Scientometrics manuscript No.
(will be inserted by the editor)

Identifying Attractive Research Fields for New Scientists

Leonidas Akritidis · Dimitrios Katsaros · Panayiots Bozanis

Received: date / Accepted: date

Abstract Prior to the beginning of a scientific career, every new scientist is obliged to confront the critical issue of defining the subject area where his/her future research will be conducted. Regardless of the capabilities of a new scholar, an erroneous selection may condemn a dignified effort and result in wasted energy, time and resources. In this article we attempt to identify the research fields which are attractive to these individuals. To the best of our knowledge, this is a new topic that has never been discussed or addressed in the literature. Here we formally set the problem and we propose a solution combining the characteristics of the attractive research areas and the new scholars. Our approach is compared against a statistical model which reveals popular research areas. The comparison of this method to our proposed model leads to the conclusion that not all trendy research areas are suitable for new scientists. A secondary outcome reveals the existence of scientific fields which although they are not so emerging, they are promising for scientists who are starting their career.

Keywords scientist · author · research area · research field · scientometrics · attractive

Leonidas Akritidis
University of Thessaly
Tel.: +030-23510-25628
E-mail: leoakr@infuth.gr

Dimitrios Katsaros
University of Thessaly
Tel.: +030-23510-25628
E-mail: dkatsar@inf.uth.gr

Panayiotis Bozanis
University of Thessaly
Tel.: +030-23510-25628
E-mail: pbozanis@inf.uth.gr

1 Introduction

One of the most important issues that a new[1] researcher has to address is the correct identification of the primary research field that will determine his/her future career. Our current experience has proved that a significant percentage of starting scientists often choose their area of interest by considering invalid parameters, including the reputation of their future mentors or supervisors, the availability of open PhD theses, or the former success of others who have managed to conduct a productive research in this specific area. Therefore, it is a common phenomenon that capable and diligent scientists are misled and engaged with scientific fields that are considered as obsolete, dead, or prohibitively competent for their current level of experience.

We firmly believe that the primary criterion for the selection of a research area is the new scientist's preferences. A research conducted in a field that is out of the interests or likes of a researcher is undoubtedly condemned. Nevertheless, this criterion is extremely hard to be modeled, since even the scientists themselves are frequently not in the position to determine whether a research area is within their own interests. Along with this notification, a sequence of questions and critical issues are posed.

Certainly the various scientific fields are not equally promising and each of them exhibits its own level of "hostility" for a new scholar. For instance, several scientific domains are considered as obsolete, as the majority of their related problems have found efficient and effective solutions. On the other hand, there are problems that can only be tackled by experienced scientists and publishing a work in such an area is relatively difficult. Apparently, new scientists are not recommended to work in such areas, since it is usually impossible to propose a solution that outperforms the existing schemes and moreover, publishing such solutions has limited probabilities due to the lack of trust by the rest of the members of the scientific community.

The identification of trendy research areas is of great interest for every scientist. Such knowledge is a valuable tool, since it can reveal the correct path for new scholars and assist them in working on modern or newly posed problems. Even the more experienced researchers could benefit from the knowledge of the most fashionable fields, as they could expand their work and develop solutions to novel problems. This is a definite advantage for the science itself.

In this paper we attempt to formally set and solve this interesting problem. Although there are exist several previous works which investigate the issue of identifying emerging topics of research, the problem of identifying attractive research areas for new scientists is new; to the best of our knowledge, there is no other work attempting to address it. In our approach we initially examine the main attributes of the problem and we study the space where the solution lies. In the sequel, we consider the most important properties of the new scientists and with that knowledge, we identify the core elements that render a research field attractive to them.

A significant parameter of our problem is the identification of the new scientists and their separation from the more experienced ones. In this work we exploit some of the most sophisticated metrics that have been proposed in the literature. We also introduce a set of Topic-Sensitive extensions which render these metrics aware

[1] In this work we also use the term *starting scientists* or *starters* to refer to new scientists

of the research field that we examine each time. These contributions are tested experimentally by employing a large dataset of scientific articles deriving from the wide areas of Engineering and Computer Science.

The rest of the paper is organized as follows: In Section 2 we state the contributions of our work and in Section 3 a study of the previous relevant articles is studied. Section 4 contains a description of the provided data and the universe where our problem is located. Furthermore, in Subsection 4.2 we formally state the problem. In Section 5 we present our proposed solution and we describe our approaches in order to confront the component issues of the problem. Section 6 contains experiments that attest our methods and finally, in Section 7 we conclude the paper with interesting notifications and findings.

2 Contributions

In this subsection we briefly present the contributions of this paper.

- We formulate the problem of identifying attractive research areas for new scientists. Initially, we provide a detailed description of the provided data and in the sequel, we formally state the problem itself along with its component issues.
- We propose a solution to the problem by taking into consideration several aspects regarding the attractiveness of a research area and the characteristics of the new scientists.
- We introduce the Topic-Sensitive extensions in order to enhance some of the existing metrics for evaluating and ranking scientists. These extensions allow us to estimate the impact of the work of an author in a particular area of research.
- We test our proposed methods by employing a large dataset containing about 1.5 million scientific articles from the wide area of Engineering and Computer Science.

3 Related Work

Although the identification of attractive research fields for new scholars has not been previously addressed, the issue of investigating emerging research areas has been studied by several previous works. The approaches proposed in these works are divided into two wider categories, the *co-word* and the *co-citation* analysis methods. The first branch includes policies which focus on directly investigating the contents of a research topic. One of the earliest relevant works is the research of [6], which employed co-word analysis and detected changes in the field of information retrieval during the period between 1987 and 1997. Furthermore, [14] introduced a co-word analysis method for measuring the latest research trends in technical documents.

The most significant problem of the co-word analysis methods is the lack of an objective mechanism which will determine the set of representative keywords from the examined documents [14],[17]. For this reason, the extraction of objective keywords from the examined documents depends highly on each analyst; this certainly introduces some bias. The requirement to eliminate bias forced the researchers to introduce more objective criterions, such as the evaluation of the increment rate of published articles with particular keywords. Several works attempted to identify emerging topics by analyzing the changes in the number of related articles [16],[24]. These studies proved that the increment rate was an effective criterion for determining the value of each keyword. Nevertheless, these works also initially require a set of pre-defined keywords before their proposed algorithms can be applied.

The second category of methods includes the works which attempt to address the problem by applying co-citation approaches. Examples of such works are [23] and [25] which examined the citation properties of several papers in order to identify emerging fields of research. Based on this analysis, they detect sets of highly cited papers; the numbers of these papers and the research area they belong to is then used to obtain the required knowledge. The major problem is that recent works cannot usually receive many with respect to the older works. This difficulty turns co-citation approaches less effective.

