


Scientific footprints in digital libraries

Claudia Ifrim¹, Xenia Koulouri², Manolis Wallace², Florin Pop¹, Mariana Mocanu¹, and Valentin Cristea¹

¹ Faculty of Automatic Control and Computers
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania 060042
claudia.ifrim@hpc.pub.ro, florin.pop@cs.pub.ro
<http://acs.pub.ro/>

²  Knowledge and Uncertainty Research Laboratory
Department of Informatics and Telecommunications
University of the Peloponnese
Tripolis, Greece 22 131
gav@uop.gr
<http://gav.uop.gr>

Abstract. In recent years, members of the academic community have increasingly turned to digital libraries to follow the latest work within their own field and to estimate papers', journals' and researchers impact. Yet, despite the powerful indexing and searching tools available, identifying the most important works and authors in a field remains a challenging task, for which a wealth of prior information is needed; existing systems fail to identify and incorporate in their results information regarding connections between publications of different disciplines. In this paper we analyze citation lists in order to not only quantify but also understand impact, by tracing the “footprints” that authors have left, i.e. the specific areas in which they have made an impact. We use the publication medium (specific journal or conference) to identify the thematic scope of each paper and feed from existing digital libraries that index scientific activity, namely Google Scholar and DBLP. This allows us to design and develop a system, the Footprint Analyzer, that can be used to successfully identify the most prominent works and authors for each scientific field, regardless of whether their own research is limited to or even focused on the specific field. Various real life examples demonstrate the proposed concepts and actual results from the developed system's operation prove the applicability and validity.

Keywords: Research Impact, Citations, Publication Medium, Digital Library, Google Scholar, DBLP

1 Introduction

Electronic and online publishing has brought about a revolution in science [32]. Access to other people's work is now faster, easier and more universal than

ever [33]. But new economic models originating from this trend have helped scientific publishing evolve into a lucrative business, and now a huge volume of scientific texts is added to existing literature daily [10]. As a result, although each individual text is more accessible than before, studying the literature and getting or maintaining a clear view of the state of the art of a specific field remains a challenging task; only the challenge has now shifted from the acquisition of access to the papers to the selection of the right papers to focus on amongst the numerous published articles.

To put this in a more specific context, it is safe to assume that almost every reader going through the pages of this journal has had to at some point in the past, or is currently trying to, identify and study the most important researchers or papers in their field; this could, for example, be the fundamental first step towards a PhD [40]. And although powerful indexing and searching tools exist, such as DBLP [21], Google Scholar [59] or ScienceDirect [60], the way to efficiently search in such a large information space is not a straightforward one.

The difficulty stems from the type of indexing that such systems apply, which is quite different from what is required for the task at hand. Papers are indexed by their titles and journals they are published in; authors are indexed by the papers they have published and the keywords they use to characterize their own research interests. But all of these (titles, journals to submit to, description of interests) are determined by the authors based on a priori preferences and are not necessarily closely related to the a posteriori information regarding the actual areas that their work has an actual impact on. For example, authors may be more inclined to choose a shorter, rather than a lengthier and more accurate title for their work in order to maximize its impact [46], or in order to maximize their perceived activity in a specific field.

An example that is close to heart for some of this paper's authors is that of Human Computer Interaction. There exist of course important conferences and journals that focus on this field, and important researchers who list it as their primary focus. Still, some of the most prominent scientist working on HCI are psychologists (who list psychology as their only expertise) and the field's seminal papers have not been published in HCI related journals. Thus, current indexing and searching systems would fail to support a user in the identification of the key papers and researchers of the field.

To overcome this, we propose herein an alternative indexing approach that focuses on papers' and researchers' impact, not as defined by themselves but rather as assessed by their "footprints" in digital libraries, i.e. the specific areas of impact as indicated from citations. Our analysis is based on the detailed examination of citation records and combines information from multiple sources; DBLP and Google Scholar are the sources considered in this work, but extension to include more sources is straight forward.

More specifically, in order to overcome the subjective nature of a paper's metadata, such as the title and keywords that are selected by the authors themselves, we base our analysis on the publication medium (specific journal or conference) which provides a more objective estimation of the broader thematic

scope [71]. Then, by examining which works cite each paper we can estimate the specific areas in which it has had an impact. This allows us to develop paper and author impact indices, that can be thematically searched, thus supporting queries that existing systems are no able to handle.

The main contributions of this paper are: the definition of a researcher's scientific footprint, a methodology to detect footprints in an objective and automated manner based on the analysis of citations, an extensible architecture that employs the notion of the footprint and is able to consider multiple information sources and the Footprint Analyzer, a preliminary implementation of the above. This paper is based on, and constitutes a combination and major extension of, paper "Agile DBLP: A Search-based Mobile Application for Structured Digital Libraries" presented at the 1st International KEYSTONE Conference (IKC 2015) [35] and paper "Extracting and visualizing research impact semantics" presented at the 9th International Workshop on Semantic and Social Media Adaptation and Personalization [71].

The remainder of this paper is organized as follows. In Section 2 we discuss existing approaches to the assessment and quantification of scientific impact and in Section 3 we review digital scientific libraries. In Section 4 we examine the types of contextual information related with citations, a notion upon which we base our definition of "footprints" in Section 5. In Section 6 we extend the notion into an algorithm, listing the steps required to generate semantic footprint indices, and in Sections 7 and 8 we present the integrated system incorporating these notions and discuss some preliminary yet indicative results. Finally, in Section 9 we list our concluding remarks.

2 Scientific impact

Scientific value and scientific recognition are subjective in their very nature, and often even random. There is no objective way by which to measure the degree of novelty or importance of a scientific proposition, and even the way it is perceived by the scientific community is not always to be trusted. A characteristic example is that of Dr. Zadeh's seminal paper on fuzzy sets [34], which was rejected and refused publication by three different journals, but has now defined not just a new subfield in applied mathematics but more importantly a whole new paradigm in scientific computing and computer engineering.

Still, a need exists to quantify the importance of scientific work; for example when wishing to comparatively assess candidates for academic positions. In order to overcome the highly subjective and unreliable nature of the related a priori information, the value of scientific work is evaluated based on the a posteriori information regarding its impact. For example, modern day Nobel prizes are decided primarily based on the actual impact candidate works have had on society and science, and the time period for the full impact of the work may be several decades; Chandrasekhar famously shared the 1983 Nobel prize in physics for work done in 1939 [15].

Although randomness and unfairness can still be claimed (potentially revolutionary works may go un-noticed and less deserving works may receive attention due to random shifts of the market's direction or simply because of an inspired title), at least objective (numerical) measures may can be defined. Despite known inherent weaknesses [27], citation counts are seen as the most trusted indications of scientific impact. An important driving factor for this is that they are quantitative and are easily and readily available in online systems such as Google Scholar.

2.1 Paper impact

Thus, a paper's impact is quantified as the count of citations it has received from the day it was published and up to the day of examination. This, of course, favors papers that were published many years ago, as they have been accumulating citations for a longer period of time. This is not seen as a weakness of the measure; it is only natural that works that have been around for a long time have had the opportunity to have a greater impact on the works of others.

Besides, it has been observed that the yearly count of citations received by a paper diminishes after a few years; so, after some time, the advantage of earlier papers is diminished.

2.2 Journal impact

Similar ideas are applied towards the evaluation of the scientific value of a publication medium, such as a journal, magazine or conference. There is, though, an important difference originating in the way to use the results of this evaluation.

