted, customer relationships, decision making, planning and forecasting can be improved, new products can be developed, and competitive advantage can be achieved.

The data that is the basis of the study and data mining, were obtained from Web of Science which is an online scientific citation indexing service maintained by Clarivate Analytics. Web of Science has known as the oldest citation resource, containing the most prestigious academic journals used for the purpose of citation analysis (Leslie and Rensleigh 2013). Citation index is the index that lists the publications published in scientific resources and the references that these publications received. There are many indexes separated by subject matter (Art and Humanities Citation Index, Science Citation Index, Emerging Sources Citation Index, Social Sciences Citation Index, Innovative Science and Technology Publications etc.). Publications and references are scanned in these indexes. These databases can be accessed via online scientific citation indexing services such as Google Scholar, Web of Science, Scopus and Cite Seer. It is a challenge to identify authors when indexing publications. To assign and identify individual researchers, organizations provided ResearcherID, OpenID and ORCID for authors. ResearcherID was introduced in January 2008 by Thomson Reuters. On the ResearcherID web site, authors are asked to link their ResearcherID profiles with their own articles. In this way, the problem of identifying authors has been solved.

ResearcherID was used to ensure that the data used in this study were reliable. However, in Akdeniz University not all researchers provided a ResearcherID. This study is therefore confined with the research areas and researchers that whose researcher ids are obtained. As of 2019 March, 1450 researchers are registered in ResearcherID database of Akdeniz University.

Several parameters (citations, publications, journals, HR metadata) were tracked to achieve the aim of the study and challenges were overcome as follows:

- Getting up-to date researcher information of the university
- Getting up-to date reliable publication information of the researchers
- Handling continues updates on citation of individual publications
- Varying publication metrics and research characteristics by research areas

- Obtaining journal info and metrics

The thesis first explains the concept of data mining, presents a review of the literature on academic performance and then it presents the methodology used in the current study. The thesis ends with a review of the main findings, discussion, implications and limitations of the study.

2. LITERATURE REVIEW

2.1. Data Mining

Data mining is a multidisciplinary sub-field of computer science and statistics. Since the concept of data mining is very comprehensive, there are many definitions in the literature. Data mining is the process of obtaining previously unknown, valid and applicable information from large data sets and using these information while decision-making (Silahtaroğlu 2013). Data mining is a new discipline that has sprung up at the confluence of several other disciplines, stimulated chiefly by the growth of large databases (Hand 2006). Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems (Chakrabarti et al. 2006).

2.2. Knowledge Discovery in Databases Process

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery (Han et al. 2011). KDD consists of 7 steps:

1. **Data cleaning**

   Data cleaning is the stage which inconsistent and noisy data is cleared from the database. The missing data should be replaced by determined new data. In the determination of the new data, many methods such as calculating average value, regression, using a constant can be used.

2. **Data integration**

   The data to be used in the application can be obtained from many different sources. This data need to be combined into one common source (Data Warehouse).

3. **Data selection**

   All of the collected data may not be required for processing. In accordance with the data mining method to be applied, useful data should be selected.

4. **Data transformation**

Data transformation is defined as the process of transforming data into appropriate form required by data mining procedure. Several methods such as min - max normalization and z-score normalization can be used according to data type.

5. **Data mining**

   At this stage, data mining techniques are applied to the prepared data. This technique may be classification, regression or clustering according to the purpose of the KDD process.

6. **Pattern evaluation**

   The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value.

7. **Knowledge presentation**

   Knowledge representation is the process of visualizing the results of data mining in a clear way. Reports and tables can be created; discriminant rules, classification rules can be supported with visuals; trees can be visualized.

2.3. Data Mining Components

Some key components are needed for the data mining process:

- **Data source:** can consist of any kind of information repository such as database, data warehouse, World Wide Web and spreadsheets.

- **Data warehouse server:** presents relevant data that is ready to be processed based on the data mining process.

- **Knowledge base:** a store of information or data that is available to draw on. Pattern evaluation module interacts with the knowledge base in order to make results more accurate and reliable.

- **Data mining engine:** The data mining engine is a key component of the data mining process that performs data mining tasks such as classification, clustering and prediction.

- **Pattern evaluation module:** uses interest measures and interacts with the data mining modules in order to focus the search towards interesting patterns (Hemalata and Vasanthakumari 2013).

- **Graphical user interface:** communicates between the user and the data mining system. This module helps the user to use the system easily and efficiently without knowing the real complexity behind the process.

2.4. Data Mining Applications

Today, data mining is used in a wide variety of industries including marketing, finance, health services, and social media. With the data extracted, customer relationships, decision making, planning and forecasting can be improved, new products can be developed, and competitive advantage can be achieved.

2.5. Data Mining Methods

There are several major data mining techniques that have been developed and used in data mining process. Data mining techniques can be separated into 3 major categories as classification, clustering and association rules by application type.

2.5.1. Classification

Classification is a supervised learning technique. Classification algorithms learn classification pattern by using training data and then uses this pattern to classify new data. The values that specify classes on the data set are called labels.

Classification algorithms can be grouped into three broad categories by their methods: decision trees, artificial neural networks and genetic algorithms.

2.5.2. Clustering

Cluster analysis is an unsupervised method for descriptive analytics. Clustering is a process that groups data with similar properties into subsets. At the beginning of the clustering process, data is not isolated and at each iteration clusters are grouped together based on their similarity to each other.

There are many methods (hierarchical methods, agglomerative clustering algorithms, partitioning clustering algorithms and graph clustering etc.) in literature for clustering analysis. Choosing the appropriate algorithm according to the data type is very important for cluster analysis.

2.5.3. Association Rules

Association rules are data mining methods that analyze the occurrence of events together (Özkan 2008). These methods reveal the association rules with certain possibilities. Association rules are the approaches that supports analyzing the historical data and identifying the association behaviors in this data.

Apriori, carma, sequence, GRI, eclat and FP-Growth are main algorithms being used in association rules analysis.

2.6. Related Works

There are several indicators that are used for the evaluation of scientific publication performance such as total number of publications, average number of citations, and number of qualified publications that are above a certain threshold. However, these indicators only assess the performance of publication and researcher on a quantitative basis in single dimensions therefore insufficient. John E. Hirsch, proposed h-index as a new indicator in 2005 to measure individual research performance in 2 dimensions namely quantity and quality (Hirsch 2005). H-index is powerful in expressing publication and citation activity for a given research area. The definition of the index is that a researcher with an index of h has published h papers each of which has been cited in other papers at least h times (Hirsch 2005). The H-index differs according to the data set. As an example, because the journals they index are different, an author's H-index indicator can be calculated differently on Google Scholar and Web of Science.

A new publication of an author does not immediately affect the h-index indicator, nor does it change the h-index indicator when a publication fails in terms of citation activity. In order to increase the H-index indicator, the number of publications and the number of citations in publications must increase proportionally. Thus, publication activity and citation activity are measured effectively.

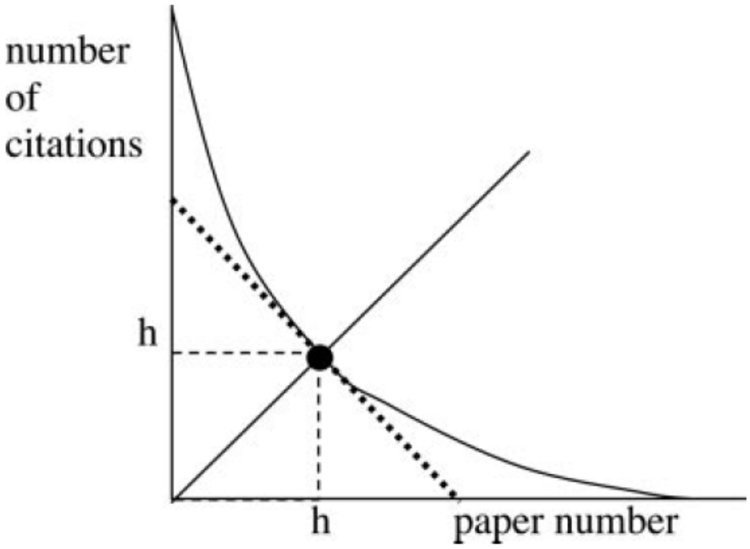


Figure 2.1. H-index representation (Hirsch 2005)

There are also negative aspects of the H-index indicator. There is a certain period of time for publications to cite. This situation poses a disadvantage for the authors who have just begun publishing. Also, the citation of previous publications may be misleading about the current publication activity of the senior authors. As the H-index cannot exceed the number of publications by definition, it may not show a small number of highly cited publications. Also, considering that the effect of publications in different research areas vary, H-index may not accurately reflect academic performance.

In 2006, Leo Egghe proposed the g-index indicator because that the H-index indicator does not show high-referenced publications (Egghe 2006). Basically, it is based on the h-index indicator, but in short g index shows that the researcher has g publications with at least g^2 references.

In order to show high-referenced publications, Chun-Ting Zhang proposed the e-index indicator in 2009. The purpose of the e-index indicator is to differentiate between different citation patterns of scientists with close h-index indicators. The e-index indicator can be explained by the formula given below (Zhang 2009).

$$e{-}index = \sqrt{total\ number\ of\ citations - minimum\ hirsch\ index\ citations\ required} \tag{2.1}$$

For example, if the total number of citations of the first 5 publications of a researcher with h-index indicator 5 is 125 (the minimum number of h-index references required for h= 5 is 5x5= 25),

$$e - index = \sqrt{125 - 25} = 10 \tag{2.2}$$

E-index is calculated as 10.

The m-index indicator has been developed for the proper comparison of academicians who are new to publishing and senior academics. When calculating the M-index indicator, h-index indicator is divided by the time between the first publication and the last publication.

Another solution to this problem was brought by Bihui Jin in 2007. The age-weighted citation rate is an age-old citation derived from the number of references to specific publications divided by the age of that article. Jin defines the AR-index indicator as the square root of the sum of the age-weighted citation of all articles contributing to the H-index indicator (Jin et al. 2007).

Table 2.1. Comparison of publication performance indexes

	Publication Activity	Citation Activity	High Referenced Publications	Seniority	Current Publication Activity
H-Index	✓	✓	X	X	X
G-Index	✓	✓	✓	X	X
E-Index	✓	✓	✓	X	X
M-Index	✓	✓	X	✓	X
AR-Index	✓	✓	✓	✓	X

There have already been many studies of academic performance in the literature. For example, Soutar et al. (2015) conducted an analysis of the research impact of 2263 marketing academics using citation metrics in the top 500 research universities. The results indicated that ranks the top 100 university marketing departments in the top 500. Patel et al. (2012) compared h- index scores for the academic performance of healthcare researchers from databases. Bar-Ilan (2008) compared the H-indexes of a list of highly-cited

Israeli researchers from Google Scholar, Web of Science and Scopus databases.

In a study conducted in 2008, in Turkey between 1990 and 2000, each year the number of publications in the SCI/SCI-E increased and it is stated that the H-index index also increased. However, those states that less number of publications that compared with other countries, it was determined that they have a much higher H-index than Turkey and Turkey ranks last compared to other countries (Umut 2008).

