University of Thessaly
Department of Electrical and Computer Engineering

Content-based Scientometrics

by Maria Markou

October 2015

CONTENT BASED SCIENTOMETRICS

A thesis submitted

by

Maria Markou

to

the Department of Electrical and Computer Engineering

University of Thessaly

in partial fulfillment of the requirements

for the degree of Master of Science

October 2015

Supervising Committee:

Manolis Vavalis, Professor

Dimitris Katsaros, Assistant Professor

Panayiotis Bozanis, Professor

Acknowledgments

I wish to express my most sincere indebtedness to the following for making the completion of the present thesis possible:

First and foremost I would like to express my gratefulness to my supervisor Manolis Vavalis for his support and valuable comments during the difficult task of conducting the research. His patience and support helped me overcome many crisis situations and finish this thesis.

I would like also to express my gratitude to Dimitris Katsaros, one of my supervising committee members, for providing me useful material for my thesis.

I owe a very important debt to the company Egritos Group – Synergasia for giving me the space to study during the working hours.

Special thanks to my closest friends Maria Zafiri and Anastasia Kaltsogianni, my English teacher, for their support and for convincing me to undertake this challenge, during the midday sessions in Grappa. I would also like to thank my classmates Argyris Varalis and Vaggelis Katsigiannakis, who have accompanied me throughout the postgraduate courses and have turned studying into fun.

Last but not least I would like to express my deepest thanks to my parents for supporting me spiritually throughout writing this thesis and to my brother, Miltos Markou, for the chill out drinks and cocktails.

Acronyms

CCA	Combined Credit Allocation
CODEN	a code classification assigned to a document or other library item consisting typically of four capital letters followed by two hyphenated groups of arabic numerals
DB	Database
DOI	Digital Object Identifier
EID	Electronic ID
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
ORCID	Open Researcher & Contributor ID
OS	Operating System
SQL	Structured Query Language
VSM	Vector Space Model
WOS	Web of Science

List of Contents

List of Figures

List of Tables

List of Equations

ABSTRACT

Accounting authorship of a scientific paper is a widely recognized as a hard problem. Attempts to solve this problem with existing conventional tools encounter insurmountable obstacles. Along these lines Nature began in 2010 (Assessing assessment, 2010) an ongoing conversation concerning the metrics to measure and assess scientific performance. This effort is not only still running but also created additional momentum. In particular, it is now apparent that the use of metrics to assess the value of scientists is unavoidable. So the quest for the best measure possible is surely justified (Count on me, 2012).

In (Nanas, Vavalis, & Houstis, 2010) novice authorship taxonomies have been proposed (Taylor & Thorisson, 2012) that ensure the clear and unambiguous declarations of authorship while heretic arguments like the one claiming that ambiguity is not entirely a bad thing in science (Zuckerman, 1968) have been also appeared in the literature from very early.

It has therefore become evident that the current scheme employed in scientometrics appears to be most probably problematic and perhaps unfair. Within this context this research aims to assess authors' participation in the recorded research activity through developing alternative assessment ways. Instead of using common quantitative metrics, the present study proposes and utilizes the developing of multi-faced-dynamic author profiles. Furthermore, Data Mining and Knowledge Management will compose an effective mechanism to support the theoretical background, the practical significance as well as the intended methodology. The design, the development and the evaluation of a software tool will also contribute to the application and evaluation of the designed author profiles and to the reliability of the obtained results.

ΠΕΡΙΛΗΨΗ

Η πιστοποίηση της συμμετοχής (πατρότητας) του συγγραφέα μιας επιστημονικής δημοσίευσης είναι ένα ευρέως διαδεδομένο και δυσεπίλυτο πρόβλημα. Οι προσπάθειες που έχουν πραγματοποιηθεί με συμβατικές μεθόδους να δώσουν λύση στο πρόβλημα αυτό, αντιμετωπίζουν ως τώρα ανυπέρβλητα εμπόδια. Το ζήτημα αυτό ανακινήθηκε ξανά το 2010 σε ένα άρθρο του γνωστού περιοδικού Nature (Assessing assessment, 2010), το οποίο αφορούσε τη μέτρηση και την αξιολόγηση των επιστημονικών επιδόσεων. Η συνεχιζόμενη αυτή προσπάθεια ωθεί την επιστημονική κοινότητα στην αναζήτηση του καλύτερου δυνατού βιβλιομετρικού δείκτη (Count on me, 2012) για την εκτίμηση της επιστημονικής αξίας.

Στη δημοσίευση (Nanas, Vavalis, & Houstis, 2010) υπήρξαν τα πρώτα βήματα (Taylor & Thorisson, 2012) για την αξιολόγηση της συμμετοχής των συγγραφέων σε επιστημονικές δημοσιεύσεις, τα οποία εξάγουν σαφή αποτελέσματα. Ενώ πρώιμες δημοσιεύσεις, όπως του (Zuckerman, 1968), ισχυρίζονται ότι η επιστημονική αμφισημία δεν είναι απαραίτητα κάτι κακό.

Ως εκ τούτου, έχει καταστεί προφανές ότι η τρέχουσα προσέγγιση που χρησιμοποιείται στην επιστημομετρία είναι κάποιες φόρες προβληματική και άδικη. Στο πλαίσιο αυτό, η παρούσα έρευνα έχει ως στόχο να αξιολογήσει τη συμμετοχή των συγγραφέων στην ερευνητική τους δραστηριότητα, μέσω της ανάπτυξης εναλλακτικών τρόπων αξιολόγησης. Η παρούσα μελέτη προτείνει την ανάπτυξη δυναμικών πολύπλευρων συγγραφικών προφίλ αντί της χρήσης κοινών ποσοτικών δεικτών. Επιπλέον, η εξόρυξη δεδομένων και η διαχείρισης γνώσης θα συνθέσουν έναν αποτελεσματικό μηχανισμό για να θεμελιώσουν το θεωρητικό υπόβαθρο, την πρακτική σημασία, καθώς και την προβλεπόμενη μεθοδολογία. Ο σχεδιασμός και η ανάπτυξη ενός λογισμικού θα συνεισφέρει στην εφαρμογή και την αξιολόγηση των δυναμικών συγγραφικών προφίλ, καθώς και στην αξιοπιστία των αποτελεσμάτων που θα επιφέρει.

1

1. INTODUCTION

Different countries and different branches produce an amount of research output which is currently evaluated by means of citation-based metrics. Indicators such as the number of papers and the number of citations determine the pattern according to which credits are allocated to co-authors of a multi-author paper.

Yet an author of a single-author paper receives the same credits as the contributors of multi-author papers rendering thus pattern of evaluation highly unjust and discriminatory. Citation-based metrics evaluate contributors as if they are the single authors of the full article. In this way, contributor gets the full impact factor score and all the citations received by this article.

To make matters worse, another indicator, that of honorific authorship, adds more injustice to the existing evaluation scheme. What is implied by the practice of honorific authorship is the granting of a byline of co-authors for purely social and political reasons. In this way, contributors with minimal involvement in the final product of work, receive the same credits as the sole conceiver, fabricator and owner of the published article.

Another issue to be taken into account is the degree to which each author is active in producing scientific work without exhibiting long pauses of inertia. Throughout a researcher's academic life, periods of low productivity ought to be considered in rendering

metrics less unjust. Thus we could create an activity/inactivity-based index to measure the time periods of high versus low productivity of each author.

Given the fact that advances in almost all scientific fields are extremely rapid it is often the case that authors may present inter-disciplinary mobility. In other words, an area that has received merit and prominent focus in the past may now appear to be outdated and of low scientific interest attracting thus minimum research. In this way researchers, in an attempt to expand an upgrade their thematic fields, may result to abandoning certain thematic areas in favor of other, trendier areas. Being able to detect the thematic field that each researcher in every institution deals with each time could provide the opportunity to introduce and enhance collaboration among researchers and institutions on that particular field.

1.1 Thesis outline and contributions

The main innovation in our study concerns the development and operational use of the author's profile. Initially, for the content-based analysis the authors' profiles and the testing paper are represented as bag-of-words models. Comparative analysis and in particular the cosine similarity coefficient is performed for fingering out the author profile that matches best a particular given paper.

In order to deal with the complexity of the overall problem that poses several vital challenges including the curse of dimensionality we will not rely only on the conventional vector-based similarity measures commonly used in Information Retrieval. We will utilize an innovative graph based structure which will be evolve on the basis of an effective bio-inspired method. This method has been already proved itself (Nanas, Vavalis, & Houstis, 2010) as a very effective tool (Markou, 2015) for filtering preferences and locating similarities in a framework similar to our multi-authored case.

1.2 Background Theory

Understanding the definitions of the terms that appear in the particular work, is of outmost importance in order to achieve a clear understanding of this thesis. Thus meanings will be provided for the following terms: bag-of-words, Vector space model, Cosine similarity and Sliding Window.

1.2.1 Bag-of-words

The bag-of-words is a common way to represent documents in matrix form as a multiset of its words. Word ordering within a document is not taken into account, instead multiplicity plays an important role (Salton & McGill, 1983).

In Information Retrieval, the bag-of-words model is used to estimate the semantic association between two documents or a document and a query by representing them as bags of words. The word frequency in documents represents the relevance of the document to a query and thus the meaning of the document can be assumed (Turney & Pantel, 2010). The bag-of-words hypothesis is the basis for applying the Vector Space Model to Information Retrieval (Salton, Wong, & Yang, 1975). The effective practical use of bag-of-words based approaches in general, but mostly for challenging problems like the one considered in our study has been investigated and alternatives have been proposed (Nanas & Vavalis, 2008)

1.2.2 Vector space model

The vector space model is a simple mathematical model that is used to represent queries and documents by a set of terms, giving the possibility to compute global similarities between them. Queries and documents are represented as vector of terms in the form of (Salton, Wong, & Yang, 1975):

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

Equation 1: Queries and documents as vector representation

where:

$w_{t,q}$ and $w_{t,j}$ represents the value of term *t*

d is the document

q is the query

To compute the similarity between the aforementioned vectors, the following similarity measures can be used: inner product, dice coefficient, cosine coefficient and Jaccard coefficient (Salton, 1989).

1.2.3 Cosine similarity

The most common way to measure similarity within the vector space model is to use the cosine coefficient, which measures cosine of the angle between two vectors in the vector space. The cosine of the angle θ between tow vectors $x=<x_1, x_2, \ldots, x_n>$ and $y=<y_1, y_2, \ldots, y_n>$ is the inner product of the vector, after they have been normalized to unit length and is calculated as follows:

$$\cos(x, y) = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|}$$

Image 1: Cosine Similarity

The cosine ranges from −1 when the vectors point in opposite directions to +1. Although the cosine similarity of two documents will range from 0 to 1, since the term frequencies cannot be negative.

1.2.4 Sliding Window

The sliding window (Nanas & Vavalis, 2008) approach is used to identify term dependencies within a document. A window is a span of contiguous words in a document's text and its size is an important parameter that defines the kind of term correlations. A small window of typically no more than three words, is called "local context" and is appropriate for identifying adjacent, syntactic correlations between terms, such as compounds. "Global context" on the other hand, is defined by a larger window (more than three terms) that may incorporate several sentences, or even the complete document.