Journals are not evaluated in order to assess which one has had the greatest overall impact on the scientific world. To the contrary, the goal is to assess the probability that an article published in a journal will make an impact in the future; readers consider this evaluation to select the journals to read and more importantly authors consider it in order to select the journals to submit to, thus maximizing the potential of their work. Therefore, the number of years that a journal has been publishing, or even the number of volumes per year or the number of articles per volume cannot be allowed to affect the evaluation.

The impact factor (IF) is the most trusted quantification of a journal's scientific potential. It is computed as the average count of citations articles published in the journal receive in the first two years after their publication; some limitations apply regarding the sources of these citations. It is clear to see that the IF is configured in a way that favors journals that publish carefully selected high quality articles, which is in accordance with the goals of journal evaluation.

2.3 Author impact

Researchers' impact (and sometimes future potential too) is also assessed based on citations. The first, most common and straightforward approach is the consideration of the cumulative number of citations for the complete list of their published work.

But given the highly competitive nature of the scientific community, it is rather expected that the prime tool to assess and compare researchers has received a lot of attention, both in the form of criticism of its objectivity and in the form of attempts to affect its outcomes. Numerous weaknesses have been identified, related to the number of years of activity, the effect of cooperation networks, self-citations, outlier works, frequency of publication etc.

In order to deal with the aforementioned weaknesses of using citation count as a metric, numerous more elaborate metrics have been designed, such as the following.

Average number of citations per paper Using the average number of citations aims to compensate for the fact that some authors publish more papers and this leads them to have higher total number of citations, where in fact each of their individual works may be cited rarely. It also compensates for differences in the duration of the career, i.e. in the number of years researchers have been publishing.

Average number of citations per author Single author papers and cooperative works do not indicate the same level of personal involvement in the work. Thus, it makes sense to distribute the count of citations for each paper equally to the contributing authors. This is not the only approach in this direction; sometimes a greater part of the contribution is assigned to the leading author, or the amount of contribution is gradually reduced based on the author's position in the author list.

Average number of citations per year Researchers that have had a longer publishing career, have inevitably produced more work. This does not necessarily indicate that their research is more important than that of younger researchers. A workaround is to average citations over the count of years, which allows for a more fair comparison of veteran and new researchers.

***h*-index and similar indices** Hirsch's *h*-index is the most well known and widely used metric after the citation count. An index of *h* indicates that *h* distinct papers of a given researcher have at least *h* citations each. Variations of the *h*-index emphasize different features, such as the number of authors in each paper [22] or the average value of the *h*-index over the years [69]. In [11] we see a partitioning of the *h*-index into *h*₁, *h*₂ and *h*₃, which help discriminate between different types of researchers such as the perfectionists and the mass producers.

The *g*-index and the *e*-index are extensions of the *h*-index. In the *g*-index the citation count is averaged [25] and in the *e*-index the square root of citations in the *h*-set beyond *h*², is considered, i.e. square root of citations beyond the minimum number of citations required to achieve an *h*-index of *h*. The *e*-index is particularly useful when comparing researchers who have the same *h*-index [66].

***s*-index** The *s*-index is based on the notion of entropy and provides a better basis for comparisons between researchers than *h*-index in the case of researchers who many citations [41].

***i*-10 index** Google Scholar's *i*-10 is the count of articles that have ten or more citations[29].

3 Digital libraries

There are various digital libraries of scientific content, that interested readers may turn to in order to search for and gain access to a specific paper; IEEEExplore, ScienceDirect, SpringerOnline to mention just a few. But these are not the focus on this work. In this work we focus on digital libraries that can be used to assess scientific importance or equivalently, as explained in the previous section, scientific impact. Therefore, libraries such as the above (that only index content of partner publishers) are not a suitable source of information.

Others exist, on the other hand, that aim to generate an author's complete list of published works, regardless of where they have been published. The following are stand out examples of this category; DBLP due its highly accurate and well curated content regarding lists of published works by each author and Google Scholar due to the extensive and all inclusive citation lists it provides. We review then both below, as they form the basis for the system presented herein.

3.1 Google Scholar

Google Scholar is a web service engine provided by Google which indexes the full text of scholarly literature across various disciplines and publishing formats. Since its release in November 2004, Google Scholar has become one of the most popular academic search engines.

Although most academic databases and search engines allow the ranking of the results by certain factors, the Google Scholar ranking algorithm is still unknown to this date. According to various studies that have tried to reverse-engineering the algorithm, Google Scholar arranges results by putting weight especially on citation counts [9] and the occurrence of search terms in the articles title [8].

Since it was first introduced in November 2004, there has been abundant literature regarding the weak and strong points of Google Scholar. So far, the studies have varied in their approaches, differing from the analysis of the user interface functionality to the content covered by the search engine. In 2010, Xiaotian Chen did an empirical study of Google Scholars coverage of scholarly journals five years after a similar study was performed. His findings showed a dramatic improvement: using the same database, Google Scholars coverage has gone from an average of 60 to a range from 98 to 100 percent [16].

On the other hand, another interesting study conducted in the same year [7] showed that Google Scholar is far easier to spam than the Google Search for Web

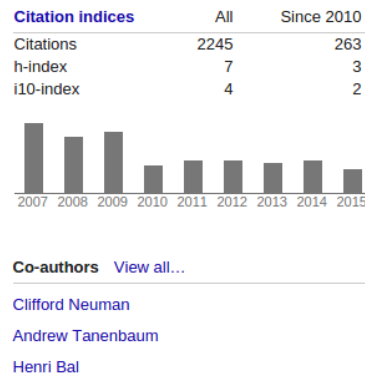


Fig. 1. Citation indices and Co-authors.

Pages. For example, it was demonstrated that Google Scholar counts references that were added to modified versions of already published articles, meaning that researchers could increase citation counts and rankings of the cited articles in order to increase their visibility on Google Scholar. Moreover, several studies have pointed out that the Google Scholar enforces the Matthew effect (sociology: "the rich get richer and the poor get poorer") by placing the high cited papers on top of the search results [9].

Google Scholar has a familiar search interface, similar to the classic Google Search. The search results are based on the terms/keywords typed in the search box. Within a Google Scholar search results, the following features are available:

- Abstract
- Cited by: Returns the list of articles that have cited the current article.
- Related articles: Returns a list of articles similar to the current article, ranked primarily by similarity but also by taking into account the relevance of each paper.
- All versions: Returns the list of all alternative sources for the current paper.
- Import citations: BibTex, EndNote, RefMan, RefWorks

Also, the following bibliometrics are available on each author profile:

- Co-authors
- Citations
- h-index
- i-10 index: The number of publications that have received at least 10 citations.

Metadata Metadata cannot be easily obtained through Google Scholar: scraping is not allowed and the data is not exposed through an API.

Information retrieval P.Jacso, in his paper "Metadata mega mess in Google Scholar" [38], discusses some of the problems that exist with Google Scholar, particularly the incorrect field detection mechanism. The author found that Google Scholar is especially "bad for metadata based searching when, beyond keywords in the title, abstract, descriptor and/or full text, the user also has to use the authors name, journal title and/or publication year in specifying the query".

Although many may argue that the "mess" can actually come from publishers and vendors, the author also pointed out that the Google Scholar developers decided not to use the metadata readily available from most of the scholarly publishers [48].

3.2 DBLP

DBLP is a digital library for computer science bibliography supported by University Trier from Germany. This project started in 1993 as a experimental server meant to test web technology, but evolved continuously, based on ad hoc solutions. The project policy is to keep the application as stable as possible. For example, URLs are only changed if they prevent an important functionality, and not because the people simply perceive them as unaesthetic.