A fuzzy logic approach was proposed by Kaptanoğlu and Özok (2010) for the evaluation of academic performance. While applying fuzzy logic approach, 3 main criteria were determined as research, education and service, and 3 different methods (Liou and Wang, Abdel Destiny and Dugdale, Chang) were tried separately for ordering fuzzy values. At the end of the study, while consistent results were obtained, it was stated that in Chang method, because of the service criterion 0, it has a different effect on the result. The study showed that the problem of academic performance evaluation can be solved as a fuzzy decision-making problem.

URAP (University Rankink by Academic Performance) research laboratory was established in 2009 at the Middle East Technical University Informatics Institute. In order to evaluate higher education institutions in line with their academic achievements, they develop scientific methods and share the results of the studies with the public (URAP 2018).

In 2016, TUBITAK published its university competency analysis report. In this report, the publications and projects of universities were examined between 2010 and 2014, and the reports based on objective data were given about the areas where the universities were competent. The indicators taken into consideration while conducting competency analysis are discussed under two main headings as “volume” and “quality” (TUBITAK 2016).

There are also several studies in the literature about the data mining methods used in this study. In a study conducted in 2013, jaccard, dice and cosine coefficient, which are similarity measurements, were compared. Cosine coefficient gave the best result in the study performed with genetic algorithm (Thada and Jaglan 2013). In another study, distance measures have been compared on k-means algorithm for text clustering. Euclidean distance was determined to be unsuitable for text clustering. The other 4 criteria showed similar performance despite differences. The Jaccard and Pearson coefficient measures

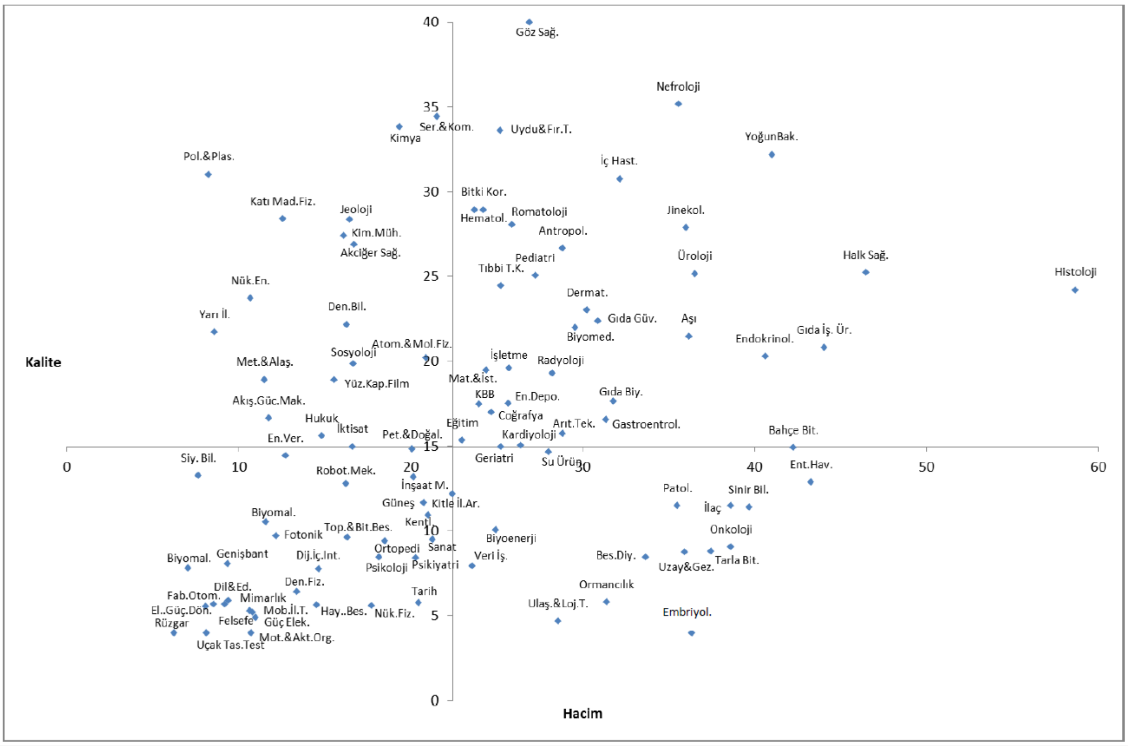


Figure 2.2. Akdeniz University research areas (TÜBİTAK 2016)

found more consistent clusters. The clustering solutions obtained with Pearson correlation coefficient and the averaged KLD divergence measure were more balanced (Huang 2008).

In 2014, an improved k-means algorithm using modified cosine distance measure was proposed. In the experiments conducted on the large data set over mahout and hadoop, better results were obtained with modified cosine distance. Better results were obtained in terms of cluster size, inter and intra cluster distances and cluster parent words (Sahu and Mohan 2014).

3. MATERIAL AND METHOD

In the process of determining the method of the study, many sources have been examined and a structure has been created to provide the most suitable conditions. In the data layer, which is the basis of the application, Microsoft SQL Server 2017 was used to store the data. Since application data is derived from many different sources, many services have been developed to keep this data up to date. Also, REST API and user interface projects are provided for processing and presenting of the data processed. In the development of these software tools, many open source software and platforms such as Angular, RabbitMQ, .NET Core, Python and Highcharts have been used.
The project can be examined in three main modules as shown in Figure. 3.3:

1. HR data integration module
   In order to keep the staff data up to date, the service is designed in the .NET core platform. This service, which runs once a day, updates the data of the staff who have started to work, quited work or change their department.

2. Web of Science data integration module
   An asynchronous structure is constructed to ensure that the Web of Science data is always up to date. Since the data to be constantly updated is very large and Web of Science web services have a request constraint, this module uses queue structure. At the implementation stage of the queue structure, open source RabbitMQ software has been used. A scheduled service posts the ResearcherID data to queue, another service consumes the queue and uses this data to update the database using the Web of Science services.

3. Data Mining & Analytic module
   REST API was developed in the .Net CORE platform for the querying and preparing the data during analytic, data mining, presentation and evaluation stages. Additionally, the REST API was developed by using python programming language because of the large number of data mining libraries and performance. Results are presented to the user with Angular single page application.

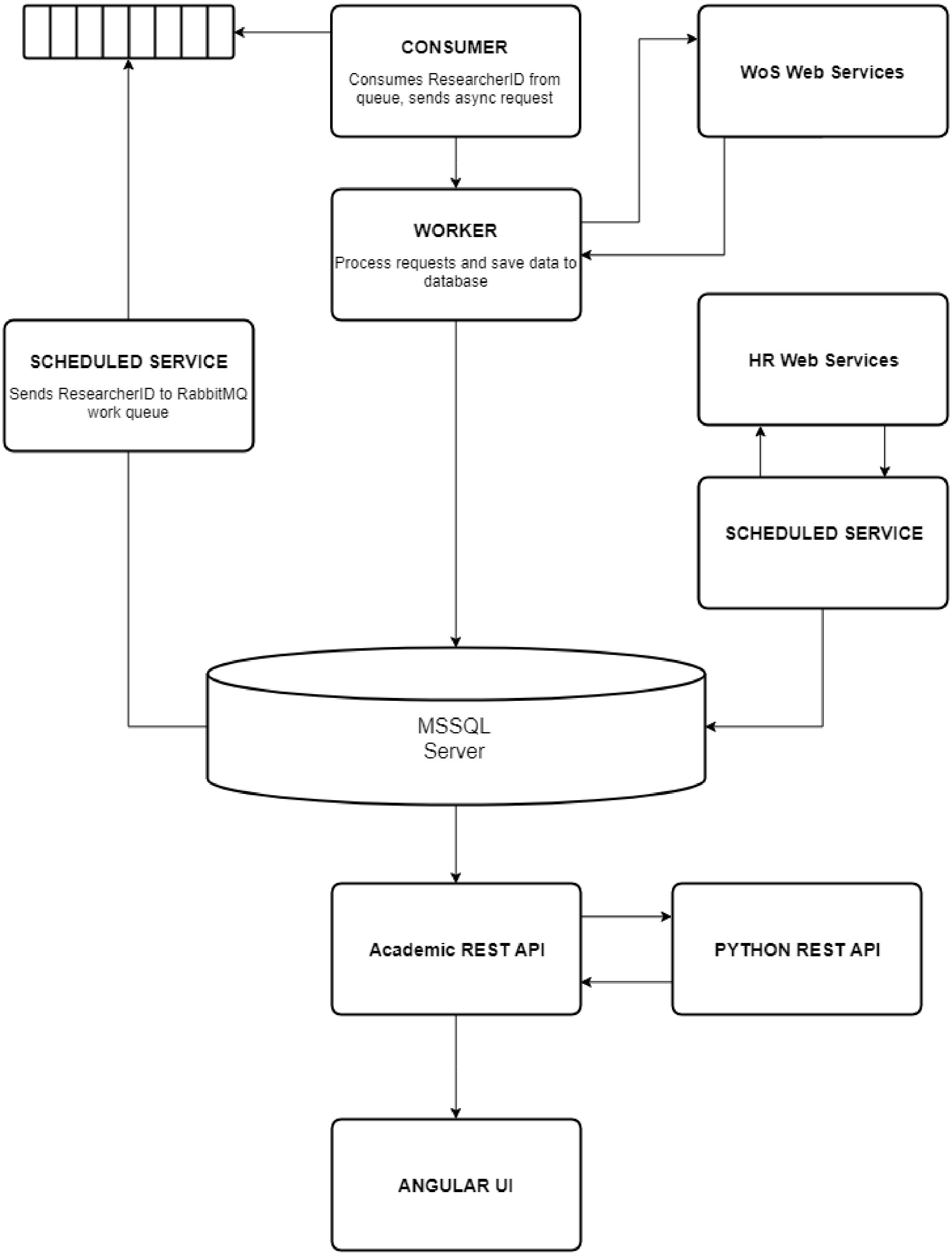

CONSUMER
Consumes ResearcherID from queue, sends async request
WoS Web Services
WORKER
Process requests and save data to database
SCHEDULED SERVICE
Sends ResearcherID to RabbitMQ work queue
HR Web Services
SCHEDULED SERVICE
MSSQL Server
Academic REST API
PYTHON REST API
ANGULAR UI

Figure 3.3. Application architecture diagram

3.1. Data Set

The key analysis reports of the academic performance system rely on the data retrieved from the Web of Science (WoS) databases. When authors publish research articles, they often supply first and last names, contact information and academic institutes. Even though such metadata can often reliably be used to evaluate and compare the performances of academic institutes, individual performance evaluation may be a challenge as names may not be unique and consistently entered throughout the career of an academicians. In order to uniquely identify authors and associate them with their publications, a ResearcherID profile was used. At Akdeniz University, administration asked academic staff to obtain a ResearcherID if they don't have one and add publications to their profile and let research office to know their ResearcherID.