2

2. LITERATURE REVIEW

Questions such as “What is a substantial contribution to a scientific paper?” and “How much credit should be entitled to an author for chasing the idea, collecting the data or managing the communication among the co-authors?” are confronted, when the issue of authorship is discussed. According to the International Committee of Medical Journal Editors an author is a person who contributes to each of the following steps (Defining the Role of Authors and Contributors, 2014):

1. has substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
2. draft the work or revise it critically for important intellectual content; AND
3. gives final approval of the version to be published; AND
4. agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

According to the above rules it is almost impossible to determine fairly the co-author’s contribution. For this reason various indicators and metrics are developed to measure scientific quality, impact or prestige.

The popular h-index (Hirsch, 2005) was introduced by the physicist Hirsch. He proposed a simple and useful way to characterize the scientific output of a researcher by counting the scientist's most cited papers and the number of citations that they have received in other publications. In particular h-index is calculated as follows: “a scientist

has index h if h of his or her N_p (published papers over n years) papers have at least h citations each and the other (N_p - h) papers have ≤h citations each". After five years Hirsch published (Hirsch J. , 2010) to amend perhaps the "most important shortcoming of the h-index". More than twenty variants (Schreiber, 2010) of the h-index have been suggested since 2005 to overcome inefficiencies of this index.

Egghe introduced g-index (Egghe, 2006) as an improvement of the h-index that lays more emphasis on the highly-cited papers. The g-index, where g is the largest rank such that the first g papers have together at least g^2 citations.

Xuan Zhen Liu and Hui Fang (Fang & Liu, 2012) formed and applied a scheme of impartial citations allocation on the basis of the contributions of each author to a paper to modify h-index and g-index.

Other simple modifications of the h-index are the h_c-index (contemporary index) (Sidiropoulos, Katsaros, & Manolopoulos, 2007), the h_m-index (Schreiber, 2008), h_{mcr}-index (Fang & Liu, 2012) and the harmonic h-index (Hagen, 2010).

The h_c-index is differentiated by adding an age-related weight to each cited article, giving less weight to older articles to make a fairer comparison between younger authors who have published a small number of significant papers but have a low h-index and those scientists who have been inactive for years and have a large h-index.

The h_m-index and the h_{mcr}-index are modifications of the h-index and have been proposed for multi-authored papers. The h_m-index considers multiple co-authorship appropriately, by counting each paper only fractionally according to the inverse of the number of authors. Another approach that takes into account multiple authorship is the h_{mc}-index. The h_{mcr}-index employs the framework of the h_m-index by replacing fractionalized counting with CCA (combined credit allocation) and makes use of the author rank in addition to the number of authors which is used in h_m-index.

The harmonic version of the h-index shared credit allocation based on the inverse of author rank. According to Hagen the harmonic h-index provides unbiased bibliometric ranking of scientific merit while retaining the original's essential simplicity, transparency and intended fairness.

In the past several scientists have proposed various approaches to quantify co-author contributions so as to achieve a better consideration of author rank. In 1963 Zuckerman (Zuckerman, 1968) described three different patterns of name ordering in multi-authored scientific articles. The first is the alphabetical order that often symbolize equality of contribution, the second first-author-out-of-sequence followed by an alphabetized group and the third type gives prime visibility to the first-author and smaller increments of visibility to each succeeding author. Marek Kosmulski (Kosmulski, 2012) remarked that any algorithm that calculates the fractional contribution of multi-authored papers solely from the author list order is inherently incorrect.

Further basic approaches to capture the multi-authoring issue are listed below:

- The normal or standard counting (Chubin, 1973) historically used by most studies, where all contributors receive full credit and others criticized this method, in particular due to the increasing inflation of the number of publications (Lindsey, 1980).
- Cole and Cole (Cole & Cole, 1974) proposed first author counting, which means that, in multi-authored articles, only the first of N authors receives the whole credit for publications and citations.
- Solla Price (de Solla Price, 1981) considers fractional counting where the publication and citation credit is equally divided among the co-authors.
- Harmonic counting has been proposed by Hodge and Greenberg (Hodge & Greenberg, 1981) and later by Cagan Sekercioglu (Sekercioglu, 2008) where the publication credit is divided up, based on the order of authors,

with the first author receiving most of the credit and subsequent authors receiving fractional credit based on their position in the author list.

All the aforementioned methods consider either the author list rank or the citations of the publications to share fairly the credits to the scientists. In the thesis a different approach is proposed. It which involves initially content-based analysis through the depiction of the authors' profiles as bag-of-words (Turney & Pantel, 2010). Former studies of Salton (Salton, Wong, & Yang, 1975) and of Jones (Jones & Furnas, 1987) considered documents as vectors represented in a document space. Given documents as vectors it is possible to compute a similarity coefficient among the documents. The constructed authors' profiles and the testing paper are treated as documents and represented as vectors. According to the study of Thada (Thada & Jaglan, 2013) the best way to calculate similarities between the authors' profiles and the testing paper is to use the cosine similarity coefficient. After the measurements the authors are ranked again according to the calculated coefficient in order to presume the participation in the testing multi-authored paper.

The main drawback of this method, is that the representation of the authors' profiles and the testing paper as bag-of-words ignore any syntactic or semantic correlations between the terms (Nanas & Vavalis, 2008). It also suffers in other terms, including efficiency and robustness. According to the second approach of this thesis, the authors' profiles and the testing paper are regarded, as network of terms (Nanas, Vavalis, & Houstis, 2010), which comprises of term correlations and leads to showing the importance of the associations between profile terms. The links between the terms are identified using the sliding window approach referred in (Nanas, Uren, & de Roeck, 2004). Two terms are linked together if they appear within the defined sliding window, which defines a span of contiguous terms. The frequency and the distance are used to calculate the weight of the link between the terms.

3

3. METHODOLOGY

This section introduces the methodological framework for constructing multi-faced-dynamic author's profiles as bag-of-words and as network of terms. More specifically, the first part addresses issues concerning data collection and text processing. Then the testing paper is validated with the two different approaches leading thus to the next part, where a hypothetical result about the contribution of each co-author in this specific paper will be presented. Finally this chapter concludes with a discussion concerning the research limitations as well as the conclusions drawn by the specific work.

3.1 Data Collection

The sampling frame comprised all the faculty authors of the Department of Electrical and Computer Engineering of the University of Thessaly including Professors, Associate Professors, Assistant Professors and their co-authors with more than two publications.

According to the research described in the particular thesis, three bibliometric sources are examined, namely Scopus, Web of Science and Google Scholar. Among them, Web of Science (WOS) and Scopus are considered two of the most widespread bibliometric databases and are frequently used for searching the literature (Chadegani, et al., 2013). More specifically, Scopus is the largest searchable abstract and citation database of research literature and selected web sources published after 1966. What is more, it is continually updated and expanding (Rew, 2010). The following table exhibits a comparison

between Web of Science (WOS) and Scopus involving a different set of criteria such as number of journals, number of records and the time period that each database covers.

Features	Scopus	Web of Science
Number of journals	18.000	12.000
	More than 57 million records	More than 90 million records
Focus	Physical sciences, health sciences, life sciences, social sciences	Science, technology, social sciences, arts and humanities
Period covered	1966-	1900-
Databases covered	100% Medline, Embase and more	Science Citation, Social Sciences Citation, Arts & Humanities Citation Indexes
Updated	daily	weekly?
Developer/Producer	Elsevier	Thomson Reuters
Citation analysis	yes	yes
Controlled vocabulary	yes - IndexTerms field	no
Export feature	yes	yes
Alerts service	yes	yes
Strengths	more versatile search tool with advantages in functionality (default, refine, format of results of citation tracker and author identification. covers 6256 unique journals, compared to WOS' 1467 greater international coverage can use "first author" as a search field in Advanced Search can search with controlled vocabulary	greater time period of coverage more options for citation analysis for institutions covers science and arts/humanities
Weaknesses	Social science coverage, esp. sociology and prior to 1966	No controlled vocabulary

Table 1: Scopus versus Web of Science (Library Guides at University of Washington Libraries, 2013)

An alternative solution is Google Scholar, a freely accessible web search engine that indexes the full text or metadata of scholarly literature. However, since Google Scholar is not a database but a search engine, there is no definition or structure for exporting the abstract and keywords of a scientific publication, which is a necessary component to conduct our research.

As our research is based on current publications issued after 1960 and according to the facts described in the table above, Scopus database is selected for the collection of our sample. The particular sample was retrieved in December 2014. The database of SciVerse contains titles, authors, abstracts, references, keywords and other information about publications for more than 57 million records (Elsevier, 2015). For this research the publication abstract plays an essential role in developing multi-faced-dynamic author profiles. Given the fact that the abstract is available in papers published after 1996, only the aforementioned papers published from 1996 until 2014 are taken into account. More specifically, 28 authors has been involved, having 917 researchers as their co-authors for the total 1128 papers. It should be mentioned that the real publication number could be much higher, due to the fact that only specific journals are enlisted in the database of Scopus.

3.2 Instrumentation

After having addressed issues concerning data collection and analysis, a reference to the instruments employed in the specific work ought to be made at this point. More specifically, it is of primary importance to be able to access all the information and resources in the right format that are necessary to proceed with the application of the similarity approaches. For this reason a data collection tool (Markou, 2015) is developed involving stages that are thoroughly discussed and presented below.

First of all, the Scopus Author ID for each faculty author is identified manually and stored into a local database. After that, the data collection tool is used to sweep over all stored Scopus Author IDs, so as to find the matching co-authors and to save their Scopus Author IDs in the same table. What follows next is retrieving all the available information about the published work of all authors. In order to achieve this, the web page of Scopus is

parsed, every time with a different query string, according to the stored Scopus Author IDs of the faculty authors and their co-authors.

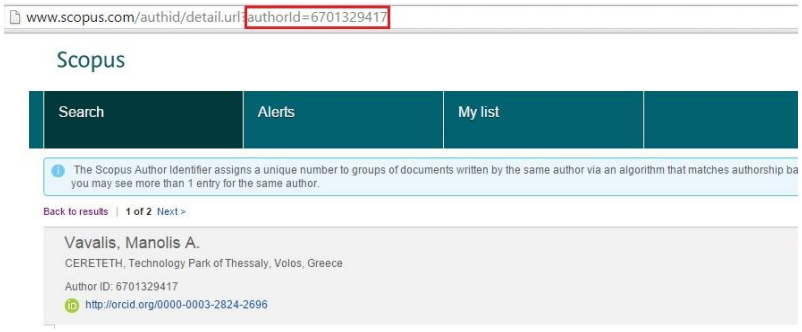

Image 2: Scopus search results

The downloaded document data contains information regarding the following issues: Citation, Author(s), document title, year, EID, source title, volume, issue, pages, citation count, source and document Type, DOI, Bibliographical information, Affiliations, serial identifiers (e.g. ISSN), pubMed id, publisher, editor(s), language of original document, correspondence address, abbreviated source title, Abstract and Keywords, Abstract, author keywords, index keywords, Fund Details, Number, acronym, sponsor, References, References, Other information, Tradenames and manufacturers, accession numbers and chemicals and conference information. The information about the authors, the published documents and their relation are stored in a database, the schema of which is described in the next chapter.