In June 2015, DBLP indexes more the 3 million publications from major information sources: VLDB³, IEEE transactions⁴ and ACM transactions⁵.

In comparison with Google Scholar or CiteSeer⁶, that crawl the web to extract metadata from publications in order to operate their journal collections, the DBLP collections are maintained with great human effort by having the data inserted manually. One of the consequences of this is that authors are disambiguated more accurately.

The complete DBLP dataset is available at <http://dblp.uni-trier.de/xml/dblp.xml>. This project has evolved from an experimental Web Server to a popular digital library service for the computer science community, but it's documentation is limited.

In the paper "DBLP - Some Lessons Learned", Michael Ley described the evolution of DBLP from the data modeling point of view. Apart from being available online, the DBLP database can be also downloaded as a large XML file (<http://dblp.uni-trier.de/xml/dblp.xml.gz>) and a schema is available as a DTD file, making it easy for researchers to integrate the data in their work (over 400 publications mention the use of DBLP for a variety of purposes [47]).

Above, you have a graphical representation of sample-information extracted from [19] in March 2012. It allows access to relatively static and limited information, showing only the publications (nodes with green color - 10 co-authored or more), and authors (orange nodes - 200 publications or more), and their interconnectivity. This interconnections show only static information, because it

³ <http://www.vldb.org/>

⁴ http://www.ieee.org/publications_standards/publications/services/journals.html

⁵ <http://dl.acm.org/pubs.cfm>

⁶ <http://citeseerx.ist.psu.edu/>

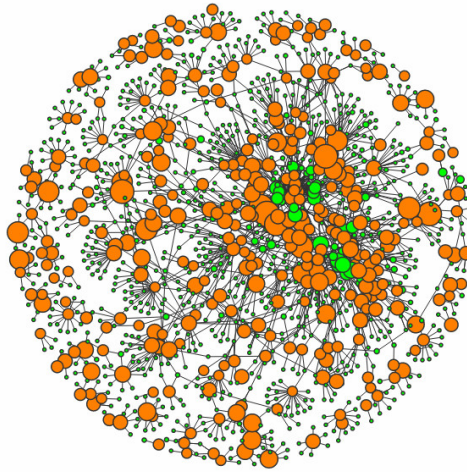


Fig. 2. DBLP Graph.

```

<dblp>
  <article key="journals/computer/TanenbaumHB06" mdate="2006-11-10">
    <author>Andrew S. Tanenbaum</author>
    <author>Jorrit N. Herder</author>
    <author>Herbert Bos</author>
    <title>Can We Make Operating Systems Reliable and Secure?</title>
    <pages>44-51</pages>
    <year>2006</year>
    <volume>39</volume>
    <journal>IEEE Computer</journal>
    <number>5</number>
  </article>
  <url>http://doi.ieeecomputersociety.org/10.1109/MC.2006.156</url>
  <url>db/journals/computer/computer39.html#TanenbaumHB06</url>
</dblp>

```

Fig. 3. Example of DBLP XML record.

is only one moment in time; it is also a limited information, because it doesn't show the entire population of the database, but only the higher values. Furthermore, it doesn't have a statistical value, as it is a sample which not relevantly chosen from the entire population.

Metadata The complet DBLP dataset is exposed in JSON and XML formats.

DBLP contains the following types of entries: article, book, in proceedings, in collection, masters/ph.d thesis, proceedings, www and provides medata like: title, author, pages, volume, journal, publisher, year. Metadata for publications are available in BibTex format and articles are identified by URIs like this: <http://dblp.unitrier.de/rec/bibtex/journals/computer/TanenbaumHB06> Access to the bibliographic metadata is available to everyone as of 2011 (relased under ODC-By).

Although the data retrieved from the DBLP database is well-formatted and the author names are disambiguated, it lacks citation references, having no indicators in assessing the relevance of the papers. Also, DBLP is limited to com-

puter science and, as pointed out by another study in [58], it does not cover all sub-fields of computer science to the same degree.

Information retrieval CompleteSearch DBLP is a tool that provides extended search capabilities for DBLP. The following features are available: phrase search, prefix search, exact word match, only first-authored papers, specify number of authors etc.

4 Citation context

In section 2 we saw how metrics based on citations are used to evaluate and quantify scientific impact. But these metrics do not make any distinction between citations, thus failing to consider contextual information. In this section we take a closer look at two different aspects of context in citations and discuss how they can be used in order to define new, more information rich citation metrics.

4.1 Role of referenced work

In each paper there are various citations, but they do not all have the same role. Whilst some citations may provide the theoretical and technical foundation of the presented work, indicating a true connection between the papers, others are used to compare results or just in the discussion of previous related, or broadly related, work. To further emphasize this, in the current paper we have split the references in two sections. The first part lists works that relevant to the work presented, whilst the second part lists works that are mentioned in the discussion but whose content is never discussed. In a conventional citation count both parts of our references would be counted as equally important, which is clearly unfair to the authors of the papers listed in the first part.

To deal with this, we should be able to first determine and then consider the role that each entry in the references has in a paper. We meet a similar concept in CiteSeer^X, the “citation context”, but that is only presented as additional information; it does not affect the calculation of the number of citations [17].

This would be a tenuous task that cannot be performed automatically considering the current state of the art in text analysis and understanding. It requires the work of human experts in the area of the considered papers who will study the papers and evaluate what the exact role of each reference is. This is not only expensive, it also introduces subjectivity, as it is up to the expert to decide the degree to which a citation reflects actual contribution. Therefore, this type of context cannot be practically considered, at least with the current state of the art.

An alternative consideration of context could focus on the part of the text where the citation is referenced. Citations, for example, in the introduction and the section on related work typically have had little or no impact on the considered work, citations in the results are typically used as benchmarks whilst citations in the description of the proposed methodology have most probably

been used as a methodological basis and are the ones with the highest true impact. This is also not possible to apply in an automated manner, and in some cases not even in a manual manner, as on the one hand not all sections in a paper can be clearly identified as related work, methodology or results, and, on the other hand, sometimes some references are not at all listed in the text which makes it impossible to determine their citation context.

The unfortunate conclusion is that, although this type of contextual information could provide very rich information regarding the nature and value of citations, it is quite improbable that it will be used in citation analysis, for the reasons explained above.

4.2 Scientific scope

In the previous subsection we discussed how we could evaluate the importance of a cited work in a paper. In this section we examine not the importance but the topic of the citation.

Many papers are monothematic. But there are also those works that are interdisciplinary and/or rely on ideas from different scientific fields. When examining a paper's references we can understand if it is interdisciplinary by examining the topic of each cited work, but it is debatable whether that would give some important insight regarding the importance of the paper. But the reverse is a lot more interesting: by examining the topics of the papers that reference a work we can see which scientific areas have been affected by that work.

Of course there is a practical question here: how could one determine in an automated and reliable manner the scientific scope of any given paper. The listed Keywords could be useful when existing, but they are not always standardized - some authors write in their own without choosing from a predetermined list - and in many publications keywords are not used at all. Paper titles could also be used, but they are often misleading. Abstract texts are less misleading, but the current state of the art in text analysis and understanding is not mature enough to provide reliable results when applied to short texts without any additional contextual information.

The publication medium (the journal, conference, edited book etc in which a paper is included) can provide reliable evidence regarding the scientific scope of the work. When a submission is considered for publication, either by a journal or by a conference, one of the first and most important checks is whether it falls within the thematic scope; scientific quality is examined secondly. Therefore, the editorial process guarantees that, for example, papers published in LNCS Transactions on Computational Collective Intelligence are related to CCI and papers included in this special issue are additionally related to keyword searching in Big Data.