In this paper we propose a score-based identification of attractive research fields for new scientists. Each research field receives a score according to numerous parameters, such as the reputation of the involved scientists, the prestige of the journals[2] which publish the related papers and the number of incoming citations. Furthermore, these parameters are considered with respect to temporal aspects which reveal the research fields which are attractive *presently.*

Regarding the issue of the evaluation of a researcher's work, there is a significant amount of work attempting to address it. The pioneering article which achieved robust results is [11], where J. Hirsch introduced *h-index*, a metric that rewards both the productivity and influence of a scientist. Motivated by the success of the h-index, several other metrics followed, such as the *SCEAS* system [20], *g-index* [7] and *f-index* [13]. In [2] a normalized version of the metric is presented, whereas in [4], a high-level study of the mathematics and performance is provided. In [8] it is attempted to minimize the gap between the lower bound of the total number of citations calculated by h-index and their real number. Additionally, in [22] two new metrics, the *contemporary h-index* and the *trend h-index* are introduced. The first takes into consideration the time that elapsed since an article was published, whereas the second takes into account the date an article received each of its citations.

Apart from the work that has been conducted towards ranking scientists, there is also a considerable research made for evaluating the prestige of a journal. Although the first relative article was published in early seventies [9], it was not before 2002 that this issue gained a remarkable attention. Bharati in [3], studied the preferences of journals for e-commerce research, whereas [12] employs citation analysis to assess journal quality and ranking. On the other hand, [15] and [21] apply scientometrics to determine the prestige of several information systems journals and scientific conferences respectively. In [18] there is a study which examines the differences across journal rankings, whereas in [5] and [19] several Hirsch-type indices for evaluating journals are proposed.

[2] In this paper we use the word *journal* to refer to a source where an article can be published. Apart from journals, the usage of this word also implies magazines, conference proceedings, digital libraries, etc.

4 Problem Formulation

In this Section we provide some necessary preliminary parameters and we describe the main characteristics of the problem. In the sequel, we state the problem formally and we identify the component issues which should be resolved before proceeding to the solution.

4.1 Preliminaries

Let us begin by introducing $P = \{p_1, p_2, ..., p_{|P|}\}$ which is the set containing all publications (also mentioned as papers, or articles) and $B = \{b_1, b_2, ..., b_{|B|}\}$ that is another set including the journals where the items of P have been published. Note that since each paper is published in exactly one journal, each entry $p_i \in P$ is mapped to a single element $b_l \in B$. Moreover, we define $A = \{a_1, a_2, ..., a_{|A|}\}$ as the set including all the authors (also mentioned as scholars, or scientists) who have contributed to the creation of the items of P and $F = \{f_1, f_2, ..., f_{|F|}\}$ which includes all the research fields involved in our problem.

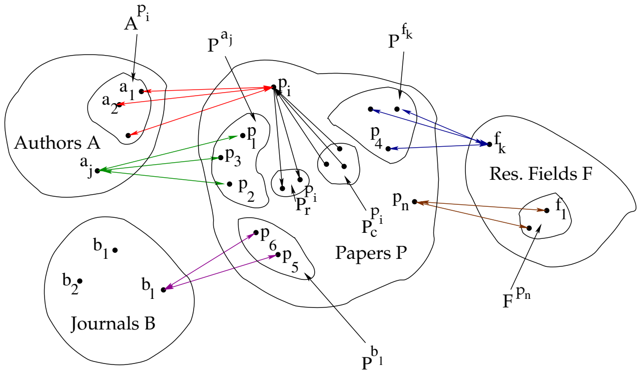

Fig. 1 Graphical representation of the examined universe

Based on the previous analysis we identify the subset $A^{p_i} \subset A$ which contains the researchers who have authored an article p_i, whereas the topic discussed in p_i is categorized to one or more research fields belonging to the subset $F^{p_i} \subset F$. Equivalently, each author a_j has published a series of papers $P^{a_j} \subset P$ and each research field f_n contains a subset of papers $P^{f_n} \subset P$.

Apart from these basic sets we also introduce the subset $P_r^{p_i} \subset P$ which contains all papers referenced by p_i, and $P_r^{p_i, f_n} \subset P_r^{p_i}$ which stores the publications referenced by p_i and also, they are classified into the research area f_n. In a similar spirit, $P_c^{p_i} \subset P$ and $P_c^{p_i, f_n} \subset P_c^{p_i}$ include the articles referring to p_i and the articles which both cite p_i and belong to the research field f_n. All introduced sets and subsets along with their connections are illustrated in Figure 1.

Finally, we use the symbol Y_i to indicate the year that the paper p_i was published in a journal b_l. Furthermore, $\Delta Y_i = Y_{now} - Y_i + 1$ is used to represent the years elapsed since the journal was published, where Y_{now} is the current year.

Symbol	Meaning
P	The set containing all papers
A	The set containing all authors
F	The set containing all research areas
B	The set containing all journals
p_i	An arbitrary paper $p_i \in P$
Y_i	The year of publication of p_i
ΔY_i	The age of publication of p_i
a_j	An arbitrary author $a_j \in A$
f_n	An arbitrary research area $f_n \in F$
b_l	An arbitrary journal $b_l \in B$
A^{p_i}	The authors who created p_i
F^{p_i}	The research areas that p_i belongs to
P^{f_n}	The papers belonging to f_n
P^{a_j}	The papers authored by a_j
P^{a_j,f_n}	The papers authored by a_j and belong to f_n
P^{b_l}	The papers published in b_l
P^{b_l,f_n}	The papers published in b_l and belong to f_n
$P_r^{p_i}$	The papers referenced by p_i
$P_r^{p_i,f_n}$	The papers referenced by p_i and belong to f_n
$P_c^{p_i}$	The papers referring to p_i
$P_c^{p_i,f_n}$	The papers referring to p_i and belong to f_n
$h_\nu^{a_j}$	A metric evaluating the work of an author a_j
$h_\mu^{b_l}$	A metric evaluating the prestige of b_l

Table 1 Summary

The quantity, the quality, the number of incoming references and some other characteristics of the publications of a researcher have been used widely to determine his/her productivity and impact. Several existing works (see Section 3) state that the activity of a researcher a_j can be evaluated by using a single value $h_\nu^{a_j}$ and they propose effective approaches towards this direction. Moreover, the characteristics of the papers published by a journal and the reputation of the involved authors can be exploited for evaluating this journal by using another metric, $h_\mu^{b_l}$. Note that the symbols ν and μ are identifiers used to differentiate the approaches that exist for evaluating a researcher's work and a journal's prestige respectively.

In Table 1 we summarize all the above notifications and in Figure 1 we illustrate the examined universe and the connections among the distinct sets of our analysis.

4.2 Problem Statement

The discussion of the previous Subsection determined the boundaries of the space where our problem lies. Our goal now is to identify the research areas F which are attractive for an author a^j, for whom the metric $h_\nu^{a_j}$ receives low values. For this purpose, for each field of research we introduce a special score S^{f_n}, which is calculated by taking into consideration the characteristics of a new scientist and an attractive research field. After that, we only have to sort the research fields by decreasing S^{f_n} order to obtain the desired knowledge.