3.3 Database

At this point a reference to the database management system employed in the particular thesis ought to be made. More specifically, Microsoft SQL Server 2012 (SQL Server 2012, 2014) is used to store and manage the sample that is exported by parsing the html search results provided by Scopus.

What is more, this section also provides a description of the basic database schema and important modifications and thus supports the approaches concerning both bag-of-words and network of terms.

3.3.1 Basic Schema

The diagram below provides a visual overview of the basic database schema, the most important tables and the relations between them. The tables called *author*, *paper* and *author_paper* have three relationships present: *author* to *paper*, *author* to *author_paper* and *paper* to *author_paper*. These relationships represent many-to-many relationship. An author can publish more than one papers and a paper can may more than one authors (co-authors). The first author is also stored in the table paper.

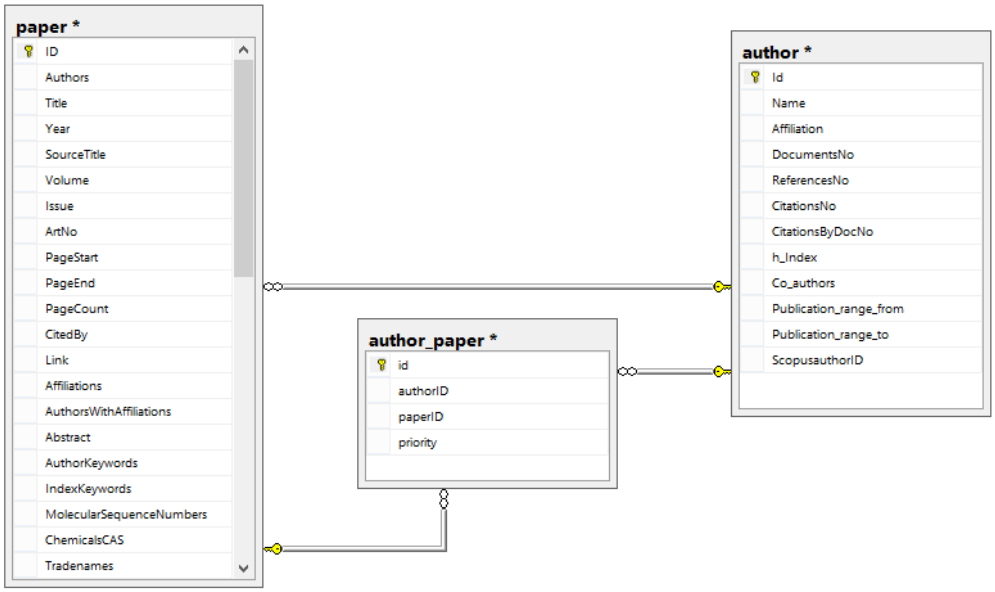

Image 3: Basic database schema

The table overviews below include additional details on the tables and columns. The table *author* stores the 28 faculty authors of the Department of Electrical and Computer Engineering and all their co-authors, according to the published papers in Scopus.

Moreover, useful information provided by Scopus, concerns the total number of documents published, references, citations, co-authors, h-index and the cited affiliation about the institution or the department of the published paper.

author	
name	**type**
Id	int
Name	nvarchar(255)
Affiliation	nvarchar(255)
DocumentsNo	int
ReferencesNo	int
CitationsNo	int
CitationsByDocNo	int
h_Index	int
Co_authors	int
Publication_range_from	int
Publication_range_to	int
ScopusauthorID	bigint

Table 2: Table author

The table *paper* stores 1.128 records with papers published by the faculty authors and their co-authors. It includes important information about the evaluation, such as title, authors, publication year, subject keywords and abstract.

paper	
name	**type**
ID	int
Authors	nvarchar(255)
Title	nvarchar(255)
Year	int
SourceTitle	nvarchar(255)
Volume	nvarchar(255)
Issue	nvarchar(255)
ArtNo	nvarchar(255)
PageStart	int
PageEnd	int
PageCount	nvarchar(255)
CitedBy	int
Link	nvarchar(255)
Affiliations	nvarchar(255)
AuthorsWithAffiliations	nvarchar(Max)

Abstract	nvarchar(Max)
AuthorKeywords	nvarchar(255)
IndexKeywords	nvarchar(255)
MolecularSequenceNumbers	nvarchar(255)
ChemicalsCAS	nvarchar(255)
Tradenames	nvarchar(255)
Manufacturers	nvarchar(255)
FundingDetails	nvarchar(255)
Refs	nvarchar(Max)
CorrespondenceAddress	nvarchar(255)
Editors	nvarchar(255)
Sponsors	nvarchar(255)
Publisher	nvarchar(255)
ConferenceName	nvarchar(255)
ConferenceDate	nvarchar(255)
ConferenceLocation	nvarchar(255)
ConferenceCode	int
ISSN	nvarchar(255)
ISBN	nvarchar(255)
CODEN	nvarchar(255)
DOI	nvarchar(255)
PubMedID	nvarchar(255)
LanguageofOriginalDocument	nvarchar(255)
AbbreviatedSourceTitle	nvarchar(255)
DocumentType	nvarchar(255)
Source	nvarchar(255)
AuthorId	int
StemmedTitle	nvarchar(Max)
StemmedAbstract	nvarchar(Max)
TitleStemmedNoPuncta	nvarchar(Max)
KeywordsStemmedNoPunct	nvarchar(Max)
AbstractStemmedNoPunct	nvarchar(Max)
coPublicationsId	int
eid	nvarchar(255)

Table 3: Table paper

The table *author_paper* defines the relationship many-to-many between the author and the published papers. The column *priority* defines the author sequence in multi-authored publications.

author_paper	
name	**type**
id	int
authorID	int
paperID	int
priority	int

Table 4: Table author_paper

3.3.2 Enhancing DB Schema for the bag-of-words approach

The particular approach is considerably supported by the creation of two tables named *nodeBoW* and *author_profilesBoW*. The table *nodeBoW* stores data about the individual terms that appeared in the text (title, keywords and abstract) of each paper.

nodeBoW	
name	**type**
id	int
term	nvarchar(Max)
weight	int
paperId	int

Table 5: Table nodeBoW

The table *author_profilesBoW* arose from the table *nodeBoW* and is created to store the most frequent unique terms of the table *nodeBoW* for each author profile. The property *weight* characterizes the frequency of each term.

author_profilesBoW	
name	**type**
id	int
term	nvarchar(Max)
weight	int
authorID	int

Table 6: Table author_profilesBoW

3.3.3 Enhancing DB Schema for the network of terms approach

Accordingly, two tables are created aiming to enhance the network of terms approach. The tables described below are added to the database schema in order to store the individual terms (*nodeNoT*) and their relations (*linkNoT*).

nodeNoT	
name	**type**
id	int
term	nvarchar(Max)
frequency	int
weight	int
paperId	int

Table 7: Table nodeNoT

The field *distance* stores the information about the distance between the terms within the sliding window, as referred in Section Sliding Window.

linkNoT	
name	**type**
id	int
nodeID1	int
nodeID2	int
weight	int
distance	int
frequency	int
paperID	Int

Table 8: Table linkNoT

The author profile is a view that results from the most frequent terms which are present in *nodeNoT* table and their dependencies stored in *linkNoT* table.

3.4 Programming Language

Python is used as a programming language to develop all the functions needed to access and process the data. Python is an object-oriented, interpreted and interactive high-level programming language. The reference implementation of Python, is a free and open-source software and has a community-based development model (Python, 2014).

In addition, Gensim (Řehůřek, 2014), a free Python library, is used to examine similarities among the author profiles and the paper at issue. Specifically, classes and functions in module *similarities.docsim* is utilized to compute cosine similarity of a dynamic query (the paper at issue) against a static corpus of documents (authors profile) in the Vector Space Model.

3.5 Visualizing tools for graphs

Mainly two visualizing tools for graphs are used to represent the network graphs generated by the author's profiles. Gephi (The Open Graph Viz Platform Gephi) is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. Gephi is open-source, free and runs on

Windows, Linux and Mac OS X. It also provides dynamic filtering used for querying the result graph (Gephi, 2014).

An alternative online tool is Fusion Tables (Google Research, 2015), an experimental data visualization web application, used to gather, visualize and share data tables powered by Google Research. Fusion table supports the representation of undirected and directed graph structures. This type of visualization illuminates relationships between nodes. Nodes are displayed as round circles and lines show the edges between them. The circle size represents the node weight.

3.6 Data Processing

3.6.1 Data cleaning and Processing

High standards in research quality of this particular work are strongly associated with and influenced by the appropriate processing of the collected data. Yet, various accuracy issues arose from difficulties in identifying author's profiles as well as from problems created by duplicate data entries. In order to deal with the ambiguity problems described above, a set of actions has to be taken. More specifically, there is a meticulous search for potential author matches and on the occurrence of duplicate publication entries in the database of Scopus, these are either deleted or merged manually in the local database.

Before creating the bag-of-words and the network of terms representation, the stored document data in the following fields: title, keywords and abstract, is undergone a certain preprocess. With the help of the data collection tool, redundant information for classification within the document data is removed. At first, the punctuation of the document data is stripped away and afterwards, using a list of common stop words, all occurrences of the document data are removed. Another important task is to reduce the number of words using stemming and keeping only the linguistic root. To achieve this goal

the Porter's algorithm (The Porter Stemming Algorithm, 2014) is applied. Thus, every word is replaced to its root form. After the preprocessing, the stemmed text, without stop words and punctuation, is stored separately in the fields, *TitleStemmedNoPunct*, *KeywordsStemmedNoPunct* and *AbstractStemmedNoPunct*.

3.6.2 Generating authors' profiles as bag-of-words

The primary aim in this work is to build the authors' profiles as bag-of-words arising from the abstract, the title and the keywords of the previously published papers, excluding the testing paper. Before measuring similarity among the authors' profiles, the following tasks have to be executed.

1. First the text that is included the title, the keywords and the abstract is split on whitespace into individual terms and the terms are then inserted into the *nodeBoW* table. At the same time, the property *weight* is calculated by counting occurrences (f_t) of each distinct term within a paper, according to the following formula, where the values 1, 0.5 and 0.01 are heuristically defined. The title and the keywords of a published paper are estimated as more influential, than the abstract and therefore they are weighted more highly.

$$w_t = \begin{cases} f_t \cdot 1, & if\ the\ term\ appears\ in\ the\ title \\ f_t \cdot 0.5, & if\ the\ term\ appears\ in\ the\ keywords \\ f_t \cdot 0.01, & if\ the\ term\ appears\ in\ the\ abstract \end{cases}$$

Equation 2: Weight of terms in bag-of-words approach

2. Moving on, the authors' profiles are generated by summarizing the 50 most frequent terms (according to their weight) from the *nodeBoW* table and inserted into the table *author_profilesBoW*. This led to a dictionary for each author that encapsulates the mapping between normalized terms and their

integer ids. The value 50 was selected, based on the experimental results in the publication "Building and Applying a Concept Hierarchy Representation of a User Profile" (Nanas, Uren, & Roeck, 2003), where the functions of unconnected (bag-of-words) and hierarchical (network of terms) profiles converge around the value 50.