Almost all conferences, journals etc, in short publications that follow an editorial process, have clearly defined scopes and often also provide potential authors with lists of relevant topics. And for those that do not have such information readily available it is quite easy to produce them manually. Especially

since this would be done only one time for each publication medium. Therefore, this can provide the basis for an automated and objective (i.e. without a human experts examining the article and providing a subjective evaluation) consideration of the scientific scope of a published paper.

Moreover, this is also interesting in the scope of citation analysis, as it provides richer insight into the impact that a paper has had.

5 Scientific Footprints

In this work we look more closely at citation records, examining scientific scope, in order to acquire deeper insight on researchers' impact. In order to better explain what type of insight we are looking at, we start by listing below details from the citation records of two highly cited authors, namely Dr. Cynthia Whissel and Dr. Theodore Simos.

5.1 Two indicative examples

Cynthia Whissell Cynthia Whissell is a professor in the Psychology Department at Laurentian University and she is a psychologist. In her own description of her research interests she lists language and the way language conveys emotion [50]. Therefore, based on studies, professional affiliation and title, as well as on her own description of herself, professor Whissell works in psychology and linguistics. It would only be natural for one to expect the impact of the work of professor Whissell to be in those fields as well.

Looking at her citation list we quickly identify paper *The dictionary of affect in language* [51] as her seminal work, as it has received by far the most citations and considerably more than her next most cited work. Looking at the title, the abstract or even the content of the paper we can determine that this specific work is also in the field of psychology and linguistics. As far as the publication medium is concerned, it is included in a book titled *Emotion: Theory, Research, and Experience*, again clearly in the field of psychology. If we examine more of her works we will reach similar conclusions; overall there is nothing to imply that the work and expertise of professor Whissell might be of interest to scientific fields other than psychology and linguistics.

But when we examine her citation list, there are some interesting surprises. For example, there is this paper:

S. Soroka, M.A. Bodet, L. Young, B. Andrew, *Campaign news and vote intentions*, Journal of Elections, Public Opinion and Parties, no. 19(4), pp. 359-376, 2009.

The title of the paper and the name of the journal point towards politics. When reading the paper we find that the content of the paper is also focused on politics. Linguistics and psychology are briefly considered in the analysis, but they are the tools, not the core subject of the work. If we study this paper more

carefully we find that the role of Whissell's paper is fundamental in the design and application of the work. In other words, Soroka et al's paper shows that professor Whissell has made through her work an impact not only in psychology and linguistics but also in political analysis.

This paper is neither an outlier nor an exception. In the citations of the aforementioned paper of professor Whissell we find evidence that it has made an impact in psychology [49] [4] [5], biology [57], affective computing [24] [64] [65] [63], artificial intelligence [44] [75] [73] [28] [72] [56] [62] [36], multimedia and image processing [37] [55] [67], speech and linguistics [20] [18] [31] [68], management [26] [1] [12] [23] [42] [70], music [14] [6], gender [45], politics [52] [74] [53], bilingualism [2] [3] and more.

Theodore Simos Theodore Simos is a professor in the Department of Informatics and Telecommunications of the University of Peloponnese. His studies include a bachelor degree in engineering and a PhD in mathematics. His teaching and research is in mathematics and on his homepage he describes himself as a researcher of mathematics [61].

At [61] we can find a list of citations for professor Simos. Almost all of the papers listed in it are published in journals and conferences in the fields of mathematics, computational chemistry and computational physics, in other words solely in theoretical and applied mathematics.

5.2 Following the footprints

What the above examples indicate is that there are different types of scientific impact. Professor Simos's work has a very deep impact in mathematics (he has more than 2000 citations in the field) but little or no impact outside that field. Professor Whissell's work on the other hand has a very broad impact in science which is not limited to psychology and linguistics. The question is, what consequences could this observation have on the design and development of digital libraries and the indexing/searching mechanisms that support them.

The answer lies with Dr. Whissell's example. We have already seen that her work is relevant to fields outside psychology. What we have not seen, because it could not be seen in the independent analysis of her citation records, is that her work is actually important in other fields. For example, researchers from the field of affective computing will be quick to identify Dr. Whissell as not only relevant but also central in the literature related to facial expression recognition.

Dr. Whissell's great impact in the field of affective computing has left its trace, or "footprint", in digital libraries. Specifically, if we were to analyze all citations found in papers of the field of affective computing, we would find a disproportionately high number of references to the work of Dr. Whissell.

The work presented herein is based on this very notion of footprint in digital libraries. Specifically, by examining citation records we aim to detect and index the footprints of all works and authors. Thus, our system will be able to handle queries not only of the form "Who are the most cited researchers of HCI" which

limit results to those working in the field but also of the form “Who are the most cited researchers *in* HCT”. This will allow for the automatic detection of the most important people (or articles) for a field and will also facilitate people, such as first year PhD students, who wish to get acquainted with a new field by reading the most important works in it.

6 Semantic citation analysis

In order for the aforementioned notions to be put into application, footprint indices are required, linking each article and each author to their respective areas of impact.

As we have already explained, most metadata linked to an article are inherently subjective and unreliable, as they are defined by the authors and most commonly are not independently reviewed and verified; the most reliable source of semantic information regarding the articles is the publication medium itself, as the editorial process involves a rigorous control of the thematic scope. For this reason our analysis is based on the examination of the publication medium.

Of course, this is not a trivial task. The way to describe the thematic scope of journals, magazines and conferences is not standardized, in many cases there is only a textual description without keywords, there may be huge differences in the breadth of the thematic scope, etc. Even the compilation of a comprehensive and universally accepted list of scientific scopes is hard to achieve. Overall, our approach for the population of the footprint indices includes a preparatory step and processing steps for articles and authors, which are further analyzed in the following sections.

6.1 Preparatory steps

The preparatory steps involve the establishment of the knowledge base that is required for the execution of the processing steps, as follows:

1. Develop a list of thematic areas
2. Compile a list of publication media (journals, magazines, conferences)
3. Assign thematic areas to each publication medium

Thematic areas. The list of scientific fields does not change; or at least it does not change often and drastically. Therefore a reasonable first step is to develop this hierarchy. Existing hierarchies exist that may be considered as a basis, as for example the one found in [30] or [13].

Wikipedia classifies sciences in 4 main fields. Natural sciences (sciences that study the laws that condition the nature), formal sciences (sciences that have a specific methodology, instead of what actually happens in reality), social sciences (which study the human and the society behavior), and applied sciences (which implement scientific knowledge for practical purposes) [76].

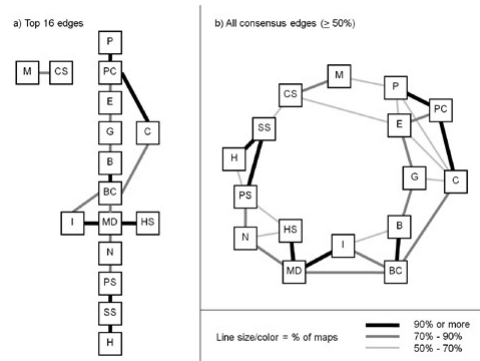


Fig. 4. Links of Scientific Fields.

Another hierarchy of sciences which we can find at Physics Portal at South Carolina State University categorizes them in three main classes: formal, natural and humanistic science. The 3 classes can be divided to 6 more subclasses: mathematics, logic, physical, biological, behavioral and social [54].