As we will see later, the main problem includes three component issues which are essential to be addressed before we proceed in the extraction of the desired information. These are the evaluation of a researcher's work, the evaluation of a journal's reputation and the classification of an article within a given taxonomy of research areas. The first two sub-problems are related to finding effective methods for computing the $h_{\nu}^{a_j}$ and $h_{\mu}^{b_l}$ metrics and the literature contains numerous satisfactory solutions for this purpose. We present some of the most important of them in Subsections 5.2 and 5.3.

Regarding the identification of the research field that an article belongs to, an algorithm for mapping each of the items of the set P to one or more entries of the set F is required. In this work we utilize a link-based classification algorithm proposed in [10]. However, the methods that we present in this work can be applied effectively regardless of the selected classification algorithm.

5 Problem Solution

In this Section we describe our proposals for solving the problem of identifying attractive research areas for new scholars. Initially we describe some of the most remarkable characteristics of the new scientists and in the sequel, we depend on these characteristics to analyze the research areas that are attractive to them. We also provide methodologies for addressing the aforementioned component subproblems.

5.1 Identifying Attractive Research Areas

The problem we discuss here concerns new scientists, that is, scientists with low $h_{\nu}^{a_j}$ values. To determine an effective solution, it is necessary that we take into consideration an accurate overview of their characteristics. Some of the most important properties of the individuals belonging to this category are the *lack of experience* and the *lack of trust*. The former, lack of experience, is connected to the fact that a new researcher is not always able to discover or even understand the open problems in some challenging research areas. Moreover, even if a problem is formulated, the scholar is not usually in the position to propose a solution that is more effective than the ones that have already been proposed by other researchers. The latter, lack of trust, means that a new researcher is not reputable and it is expected that his projects will be treated with caution by the rest of the members of the scientific community.

Concerning the research fields, we determine two significant properties: *popularity* and *attractiveness for new scientists*. The former is mainly connected to the number of published articles and the number of scientists dealing with this particular research field. Regarding the latter, our research has shown that not all popular research topics are suitable for them and that additional properties must be considered. We shall discuss these properties shortly, since one of the primary goals of this work is to provide evidence supporting this claim.

To quantify the aforementioned properties and construct a model for evaluating each scientific field, we performed an enquiry among our colleagues. In particular, we have prepared a Web interface and we have asked from other PhD candidates to determine the reasons which render an area of research attractive, and the motivations that led them choose the subjects of their dissertations. The enquiry was answered by 141 new scientists from multiple departments of several universities and its conclusions proved that the most significant attributes that render a research field attractive for a new researcher are:

- *Number of recent articles:* Among all the enquiry answerers, a remarkable percentage of 62% agrees that the number of articles dealing with multiple problems from the same research area is a strong indication about the area's attractiveness and popularity. However, this parameter alone is not sufficient; the articles should also be recent, unless we desire to identify obsolete research fields which were once trendy. Recency is related to the time that has elapsed since a given date. In this work we assume that a paper is recent if it was published up to Y years before the current date and in Section 6 we are conducting experiments by examining different values of recency (i.e. we set $Y = 1, 2$ or 3 years).
- *Impact of articles:* To characterize a research area as attractive for a new scholar the number of recent publications is not adequate; the matter of the impact of these papers is equally important. The impact an article has in the scientific community can be evaluated by applying citation analysis methods which are based on the information provided by the inter linkage of the research papers. Such information includes the number of citations each paper acquired, their age, the publishing journal etc. Furthermore, the number of recent citations received by an entire research field, partially reveals its current popularity. This parameter was verified by the 68% of our enquiry answerers.
- *Reputation of the publishing journals:* Publication in prestigious journals has significant influence on promotion decisions, tenure and peer recognition. When an article is published in a reputable journal, it is expected that it will gain the attention of a large number of other scientists. Indeed, our enquiry confirmed that a percentage of 64% of new scientists will probably make an effort to propose a more effective methodology to confront the problem that the paper in question studies. In other words, other scientists are being attracted by the content of the papers which are published in high-level journals, since a more efficient approach to the same problem may result in a publication by a journal of equal or higher reputation. Furthermore, it is a common strategy for many new scientists to watch and study the articles published in the most important journals in order to determine the object of their future research. Consequently, the more articles from the same research areas are published in reputable journals, the more attractive this research area is for new authors.
- *Influence of the contributing authors:* In our effort to identify the attractive research areas for new scientists, we also examine the reputation of the authors who have published the most recent and influential works. When a high-level scientist deals with a problem and proposes an effective solution, it is expected that his/her work will be published in a top-quality journal. This is due to his/her high level of expertise and the trust he/she enjoys by the other members of the scientific community. Nonetheless, this does not make the research area the paper belongs to attractive for a starting scientist. Instead, we believe that this matter is detrimental to an author of low reputation who is usually not able to propose a more effective solution. 42% of our enquiry participants stated that they examine the previous experience and a paper author and they are influenced by the qualitative publications of other new scientists.

Based on the aforementioned enquiry and the parameters we discussed above, we conclude that popularity is not the only parameter affecting the new scholars during the selection of their area of research. Other characteristics such as the *impact* of the published articles, the *reputation* of the publishing journals and the *popularity among the other new scholars* must be considered when searching for attractive fields of research for new scientists.

Now we summarize the above notifications by characterizing a field of research as popular for a specific year Y, if its corresponding publications are:

$$\begin{gathered}[Multitudinous]\ AND\ [Influential] \\ AND\ [Authored\ by\ multiple\ distinct\ scientists] \end{gathered} \tag{1}$$

Consequently, the more publications a research field has, the more popular it is. Additional indications of popularity are the number of incoming citations and the number of the distinct authors dealing with the problems of the research field in question. Based on these properties we introduce the following scoring formula which determines the popularity of a research field:

$$S_{1,Y}^{f_n} = |P_Y^{f_n}| + \sum_{i=1}^{|P^{f_n}|} |P_{c,Y}^{p_i}| + \sum_{i=1}^{|P_Y^{f_n}|} |A^{p_i}| \tag{2}$$

The criteria which render a research area attractive for new scientists are different. According to our discussion, a topic is suitable for new scholars if the papers which are relevant to it are:

$$\begin{gathered}[Multitudinous]\ AND\ [Recent]\ AND\ [Influential] \\ AND\ [Published\ in\ reputable\ journals] \\ AND\ [Authored\ by\ new\ scientists] \end{gathered} \tag{3}$$

Now the parameters of 3 provide a qualitative solution to the problem of identifying attractive research areas for new scientists. In order to quantify our solution we must determine numerically the attractiveness of each scientific area and the following equation fulfils our goal:

$$S_{2,\nu,\mu}^{f_n} = \sum_{i=1}^{|P^{f_n}|} \frac{|P_c^{p_i}| h_\mu^{b_l}}{(\Delta Y_i)^\delta} \Big(\sum_{j=1}^{|A^{p_i}|} \frac{\lambda}{h_\nu^{a_j}} \Big) \tag{4}$$

where λ is a constant quantity used to assign the second sum a meaningfully large value, and δ is a parameter which determines the rate at which a publication becomes "old". A typical value for this parameter is $\delta = 1$.