3. The gensim function *doc2bow* is used to convert the collection of terms to a bag-of-words representation and this resulted in the creation of a sparse vector with tuples (term_id, term_weight).
4. Steps 2 and 3 are repeated and thus a sparse vector for the testing paper is created.
5. The gensim function *similarities.docsim.Similarity* is used to compute cosine similarity of the testing paper (dynamic query) against the authors' profiles, which contributed to the paper (a static corpus of documents).

The values of the cosine coefficient falls between of 0 and 1, since the term weight cannot be negative.

3.6.3 Generating authors' profiles as network of terms

The preprocessing tasks discussed in the previous section, constitute a prerequisite for conducting the second approach that represents authors' profiles as weighted network of terms. At this point, a brief outline of the stages involving the specific process will be presented.

1. At first, the table *nodeNoT* is filled with data following the paradigm of *nodeBoW* table. In this case the property *weight* is calculated according to the calculation method described in "Nootropia: A User Profiling Model Based on a Self-Organising Term Network" (Nanas, Uren, & de Roeck, Nootropia: A User Profiling Model Based on a Self-Organising Term

Network, 2004). The network terms within the table *nodeNoT* are weighted, by means of the following equation:

$$RelDF_t^D = w_t^D = \frac{1}{20} - \frac{n}{N}$$

Equation 3: Weight of terms in network of terms approach

where:

- *t* is the term in the publication
- *n* is the number of publications that contains the term *t*
- *N* is the total number of publications
- The value *20* is defined heuristically, as defined in the aforementioned paper

Only the 50 first terms with the highest weight are taken into account for generating the authors' profiles.

2. Then, table *linkNoT* is filled with data using the sliding window approach. The particular approach as found in the chapter Multi-topic Profile Representation and Document Evaluation about network initialization in (Nanas, Uren, & de Roeck, 2004) is used, having selected a window size of seven. According to this theory, two terms are considered to be linked, if they appear at least once in the window of seven consecutive words. The property *distance* is updated with *d=1* when two extracted terms appear next to each other, whereas if *m* words intervene between them, the *distance* is *d=m + 1*. The property *weight* between two links is calculated as follows:

$$w_{ij} = \frac{fr_{ij}^2}{fr_i \cdot fr_j} + \frac{1}{d}$$

Equation 4: Weight of links in network of terms approach

where:

- $0 < w_{ij} \leq 1$
- fr_{ij} is the number of times t_i and t_j co-occur within the sliding window
- fr_i and fr_j are respectively the number of occurrences of t_i and t_j in the user specified documents
- d_{ij} is the average distance between t_i and t_j, within the sliding window

4

4. IMPLEMENATION AND RESULTS

For the purpose of the specific assignment only a small sample is selected (three authors) in order to present more clear results. What shall be examined at this point is the extent of contribution of each author to each paper, so that an estimation can be made regarding the amount of similarity between the author's network of terms and that of the testing paper.

For this reason, the following authors are selected: Akritidis, L., Katsaros, D. and, Bozanis P., who have co-authored 8 publications. More specifically, the author profile of Katsaros, D. is generated out of 50 publications, the author profile of Bozanis, P. is generated out of 27 publications and the author profile of Akritidis, L. is generated out of 10 publications. In order to interpret the results of this research, it ought to be mentioned at this point that Akritidis, L., during his publishing work, has been supervised by Bozanis, P. and Katsaros, D.

4.1 Authors' profiles as bag-of-words

The first attempt generated the authors' profiles including all the terms that appeared in publications in which the authors have contributed. Within the second attempt, only the 50 most frequent terms with the highest weight are taken into consideration. A visualization of the bag-of-words for the selected authors' profiles of Katsaros, D., Bozanis, P. and Akritidis, L. can be evident in the following images.

The author profile of Katsaros D. is shown to be mostly relative to the topics sensor or wireless network, cloud computing, distributed systems, web and data mining.

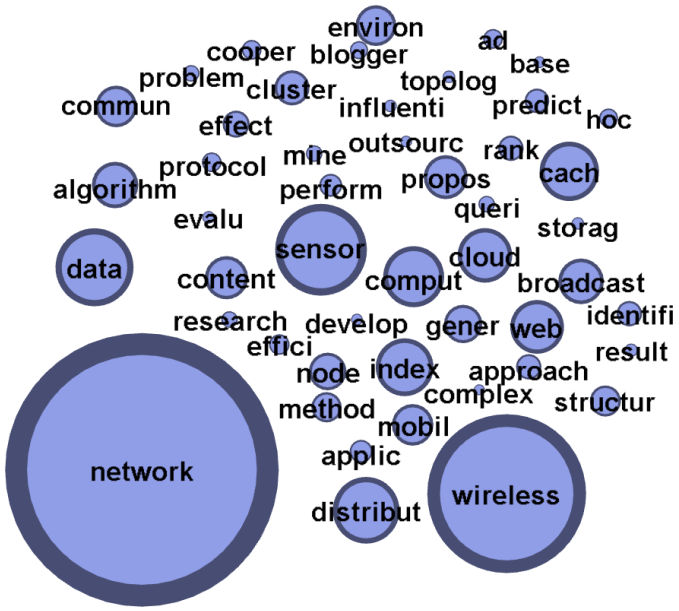

Image 4: Bag-of-words profile of Katsaros D.

The author profile of Bozanis P. is more closely related to the topics data structures, indexing, ranking, search engines and web. A lot of terms contained in the profile of Bozanis are insignificant and should have been removed in the preprocessing phase. Words like *result*, *method*, *present*, *set* and *solut* should have been added in the stop words list.

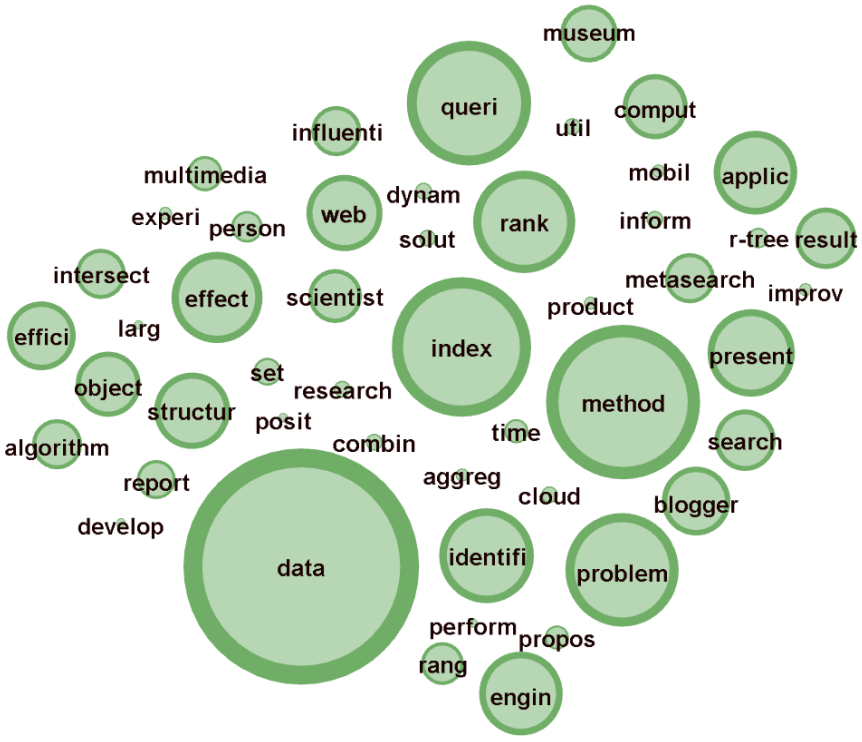

Image 5: Bag-of-words profile of Bozanis P.

Finally the author profile of Akritidis L. is strongly linked to the topics scientist ranking, search engines, influential bloggers, metasearch and web. Terms with strong weight, like *identifi* and *method* are also considered as stop words.

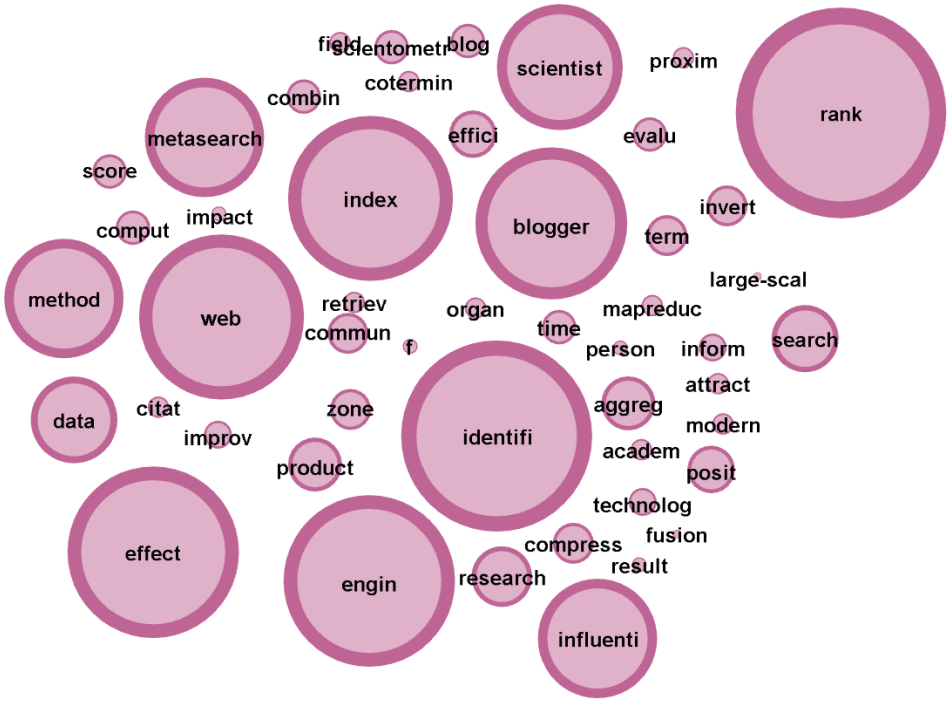

Image 6: Bag-of-words profile of Akritidis L.

Yet a drawback is encountered concerning the particular experiment. The results drawn from this work may lack a certain amount of credibility, as in many articles (40%) the field *keywords* is missing and thus it could not be taken into account. The field keywords contributed to a great extent to the formation of the results due to its substantial weight among the three factors (title, keywords and abstract) that are responsible for the generation of authors' profiles.

Table below presents detailed results including the first and the second attempts concerning the 8 selected publications. All the publications, apart from one, have the same author sequence, Akritidis L., Katsaros D., Bozanis P. and none of them is alphabetically ordered, but sequenced according to each author's contribution.