We are going to choose as our basis the classification we found in paper has been visually "mapped" according to various branches of sciences [43]. According this paper sciences are divided in more main classes which is comfier and more useful in our paper, in order to categorize the list of means of publication, according to this hierarchy. The hierarchy that we are going to use contains 16 main divisions each of ones has its own smaller classes (subdivisions). The 16 main classes are: mathematics, physics, chemistry, physical chemistry, biochemistry, computer sciences, engineering, earth science (geoscience), infectious disease, medical specialties, brain research, health services, psychology, social sciences and humanities.

We can observe that these 16 fields are dependant. There are areas where they can be overlapped in the map of sciences or there can be sciences that they are linked by different ways. Fields that can be enlisted in more than one classes are name interdisciplinary fields. In the map of sciences, the edges between the scientific fields demonstrate the links of sciences.

Publication media. We have used DBLP metadata in order to acquire a first list of previous and running journals and conferences, knowing that although this list is long it is far from complete. A comprehensive list of publication media is not easy to establish. Moreover, the list is not static as some conferences disappear whilst new ones appear every year; there are similar changes to the list of journals, but they are less frequent and thus easier to tackle.

Therefore the pre-processing step regarding the acquisition of publication media is not meant to produce a complete and finalized list. As we will explain in the following, the list of media can be updated during the processing steps;

the role of this pre-processing step is to facilitate the initiation processing steps by dealing with the problem of cold start.

Medium to area assignments. Although the DBLP metadata are carefully curated, they do not contain semantic information regarding the thematic scope of the included publication media, other than their title. This title is often, but not always, enough to have a rough idea of the thematic coverage.

In order to overcome this, we follow a semi-automatic approach to the identification of the links between publication media and the thematic areas: we start by applying an automated string matching approach to identify probably links (differences between UK and US spellings, synonyms etc limit the success of this step) and continue with a manual step during which the journal's or conference's site is examined and the description of the scope is used to determine which thematic areas are most and truly associated with it.

The initial step is to collect all the publication media that are shown up in the record of the list or references. Afterwards, we checked each of their names and we studied the subjects which they were dealing and that they are published each year. In that way, we choose which sciences and which scientific field each publication deals with. So we took into consideration the hierarchy of sciences that we created and we exhibited in section 6, in order to match the correspondences of sciences that we found in our previous step, in our hierarchy.

6.2 Processing steps, for each work

For each considered article, we examine the list of citations as follows:

1. Acquire the list of citations
2. Identify the thematic area of each citing work
3. Aggregate findings and populate the article footprint index

List of citations. We use Google Scholar to acquire comprehensive lists of citation for the articles that we examine. Of course, since the database is not curated, there are numerous errors (false positives, incorrectly assigned fields, damaged titles, repetitions etc).

In a parallel work (which is not yet completed and is expected to be sent for publication in 2016) we are examining the validity of Google Scholar by comparing its automated results with manually established lists of citations. Our early findings indicate that, although error rates are high (often exceeding 20%), the deviation is small. Thus citations retrieved by Google Scholar are a relatively reliable source given that the error rate is similar for different articles and authors.

The acquisition of the list is non-trivial, as the system does not provide an open version of the data or a freely accessible API. Quite the contrary, there are provisions to block repeated access in order to prohibit robotic crawling. Our



Fig. 5. Google Scholar.



Fig. 6. HTML source.

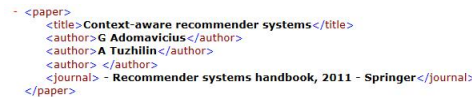


Fig. 7. XML document.

approach includes an automated tool that queries the website and processes the HTML results in order to extract citation data.

Different fields are identified, allowing for the establishment of a structured XML document as seen in figure 7. The acquired list is then trimmed, to remove self citations and repetitions, to the extent that they can be identified in an automated manner.

Limitations set by Google results in an upper threshold to the frequency of access and therefore to the speed of crawling. What is permitted is enough for the system presented herein to work as a proof of concept, but a more open access to the data will be required for a full scale application.

Thematic area of each citation. In earlier sections we have explained that we will use the publication medium to identify the thematic scope, we have developed the lists of publication media and scopes and established the associations between publication media and thematic areas. In the previous paragraph we also saw how the publication medium is acquired as a separate field in an XML document.

Unfortunately, this does not imply that this is a simple task. The publication medium included in the XML document comes from the uncurated Google Scholar system, whilst the entries in the publication media to thematic scope association table come from the DBLP system. The titles between the two do not

match an a manual matching step is required. This is a highly time consuming step for the first runs, but as more links (between Scholar titles and DBLP titles) are established the need for manual intervention diminishes.

In the end, each citing article is associated to a publication medium, and by extension to its thematic areas.

Aggregated impact for each work. Conventionally all citations associated with a published work are considered equally and uniformly, and overall impact is given as the count of citations. Given the additional thematic information that now becomes available, a rising question is the validity of considering uniformly references that have been published in a publication medium with an impact on a single science and references whose publication medium influences more than a single science. Our approach is a variable weighting factor for the two cases. In case that the papers influence a single scientific field weight will be equal to 1, whereas the weight will be distributed uniformly when multiple fields are impacted.

The reasoning behind this is not so much that impact is distributed between the different fields, but rather that it is not possible to safely detect the field of impact without firsts performing a deeper semantic analysis, for example by reading the full text of the paper. Thus the reduced weight denotes the reduced confidence regarding the association between the considered citation and the actual field of impact. Distributing the weight differently allows for the overall counts to be unaffected; this is fair, as otherwise works published in interdisciplinary journals and conferences would have a greater (or smaller) weight in the assessment of impact and there is no evidence to support this.

The aggregated impact for each work is given as the sum of weights, for each scientific field; as expected the impact is not calculated as a single number but rather as an array of numbers, one per field.

6.3 Processing steps, for each author

For each considered article, we examine the list of public works as follows:

1. Acquire the list of published works
2. Identify the impact of each work
3. Aggregate results and populate the author footprint index

List of published works. The list of published works is extracted from DBLP, using the provided APIs. We do this knowingly that DBLP is limited to computer science, and thus this first implementation of our approach will serve as a proof of concept. A latter implementation will extend to other fields.

Impact for each work. For each work attributed to the author by DBLP, the impact is readily available in the article footprint index, through the process

of paragraph 6.2. In most cases there are no differences between article titles as they are reported in DBLP and Google Scholar, so querying the index is straightforward.

Aggregated impact for each author of results. In the conventional approach, an author's citation count is calculated as the sum of citations for all of the author's published works. By extension, in our work we calculate an author's impact in each field as the sum of the impact values for that field for all of the author's works. Thus, the aggregated impact for the author is a vector calculated as the sum of the impact vectors of all of the author's published works, as calculated above.

7 Integrated System

In order to automate the steps described in section 6 and to have a consistent and accurate dataset to analyze we developed a prototype, named Footprint Analyzer, that aggregates information from DBLP and Google Scholar. Other digital libraries could (and will in the future) also be considered as sources, but the integration with more sources is outside the scope of this paper and this feature is not included in the current version of the Footprint Analyzer.

Footprint Analyzer is a web application implemented in NodeJS. For a given author name, we extract all the important elements that define a person's scientific interests that we can collect from Google Scholar and DBLP: frequent keywords, conferences attended, collaborative colleagues (co-authors), top 5 most important co-authors and published work.