The first phase of this study was getting researcher information from human resources department. Department hierarchy has been constructed in database and staff populated with some information such as name, surname, title, ResearcherID, date of start and date of dismissal according to their department

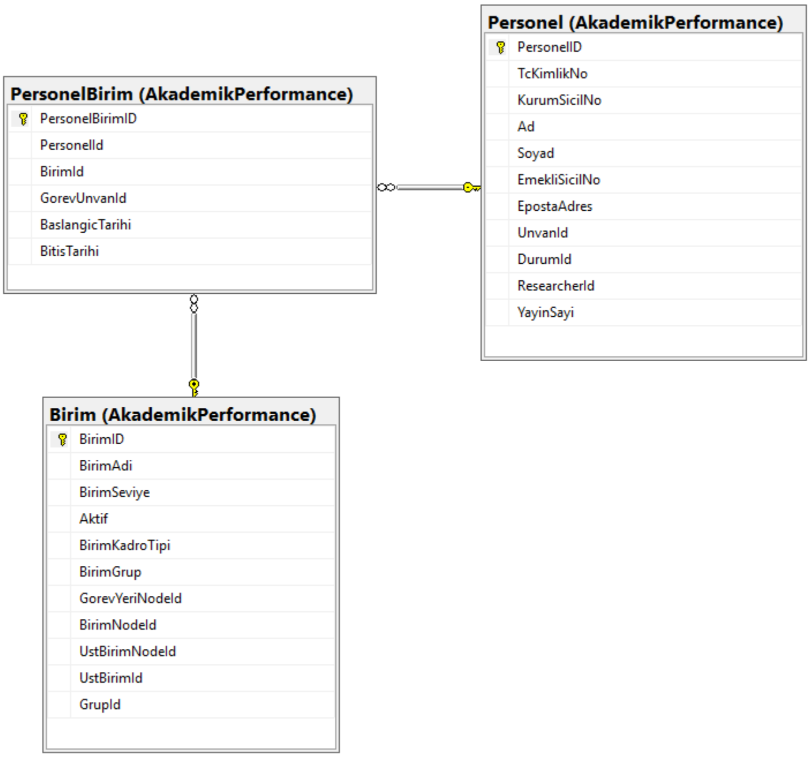


Figure 3.4. Department hierarchy and staff diagram

Subsequently, ResearcherID information was used for obtaining data from Web of Science Web Services Expanded. Web of Science provides SOAP-based APIs which comply JAX-WS, WSDL 1.1, SOAP 1.1 standards (Web of Science). Publications of researchers were obtained through Web of Science API's search method using ResearcherID as author identifier. The response of the search method returns list of publications of researcher and consists of metadata as follows:

- WoSUID (WoS Unique Identifier)
- Title
- Abstract
- Publication date
- Database edition
- Document type
- Language
- Page count
- Journal info
- Subjects
- Headings
- Keywords
- Keyword+ (keyword that WoS assigned)
- Contributors
- Organizations
- Number of citations
- Number of references

Based on the API response a database schema has been designed to prevent data repetition, and established data consistency (Figure. 3.5). After database populated with publications of researchers, publications that cites to these publications and referenced by these publications were obtained with using Web of Science unique identifier. InCites API provides information for publications:

- Average number of citations to articles of the same document type from the same journal in the same database year

- Citation impact normalized for journal, year and document type subject
- Average number of times articles from a journal published in the past two years have been cited in the JCR year
- The harmonic mean of citation rate values for all research fields to which an article is assigned
- The percentile in which the paper ranks in its category and database year, based on total citations received by the paper
- Citation impact normalized for subject, year and document type
- Publication has at least two different countries among the affiliations of the co-authors
- Indicates that more than one institution has contributed to the document
- Papers that list their organization type as corporate for one or more of the co-authors affiliations
- For each publication, these information has been obtained and populated.

Journal is the one of the most important criteria for publication. In order to evaluate journal's impact on publication, some metrics were gathered from Journal Citation Reports. Since we do not have access to Journal Citation Reports services, metrics were gathered as excel file by yearly. Then transferred to the database from the excel files. Metrics consists of:

- Journal name
- Number of citations
- Web of Science document count
- Impact factor (5 years)
- Eigen factor
- Impact factor
- Quartile

Journal Impact Factor is a publication-level metric introduced by Eugene Garfield, the founder of the Institute for Scientific Information, in 1999. The Journal impact factor shows average number of citations to articles published recent 2 years in that journal. JIF has a simple formula:

$$JIF_{year-1} = \frac{citations_{year-1} + citations_{year-2}}{publications_{year-1} + publications_{year-2}} \tag{3.3}$$

JIF Quartile score represents journal's percentile in their own category. Due to InCites calculates Journal Impact Factor with considering Web of Science research areas, a publication may has multiple Quartile score. In such a case, highest Quartile score is considered.

$$Z = \frac{Journal's\ rank\ in\ category}{Total\ number\ of\ journals\ in\ category} \tag{3.4}$$

Table 3.2. JIF Quartile calculation

Q1	0.00 < Z < 0.25
Q2	0.25 < Z < 0.50
Q3	0.50 < Z < 0.75
Q4	0.75 < Z

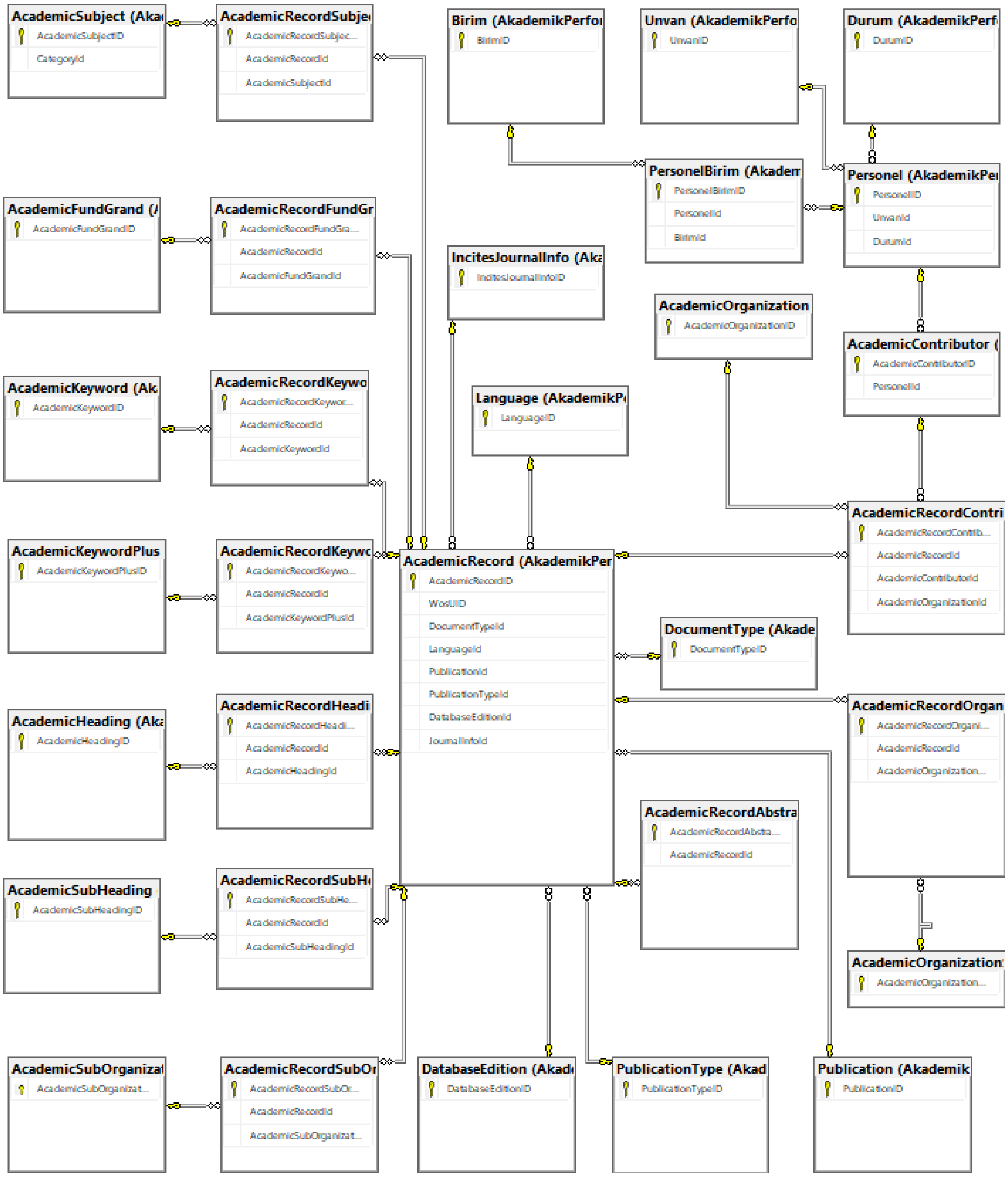


Figure 3.5. Database diagram

3.2. Determination of Research Fields

Nowadays, the studies are grouped in different research areas, but the group numbers and labels vary. For example, Web of Science uses 5 broad categories; arts & humanities, life sciences & bio medicine, physical sciences, social sciences, engineering and technology. According to Comte, Science has 5 branches which are earth & space, social sciences, life sciences, physical sciences and formal sciences. Essencial Science Indicators(ESI), which is an analytical tool, ranks authors, institutions, countries and journals in 22 broad fields.

According to the Web of Science categorization scheme, when the publications in 5 main fields are examined, there are big differences. Based on the Web of Science categorization schema, the data obtained from InCites show a large difference in the number of publications, the number of researchers and the number of citations. More than half of publications and citations are in the life sciences field. On the contrary, the publications in the field of art do not constitute one percent of the total publications.

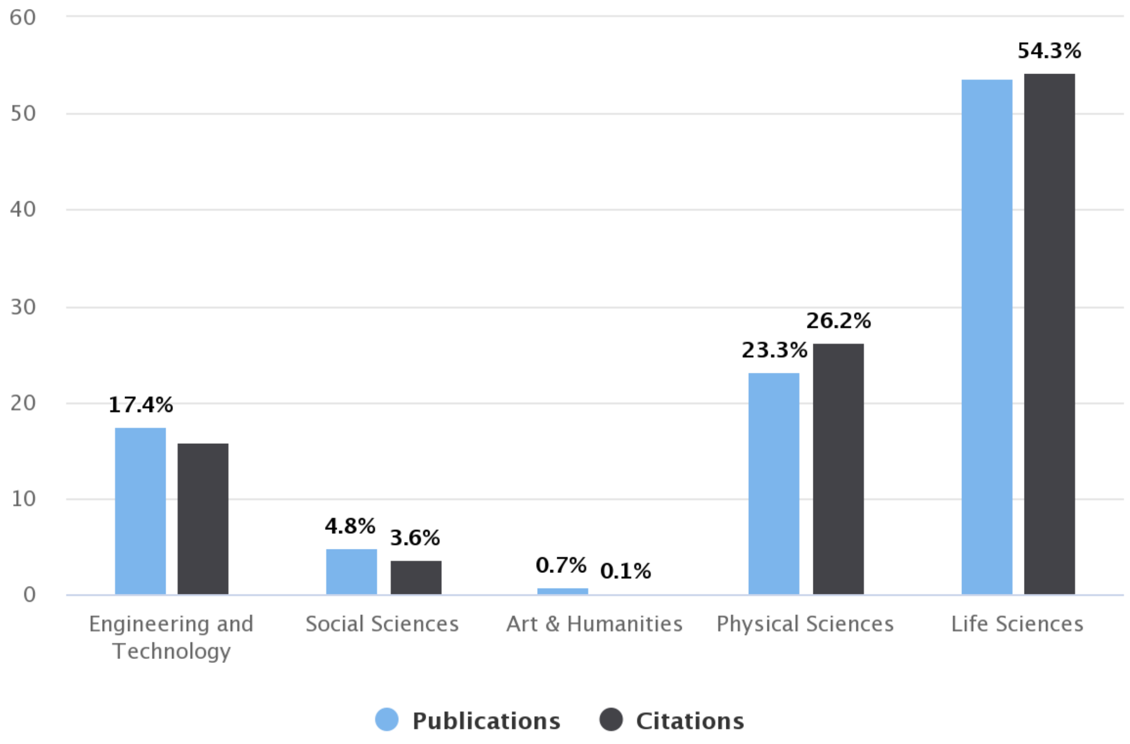


Figure 3.6. Publication and Citation distrubiton

Incites data also shows that there are differences between quartiles. Average citations and quartiles are proportional in each research area except art & humanities.