The following table exhibits the results of the bag-of-words approach. The first three columns refer to the authors' profiles generated by all the terms, whereas the last three involve the pruned profiles with the 50 most frequent terms.

No	Testing Paper	Katsaros [all]	Bozanis [all]	Akritidis [all]	Katsaros [50]	Bozanis [50]	Akritidis [50]
1	Title: **Effective ranking fusion methods for personalized metasearchengines** Authors: **Akritidis L., Katsaros D., Bozanis P.** Year: **2008**	0,3189781	0,45701018	0,4384447	0,29483742	0,54933745	0,51585257
2	Title: **Modern web technologies** Authors: **Akritidis L., Katsaros D., Bozanis P.** Year: **2009**	0,24298592	0,20900258	0,19682479	0,28868061	0,19830015	0,18936428
3	Title: **Identifying influential bloggers: Time does matter** Authors: **Akritidis L., Katsaros D., Bozanis P.** Year: **2009**	0,24577153	0,27562955	0,32749596	0,24374573	0,31554538	0,42587274
4	Title: **The f index: quantifying the impact of coterminalcitations on scientists ranking** Authors: **Katsaros D., Akritidis L., Bozanis P.** Year: **2009**	0,2499067	0,28602719	0,31480718	0,27156553	0,29784137	0,37601587
5	Title: **Identifying the productive and influential bloggers in a community** Authors: **Akritidis L., Katsaros D., Bozanis P.** Year: **2011**	0,34602717	0,33746785	0,40501225	0,28348947	0,32875162	0,43574214
6	Title: **Effective rank aggregation for metasearching** Authors: **Akritidis L., Katsaros D., Bozanis P.** Year: **2011**	0,29837406	0,43171772	0,43143049	0,27798986	0,51521379	0,51134902
7	Title: **Identifying attractive research fields for new scientists** Authors: **Akritidis L., Katsaros D., Bozanis P** Year: **2012**	0,21627975	0,19430989	0,2034854	0,21749556	0,26994324	0,25441185
8	Title: **Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: A simulation-based evaluation** Authors: **Akritidis L., Katsaros D., Bozanis P** Year: **2012**	0,28807455	0,4009667	0,41577595	0,27541146	0,50902504	0,52750033

Table 9: Bag-of-words similarity results

Below, the results are visualized as diagrams. The first diagram shows the author profile including all terms. What can be evident from the particular diagram is that the dominant profile is that of Akritidis, L. Excluding papers one and two, in all the other papers Akritidi's profile shows a very close similarity to the measured paper. As far as the second paper is concerned, the author profile of Akritidis, L. bares significantly less resemblance to the testing paper. However, Katsaros', D. profile seems to shares more common terms with the bag-of-words of the second paper. Another issue is that the fourth paper should have more similarities with the author profile of Katsaros, D., as it is him that contributed to the specific paper the most, according to the authors' order sequence. Yet, instead of that, Akritidis, L participation to the paper appears to be of larger extent. This could be attributed to the fact that Katsaros', D. profile is generated out of 50 papers and is thus more complex than the other ones. This can lead to the assumption that the author profile of Katsaros, D. can be described as a multi-topic profile. A possible way to overcome this problem is to apply clustering methods, so as to group the different topics into classes. This could be achieved using methods like "Feature Selection and Transformation Methods for Text Clustering", "Distance-based Clustering Algorithms", "Word and Phrase-based Clustering", "Probabilistic Document Clustering and Topic Models", "Online Clustering with Text Streams", "Clustering Text in Networks" or "Semi-Supervised Clustering" (Aggarwal & Zhai, 2012). Overall the result can be described as satisfactory, as it is in accordance with our expectations.

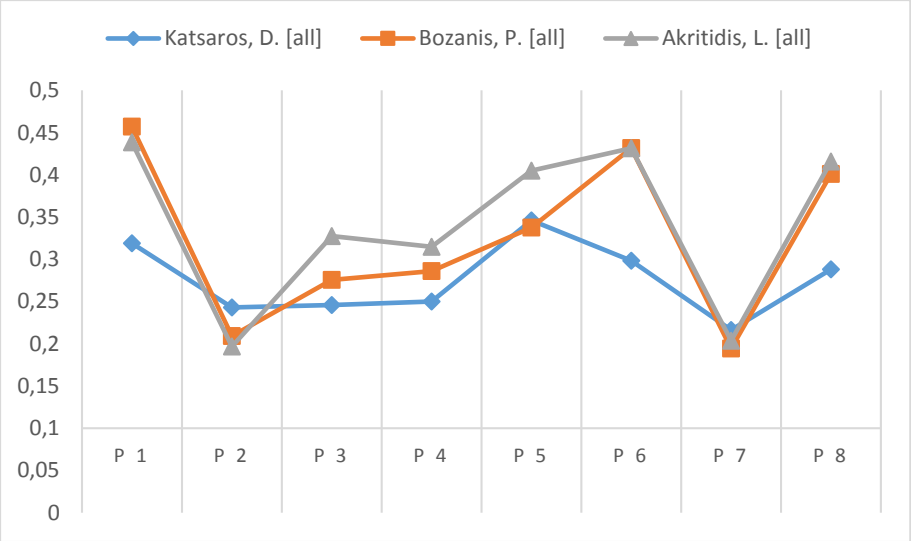

Image 7: Similarity chart including all terms

The results in the second diagram that represents the pruned authors' profiles, are similar to the ones exhibited above, apart from minor differences. More specifically, in paper five the similarity order is the same like the author sequence in the testing paper. Furthermore, in paper seven a slight difference is shown, that highlights the profile of Bozanis, P. as more resembling.

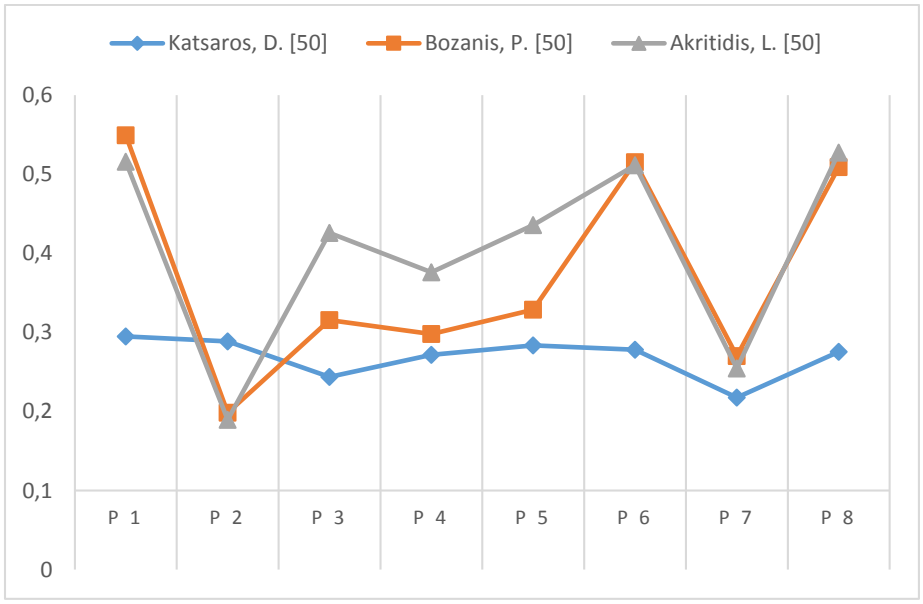

Image 8: Similarity chart including 50 most common terms

4.2 Authors' profiles as network of terms

The second approach represents the authors' profiles as network of terms by taking into account the correlations between the terms, which are ignored in the bag-of-words approach. In other words, the network of terms, incorporating term dependencies, represents the author profile.

What ought to mention at this point that all papers that are examined involve the same topic. What becomes evident from the results of this approach is that both the networks of terms of Katsaros, D. and Bozanis, P. are more complex than this of Akritidis, L. In particular, the network of terms of Katsaros, D. counts 14.908 relations and that of Bozanis, P. counts 9.569, while the dependencies of Akritidis, L. are only 3.112. For this reason, this could be an indication that the author profiles of Katsaros, D. and Bozanis, P. comprise multiple topics of interest and this could account for the fact that the author profile of Katsaros, D. seems to have the lowest contribution.

A way to deal with the drawbacks caused by multi-topic profiles can involve the formulation of a separate hierarchy for each general topic found in the author profiles. There are also methods for the automatic construction of hierarchical networks that explicitly capture topic-subtopic relations between terms. Further insight in the specific topic shall be offered in the Further Research section.

The following figure visualizes the network of terms of Katsaros, D., pruned by 50 terms with the highest weight. What becomes evident that the most common terms in the profile of Katsaros, D are the following: network, wireless, sensor, data, distribute, web, index, cache, cloud and broadcast

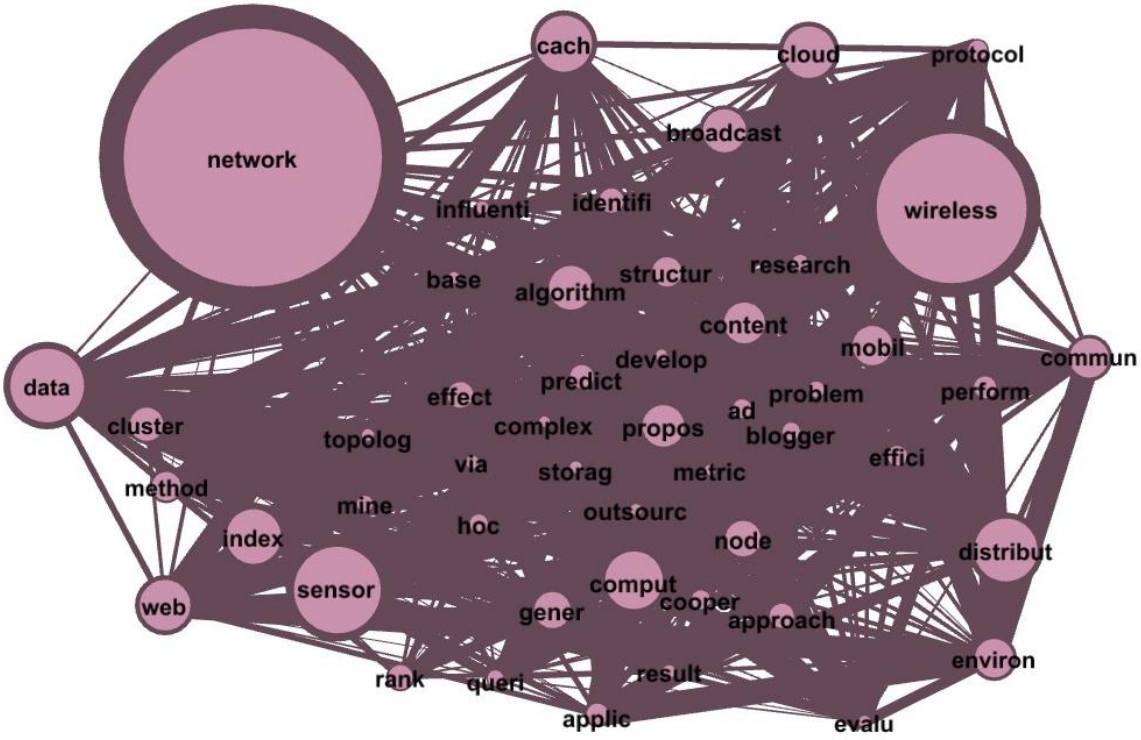

Image 9: Network of terms profile of Katsaros, D.