For citation tracking we use data from Google Scholar. Google's algorithms allow it to retrieve a large number of citations results, which we then clean as they often include duplicates and citations that simply aren't real.

Footprint Analyzer exposes an interface that allows the user to enter an author's name or the title of an article. To control the results, a user can also specify years of publishing activity or restrict to one thematic area.

We populate our database with the list of conferences and their assigned articles, using the list that we retrieve from DBLP.

For each conference we have assigned thematic areas from our nomenclator defined in section 6.

Results are loaded as soon as a user runs a query. They can be sorted by clicking on each label heading of the table of results (footprint index, author, title, year of publication, publication media). The user can uncheck any entry in the results that is irrelevant, a case in which the footprint index is automatically updated.

In order to calculate the footprint index of each author we followed the steps described in section 6.

The main focus of our work is the analysis of an author's impact, but is also possible to use Footprint Analyzer to analyze the impact of a specific article.

The power of any data-mining project lies in the amount of data that can be processed to provide meaningful and statistically relevant information. Merging information from two digital libraries is necessary in order to obtain an accurate scientific footprint of each publication or author; the consideration of more sources in the future can further enhance the accuracy and reliability of the results.

Once extracted, the metadata is stored in a MySQL database. In the current status of the Footprint Analyzer the data inserted in the database is based on XML responses from DBLP and BibTex format for Google Scholar.

7.1 Architecture

The main tasks performed by the implemented Footprint Analyzer prototype are the following:

1. Acquire published authors records
2. Analyze authors domains of interest based on keywords from their publications
3. Establish links between publication media and our nomenclator of thematic areas
4. Analyze citations of each article
5. Remove self-citations and repetitions
6. Retrieve the distributed knowledge from different digital libraries by using exposed APIs and lightweight web-crawling methods
7. Calculate footprint index of an author or journal

For digital libraries that don't expose an API, the following principles of web-crawling are employed by the system:

1. Starts with a set of seeds (i.e., author names, titles) which represent the list of URLs to visit (used in constructing URL requests)
2. Crawler starts fetching pages
3. Result pages are parsed to find link tags that might contain other useful URLs to fetch (e.g., publications, profile pages, etc)
4. New URLs are constructed (child URLs from the initial parent URLs)
5. Continue until all necessary info has been retrieved

Fig. 8 describes the main components of the Footprint Analyzer and the way they interact.

DBLP Module - sends the query to DBLP based on parameters received from the user (e.g. author name, years, etc.), parses the response received from DBLP, saves the information in the MySQL database and sends to **Google Scholar Module** the list of published works for citation tracking.

Google Scholar Module - retrieves and saves in our database information regarding citations of articles returned by **DBLP Module**.

Thematic Areas Mapper - retrieves the list of conferences and published articles and assigns them to thematic areas.

Footprint Index Module - calculates the footprint index for an author or an article as described in section 6 based on results from our database.

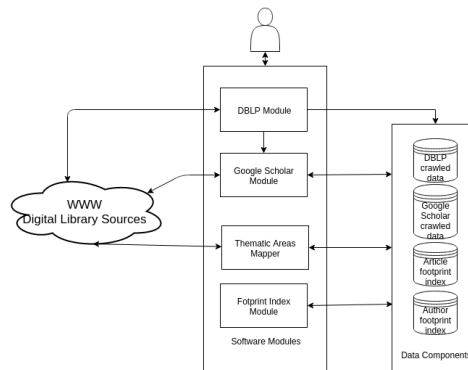


Fig. 8. Footprint Analyzer Prototype Architecture.

7.2 Problems encountered

As part of our project to automatically retrieve and analyse scholarly literature, Google Scholar and DBLP were considered as data sources.

Google Scholar "Google Scholar is a freely available crawled service with a familiar interface similar to Google Web Search" [58]. Google indexes articles from most major academic repositories and publishers. The results retrieved from a Google Scholar search are an important aspect of the research tool we are considering.

Many scholarly publishers, databases, and products offer APIs (application programming interface) to allow users with programming skills to more powerfully extract data to use for a variety of purposes. Some APIs allow programmatic bibliographic searching of a citation database, others allow extraction of statistical data, while others allow dynamic querying and posting of blog content.

Although being probably one of the most useful tools on the web today for academics, Google Scholar does not provide any API or other automated data extraction means. There is a lot of complaining on the web about Google's failure to provide an API for web search, which leaves people writing custom scrapers in Python, Perl, R, etc. As first part of our work, we scrap Google Scholar pages for the extraction and processing of information about authors articles, citations, years of publishing, etc.

Technical details and limitations

First of all, Google's Terms of Service do not allow "the sending of automated queries of any sort without express permission in advance from Google" [48]. In agreement with Google Scholar terms of service which prohibits the automatic querying, only the displayed result page can be processed.

In February 2013 Google Scholar reduced the maximum number of results per page from 100 to 20 results and if an excessive number of queries is detected, Google Scholar will refuse to accept further queries from the requesters IP address.

Since there are no APIs for Google Scholar, we have to parse directly the HTML of Google Scholars response page about the requested author/requested article. Starting from the data retrieved we will make some computations and statistic analysis.

We started by trying to retrieve information from Google Scholar using a custom parser. The plan is to make a query to Google Scholar using a cookie and to be able to retrieve the HTML file with the results.

Generating a cookie is mandatory if we want access to the BibTeX files provided in the search results - the BibTeX entries are not displayed by default and are only showed if they are manually enabled in the Google Scholar search settings. We will need to parse the BibTeX entries as well in order to retrieve the list of author names of a certain publication.

After analyzing the structure of the webpage, we consider the following:

- Number of results retrieved - `gs_ab_md` CSS class
- Titles - `gs_rt` CSS class
- Number of citations - `gs_fl` CSS class
- BibTeX entry

All this information has been extracted by means of mechanisms for regular expressions.

DBLP A record from DBLP data set can be:

- article
- book
- incollection
- inproceedings
- proceedings
- mastersthesis, phdthesis , www

For each record we have one or more of the following metadata fields: title, booktitle, author, editor, pages, year, publisher, address, journal, volume, number, month, cdrom, url, ee, cite, note, crossref, series, isbn, school, chapter.

We can identify the following links:

- a publication is linked to the authors and editors
- a paper is linked to the journal, proceedings or book in which it was published
- citation links are created for each non-empty "cite" element in a publications record.

On <http://dblp.uni-trier.de/>, DBLP provides a primitive form to search for persons inside the bibliography. The Footprint Analyzer accesses DBLP data records by retrieving author information, parsing it and then requesting additional data records based on the processed data.

7.3 Software licensing

As part of our future work, once the Footprint Analyzer has reached the desired level of maturity and stability, we plan to release the full integrated code under a GPLv3 license. Early versions of parts of the code (the module extracting citation information from Google Scholar and the module processing citation lists to extract footprint information) are already released under a GPLv3 license at GitHub [39].

8 Experimental results

As an experimental result we will analyze the author **Cynthia Whissell**, author mentioned also in the example described in section 5.

When we query DBLP on author **Cynthia Whissell** published work we receive two records:

- The Times and the Man as Predictors of Emotion and Style in the Inaugural Addresses of U.S. Presidents. *Computers and the Humanities* 35(3): 255-272 (2001)
- Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities* 30(3): 257-265 (1996)

The first article has 16 citations and the second one 43. After retrieving the lists of citations for each record and removing irrelevant records, the footprint index module has all the data in order to calculate the footprint index for our author.

Our thematic areas mapper identified three thematic areas for our author: psychology, social science and computer science.