To compute the $S_{2,\nu,\mu}^{f_n}$ scores we must initially map each article to the corresponding research field. It is also required to calculate the values of the $h_\nu^{a_j}$ and $h_\mu^{b_l}$ metrics, which indicate the reputation of the scientist who authored each paper and the prestige of the journal which published it, respectively. In the sequel, we iterate over all publications belonging to the research area f_n and evaluate the desired scores by considering the number of citations each of these publications acquired.

Equation 4 can be further enhanced by taking into consideration that an area could be attractive for a starting scholar, if the papers mapped to it receive recent citations. This reveals that the problems described in those works although they are old, still affect the scientific community. The following scoring formula incorporates this intuitive criterion:

$$S_{3,\nu,\mu}^{f_n} = \sum_{i=1}^{|P^{f_n}|} \frac{h_\mu^{b_l}}{(\Delta Y_i)^\delta} \Big(\sum_{x=1}^{|P_c^{p_i}|} \frac{1}{(\Delta Y_x)^\delta} \sum_{j=1}^{|A^{p_i}|} \frac{\lambda}{h_\nu^{a_j}} \Big) \tag{5}$$

Notice that the usage of the time interval in the denominator of the first sum of 4 and 5 denotes that we are mainly interested for research areas which attracted multiple publications recently. In addition, the placement of the $h_\nu^{a_j}$ metric in the denominator of the second sum reveals our goal to reward the publications authored by new scientists. Finally, the selection of placing $h_\mu^{b_l}$ in the numerator is justified by our intention to highlight the articles that have been published in prestigious journals.

5.2 Researchers Evaluation

The proposed solution requires the existence of a mechanism that evaluates the scientific work of a scholar. In this Subsection we describe some of the most important metrics that have been presented for this purpose. In addition, we introduce a set of extensions which can be attached to these metrics to facilitate topic-sensitive evaluation.

Although several scientists argue about the usefulness or the correctness of judging an author's work by using scalar values [12], [4]), it is the only methodology that has been proposed so far and moreover, it is widely used by other researchers.

5.2.1 Existing Approaches

The first and most popular metric for evaluating the contribution of a scientist is *h-index*, defined as follows:

Definition. A researcher a_j has h-index $h_1^{a_j}$, if $h_1^{a_j}$ of his/her $|P^{a_j}|$ articles have received at least $h_1^{a_j}$ citations each and the rest $(|P^{a_j}| - h_1^{a_j})$ articles have received no more than $h_1^{a_j}$ citations.

This metric calculates how broad the research work of a scientist is, since it accounts for both productivity and impact. Consequently, a researcher not only has to publish numerous articles, but also these works should be rewarded by being referenced by multiple papers.

Two interesting generalizations of h-index are the *contemporary* and the *trend* h-indices, introduced in [19]. Both of these metrics take into account several temporal characteristics of the research activity of a scientist. In particular, the *contemporary* h-index is sensitive to the time that has elapsed since an article was published and can detect scientists who contributed a number of significant articles that produced a large h-index, but now they are rather inactive or retired. This metric assigns higher rankings to the authors who are currently active, or the new scientists who have currently published a small number of works but are expected to contribute a large number of significant works in the near future.

On the other hand, the *trend* h-index incorporates the idea to assign scores to each paper by taking into account the year when an article acquired a particular citation, i.e., the age of each citation. This metric identifies the scientists whose works are referenced until now. If an old article still receives multiple citations, then it is an indication that the ideas it conveys continue to influence other researchers.

5.2.2 Topic-Sensitive Extensions

Often, many scientists contribute knowledge to more than one scientific fields and publish projects in multiple adjacent areas of research. Therefore, it is possible for a scientist to be distinguished in some research fields, whereas in others, the impact of his/her works to be limited. For instance, a scholar may have authored broadly acceptable articles regarding "Fiber optics", but his/her publications that are relevant to "Performance Analysis" not to be equally influential.

The existing metrics are not sensitive to this concept; they take into account all the publications of an author and provide a single value indicating the productivity and/or impact. For this reason, we introduce here a set of *Topic-Sensitive (TS)* extensions, which can be applied to all three previous approaches. The idea is to divide the works of a scientist according to the research field they belong to and then compute multiple metric values, one for each research field. The *Topic Sensitive h-index (TSh-index)* incorporates this idea:

Definition. A researcher a_j has TSh-index $h^{a_j}_{1,f_n}$ for the research field f_n, if $h^{a_j}_{1,f_n}$ of his/her $|P^{a_j}|$ articles that discuss a topic belonging to f_n, have received at least $h^{a_j}_{1,f_n}$ citations each and the rest $(|P^{a_j}| - h^{a_j}_{1,f_n})$ articles have received no more than $h^{a_j}_{1,f_n}$ citations.

This metric calculates how broad the research work of a scientist is for a specific research area and identifies the scientists who are experts and reputable in a particular field of expertise.

Now let us examine how the time-variants of the h-index can be extended by applying the topic sensitivity approach. Regarding the contemporary h-index, we convert the scores presented in [19] to the ones of equation 6:

$$S^{p_i,f_n}_c = \gamma \frac{|P^{p_i,f_n}_c|}{(\Delta Y_i)^\delta} \tag{6}$$

That is, instead of evaluating all the articles of an author, we take into consideration only the papers belonging to the area of research for which we desire to rank a scientist. These scores S^{p_i,f_n}_c are used to phrase the definition of the *contemporary TSh-index*:

Definition. A researcher a_j has contemporary TSh-index $h^{a_j}_{2,f_n}$ for the research field f_n, if $h^{a_j}_{2,f_n}$ of his/her $|P^{a_j}|$ articles that discuss a topic belonging to f_n, get a score of $S^{p_i,f_n}_c \geq h^{a_j}_{2,f_n}$ and the rest $(|P^{a_j}| - h^{a_j}_{2,f_n})$ articles get a score of $S^{p_i,f_n}_c < h^{a_j}_{2,f_n}$.

Similarly to the original contemporary h-index, this metric rewards the scholars who are currently active, or the new scientists who have currently published only a small number of influential works. The difference is that this procedure is performed on a per-topic level and one scientist can be assigned different rankings according to the research of area that we examine each time.

The trend h-index can be also extended by adopting an identical approach. Therefore, the original scores of [19] are modified according to the equation 7:

$$S_t^{p_i,f_n} = \gamma \sum_{n=1}^{|P_c^{p_i,f_n}|} \frac{1}{(\Delta Y_n)^\delta} \tag{7}$$

Based on these modified scores $S_t^{p_i,f_n}$, the definition of the *trend TSh-index* follows:

Definition. A researcher a_j has trend TSh-index $h_{3,f_n}^{a_j}$ for the research field f_n, if $h_{3,f_n}^{a_j}$ of his/her $|P^{a_j}|$ articles that discuss a topic belonging to f_n, get a score of $S_t^{p_i,f_n} \geq h_{3,f_n}^{a_j}$ and the rest $(|P^{a_j}| - h_{3,f_n}^{a_j})$ articles get a score of $S_t^{p_i,f_n} < h_{3,f_n}^{a_j}$.