For display purposes, the selected term *network* is one of the most common in the profile of Katsaros, D., that counts 1.170 relations. The tightest dependencies are wireless networks, sensor networks, ad hoc networks, clustering network and content distribution networks.

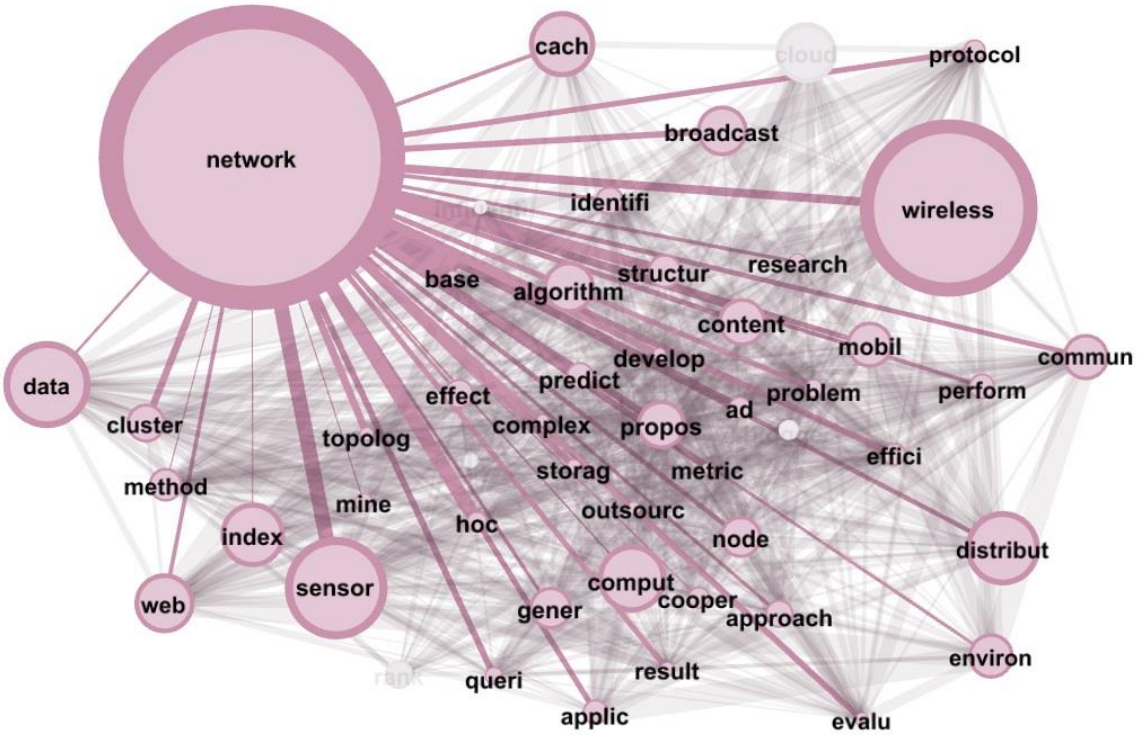

Image 10: Network of terms profile of Katsaros, D. with selected term

The network of terms of Bozanis, P., reduced by the 50 most common terms, is illustrated below. The most frequent terms in the author profile of Bozanis, P. are the following: data, method, problem, query, rank and index.

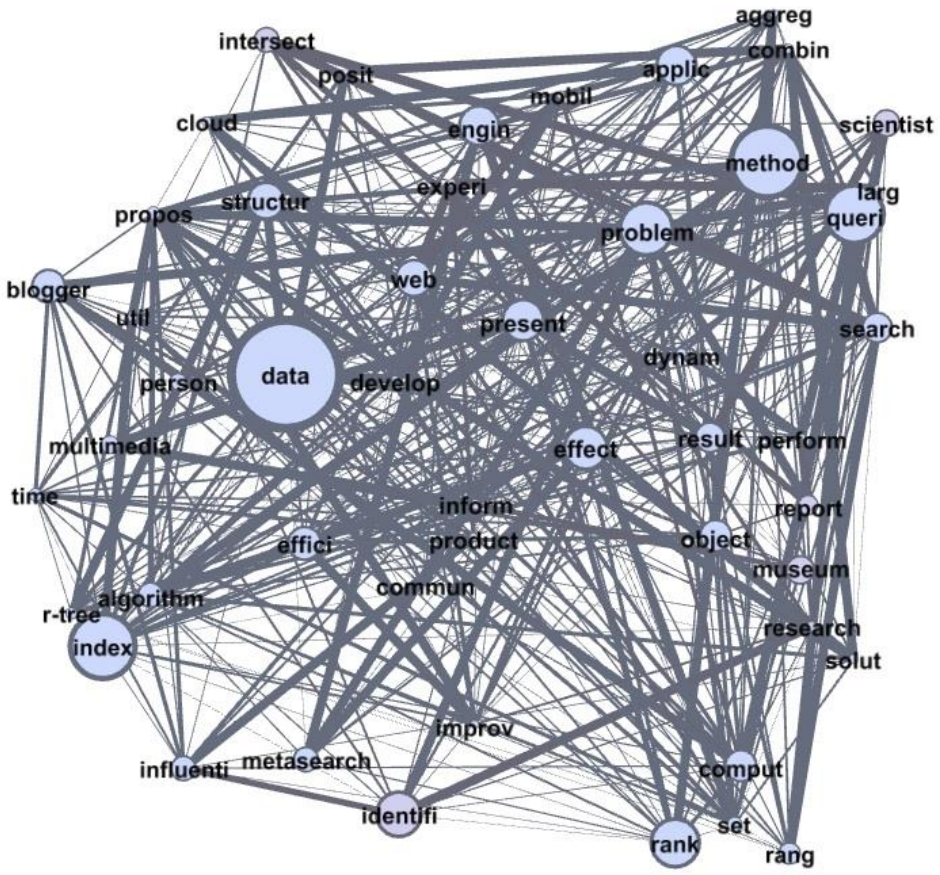

Image 11: Network of terms profile of Bozanis, P.

The image below shows the network of graph of Bozanis, P. What one can detect in the particular image is that the selected term *data* has relations to the topics data structure, data query, spatial data, index data structure, data storage and items.

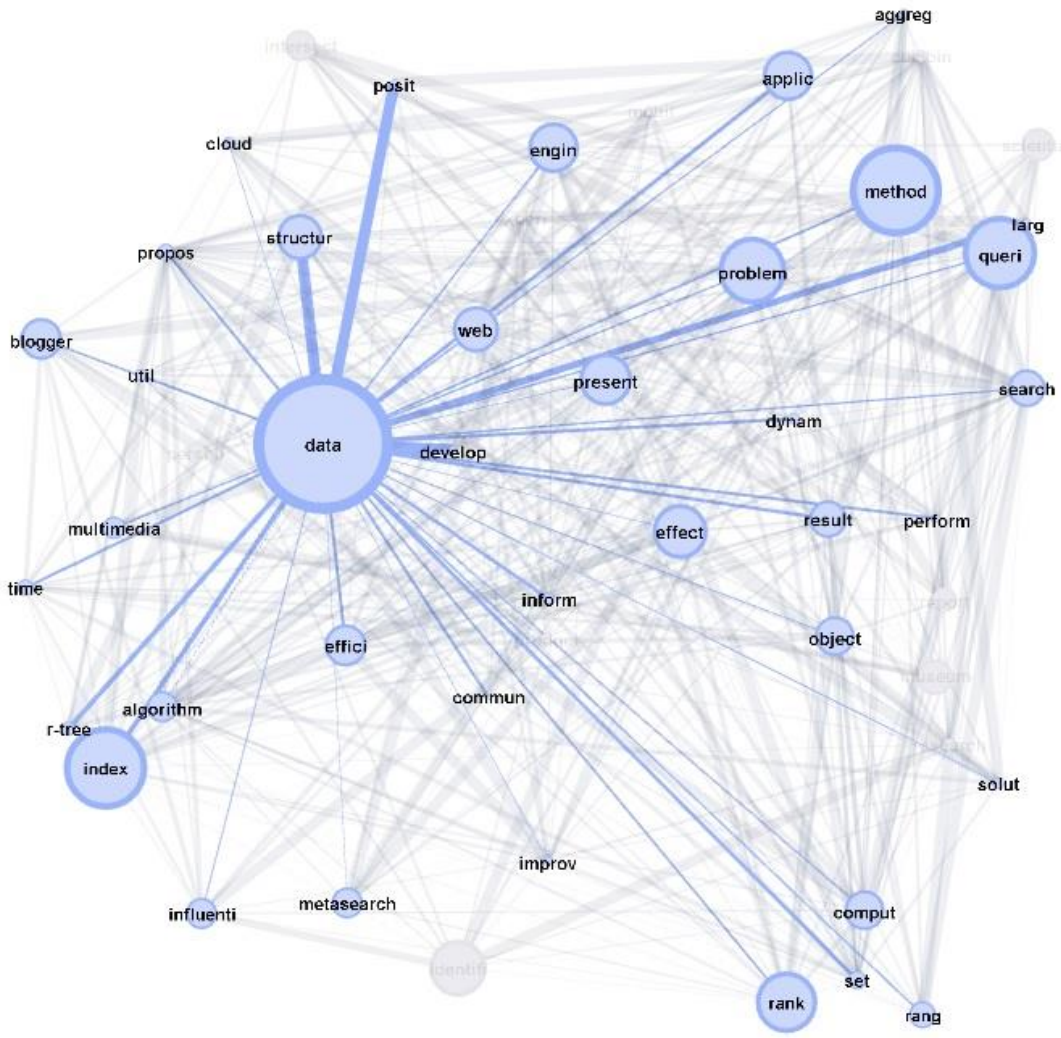

Image 12: Network of terms profile of Bozanis, P. with selected term

The following figure shows the 50 most common terms in the author profile of Akritidis, L., which is illustrated as a network of terms. The most significant terms within the network are the following: identification, rank, index, web, engine, method, scientist, blogger and effect.