In this section, cluster analysis has been used in order to group Web of Science categories and find optimum number of research areas. Each publication assigned to one or more categories by Web of Science (Table 3.4). It is possible to make a deduction according to categories of publications. For example; it can be said that mathematics and mechanics

Table 3.3. InCites Data

Category	Quartile	# of publications	# of citations	avg citation	% docs cited
Engineering and Technology	Q1	2,429,314	63,584,260	26.1738	93.3679
	Q2	1,333,544	16,978,585	12.7319	85.6622
	Q3	935,501	6,491,368	6.9389	68.2629
	Q4	914,925	2,795,501	3.0554	41.4337
Social Sciences	Q1	526,086	12,111,46	23.0217	70.8107
	Q2	408,437	4,929,399	12.0689	65.9522
	Q3	311,102	2,535,155	8.1489	60.1770
	Q4	308,199	1,098,474	3.5641	41.7107
Art & Humanities	Q1	77,415	216,494	2.7965	23.6815
	Q2	57,606	160,819	2.7917	30.8857
	Q3	47,613	132,348	2.7796	38.2522
	Q4	55,344	108,544	1.9612	34.1338
Physical Sciences	Q1	3,544,374	109,165,791	30.7997	94.2362
	Q2	1,930,814	27,233,660	14.1047	88.5676
	Q3	1,103,921	8,454,993	7.6590	79.3266
	Q4	908,806	3,872,512	4.2610	61.6276
Life Sciences	Q1	9,210,795	220,987,720	23.9922	62.5374
	Q2	3,816,193	53,423,280	13.9991	73.7076
	Q3	2,396,845	24,119,084	10.0628	75.8171
	Q4	1,856,043	10,331,769	5.5665	62.8506

has a relationship. Likewise, agriculture and plant sciences. Frequency of categories that appear together may show their interest level. With this approach, categories that more appears together should be into same cluster.

With this objective in mind, Total of 238,267 random and unique publications (1000 publications per WoS category have been obtained to establish well-balanced data set. Subsequently, symmetric adjacency matrix has been created with calculating binary com-

Table 3.4. Sample publications and their research areas

Publications	**Research Areas**
Publication 1	Engineering(A), Mathematics(B), Mechanics(C)
Publication 2	Mathematics(B), Mechanics(C), Physics(D)
Publication 3	Engineering(A), Materials Science(E)
Publication 4	Construction & Building technology(F), Engineering(A), Materials Science(E)
Publication 5	Chemistry(G), Electrochemistry(H)
Publication 6	Dermatology(I), Surgery(J)
Publication 7	Transplantation(K), Urology & Nephrology(L)
Publication 8	Agriculture(M), Plant Sciences(N)
Publication 9	Dermatology(O), Surgery(J)
Publication 10	Business & Economics(P), Social Sciences(R)

binations of categories assigned to publications (Table 3.5). Every time two categories appear together, value of the related cell has been increased by 1.

At this part of the study, hierarchical clustering, which is one of the distance based clustering techniques, has been used. Hierarchical clustering has a simple algorithm. At first, it is considered that each observation as a separate cluster. Then, following steps are applied:

```
repeat
    identify the two clusters that are closest
    merge these clusters
    calculate new distances with linkage criteria
until all clusters merged
```

Linkage criteria is one of the most important criteria that affects the accuracy of results of cluster analysis report. Linkage criteria determines from where distance is computed. Commonly used 3 linkage criteria are:

Table 3.5. Sample adjacency matrix calculated from Table 3.4

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R
A		1	1		2	1											
B	1		2	1													
C	1	2		1													
D		1	1														
E	2					1											
F	1				1												
G								1									
H							1										
I										1							
J									1						1		
K												1					
L											1						
M																	
N																	
O									1								
P																	1
R																1	

- Single linkage: between the two most similar parts of a cluster
- Complete linkage: between the two less similar parts of a cluster
- Average linkage: center of the clusters

In this study, complete linkage was used for the exact separation of research areas. As a result of the process, a similar output to the Web of Science categorization scheme was obtained. However, some categories differ because of interdisciplinary works. For example, radiology, nuclear medicine and medical imaging categories are in computer science cluster. This is due to the studies conducted between image processing sub-field of computer science and medical imaging field. Examples can be reproduced but these are exceptions and can be ignored.

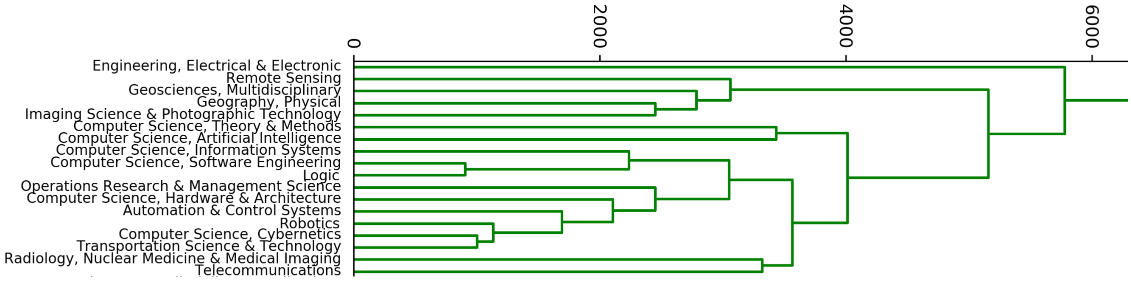


Figure 3.7. Computer science part of dendogram

As a result of cluster analysis, it was decided to evaluate the publications according to 5 research areas; life sciences, art & humanities, technology, social sciences and physical sciences.

3.3. Clustering of Researchers

Increasing collaboration and determining the areas of strength require correct clustering of scientist within a research organization. Grouping of researchers based on department/program may not be the best approach as science is increasingly becoming more interdisciplinary (Porter and Rafols 2009). Therefore, the aim of this process is to objectively group researchers using Web of Science categories. By grouping data using well established data mining techniques, it may be possible to identify collaboration potential between researchers by allowing them to find each other through a developed software.

Term Frequency (TF) is often used in information retrieval and text mining. It shows how frequent a term occurs in a document. In our approach, for each researcher, publications are assigned to 252 Web of Science categories and counted (Table 3.6). Subsequently, a 252-dimensional vector was created for each investigator and each entry is populated with TF-IDF value.

Term Frequency measures the frequency of the category in researcher's publications (Salton and Buckley 1988). Since the number of publication of each researcher is different, the importance of the category for each researcher has to be normalized. In order to normalize and get a relative importance of the research field for the researcher, the publication count in each research area is divided by the total number of publications of the researcher as given below.

Table 3.6. Sample researcher category selection

Researcher	Web of Science Category	Number of Documents
Researcher 1	Hematology	1
	Radiology, Nuclear Medicine & Medical Imaging	1
	Oncology	1
	Medicine, General & Internal	1
	Surgery	1
	Pathology	1
	Geriatrics & Gerontology	1
	Medicine, Legal	6
	Anatomy & Morphology	1
	Gerontology	1
Researcher 2	Food Science & Technology	1
	Agriculture, Multidisciplinary	1
	Plant Sciences	2
	Horticulture	1
Researcher 3	Mathematical & Computational Biology	2
	Multidisciplinary Sciences	2
	Mathematics, Applied	1
	Computer Sciences, Interdisciplinary Applications	1
	Biology	1

$$TF(t) = \frac{\#\, of\ times\ category\ t\ appears\ in\ researcher}{Total\ number\ of\ publications\ of\ researcher} \tag{3.5}$$

On the other hand, the Inverse Data Frequency (IDF) measures the importance of a term with respect to all terms (Robertson 2004). In the application of IDF to research areas, all categories are initially considered equally important. As some common categories such as multidisciplinary sciences, engineering, interdisciplinary computer science and others occur more frequent, the IDF reduces their importance. Meantime, less frequent research area terms become more important with IDF. The logarithmic function is used to calculate the weight of the rare research areas in the entire data set. In that way, the research areas that occur rarely in the data set have a high IDF score.

$$IDF(t) = \log(\frac{Total\ number\ of researcher}{Number\ of\ researcher\ who\ has\ category\ t}) \tag{3.6}$$

After the term frequency and inverse data frequency values are calculated, the multiplication of these values gives the TF-IDF weight.

In this study, k-means method which is one of the partitional clustering methods was used. The mechanism of this method is to minimize the distances of each point to the mid point of the cluster at every iteration. After each iteration, the mid point and the labels of the points are readjusted for optimum cluster (Özkan 2008). Before the K-Means algorithm is applied, the optimum number of sets K should be determined. The approach to determine optimum K is basically calculated by selecting different K values and calculating the total distance of points to their assigned mid points. The process continues until significant changes drop below a set threshold. To determine the set threshold, elbow method is used. Elbow method examines the percentage of variance as a function cluster count. Elbow method consists of following steps (Bholowalia and Kumar 2014):

```
set k=1, WCSS(k)=within cluster sum of errors;
do while:
  set k++;
  WCSS(k)=within cluster sum of errors;
  $\delta = WCSS(k) - WCSS(k-1)$
  if delta < threshold (when sharp drop)
    break
end
```

Within a cluster sum of errors are calculated as follows:

$$WCSS = \sum_{1}^{n}(y_i - x_i)^2 \tag{3.7}$$

where y_i centroid for the observation x_i.

Another factor affecting the performance of clustering is the selection of the distance function. In this study, the following 3 different distance functions are applied:

3.3.1. Euclidean Distance

$$d(\vec{u}, \vec{v}) = ||\vec{u} - \vec{v}|| = \sqrt{\sum_{1}^{n}(u_i - v_i)^2} \tag{3.8}$$

3.3.2. Cosine Distance

$$d(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{||\vec{u}|| \, ||\vec{v}||} = \frac{\sum_{1}^{n} a_i b_i}{\sqrt{\sum_{1}^{n} a_i^2}\sqrt{\sum_{1}^{n} b_i^2}} \tag{3.9}$$

WCSS values calculated using euclidean distance are shown in the Figure 3.8. Accordingly, the k value was first determined as 5. However, when the clusters were examined, a very unbalanced distribution has been observed. 578 of 656 researchers are in same cluster. This comprehensive cluster is usually composed of researchers working in the medical sciences and natural sciences. Since such clustering is undesirable, re-clustering has been performed by increasing the number of clusters. Nevertheless, a balanced distribution could not be achieved.

WCSS values calculated using cosine distance are shown in the Figure 3.8. The k value was determined as 11 and clustering was performed. The results are shown in Table 3.6.

The number of researchers in clusters balanced and ranged from 29 to 136. Each cluster has some significant research categories:

- **C1 (astronomy & astrophysics, geosciences)**

  The majority of the studies in this cluster are about physics and astronomy. There are also studies on geology and archaeology.