ν	**Symbol**	**Meaning**
1	$h_1^{a_j}$	h-index
2	$h_2^{a_j}$	contemporary h-index
3	$h_3^{a_j}$	trend h-index
$1, f_n$	$h_{1,f_n}^{a_j}$	Topic-Sensitive h-index
$2, f_n$	$h_{2,f_n}^{a_j}$	contemporary TSh-index
$3, f_n$	$h_{3,f_n}^{a_j}$	trend TSh-index

Table 2 Summary of metrics for evaluating the work of a scientist

In contrast to the contemporary TSh-index which is sensitive to the age of each publication, this metric takes into consideration the year that each article received its citations. We anticipate that this approach will rank higher the authors whose work in a specific scientific field is considered pioneering (since it still attracts references) and could set a new line of research.

In Table 2 we summarize all the metrics that we have previously discussed, including the Topic-Sensitive extensions. The left column denotes the value that ν receives for each case; the middle column contains the corresponding symbol for each metric, whereas in the last column we record its respective name.

5.3 Journals Evaluation

The third issue that is related to our problem regards the matter of determining an effective mechanism in order to evaluate the reputation of a journal. In this Subsection we provide reviews of some of the most popular metrics for ranking journals. One of the most popular journal evaluation metrics is the *impact factor*, [1] defined as follows:

Definition. In a given year, the impact factor of a journal is the average number of citations received by each paper published in that journal during the two preceding years.

The impact factor for a journal is computed at an annual basis and it is sensitive to the total number of citations received by each published paper. Similarly to ranking scientists, the original h-index metric can also be utilized to rank journals and a definition adjacent to the one provided in Subsection 5.2.1 can be phrased:

Definition. A journal b_l has h-index $h_1^{b_l}$, if $h_1^{b_l}$ of his/her $|P^{b_l}|$ articles have received at least $h_1^{b_l}$ citations each, and the rest $(|P^{b_l}| - h_1^{b_l})$ articles have received no more than $h_1^{b_l}$ citations.

Ranking journals by using h-index is not as robust as ranking scientists since this metric awards both productivity and influence of an author. Nevertheless, in the case we study it holds that different journals publish different numbers of articles. For instance, a journal which publishes four issues yearly usually contains more articles than an annual conference. Therefore, the employment of the plain h-index metric is rather unfair for journals publishing a small number of articles. Another drawback of the original h-index is that it ignores the fact that a journal may be older than another.

To address this last problem, the authors of [19] define a subset $P_Y^{b_l} \subset P$ including all the papers published by the journal b_l during the year Y. Based on this subset they introduce a metric, *yearly h-index*, which evaluates the prestige of b_l on a per year basis. Its definition is phrased below:

Definition. A journal b_l has yearly h-index $h_{2,Y}^{b_l}$, if $h_{2,Y}^{b_l}$ of its $|P_Y^{b_l}|$ articles published during the year Y have received at least $h_{2,Y}^{b_l}$ citations each and the rest $(|P_Y^{b_l}| - h_{2,Y}^{b_l})$ articles received no more than $h_{2,Y}^{b_l}$ citations.

However, this metric is not sensitive to the first problem. For this reason, a normalized version with respect to the number of the published articles is required. Its formal definition is given below:

Definition. A journal b_l for the year Y has normalized h-index $h_{3,Y}^{b_l} = h_{2,Y}^{b_l}/|P_Y^{b_l}|$, if $h_{2,Y}^{b_l}$ of its $|P_Y^{b_l}|$ articles published during the year Y have received $h_{2,Y}^{b_l}$ citations each, and the rest $(|P_Y^{b_l}| - h_{2,Y}^{b_l})$ articles have received no more than $h_{2,Y}^{b_l}$ citations.

Similarly to the yearly h-index, the normalized h-index confronts the problem of the different journal ages, since it operates on an annual basis. Furthermore, it overcomes the issue of different number of publications by dividing the yearly h-index by the number of the articles a journal published during a specific year Y.

μ	**Symbol**	**Meaning**
1	$h_1^{b_l}$	h-index for journals
$2, Y$	$h_{2,Y}^{b_l}$	yearly h-index for the year Y
$3, Y$	$h_{3,Y}^{b_l}$	normalized h-index for the year Y
4	$h_4^{b_l}$	contemporary h-index for journals
5	$h_5^{b_l}$	trend h-index for journals
$6, Y$	$h_6^{b_l}$	Impact Factor for the year Y

Table 3 Summary of metrics for evaluating a journal

The contemporary and trend h-indices can also be applied to evaluate the prestige of a journal. Notice that for these two metrics there is no significant difference between the author and the journal versions; consequently, we apply identical definitions. Finally, in Table 3 we summarize the metrics which can be used to evaluate a journal.

6 Experiments

To conduct a thorough experimental analysis of the proposed methods, it is required that we construct or select an existing taxonomy of research fields. Furthermore, it is essential that we obtain a dataset of research articles which must be large enough to provide reliable results. For each paper of our dataset we need to acquire all the accompanying metadata including the authors, the year of publication, its keywords, the publishing journal, its references and if supported, its classification into one or more research fields of our employed taxonomy.

Apparently, a percentage of the articles of the dataset must support the given taxonomy. This is necessary in order to train the model of our classification algorithm.

6.1 Dataset and Taxonomy Characteristics

To the best of our knowledge, there are not any publicly available datasets satisfying all the aforementioned requirements. The strict policy applied by the digital libraries in order to protect their records, prevents us from accessing their databases. Nonetheless, *CiteSeerX*[3], a scientific digital library and search engine, allows its users to access its records[4] and provides a harvest mechanism[5] for retrieving the entire database and the full text of the articles. At the time we downloaded this database[6], *CiteSeerX* was containing 1,634,136 research articles. The majority of these papers are related to the wide fields of Engineering, Mathematics and Computer Science. From these papers we have removed some duplicate articles and some which were not accompanied by the desired meta-data (i.e. authors, journal or date of publication). At the end of this filtration process, our dataset was comprised of 1,429,398 distinct articles.

After the elimination of the problematic articles (i.e. duplicate entries and entries missing the required meta-data), we applied the link-based classification algorithm introduced in [10]. According to this method, the category of each paper depends on the category of its neighboring (i.e. citing) articles. Moreover, before applying the algorithm, it is required that we determine the set of categories (the taxonomy) where the items of our collection will be classified.

Regarding the taxonomy structure, we considered a number of existing propositions. For instance, *Google Scholar*[7], is a vertical search engine designed to facilitate searching for articles and authors. It employs a classification model that categorizes the articles into nine generic research fields. However, the search engine classifies articles belonging into different research areas to the same category (i.e. papers regarding Mathematics and Computer Science are all classified into the same category). Apart from this notification, we firmly believe that these nine categories are not adequate to provide satisfactory information. We need a more precise mechanism that divides the main research fields into multiple levels of smaller research fields.