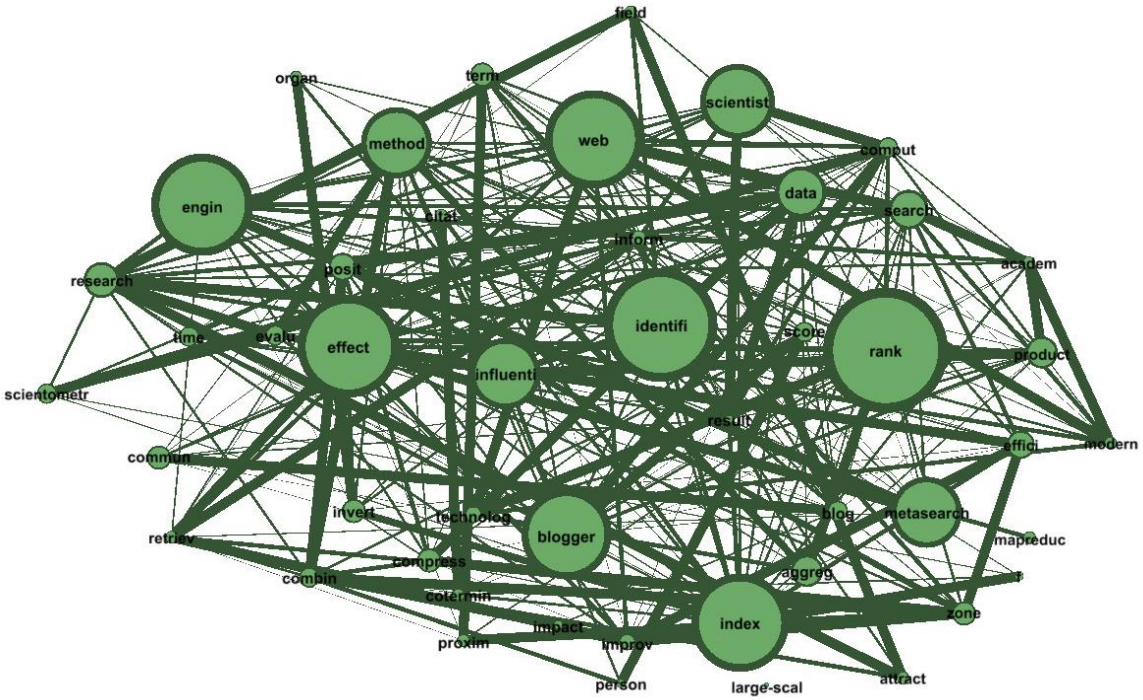

Image 13: Network of terms profile of Akritidis, L.

The network of Akritidis, L. contains 3.112 dependencies. One of the most common is the term *index*, which is strongly related to terms like "f", inverted, information, productivity and spatial.

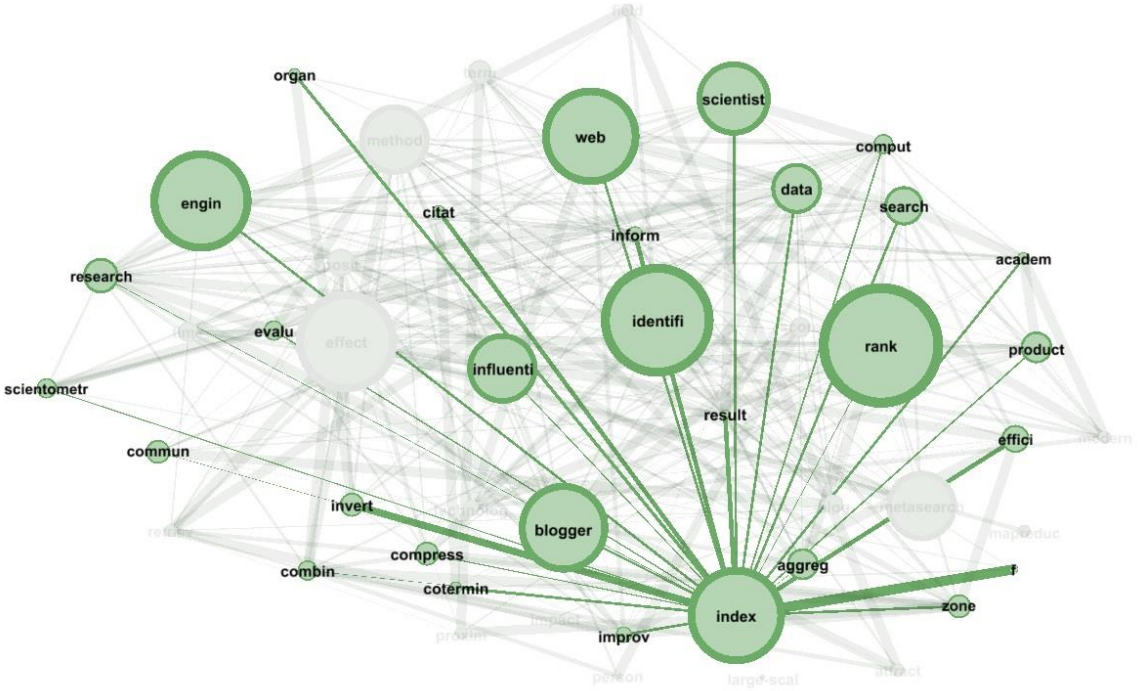

Image 14: Network of terms profile of Akritidis, L. with selected term

5

5. DISCUSSION

The final chapter of this thesis provides a brief overview of our research, including the statement of the problem concerning multi-authorship and the research methods involved. More specifically the majority of this chapter is devoted to future research that will take place upon the completion of the specific research.

5.1 Conclusions

This part focuses on a brief presentation of the methodologies used, in order to construct dynamic authors' profiles, based on their published scientific papers. In particular, two approaches are examined.

According to the first approach, authors' profiles are represented as bag-of-words. For this reason publication metadata, such as title, keywords and abstract are taken into consideration. Experiments are conducted, by means of calculating the cosine coefficient, so as to measure the degree of similarity between the testing paper and the authors' profiles, who participate in the particular paper. Despite the promising results, implementing author profile clustering may lead to further improvements.

The second approach suggests a methodology that represents author profiles as network of terms, instead of bag-of-words. According to the sliding window approach, term dependencies are identified and weighted, and thus syntactic and semantic correlations between terms are taken into account. The network of terms representation allows the author profile to focus on the most relevant term combinations.

Based on the results of our research, the network of terms approach can be considered a more efficient way to represent an author profile. Yet, at this point it should be mentioned that this thesis has a limited spectrum as it is based on a small scale research. As a result, further experiments need to be conducted so as to achieve improvements in the evaluation of similarity measures between networks of terms. Thus this thesis can stimulate further research in this field.

5.2 Further Research

At this point further insight will be offered in terms of the future research that shall take place upon the completion of the particular experiment. More specifically this research will involve changing the sampling frame and by doing so it will be possible to verify our results more accurately. Inspired by the journal article "Collective credit allocation in science" (Shena & Barabási, 2014) the proposed approach can be validated by means of a sample comprising the Nobel prize-winning publications. In this case the Nobel Prize committee has already decided the Nobel laureates and thus where the main credit goes. Therefore, the next step is to apply the described approaches to the Nobel prize-winning publications in Physics, Chemistry and Medicine. In particular, the selected sample would contain 25 papers in Physics, 24 papers in Chemistry and 14 paper in Medicine. Papers in Economics should be excluded from the data sample, because they are single authored papers. The validation process will show if the results coincide with the decision of the Nobel Prize committee. This experimental run requires the construction of profiles of all authors, who participate in the writing of the selected articles. One issue that might be challenging is the credit allocation in papers of Physics, where it is usual to encounter "hyperauthorship" (Cronin, 2001), which means a listing of a large number of contributors on scientific papers.

In order to accomplish the task of gathering the data sample, Scopus has to be combined with other bibliometric databases. Among difficulties that one may encounter during the data collection involves dealing with the author redundancy. The solution to this

problem could be offered by the use of the Open Researcher & Contributor ID (ORCID), which is an initiative to solve the author name ambiguity problem, instead of the Scopus Author ID. The particular approach is described in "Scientists: your number is up" (Butler, 2012) and "Open Researcher & Contributor ID (ORCID): Solving the Name Ambiguity Problem" (Wilson & Fenner, 2012).

Another issue is that, over time, the level of interest in each topic may vary, as new topics of interest can emerge and a previously interesting topic may wane and even become obsolete. To represent the dynamic aspect of the author's profile, the time factor needs to be considered. More specifically, each subnetwork will be arranged on the basis of the paper's publication date. This process will lead to a constant update of author's profile, which will be adapted to the context of every new paper.

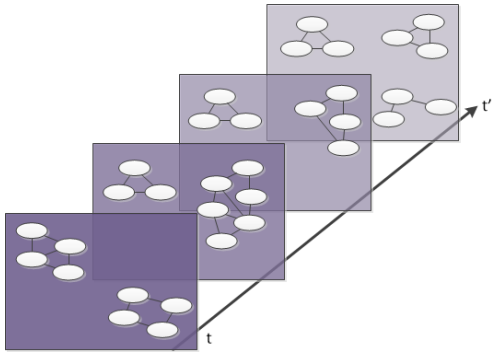

Image 15: Author profile with time factor

Due to the complex structure of networks, further experiments of representation and similarity methods are needed. For this reason, future research should be based on the methodology discussed in the paper "Nootropia: A User Profiling Model Based on a Self-Organising Term Network" (Nanas, Uren, & de Roeck, Nootropia: A User Profiling Model Based on a Self-Organising Term Network, 2004), which identifies similarities among weighted networks of multiple topics.

In order to form clusters of topics of interest within the author profile and thus dividing the network of terms into separate hierarchical subnetworks, the terms have to be

ordered according to decreasing weight. The image below illustrates an author profile with two different research areas (subnetworks) and a small number of common terms.

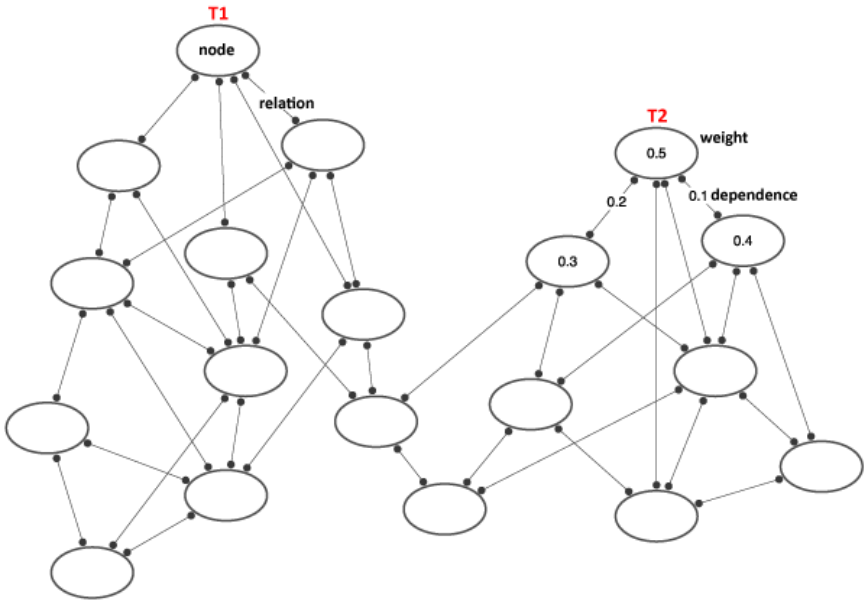

Image 16: Author profile with two topics of interest

The subnetworks are identified by the terms T1 and T2, also called “dominant”, which are strictly related only to terms with lower weight. The number of subnetworks within the author profile is named “breadth”. Furthermore, the “size” of a subnetwork is determined by the number of terms that are connected with the “dominant” term.