For the first article we have 3 citations mapped on social science thematic area, 13 citations in psychology and none in computer science.

For the second article of our author we marked as irrelevant 7 records out of 43 citations (we preprocess the list of citations received from Google Scholar). For this article, the citations were categorized as follow: 8 citations in social science, 21 in psychology and 7 in computer science.

9 Conclusions

In this paper we have presented the notion of a footprints in the academic world, defined as the traces of impact that an article, or an author, has had in different scientific fields. In order to avoid subjectivity in the estimation and quantification of the footprint we have opted to avoid author defined parameters and have instead focused our analysis on the journal or conference where a citing article has been published. This provides an objective and reliable indication of thematic scope, which allows us to see, in a semi-automated manner, which scientific areas have been affected by an author's work.

These steps allow us to develop semantic footprint indices, associating articles and authors with the fields where they have made an impact. Querying such indices we can provide semantic services that were unimaginable before, such as the automated identification of the most prominent works and authors for each field, regardless of their originally intended scope.

In our paper we have presented various real life examples to support the validity of our approach, and have also presented a preliminary integrated implementation, the Footprint Analyzer, based on information acquired from DBLP and Google Scholar. Constraints set by Google Scholar and DBLP limit the extent of our experimentation, but the presented results suffice for a proof of concept. Besides, the architecture makes it straightforward to link more academic sources, particularly those that provide structured exposed APIs.

Of course our work is not complete. As part of our future work we would like to develop a more complete hierarchy of scientific fields which includes the whole spectrum of scientific fields in a greater depth and making an automatic match of journals with the multiple scientific fields. As we intend to use Apache Solr ⁷ in order to index our database and to create a section for manual match the papers that have inconsistencies in titles or author names and to mark them as duplicates.

Acknowledgments

This work has been supported by COST Action IC1302: Semantic keyword-based search on structured data sources (KEYSTONE).

This work has been partially funded by the Sectoral Operational Programme Human Resources Development 20072013 of the Ministry of European Funds through the Financial Agreement POSDRU/187/1.5/S/155536 and partially supported by "DataWay - Real-time Data Processing Platform for Smart Cities: Making sense of Big Data" grant of the Romanian National Authority for Scientific Research and Innovation, CNCS UEFISCDI, project number PN-II-RU-TE-2014-4-2731.

References

1. A.A. Armenakis, S.G. Harris, *Reflections: our Journey in Organizational Change Research and Practice* Journal of Change Management, Vol. 9, No. 2, pp. 127-142, 2009.
2. J. Altarriba, *Cognitive approaches to the study of emotion-laden and emotion words in monolingual and bilingual memory*, Bilingual Education and Bilingualism, vol 56, pp. 232-256, 2006.
3. J. Altarriba, *Emotion, Memory, and Bilingualism*, Foundations of Bilingual Memory, pp. 185-203, 2014.
4. J. Altarriba, L.M. Bauer, C. Benvenuto, *Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words*, Behavior Research Methods, 1999.

⁷ [urlhttp://lucene.apache.org/solr/](http://lucene.apache.org/solr/)

5. J. Altarriba, L.M. Bauer, *The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words*, The American journal of psychology, 2004.
6. M. Barthelet, G. Fazekas, and M.B. Sandler, *Music Emotion Recognition: From Content- to Context-Based Models*, International Symposium on Computer Music Modeling and Retrieval (CMMR), pp.228-252, 2012.
7. J. Beel and B. Gipp, *Academic search engine spam and google scholars resilience against it*, Journal of electronic publishing, 13(3), 2010.
8. J. Beel and B. Gipp, *Google scholars ranking algorithm: The impact of citation counts (an empirical study)*, In Research Challenges in Information Science (RCIS), Third International Conference on, pages 439-446. IEEE, 2009.
9. J. Beel and B. Gipp., *Google scholars ranking algorithm: an introductory overview*, In Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI09), volume 1, pages 230-241, 2009.
10. L. Bornmann and R. Mutz, *Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references*, Journal of the Association for Information Science and Technology, 66(11), pp. 2215-2222, 2015.
11. L. Bornmann, R. Mutz and H.D. Daniel, *The h index research output measurement: Two approaches to enhance its accuracy*, Journal of Informetrics no. 4(3), pp. 407-414, 2010.
12. K. Byron, *Carrying too heavy a load? The communication and miscommunication of emotion by email*, Academy of Management Review, vol. 33, pp. 309-327, 2008.
13. ScienceDirect, *Classification of topics*, <http://www.sciencedirect.com/>, retrieved December 2016.
14. S. Canazza, G. De Poli, A. Roda, and A. Vidolin, *An abstract control space for communication of sensory expressive intentions in music performance*, Journal of New Music Research, vol. 32(3), pp. 281-294, 2003
15. S. Chandrasekhar, *An Introduction to the Study of Stellar Structure*, Chicago, Ill., University of Chicago Press, 1939.
16. X. Chen, *Google scholars dramatic coverage improvement five years after debut*, Serials Review, 36(4):221-226, 2010.
17. *CiteSeer^X*, <http://citeseerx.ist.psu.edu/>, retrieved December 2016.
18. C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, *Fear-type emotion recognition for future audio-based surveillance systems*, Speech Communication, vol. 50(6), pp. 487-503, 2008.
19. Visualisation of DBLPs Bibliography Samples, <http://suksant.com/tag/dblp/>, retrived December 2015.
20. R. Cowie, R.R. Cornelius, *Describing the emotional states that are expressed in speech*, Speech Communication, Volume 40, Issues 12, pp. 5-32, 2003.
21. *DBLP* <http://dblp.uni-trier.de>, retrieved December 2016.
22. P.D. Batista, M.G. Campiteli, O. Konouchi and A.S. Martinez, *Is it possible to compare researchers with different scientific interests?*, Scientometrics, vol. 68, no. 1, pp. 179-189, 2006.
23. S. Djamasbi, D.M. Strong, *The effect of positive mood on intention to use computerized decision aids*, Information & Management, Volume 45, Issue 1, pp. 43-51, 2008.
24. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis, *The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data*, Proc. 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 2007