- **C2 (urology & nephrology, pediatrics)**

  Urology & nephrology and pediatrics are the most popular subjects in this cluster.

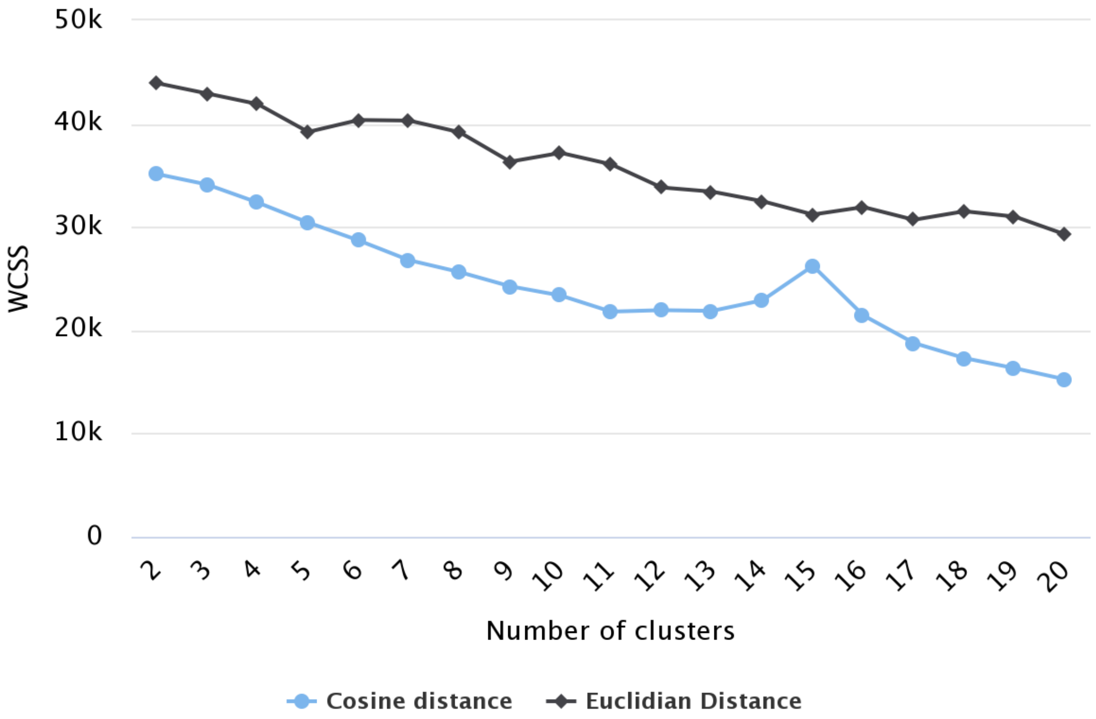


Figure 3.8. WCSS with using euclidian and cosine distance.

Urology & nephrology and pediatrics are close to each other because of the effective publications of 3 researchers working in pediatric nephrology department.

- **C3 (biochemistry & molecular biology)**

  In this cluster, studies on biology, biochemistry and cell biology are dominant.

- **C4 (education & educational research)**

  This cluster includes studies on social sciences. It has been observed that gerontology is a common area with both social sciences and medical sciences.

- **C5 (nursing)**

  The main topics are related to nursing, oncology and management. There are publications on the care of cancer patients and hospital management.

- **C6 (veterinary sciences)**

  This section includes studies on animals, insects and fish. Emergency medicine, which is a sub-field of medical sciences, is also included in this cluster because of animal origin injuries.

- **C7 (surgery)**

  Surgical studies in the field of medicine are included in this cluster. Due to its close relationship with Pathology and Medical Imaging they are grouped together.

- **C8 (mathematics, dermatology)**

  This was the most surprising cluster. However, at Akdeniz, there were enough number of people doing cross-field research in both fields.

- **C9 (materials sciences)**

  This cluster is entirely based on materials science and its footsteps in both mechanical and civil engineering.

- **C10 (obstetrics & gynecology)**

  As a subset of medical sciences concentrated on gynecology and women health.

- **C11 (food science & technology, plant sciences)**

  Food and plant sciences were under the life sciences category according to WoS. As school of Agronomy has a large footprint at Akdeniz, there are separated and clustered together.

Table 3.7. Clusters by hiearchical clustering

Clusters	Research Areas	Departments
C1 (82)	Astronomy & Astrophysics (199) Geosciences, Multidisciplinary (166) Physics, Nuclear (109) Physics, Multidisciplinary(87) Environmental Sciences(80)	Faculty of Science (19) Faculty of Engineering (18) Faculty of Literature (13) Faculty of Economis and Administritive Sciences (5) Teknik Bilimler (4)
C2 (37)	Urology & Nephrology (258) Pediatrics (257) Rheumatology (146) Infectious Diseases (96) Genetics & Heredity(85)	Faculty of Medicine (29) Faculty of Health Sciences(2) Faculty of Sport Sciences (2) Faculty of Science (1) Faculty of Dentistry (1)
C3 (70)	Biochemistry & Molecular Biology (415) Endocrinology & Metabolism (278) Medicine, Research & Experimental (256) Cardiac & Cardiovascular Systems (227) Neurosciences (213)	Faculty of Medicine (42) Graduate School of Health Sciences (8) Faculty of Science (7) Faculty of Engineering (3) Faculty of Health Sciences (2)
C4 (44)	Education & Educational Research (147) Social Sciences, Interdisciplinary (26) Sport Sciences (12) Linguistics (10) Computer Sciences, Interdisciplinary (10)	Faculty of Education (14) Faculty of Sport Sciences (9) Faculty of Economics and Administrative Sciences (6) Faculty of Applied Sciences (5) Faculty of Literature (4)
C5 (50)	Nursing (119) Hospitality, Leisure, Sport & Tourism (75) Management (61) Oncology (42) Medicine, General & Internal (31)	Faculty of Nursing (16) Faculty of Economics and Administrative Sciences (8) Faculty of Tourism (5) Faculty of Applied Sciences (4) Faculty of Medicine (3)

Contunation of **Table 3.7.**

C6 (29)	Veterinary Sciences (128) Agriculture, Dairy & Animal Science (76) Fisheries (46) Entomology (28)	Faculty of Agriculture (11) Faculty of Aquaculture (10) Faculty of Medicine (2) Graduate School of Natural and Applied Sciences (2)
C7 (83)	Surgery (417) Radiology, Nuclear Medicine & Medical Imaging (280) Urology & Nephrology (261) Clinical Neurology (222) Transplantation (213)	Faculty of Medicine (56) Faculty of Dentistry (12) Faculty of Literature (3) Faculty of Engineering (3) Faculty of Science (3)
C8 (53)	Mathematics, Applied (280) Dermatology (237) Mathematics (221) Engineering, Electrical & Electronic (169)	Faculty of Science (19) Faculty of Engineering (14) Faculty of Medicine (6)
C9 (31)	Materials Science, Multidisciplinary (109) Engineering, Multidisciplinary (84) Engineering, Mechanical (77) Engineering, Civil (76) Mechanics (58)	Faculty of Engineering (20)
C10 (36)	Obstetrics & Gynecology (486) Reproductive Biology (290) Ophthalmology (140) Pathology (137) Cell Biology (125)	Faculty of Medicine (34) Faculty of Engineering(1) Faculty of Nursing (1)
C11 (136)	Food Science & Technology (496) Plant Sciences (449) Agronomy (320) Environmental Sciences (237) Biotechnology & Applied Microbiology (205)	Faculty of Agriculture (55) Faculty of Engineering(27) Faculty of Science (16)

4. RESULTS AND DISCUSSION

The main purpose of this study is to evaluate the academic publication performance data of a university. To achieve this objective, some indicators have been determined and presented. Distribution of academic titles per academic unit may be shown in Fig. 4.9. Young universities and departments tend to have higher ratio of Assistant Professors then Professors and Associate Professors. As academic units age, the ratio of Professors increase significantly which in turn may impact academic output performance.

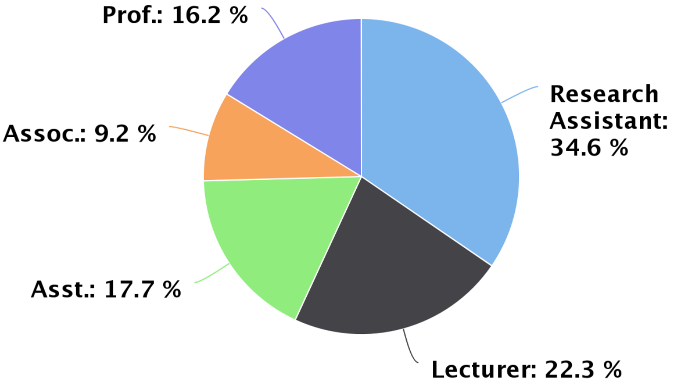


Figure 4.9. Distribution of titles of academic staff at Akdeniz Univesity

Fig. 4.10 show yearly publication performance of any selected unit. Such information may be used to follow trends over time and in cases it drops below a certain control limit for a period of time then cause and effect may be investigated for improvement.

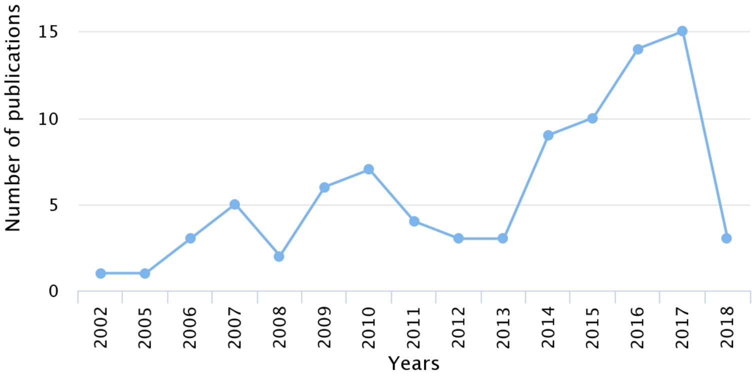


Figure 4.10. Publication count by year of an academic unit

Fig. 4.11 shows the number of publications index at various databases in a pie chart for academic units. It may be expected that while the publications for an Engineering Faculty appear primarily in SCI and ISTP, in Social Sciences fields, the large chunk of publications is expected to appear in AHCI and SSCI. By comparing the ratio between SCI to ISTP to SSCI and AHCI, relative performances of academic units within a university may be obtained.

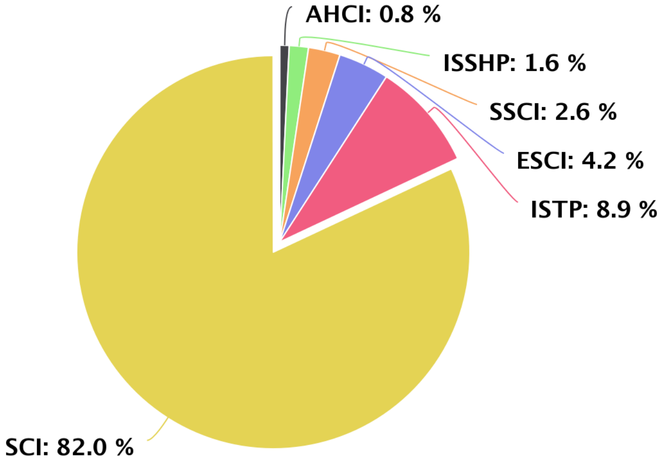


Figure 4.11. Distribution by indexes of publications at Akdeniz University

In Fig. 4.12, for the whole university and each school, academic department and staff we plot:

- Number of citations
- Number of publications
- Average citation per publication
- Average citation per academic staff
- Average publication per academic staff

As shown in Fig. 4.13, Word Cloud has been generated for research areas for the whole university, school, department and academic staff using the corresponding frequencies of publications. However as some publications have higher citations than others, the citation count was used as a multiplier in keyword frequency count in Word Cloud.