[3] http://citeseerx.ist.psu.edu/

[4] http://citeseerx.ist.psu.edu/about/metadata

[5] http://citeseerx.ist.psu.edu/oai2

[6] August 16th, 2010

[7] http://scholar.google.com

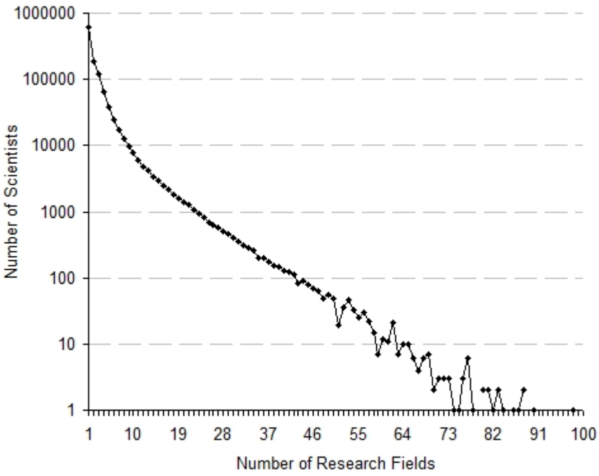

Fig. 2 Number of Authors vs Number of Research Areas

IEEE and *ACM* utilize a common taxonomy structure to categorize the articles they publish. That structure if far more informative than that of *Google Scholar's*, since it divides the generic term "Computer Science" into a large number of levels and sub-levels of research fields and furthermore, the classification is hierarchical. It consists of 11 first-level research fields divided into 81 second-level and 276 third-level classes. Our dataset consists of 744,760 articles supporting this taxonomy, whereas the rest 684,638 do not.

In our experiments we focus on research areas and articles which are related to the Computer Science and we employ the aforementioned taxonomy structure. However, the ideas and the concepts we describe here can also be used with other taxonomies with no additional effort.

6.2 Identifying Reputable Scientists

In this Subsection we apply the current state-of-the-art approaches for ranking scientists, as well as our proposed Topic-Sensitive extensions. Notice that all the metric values we present in this work have been calculated by using our test dataset; for other collections of papers these values can vary significantly. The articles of our dataset were authored by 1,209,316 scholars, a value which is translated to about 1.18 articles per author. However, the vast majority of them (about 70%) has published only once.

Figure 2 illustrates the distribution of authors with respect to the number of research areas their papers belong to. The vertical axis of this graph is in logarithmic scale. From this representation we conclude that a significant percentage of 54.4% of the authors have dealt with only one field. Only 18.9% of the authors of our dataset have published articles in more than three areas of research.

Table 4 contains rankings of the top-15 scientists of our dataset, according to three popular scientometrics. The h-index metric has been used for the left ranking, contemporary h-index determines the middle ranking, whereas trend h-index determines the right ranking. The third column of these rankings represents

Author	$\lvert P^{a_j}\rvert$	$h_1^{a_j}$
H. Garcia-Molina	328	46
J. Ullman	195	40
S. Shenker	170	40
P. Hanrahan	113	36
D. Estrin	142	36
C. Faloutsos	246	35
D. E. Culler	116	35
D. J. DeWitt	163	34
J. Widom	130	34
J. Han	290	34
C. Papadimitriou	253	34
W. B. Croft	201	34
R. Agrawal	147	34
T. Anderson	138	34
R. Fagin	114	34

Author	$h_2^{a_j}$
H. Garcia-Molina	51
Philip S. Yu	35
B. Forouzan	33
D. E. Culler	31
P. Hanrahan	31
S. Shenker	30
D. Estrin	30
T. Anderson	29
R. Motwani	29
M. Abadi	29
R. Kumar	28
M. D. Hill	27
W. B. Croft	27
J. Ullman	27
P. A. Bernstein	26

Author	$h_3^{a_j}$
S. Shenker	53
H. Garcia-Molina	51
A. K. Jain	47
J. Han	46
J. Widom	44
D. J. DeWitt	42
M. Stonebraker	42
M. D. Hill	42
J. Ullman	41
B. Shneiderman	41
R. Motwani	41
C. Faloutsos	39
P. Hanrahan	39
D. Estrin	39
T. Anderson	38

Table 4 Authors rankings (all research areas) according to h-index (left), contemporary h-index (center), trend h-index (right).

the total number of publications of a particular author, whereas the last column indicates the value the metric receives.

The scientist with the widest impact according to h-index is *H. Garcia-Molina* with 328 publications and $h_1^{a_j} = 46$, followed by *J. Ullman* (195 papers and $h_1^{a_j} = 40$) and *S. Shenker* (170 papers and $h_1^{a_j}$=40). Regarding the ranking according to the contemporary h-index, *H. Garcia-Molina* is again the top-scientist since his works not only are numerous and receive many citations, but also are recent. Recall that this metric is sensitive to the age of each publication and the score each article receives decays as time elapses. However, *J. Ullman*, the second most reputable scientist according to h-index, is ranked in the 14^{th} position and *S. Shenker* is ranked sixth. The second best performing scientist according to $h_2^{a_j}$ is *Philip S. Yu*, who does not appear in the top-15 h-index based ranking.

In contrast to the contemporary h-index, Trend h-index $h_3^{a_j}$ is sensitive to the age of each citation. The top-level scientist according to it is *S. Shenker* who is apparently the author whose works are still being referenced by the recent publications. *H. Garcia-Molina* is ranked second in this occasion, whereas *J. Ullman* is located in the ninth position of the Table.

Now let us study the rankings constructed by our proposed Topic-Sensitive extensions. Recall that these metrics are not applied in the entire set of an author's publications, but it is required that we isolate the papers which are mapped to a specific field of research. In Table 5 we present the ten most highly-ranked scholars according to TSh-index, for four different research areas: *Language Classification*, *Network Architecture and Design*, *Information Search and Retrieval* and *Database Applications*. The second column of each ranking denotes the number of publications which are both authored by a specific scientist and are mapped to the examined research field. The third column records the value that the applied metric receives.