In order to evaluate the similarity between the testing paper and the author profile, a directed spreading activation model is used. Terms that appear both in the author profile as well as in the testing paper, are immediately activated. These terms activate sequentially other terms that are directly linked together and are higher in the hierarchy.

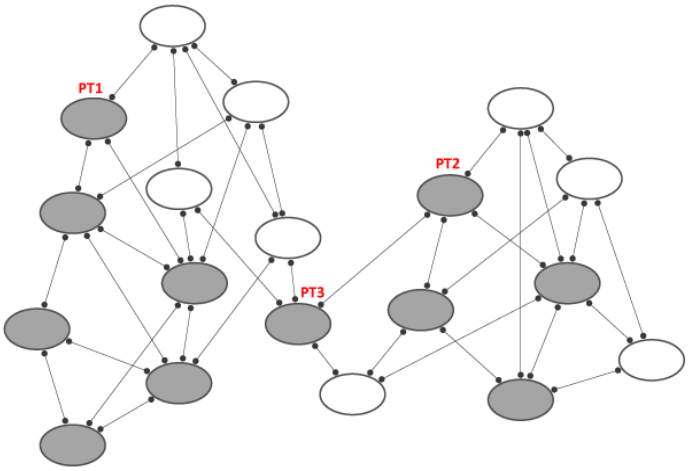

Image 17: Activated testing paper terms

The testing paper *P* has an initial energy (activation) that is equal to 1 and is stored in the corresponding terms (activated terms), which are also included in the author profile. The amount of energy that is transferred between two activated terms is proportionate to the weight of the relation between them. The terms PT1, PT2 and PT3 are dominant, which means that there are 3 different topics in this scientific publication and thus the paper breadth *b* equals to 3. The size of the corresponding subnetwork is equal to the number of the activated terms that share energy (the dominant terms are excluded). Therefore, the size of the testing paper *p* is the total number of the activated terms that transfer energy, in the example above $p=8$. The total number of activated terms *a* is equal to $b+p$.

If and only if, an activated term t_i is directly linked to another activated term t_j with larger weight, then an amount of energy E_{ij} is transferred from t_i to t_j through the corresponding relation. E_{ij} is calculated as follows:

$$E_{ij} = \begin{cases} E_i^c \cdot w_{ij} & if \sum_{k \in A^h} w_{ik} \leq 1 \\ E_i^c \cdot \left(\frac{w_{ij}}{\sum_{k \in A^h} w_{ik}} \right) & if \sum_{k \in A^h} w_{ik} > 1 \end{cases}$$

Equation 5: Amount of energy that is transferred from term t_i to term t_j

where:

- t_i, t_j are terms directly linked to each other
- E_i^c is the current energy of term t_i
- w_{ij} is the weight of the relation between t_i and t_j
- A^h is the set of activated terms, which are higher in the hierarchy that t_i is linked to

The current energy of the term t_i is calculated in the formula below.

$$E_i^c = 1 + \sum_{k \in A^l} E_{ik}$$

Equation 6: Current energy

where:

- A^l is the set of activated terms, which are lower in the hierarchy that t_i is linked to

The final energy E^f is also calculated with an equation that is defined in the forenamed paper.

$$E_i^f = E_i^c - \sum_{k \in A^h} E_{ik}$$

Equation 7: Formula for calculating the final energy

The similarity score S_P is then based on the final energies of activated terms (E^f) and it is calculated as the weighted sum of the final activation of terms with the following equation.

$$S_P = \frac{\sum_{i \in A} w_i \cdot E_i^f}{\log(NT)} \cdot \log\left(1 + \frac{b+p}{b}\right)$$

Equation 8: Similarity of a testing paper

where:

- E_f is the final energies of the author profile
- A is the set of activated author profile terms
- NT is the number of terms in the testing paper
- w_i is the weight of an activated term t_i
- b is the number of dominant terms within the testing paper
- p is total number of terms in the testing paper, that share energy

When the size of subnetworks is large, the similarity score of the testing paper increases. The opposite happens when the terms are isolated, such as the term PT3 in image 17.

All the aforementioned issues will consist the basis upon which our future research will be conducted.

6

6. References

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms. In C. C. Aggarwal, & C. Zhai, *Mining Text Data* (pp. 77-128). New York: Springer.

Assessing assessment. (2010). *Nature, 465*(7300), 845.

Butler, D. (2012). Scientists: your number is up. *Nature*, 564.

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., & Ebrahim, N. A. (2013). A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases. *Asian Social Science*, 18-26.

Chubin, D. (1973). On the Use of the "Science Citation Index" in Sociology. *The American Sociologist*, 187-191.

Cole, J. R., & Cole, S. (1974). Social Stratification in Science. *Administrative Science Quarterly*, 264-266.

Content Coverage Guide. (2014, 12 20). Retrieved from Elsevier: http://www.elsevier.com/__data/assets/pdf_file/0007/69451/sc_content-coverage-guide_july-2014.pdf

Count on me. (2012). *Nature, 487*(7415), 177.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, 558–569.

de Solla Price, D. J. (1981). Multiple Authorship. *Science*, 986.

Defining the Role of Authors and Contributors. (2014, 10 10). Retrieved from International Commitee of Medical Journal Editors: http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 131–152.

Elsevier. (2015, 1 12). Retrieved from Content: http://www.elsevier.com/solutions/scopus/content

Fang, H., & Liu, Z. X. (2012). Fairly sharing the credit of multi-authored papers and its application in the modification of h-index and g-index. *Scientometrics*, 37-49.

Fang, H., & Liu, Z. X. (2012). Modifying h-index by allocating credit of multi-authored papers whose author names rank based on contribution. *Journal of Informetrics*, 557–565.

Gephi. (2014, 1 17). Retrieved from Features: http://gephi.github.io/features/

Google Research. (2015, 1 28). Retrieved from About Fusion Tables: https://support.google.com/fusiontables/answer/2571232?hl=en#host

Hagen, N. T. (2010). Harmonic publication and citation counting: sharing authorship credit equitably – not equally, geometrically or arithmetically. *Scientometrics*, 785–793.

Hirsch, J. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 741-754.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*, 16569–16572.

Hodge, S. E., & Greenberg, D. A. (1981). Publication Credit. *Science*, 50.

Jones, W. P., & Furnas, G. W. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the Association for Information Science and Technology*, 420-442.

Kosmulski, M. (2012). The order in the lists of authors in multi-author papers revisited. *Journal of Informetrics*, 639–644.

Library Guides at University of Washington Libraries. (2013, 7 14). Retrieved from Scopus vs. Web of Science: http://guides.lib.uw.edu/c.php?g=99232&p=642081

Lindsey, D. (1980). Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship. *Social Studies of Science*, 145-162.

Markou, M. E. (2015, 09 14). *GitHub*. Retrieved from mmarkou: https://github.com/mmarkou/

Nanas, N., & Vavalis, M. (2008). A "Bag" or a "Window" of Words for Information Filtering? *5th Hellenic conference on Artificial Intelligence: Theories, Models and Applications* (pp. 182 - 193). Berlin, Heidelberg: Springer-Verlag.

Nanas, N., Uren, V. S., & de Roeck, A. (2004). Nootropia: A User Profiling Model Based on a Self-Organising Term Network. *Artificial Immune Systems, 3239*, 146-160.

Nanas, N., Uren, V., & Roeck, A. (2003). Building and Applying a Concept Hierarchy Representation of a User Profile. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 198-204). New York: ACM.

Nanas, N., Vavalis, M., & Houstis, E. (2010). Personalised news and scientific literature aggregation. *Information Processing and Management: an International Journal, 46*(3), 268-283.

Python. (2014, 6 7). Retrieved from Python 3.5.0 documentation: https://docs.python.org/3/

Řehůřek, R. (2014, 7 24). *Gensim topic modelling for humans*. Retrieved from API Reference: https://radimrehurek.com/gensim/apiref.html

Rew, D. A. (2010). SCOPUS: Another step towards seamless integration of the world's medical literature. *European Journal of Surgical Oncology (EJSO)*, 2–3.

Salton, G. (1989). *Automatic Text Processing: The transformation, analysis, and retrieval of information by computer.* Boston: Addison-Wesley Longman Publishing Co.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill Book Company.

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 613-620.

Schreiber, M. (2008). A modification of the h-index: the hm-index accounts for multi-authored manuscripts. *Journal of Informetrics*, 211-216.

Schreiber, M. (2010). Twenty Hirsch index variants and other indicators giving more or less. *Annals of Physics*, Berlin.

Sekercioglu, C. H. (2008). Quantifying Coauthor Contributions. *Science*, 371.

Shena, H.-W., & Barabási, A.-L. (2014). Collective credit allocation in science. *PNAS*, 12325–12330.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized h-index for Disclosing Latent Facts in Citation Networks. *Scientometrics*, 253-280.

SQL Server 2012. (2014, 5 14). Retrieved from Product Documentation: https://technet.microsoft.com/en-us/library/hh995091(v=sql.10).aspx

Taylor, M., & Thorisson, G. A. (2012). Fixing authorship – towards a practical model of contributorship. *Research Trends*(31), 3-6.

Thada, V., & 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*, 202-205.

The Porter Stemming Algorithm. (2014, 5 10). Retrieved from The Porter Stemming Algorithm: http://tartarus.org/~martin/PorterStemmer/

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 141-188.

Wilson, B., & Fenner, M. (2012). Open Researcher & Contributor ID (ORCID): Solving the Name Ambiguity Problem. *EDUCAUSE*, 54-55.

Zuckerman, H. A. (1968). Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity. *American Journal of Sociology, 74*(3), 276-291.

Zuckerman, H. A. (1968). Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity. *American Journal of Sociology*, 276-291.

Accounting authorship of a scientific paper is a widely recognized as a hard problem. Attempts to solve this problem with existing conventional tools encounter insurmountable obstacles. Along these lines Nature began in 2010 (Assessing assessment, 2010) an ongoing conversation concerning the metrics to measure and assess scientific performance. This effort is not only still running but also created additional momentum. In particular, it is now apparent that the use of metrics to assess the value of scientists is unavoidable. So the quest for the best measure possible is surely justified (Count on me, 2012).

In (Nanas, Vavalis, & Houstis, 2010) novice authorship taxonomies have been proposed (Taylor & Thorisson, 2012) that ensure the clear and unambiguous declarations of authorship while heretic arguments like the one claiming that ambiguity is not entirely a bad thing in science (Zuckerman, 1968) have been also appeared in the literature from very early.

It has therefore become evident that the current scheme employed in scientometrics appears to be most probably problematic and perhaps unfair. Within this context this research aims to assess authors' participation in the recorded research activity through developing alternative assessment ways. Instead of using common quantitative metrics, the present study proposes and utilizes the developing of multi-faced-dynamic author profiles. Furthermore, Data Mining and Knowledge Management will compose an effective mechanism to support the theoretical background, the practical significance as well as the intended methodology. The design, the development and the evaluation of a software tool will also contribute to the application and evaluation of the designed author profiles and to the reliability of the obtained results.