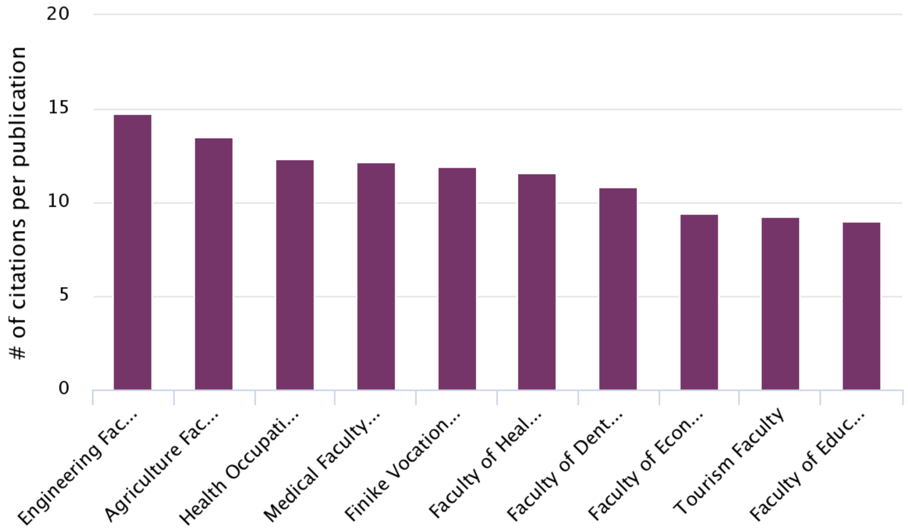


Figure 4.12. Number of citations per publication for all academic units

Such word cloud shown in Fig. 4.13 helps administrators to understand quickly what the academic unit focuses and produce strategies to effectively use resources.

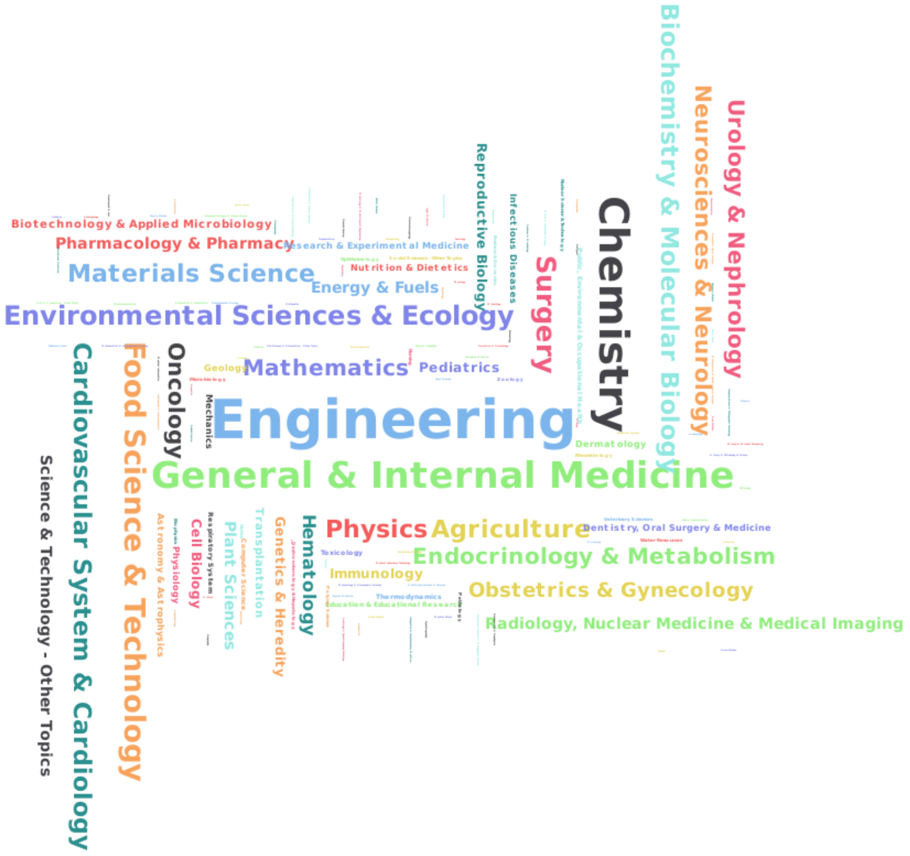


Figure 4.13. Word cloud of research areas

Quality of a publication may be assessed by JIF quartile value. JIF quartile distribution of publications for an academic unit or staff as shown in Fig. 4.14 and Fig. 4.15 will indicate the quality and impact of research that is carried out. Ideally, it is desired to have most of the publications appear at Q1 and Q2 journals for a given research area.

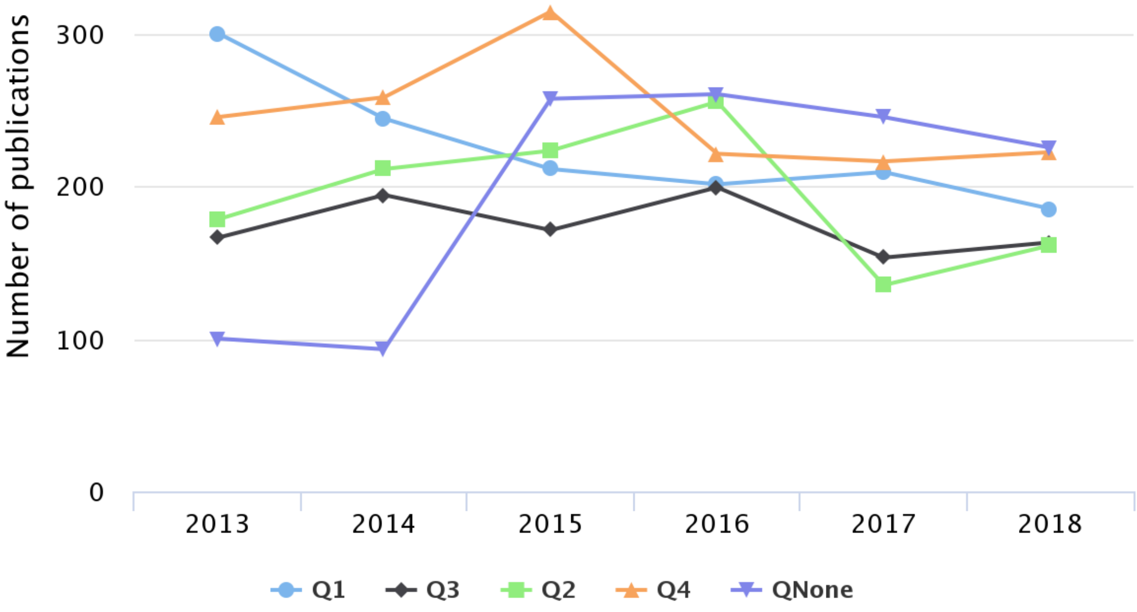


Figure 4.14. Q values chart for departments

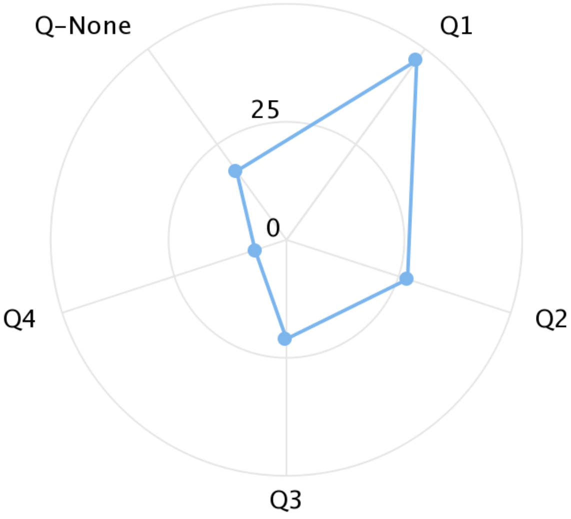


Figure 4.15. Q values chart for researchers

Performance plots obtained for the university as a whole can also be retrieved easily for each faculty and department by simply selecting the unit from the tree of academic organization as shown in Fig. 4.16.

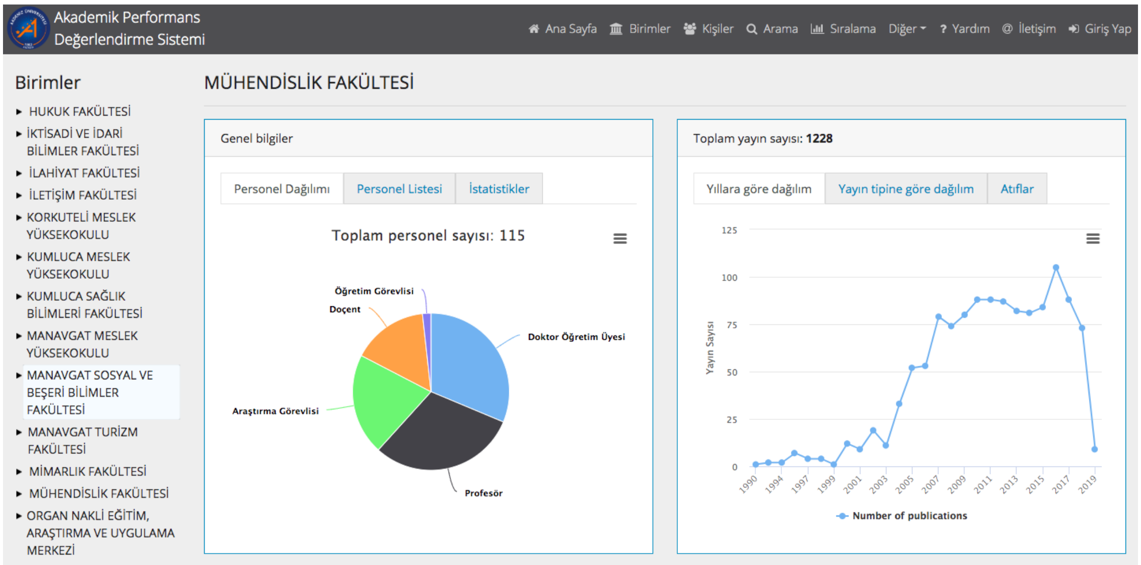


Figure 4.16. Department-based filtering of academic performance criteria

Network graphs were prepared for departments and research fields by examining the joint publications of the researchers. The colors of the nodes show the research areas and the thickness of the connection between the nodes shows strength of collaboration as shown in Fig. 4.17.