We shall discuss the *Information Search and Retrieval* research field, however, the conclusions we extract from this discussion can be generalized and are valid for the other fields too. The author who is ranked first in that particular field is *W. B. Croft* who has authored 153 relevant articles and has $h_{1,f_n}^{a_j} = 33$. Notice that this author is ranked 12^{th} according to the plain h-index metric, and has authored in total 201 works. Nonetheless, when TSh-index is applied, only 153 of these works

Language Classifications

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{1,fn}$
C. Chambers	31	20
G. L. Steele, Jr	60	19
S. P. Jones	73	16
P. Wadler	35	16
M. Felleisen	43	15
K. Kennedy	39	14
N. Wirth	39	14
D. Ungar	35	14
B. Liskov	35	13
M. Wand	25	12

Network Architecture and Design

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{1,fn}$
D. Estrin	86	27
H.Balakrishnan	60	24
S. Shenker	67	22
D. E. Culler	42	20
N. H. Vaidya	106	20
Lixia Zhang	63	19
F. Floyd	30	19
I. F. Akyildiz	83	17
R. Morris	32	17
J. A. Stankovic	69	16

Information Search and Retrieval

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{1,fn}$
W. B. Croft	153	33
Cheng Xiang Zhai	83	19
G. Salton	123	18
C. Buckley	57	18
J. Callan	70	18
Wei-Ying Ma	112	18
H. Garcia-Molina	61	17
S. T. Dumais	61	17
S. E. Robertson	58	16
S. Lawrence	27	16

Database Applications

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{1,fn}$
Jiawei Han	192	34
Philip Yu	138	19
Jian Pei	96	18
R. Agrawal	32	17
M. J. Zaki	90	17
H.-P. Kriegel	96	16
E. Keogh	53	16
R. Srikant	22	14
R. T. Ng	44	14
Ke Wang	53	13

Table 5 Authors ranking according to TSh-index for various research areas.

are considered. A similar notification can also be made for *H. Garcia-Molina* who has authored in total 328 articles, but only 61 of them are related to the field of *Information Search and Retrieval*.

Table 6 contains author rankings for the aforementioned areas of research according to the Trend TSh-index. This metric rewards scholars for a particular research field, if their works continue to be cited until presently. *W.B. Croft* is still on the top of the list for the *Information Search and Retrieval* research field, However, *Wei-Ying Ma* has climbed in the second place (he was sixth according to TSh-index), whereas *G. Salton* is no longer among the top-10 authors. This observation leads to the conclusion that the works of the latter author do not receive many recent citations; potentially the problems discussed in those works have been addressed, or the topics are outdated.

6.3 Identifying Prestigious Journals

We continue our processing by attempting to detect the prestigious journals, since this information is valuable for identifying the attractive research fields. Recall that if a large number of articles associated with a particular scientific area is published in reputable journals, then this area becomes attractive for other scholars.

In Subsection 5.3 we have described some of the most important metrics for evaluating scientific journals. Due to limited space we focus primarily on the h-index for journals and the impact factor. In Table 7 we present the ranking of the journals we encountered in our dataset according to this metric. As previously, the rankings presented here should not be treated as representations of the value

Language Classifications

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{3,fn}$
C. Chambers	31	17
P. Wadler	35	14
G. L. Steele, Jr	60	13
M. Felleisen	43	13
S. P. Jones	73	12
Krishnamurthi	26	12
D. Ungar	35	11
D. Grove	22	11
B. G. Ryder	32	11
D. F. Bacon	25	11

Network Architecture and Design

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{3,fn}$
D. Estrin	86	30
H. Balakrishnan	60	24
D. E. Culler	42	23
N. H. Vaidya	106	22
S. Shenker	67	21
Lixia Zhang	63	20
R. Morris	32	20
I. F. Akyildiz	83	18
M. Srivastava	64	18
R. Govindan	48	18

Information Search and Retrieval

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{3,fn}$
W. B. Croft	153	28
Wei-Ying Ma	112	23
Cheng Xiang Zhai	83	21
S. T. Dumais	61	20
J. Callan	70	19
S. E. Robertson	58	18
H. Garcia-Molina	61	17
C. Buckley	57	17
A. Spink	76	16
Jiawei Han	44	16

Database Applications

Author	$\lvert P^{a_j}_{fn} \rvert$	$h^{a_j}_{3,fn}$
Jiawei Han	192	34
Philip Yu	138	22
Jian Pei	96	21
M. J. Zaki	90	19
R. Agrawal	32	17
C. Faloutsos	76	17
G. Karypis	32	16
E. Keogh	53	16
H.-P. Kriegel	96	15
Ke Wang	53	14

Table 6 Authors ranking according to Trend TSh-index for various research areas.

Journal Name	$h^{b_l}_1$	$\lvert P^{b_l}_c \rvert$
Communications of the ACM	10741	122
International Conference on Computer Graphics and Interactive Technology	8812	111
International Conference on Management of Data	2632	92
IEEE Trans. on Pat. Analysis and Machine Intelligence	3619	90
Journal of the ACM	2752	85
Applications, Technologies, Architectures, and Protocols for Computer Communication	1377	76
Conference on Human Factors in Computing Systems	7557	76
Artificial Intelligence	1987	73
ACM Computing Surveys	1300	72
IEEE Transactions on Software Engineering	3043	71
Very Large Data Bases	2406	70
International Symposium on Computer Architecture	1491	69
ACM Conference on Research and Development in Information Retrieval	2252	68
Symposium on Principles of Programming Languages	1188	67
Conference on Programming Language Design and Implementation	772	66

Table 7 Journals Ranking according to h-index

of a journal; it is possible that multiple papers from a journal are missing and the same could also be valid for their citations.

Table 8 illustrates the ranking of the journals for 2009 according to the impact factor. Notice that the only journal which is common in these two rankings is *Applications, Technologies, Architectures, and Protocols for Computer Communication*. This is an indication that a per-year journal evaluation leads to significantly different results than an all-year evaluation process.

Journal Name	$h^{b_l}_{6,2009}$	$\|P^{b_l}_c\|$
ACM Symposium on Operating Systems Principles	7.29	175
Web Search and Web Data Mining	5.27	137
ACM Computing Surveys (CSUR)	4.71	146
International Symposium on Computer Architecture	4.46	370
Proceedings of the 6th USENIX Conference on File and Storage Technologies	3.90	82
Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation	3.80	114
Applications, Technologies, Architectures, and Protocols for Computer Communication	3.67	588
Computational Linguistics	3.41	218
Internet Measurement Conference	3.38	243
Conference on Programming Language Design and Implementation	3.33	276

Table 8 Journals Ranking for 2009 according to impact factor

6.4 Popular Research Areas

In this Subsection we are based on our dataset to present the research areas which are the most popular. Recall that a research field is considered as popular in case many relevant articles are published and these articles have significant impact on the scientific community. Finally, the number of authors dealing with its problems is another indication of popularity.

Figure 3 illustrates the 35 most popular research fields in the last three years. The left part of the diagram depicts the number of relevant articles for each area, the middle part determines their popularity according to the number of incoming citations, whereas the right part reveals the number of distinct authors addressing problems which are relevant to the respective area.

Let us study the data displayed in these diagrams. The area which attracted the most publications in all three years is *Network Architecture and Design*; 12,992 articles of 2009 were mapped to this category. The second most popular area for 2008 and 2009 is *Model Development.* However, the second most popular field of research in 2007 was *Design Methodology.*

Regarding the number of incoming references, *Network Architecture and Design* is again the most popular field for 2009. Nevertheless, the area of *Non-numerical algorithms and problems* occupied the first position in 2007 and 2008. Other top-ranked research fields according to the number of in-links is *Learning* and *Information Search and Retrieval.* Although these fields had fewer papers than *Model Development* and *Design Methodology*, these papers attracted much more citations. This indicates that these papers affected more scientists.