25. L. Egghe, *Theory and practise of the g-index* Scientometrics no. 69(1), pp. 131-152, 2013.
26. K. Elsbach, P.S. Barr, *The effects of mood on individuals' use of structured decision protocols*,
27. A. Figa-Talamanca, *Strengths and weaknesses of citation indices and impact factors*, Quality assessment in higher education, pp. 83-88, 2007.
28. N. Fragopanagos and J.G. Taylor, *Emotion recognition in humancomputer interaction* Neural Networks, no. 18(4), pp. 389-405, 2005.
29. Google Scholar Blog, *Google Scholar Citations Open To All*, Google, 16 November 2011, <http://googlescholar.blogspot.gr/2011/11/google-scholar-citations-open-to-all.html>, retrieved December 2016.
30. W. Glanzel and A. Schubert, *A new classification scheme of science fields and subfields designed for scientometric evaluation purposes*, Scientometrics, no.56(3), pp. 357-367, 2003.
31. M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J.Hirschberg, S. Kajarekar, *Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection*, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, 2006
32. S. Harnad, *Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge*, Public-Access Computer Systems Review 2 (1), pp. 39 - 53, 1991.
33. S. Harnad, *The Optimal and Inevitable Outcome for Research in the Online Age*, CILIP Update, pp. 46-48, 2012.
34. Zadeh, *Fuzzy sets*, Information and Control 8 (3): 338353, 1965
35. C. Ifrim, F. Pop, M. Mocanu and V. Cristea, *Agile DBLP: A Search-based Mobile Application for Structured Digital Libraries*, J. Cardoso, G.J. Houben, F. Guerra, A.M. Pinto, Y. Velegrakis (Eds), Proceedings of the 1st KEYSTONE Conference, Lecture Notes in Computer Science, Springer, 2015
36. S. Ioannou, G. Caridakis, K. Karpouzis, S. Kollias, *Robust Feature Detection for Facial Expression Recognition*, EURASIP Journal on Image and Video Processing, Volume 2007, Issue 2, 2007.
37. J.J.M. Kierkels, M. Soleymani and T. Pun, *Queries and tags in affect-based multimedia retrieval*, Proceedings of the IEEE international conference on Multimedia and Expo, IEEE Press, Piscataway, NJ, USA, pp. 1436-1439, 2009.
38. P. Jacs, *Metadata mega mess in Google Scholar*. Online Information Review 34.1 (2010): 175-191, 2010.
39. *Knowledge and Uncertainty Research Laboratory GitHub* <https://github.com/gavlab-gr/>, retrieved December 2016.
40. S.K. Foss, *Destination dissertation: A traveler's guide to a done dissertation*, Rowman and Littlefield Publishers, 2007.
41. Z.K. Silagadze, *Citation entropy and research impact estimation*, Acta Physica Polonica, B41, pp. 23252333, 2009.
42. J. Keyton and F.L. Smith, *Distrust in Leaders: Dimensions, Patterns, and Emotional Intensity* Journal of Leadership & Organizational Studies, no. 16(1), pp. 6-18. 2009.
43. R. Klavans, and K. W. Boyack, *Toward a consensus map of science.*, Journal of the American Society for information science and technology, vol. 60, no. 3, pp. 455-476, 2009
44. T. Kostoulas, I. Mporas, O. Kocsis, T. Ganchev, N. Katsaounos, J.J. Santamaria, S. Jimenez-Murcia, F. Fernandez-Aranda, N. Fakotakis, *Affective Speech Interface*

- in Serious Games for Supporting Therapy of Mental Disorders*, Expert Systems with Applications, Vol. 39, Issue 12, pp. 11072-11079, 2012.
45. S.L. Dubois, *Gender differences in the emotional tone of written sexual fantasies*, The Canadian Journal of Sexuality, no. 6(4), pp. 307-315, 1997.
 46. A. Letchford, H.S. Moat and T. Preis, *The Advantage of Short Paper Titles*, Royal Society Open Science (The Royal Society, 2015), 150266 <http://dx.doi.org/10.1098/rsos.150266>
 47. M. Ley et al. *Dblp-some lessons learned*. Proceedings of the VLDB Endowment, 2(2):14931500, 2009.
 48. Google Scholars Ghost Authors, <http://lj.libraryjournal.com/2009/11/industry-news/google-scholars-ghost-authors/>, retrived December 2015
 49. J.M. Bartunek, D.M. Rousseau, J.W. Rudolph and J.A. DePalma. *On the receiving end sensemaking, emotion, and assessments of an organizational change initiated by others*, Journal of Applied Behavioral Science, vol. 42, issue 2, pp. 182-206, 2006.
 50. C.M. Whissell, *Homepage*, <http://laurentian.ca/faculty/cwhissell>, retrieved December 2016.
 51. C.M. Whissell, *The dictionary of affect in language*, in R. Plutchik and H. Kellerman (Ed.), *Emotion: Theory, Research, and Experience*, pp. 113131, New York, Academic Press, (1989).
 52. S.N. Soroka, M.A. Bodet, L. Young, B. Andrew, *Campaign news and vote intentions*, Journal of Elections, Public Opinion and Parties, no. 19(4), pp. 359-376, 2009.
 53. S.N. Soroka, *Negativity in Democratic Politics: Causes and Consequences*, Cambridge University Press, 2014.
 54. Physics Portal at South Carolina State University, *The Branches of Science*, <http://www.cnrt.scsu.edu/~psc152/A/branches.htm>, retrieved December 2016.
 55. A. Raouzaïou, N. Tsapatoulis, K. Karpouzis and S. Kollias,
 56. A. Reyes and P. Rosso, *Making objective decisions from subjective data: Detecting irony in customer reviews* Decision Support Systems, no. 53(4), pp. 754-760, 2012.
 57. P. Richards, M.A. Persinger and S.A. Koren, *Modification of semantic memory in normal subjects by application across the temporal lobes of a weak (1 microT) magnetic field structure that promotes long-term potentiation in hippocampal slices*, Electro- and Magnetobiology, vol. 15(2), pp. 141-148, 1996,
 58. Philipp Mayr and Anne-Kathrin Walter, *Studying journal coverage in google scholar*, Journal of Library Administration, 47(1-2):8199, 2008.
 59. Google, *Scholar*, <https://scholar.google.com/>, retrieved December 2016.
 60. *ScienceDirect* <http://sciencedirect.com/>, retrieved December 2016.
 61. Theodore Simos, *Homepage*, https://users.uop.gr/~simos/CV_Simos_EV.pdf, retrieved December 2016.
 62. S. Scherer, F. Schwenker, and G. Palm, *Classifier fusion for emotion recognition from speech, invited book chapter contribution*, Advanced Intelligent Environments, no. 5, pp. 95-117, 2009.
 63. M. Schroder, *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions*, Affective dialogue systems, Springer, 2004.
 64. B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll, *Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies* IEEE Transactions on Affective Computing (TAC), vol. 1, no. 2, pp. 119-131, 2010.
 65. Z. Send, M. Pantic and T.S. Huang, *Emotion recognition based on multimodal information*, Affective Information Processing, Springer Verlag, London, pp. 241-266, 2009.
 66. C.T. Zhang, *The e-index, complementing the h-index for excess citations*, PLoS ONE, no 5(5), 2009.

67. N. Tsapatsoulis, K. Karpouzis, G. Stamou, F. Piat and S. Kollias,
68. B. Vlasenko, D. Prylipko, R. Bock, A. Wendemuth, *Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications*, Computer Speech & Language, Volume 28, Issue 2, pp. 483-500, 2014.
69. O. von Bohlen und Halbach, *How to judge a book by its cover? How useful are bibliometric indices for the evaluation of "scientific quality" or "scientific productivity"?*, Annals of Anatomy no. 193(3), pp. 191196, 2011
70. K.W. Mossholder, R.P. Settoon, S. G. Harris, A.A. Armenakis, *Measuring emotion in open-ended survey responses: An application of textual data analysis*, Journal of Management, no. 21(2), pp. 335-355, 1995.
71. M. Wallace, *Extracting and visualizing research impact semantics*, Proceedings of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization, Corfu, Greece, 2014
72. J. Yi and W. Niblack, *Sentiment Mining in WebFountain*, Proceedings of 21st International Conference on Data Engineering. pp. 1073 - 1083, 2005.
73. J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, *Sentiment Analyzer: Extracting Sentiments About A Given Topic Using Natural Language Processing Techniques*, in Proceedings of the IEEE International Conference on Data Mining (ICDM), 2003.
74. L. Young, S.N. Soroka, *Affective news: The automated coding of sentiment in political texts*, Political Communication, vol. 29, pp. 205-231, 2012.
75. Z. Zeng, M. Pantic, G.I. Roisman and T.S. Huang, *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions*, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 31(1), pp. 39-58, 2009.
76. WikiPedia, *Branches of science*, https://en.wikipedia.org/wiki/Branches_of_science, retrieved December 2016.