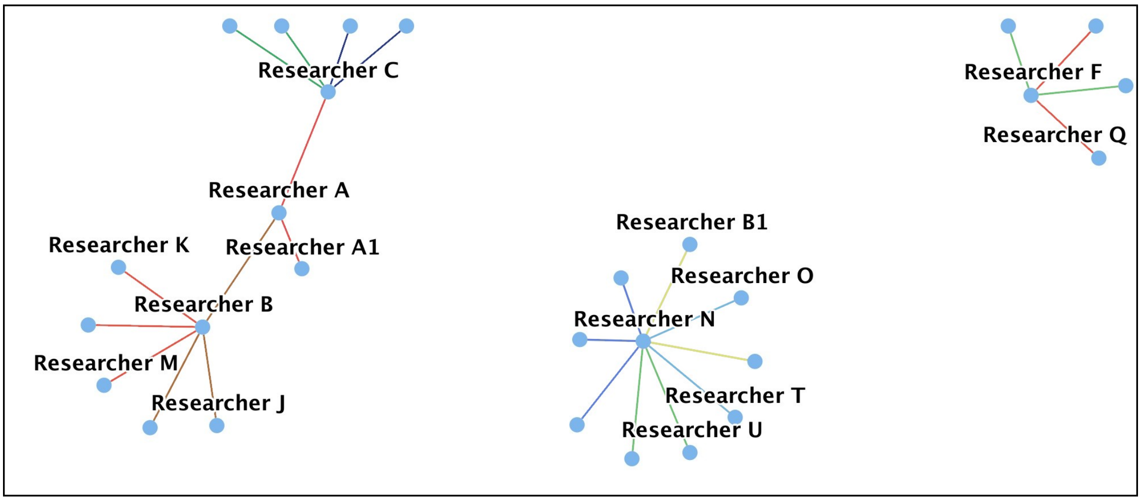
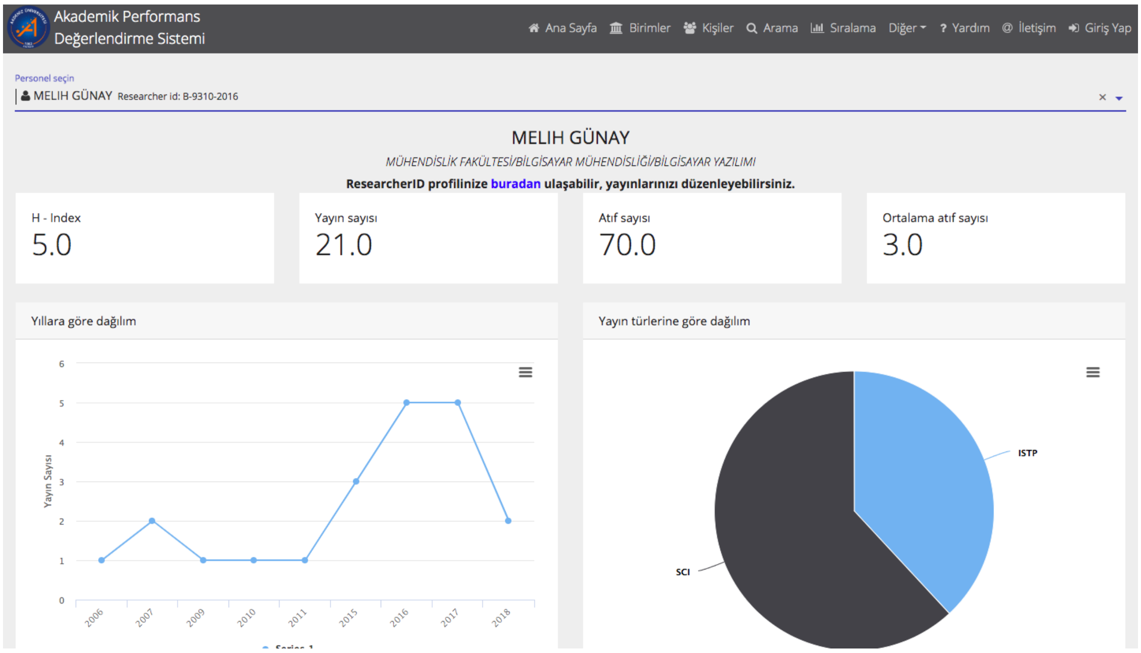


Figure 4.17. Researcher network graph

Performance metrics that were reported for unit is customized to view the publication performance of an individual researcher in a separate module as shown in Fig. 4.18. In this page, mostly cited publications, H-Index of the researcher, Publication Count and Total Number of Citations of publications are shown. The publication indexed database distribution is also plotted in a pie chart. Yearly publication count, citation per publication and their averages are also plotted per researcher within the staff performance page.



Personel seçin
MELIH GÜNAY Researcher id: B-9310-2016

MELIH GÜNAY

MÜHENDİSLİK FAKÜLTESİ/BİLGİSAYAR MÜHENDİSLİĞİ/BİLGİSAYAR YAZILIMI

ResearcherID profilinize buradan ulaşabilir, yayınlarınızı düzenleyebilirsiniz.

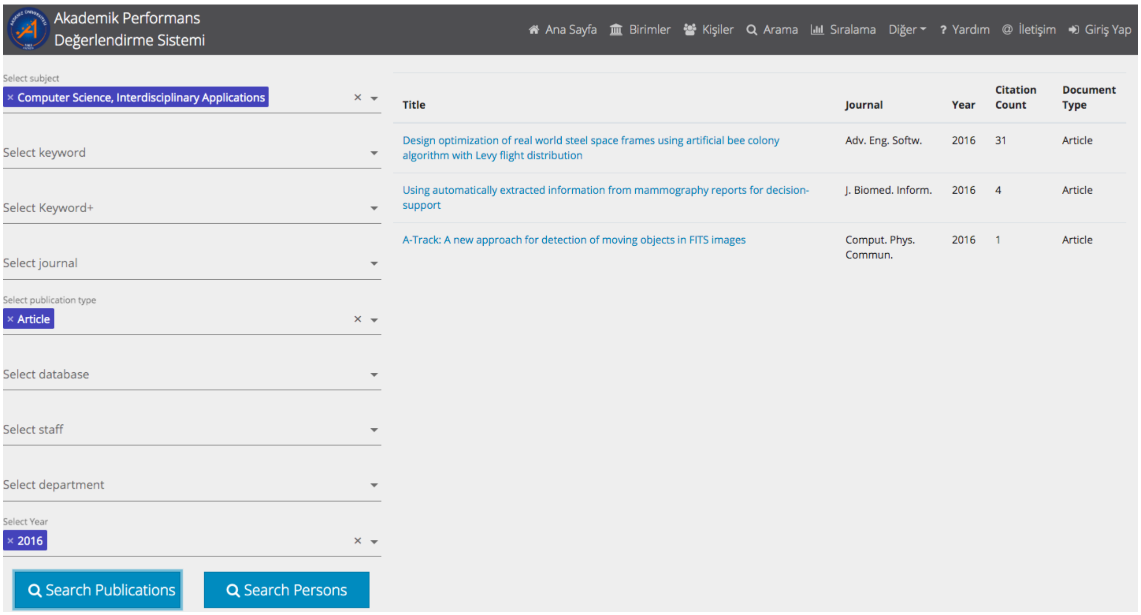
H - Index	Yayın sayısı	Atıf sayısı	Ortalama atıf sayısı
5.0	21.0	70.0	3.0

Figure 4.18. Academic staff performance profile page

Custom search page has been designed for users to search publications, researchers and departments (Fig. 4.19). This page allows researchers and publications to be filtered by keywords, subjects, researchers, departments, years, journals, database, publication types and q values. The result set returned may be sorted by the user by selecting the table column header.

In addition, for whole university or each academic unit, a ranking list that includes bibliometric indicators is tabulated for the followings items:

- H-Indexes, M-Indexes, G-Indexes of Academic Researchers
- Citation Count of Academic Researchers
- Most frequent Keywords of publications
- Mostly cited publications of researchers



Select subject: Computer Science, Interdisciplinary Applications

Select keyword

Select Keyword+

Select journal

Select publication type: Article

Select database

Select staff

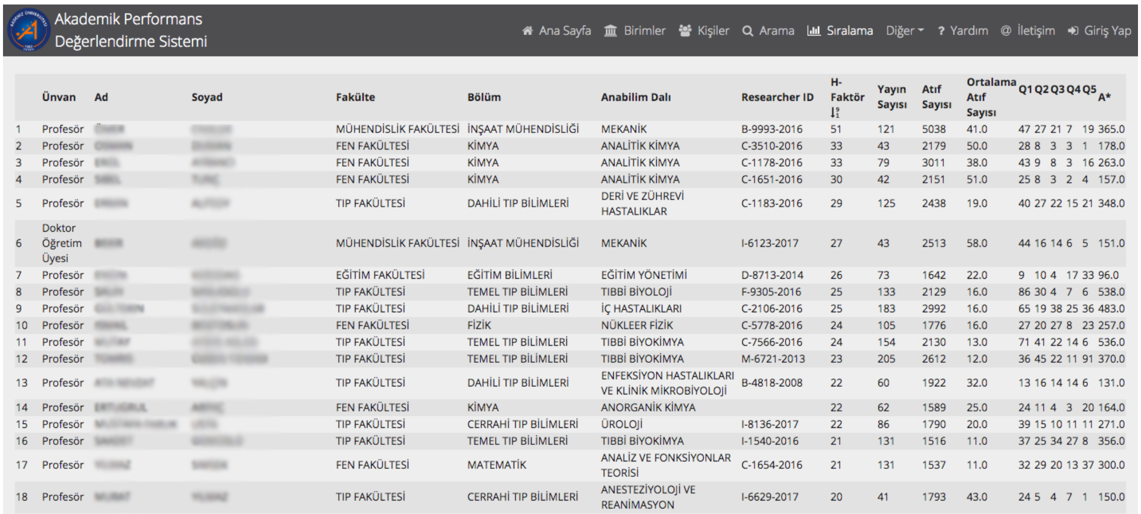
Select department

Select Year: 2016

Title	Journal	Year	Citation Count	Document Type
Design optimization of real world steel space frames using artificial bee colony algorithm with Levy flight distribution	Adv. Eng. Softw.	2016	31	Article
Using automatically extracted information from mammography reports for decision-support	J. Biomed. Inform.	2016	4	Article
A-Track: A new approach for detection of moving objects in FITS images	Comput. Phys. Commun.	2016	1	Article

Figure 4.19. Custom search page

Knowing high performing academic researchers, publications and research areas, university administrator may support people and research topics where the university is more effective.



	Ünvan	Ad	Soyad	Fakülte	Bölüm	Anabilim Dalı	Researcher ID	H-Faktör ↓	Yayın Sayısı	Atıf Sayısı	Ortalama Atıf Sayısı	Q1	Q2	Q3	Q4	Q5	A*
1	Profesör	[redacted]	[redacted]	MÜHENDİSLİK FAKÜLTESİ	İNŞAAT MÜHENDİSLİĞİ	MEKANİK	B-9993-2016	51	121	5038	41.0	47	27	21	7	19	365.0
2	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	KİMYA	ANALİTİK KİMYA	C-3510-2016	33	43	2179	50.0	28	8	3	3	1	178.0
3	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	KİMYA	ANALİTİK KİMYA	C-1178-2016	33	79	3011	38.0	43	9	8	3	16	263.0
4	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	KİMYA	ANALİTİK KİMYA	C-1651-2016	30	42	2151	51.0	25	8	3	2	4	157.0
5	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	DAHİLİ TIP BİLİMLERİ	DERİ VE ZÜHREVİ HASTALIKLAR	C-1183-2016	29	125	2438	19.0	40	27	22	15	21	348.0
6	Doktor Öğretim Üyesi	[redacted]	[redacted]	MÜHENDİSLİK FAKÜLTESİ	İNŞAAT MÜHENDİSLİĞİ	MEKANİK	I-6123-2017	27	43	2513	58.0	44	16	14	6	5	151.0
7	Profesör	[redacted]	[redacted]	EĞİTİM FAKÜLTESİ	EĞİTİM BİLİMLERİ	EĞİTİM YÖNETİMİ	D-8713-2014	26	73	1642	22.0	9	10	4	17	33	96.0
8	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	TEMEL TIP BİLİMLERİ	TIBBİ BİYOLOJİ	F-9305-2016	25	133	2129	16.0	86	30	4	7	6	538.0
9	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	DAHİLİ TIP BİLİMLERİ	İÇ HASTALIKLARI	C-2106-2016	25	183	2992	16.0	65	19	38	25	36	483.0
10	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	FİZİK	NÜKLEER FİZİK	C-5778-2016	24	105	1776	16.0	27	20	27	8	23	257.0
11	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	TEMEL TIP BİLİMLERİ	TIBBİ BİYOKİMYA	C-7566-2016	24	154	2130	13.0	71	41	22	14	6	536.0
12	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	TEMEL TIP BİLİMLERİ	TIBBİ BİYOKİMYA	M-6721-2013	23	205	2612	12.0	36	45	22	11	91	370.0
13	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	DAHİLİ TIP BİLİMLERİ	ENFEKSİYON HASTALIKLARI VE KLİNİK MİKROBİYOLOJİ	B-4818-2008	22	60	1922	32.0	13	16	14	14	6	131.0
14	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	KİMYA	ANORGANİK KİMYA		22	62	1589	25.0	24	11	4	3	20	164.0
15	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	CERRAHİ TIP BİLİMLERİ	ÜROLOJİ	I-8136-2017	22	86	1790	20.0	39	15	10	11	11	271.0
16	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	TEMEL TIP BİLİMLERİ	TIBBİ BİYOKİMYA	I-1540-2016	21	131	1516	11.0	37	25	34	27	8	356.0
17	Profesör	[redacted]	[redacted]	FEN FAKÜLTESİ	MATEMATİK	ANALİZ VE FONKSİYONLAR TEORİSİ	C-1654-2016	21	131	1537	11.0	32	29	20	13	37	300.0
18	Profesör	[redacted]	[redacted]	TIP FAKÜLTESİ	CERRAHİ TIP BİLİMLERİ	ANESTEZİYOLOJİ VE REANİMASYON	I-6629-2017	20	41	1793	43.0	24	5	4	7	1	150.0

Figure 4.20. Ranking list

There are a few other software tools used for research evaluation. InCites Benchmarking & Analytics, the most known of these tools, is a customized, web-based research evaluation tool. InCites B&A allows user to analyze institutional productivity, monitor collaboration activity, identify influential researchers, showcase strengths, and discover areas of opportunity.

The differences and common points of this study with InCites are shown in Table 4.8.

Table 4.8. Comparison of developed software and InCites

	Developed Software	InCites
Ranking by indicators	✓	X
Reporting	✓	✓
Department-based filtering within the university	✓	X
Advanced Filtering	✓	✓
Source Data	Web of Science	Web of Science
Fee	Free (Open Source)	Paid

5. CONCLUSION

With this study, an original software for the evaluation of academic publication performance was implemented for the Higher Education Institutes. Research performance for an academic unit or staff may be accessible in real time. Since the system is integrated with the Web of Science, processed data is reliable.

Academic Performance Evaluation System includes a number of features such as:

- Determination of the contribution of a department to the university
- Determination of the contribution of the researchers to the their departments
- Finding specialized researchers, departments, or publications in a specific research area
- Finding the potential collaborators for research within the system
- Determination of interdisciplinary studies
- Determination of studies with international cooperation
- Obtaining research performance by subject area of academic units and researchers in 5 broad categories namely; Life Science and Biomedicine, Art and Humanities, Physical Sciences, Social Sciences, Engineering and Technology.
- Ranking of researchers using indicators such as H-index, g-index, m-index, number of publications and number of citations.

As publication performance is a good indicator of quality of research and impact, it may be used to determine the effectiveness of academicians and universities. Based on the publication performance, universities may develop an objective method for promotions, appointments and resource allocations. Also, higher education councils of governments may use such data to develop policy and implement a publication-based incentive system for promotion of scientific research. If such academic data made public by the university, industry, research centers, academics and students may find partners for research in desired topics of interest. The software proposed here addresses a significant need.

6. REFERENCES

Adachi, T. and Kongo, T. 2015. Further axiomatizations of Egghe's g-index. *Journal of Informetrics*, 9 (4): 839-844.

Adriaanse, L. S. and Rensleigh, C. 2013. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31(6): 727-744.

Alaşehir, O., Çakır, M.P., Acartürk, C., Baykal N. and Akbulut, U. 2014. URAP-TR: A national ranking for Turkish universities based on academic performance. *Scientometrics*, 74(2): 257-271.

Bar-Ilan, J. 2008. Which h-index? – A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2): 257-271.

Bholowalia, P. and Kumar, A. 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, 105 (9): 17-24.

Chakrabarti, S. Ester, M. Fayyad, U. Gehrke, J. Han, J. Morishita, S. Piatetsky-Shapiro, G and Wang, W. 2006. Data mining curriculum: A proposal (Version 1.0). *Intensive Working Group of ACM SIGKDD Curriculum Committee*. 140: 1-10.

Clarivate Analytics: https://clarivate.com/products/web-of-science [Last access date: 2019-04-30].

Curado, C., Henriques, P. L., Oliveira, M. and Matos, P. V. 2016. A fuzzy-set analysis of hard and soft sciences publication performance. *Journal of Business Research*, 69 (11), 5348-5353.

Egghe, L. 2006. Theory and practice of the g-index. *Scientometrics*, 69(1): 131–152.

Essential Science Indicators Journal Category Scope Notes: http://help.incites.clarivate.com/inCites2Live/8300-TRS.html [Last access date: 2019-04-30].

Garfield, E. 2016. The History and Meaning of the Journal Impact Factor. *JAMA*, 295 (1), 90-93.

Han, J., Pei, J. and Kamber, M. 2011. Data mining: concepts and techniques. Elsevier, Waltham, 703 p.

Hand D.J. 2006. Data Mining. *Encyclopedia of Environmetrics*, 2 (1), 99-115.

Hemalata, K. and Vasanthakumari, G. 2013. Implementation of Object Oriented Approach To Sequential Pattern Mining From Multidimensional Sequence Data. *International journal of modern engineering research*, 1 (1): 84-89.

Highsoft A.S. https://www.highcharts.com, [Last access date: 2019-04-30].

Hirsch, J. E. 2005. An Index to Quantify an Individual's Scientific Research Output. Proceedings of the National Academy of Sciences of the United States of America 102.46, pp. 16569–16572. 27 March, PMC, Web.

Hirsch, J.E. 2007. Does the h index have predictive power?. *PNAS*, 104 (49): 19193-19198.

Huang, A. 2008. Similarity measures for text document clustering. Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), pp. 9-56, 14 April, Christchurch, New Zealand.

Jin, B., Liang, L., Rousseau, R., and Egghe, L. 2007. The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52 (6): 855–863.

Kaptanoglu, D. and Özok, A.F. 2010. Akademik performans değerlendirmesi için bir bulanık model. *İTÜDERGİSİ*, 5 (1): 193-204.

Kongo, T. 2014. An alternative axiomatization of the Hirsch index. *Journal of Informetrics*, 8 (1): 252-258.

Larsen, P., Von Ins, M. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84 (3): 575-603.

Leslie, S.A and Rensleigh, C. 2013. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31 (6): 727-744.

Özkan, Y. 2008. Veri madenciliği yöntemleri. Papatya Yayıncılık Eğitim, İstanbul, 240 p.

Panczyk, M., Woynarowska-Soldan, M., Belowska, J., Zarzeka and A., Gotlib, J. 2015. Bibliometric Evaluation of Scientific Literature in the Area of Research in Education Using Incites Database of Thomson Reuters. Proceedings of INTED2015 Conference, pp. 487-496, 2-4 March, Madrid, Spain.

Pathel, V.M., Ashrafian, H., Almoudaris, A., Makanjuola, J., Bucciarelli-Ducci, C., Darzi, A. and Athanasiou, T. 2012. Measuring Academic Performance for Healthcare Researchers with the H Index: Which Search Tool Should Be Used?. *Medical Principle and Practice*, 22 (2): 178-183.

Porter, A. and Rafols, I. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81 (3): 719-745.

Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 60 (5): 503-520.

Rørstad, K. and Aksnes, D.W. 2015. Publication rate expressed by age, gender and academic position–A large-scale analysis of Norwegian academic staff. *Journal of Informetrics*, 9 (2): 317-333.

Rosenstreich, D. and Wooliscroft, B. 2009. Measuring the impact of accounting journals using Google Scholar and the g-index. *The British Accounting Review*, 41 (4): 227-239.

Sahu, L. and Mohan, B.R. 2014. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop. 2014 9th International Conference on Industrial and Information Systems (ICIIS), pp. 1-5, 15-17 December, Gwalior, India.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24 (5): 513-523.

Scientific and Technological Research Council of Turkey, http://tubitak.gov.tr/ [Last access date: 2019-04-30].

Silahtaroğlu, G. 2008. Veri madenciliği. Papatya Yayınları, İstanbul, 304.

Soutar, G.N., Wilkinson, I. and Young, L. 2015. Research performance of marketing academics and departments: An international comparison. *Australasian Marketing Journal*, 23 (2): 155-161.

Thada, V. and Jaglan, V. 2013. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2 (4): 202-205.

Umut, A. L. 2008. Bilimsel yayınların değerlendirilmesi: h-endeksi ve Türkiye'nin performansı. Bilgi Dünyası, 9 (2): 263-285.

University Ranking by Academic Performance: http://www.urapcenter.org [Last access date: 2019-04-30].

Web of Science: https://clarivate.com/products/web-of-science [Last access date: 2019-04-30].

Web of Science Research Areas: https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html [Last access date: 2019-04-30].

Zhang, C.T. 2009. The e-Index, Complementing the h-Index for Excess Citations. *PlosOne*, 4 (5): 1-4.

CURRICULUM VITAE

TAHA YİĞİT ALKAN

tahayigitalkan@windowslive.com

EDUCATION

Master of Science 2016-2019	Akdeniz University Institute of Natural and Applied Sciences, Department of Computer Engineering, Antalya
Bachelor of Science 2011-2016	Süleyman Demirel University Faculty of Engineering, Department of Computer Engineering, Isparta

WORK EXPERIENCE

Research Assistant 2018-	Akdeniz University Institute of Natural and Applied Sciences, Department of Computer Engineering, Antalya
Software Developer 2016-2018	Süleyman Demirel University Department of Information Technologies, Isparta

PUBLICATIONS

Papers delivered in International Conferences and Printed as a Proceedings

1- Alkan, T.Y., Özbek, F. and Günay, M. (2019). Evaluation of researcher performance of academic units at Akdeniz University. Conference on Artificial Intelligence and Applied Mathematics 2019.