The third part which determines the popularity of a research field according to the equation 2 is the number of authors publishing articles that are relevant to this particular field. The right diagram of Figure 3 indicates that *Network Architecture and Design* was the most popular area for 2009. However, in the previous two years the field of research of *Language Classifications* was attracting more scientists. *Non-numerical algorithms and problems*, *Model Development*, and *Design Methodology* are the next three highest ranked scientific topics.

In Figure 4 we illustrate the value of the $S^{f_n}_{1,Y}$ score for the 20 most popular research areas of 2007, 2008, and 2009. *Network Architecture and Design* has been the most popular topic of research during 2008 and 2009. On the other hand, *Non-numerical algorithms and Problems* and *Language Classifications* were the

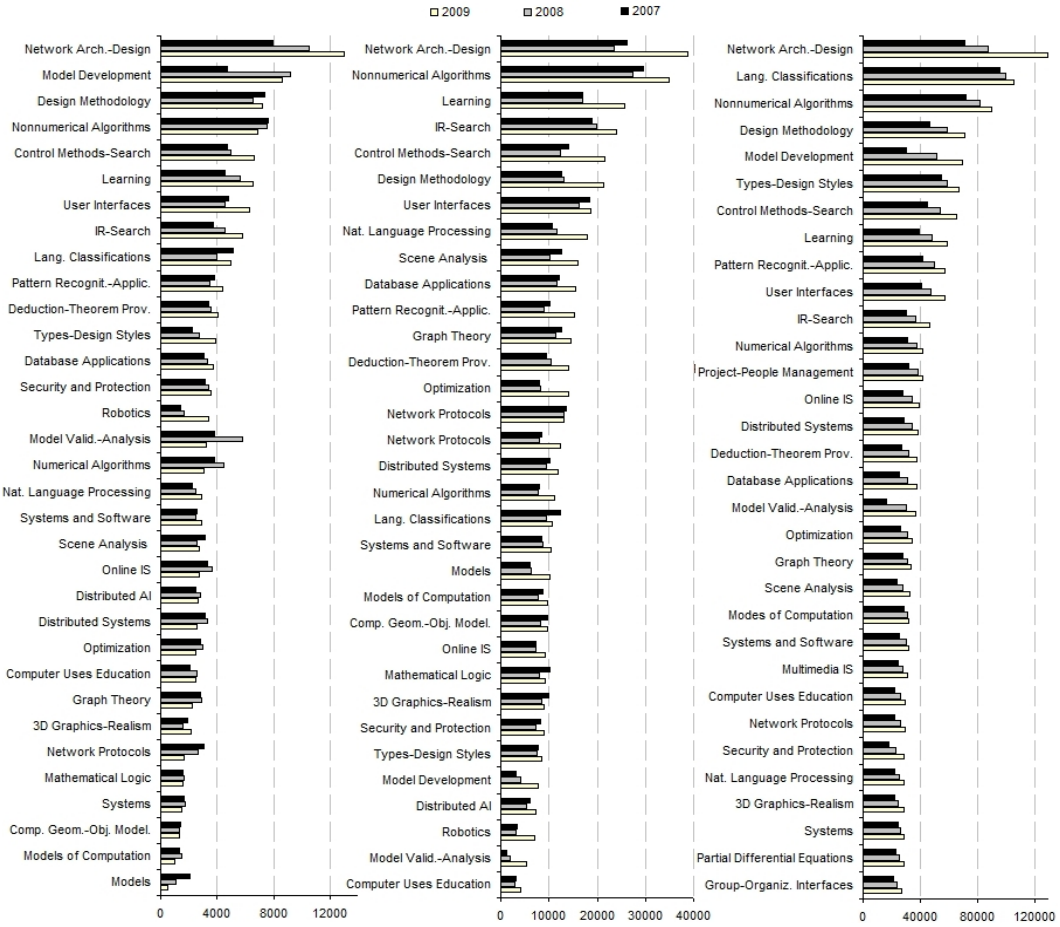

Fig. 3 Popular Research Fields in the last 3 years by number of published papers (left) number of incoming citations (center), and number of distinct authors (right)

most widespread scientific areas of 2007. This notification leads to the conclusion that in the past two years, there has been a significant increase in the research conducted towards *Network Architecture and Design*; this increase has rendered this area as the most popular in 2008 and 2009. The top-5 popularity ranking of Figure 4 also includes *Design Methodology* and *Control Methods and Search.*

Finally, the reader should notice that although *Learning* and *Information Search and Retrieval* are the third and fourth most cited research areas (middle diagram of Figure 3), they are not among the most popular. This is a strong indication that popularity is a generic metric which keeps plenty of useful information hidden.

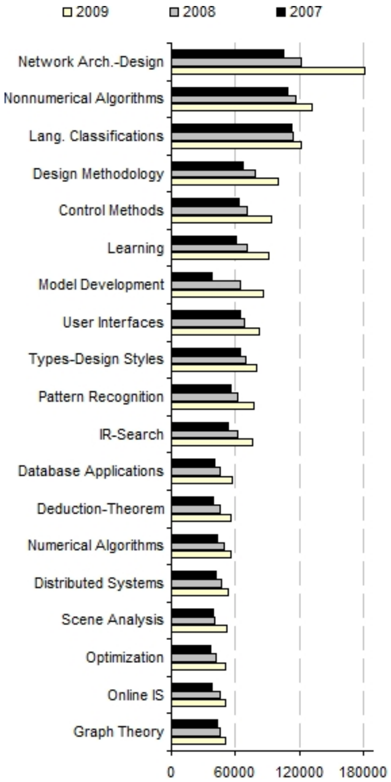

Fig. 4 The 20 most popular research fields according to $S^{f_n}_{1,Y}$ in the 3 last years

6.5 Attractive Research Areas for New Scientists

In this Subsection we present the research areas which according to the discussion of Subsection 5.1 are the most attractive for new scientists. In the following discussion we attempt to experimentally verify whether the popular research areas are all suitable for new scientists. In addition, we shall try to identify other topics which although they are not so popular as others, they could prove themselves promising for this class of scientists.

Recall that the scores of the equations 4 and 5 depend on both $h^{a_j}_{\nu}$ and $h^{b_l}_{\mu}$ metrics which evaluate the work of an author a_j and the prestige of a journal b_l respectively. However, since the number of possible combinations of these two metrics is quite large, we only provide results for some representative cases.

Initially we attempt to identify the research fields which are attractive for new scientists according to $S^{f_n}_{2,\nu,\mu}$. In Tables 9 and 10 we record four different such rankings for various combinations of author and journal evaluation metrics. The left ranking of Table 9 is produced by using h-index for both authors and journals ($\nu = 1$, $\mu = 1$), whereas the right ranking is constructed by employing the trend h-index for authors and the plain h-index for journals ($\nu = 3$, $\mu = 1$). Regarding the lists of Table 10, the left one shows the 15 most attractive research fields in case the Topic-Sensitive h-index is used to evaluate the work of a researcher and plain h-index is used to determine the prestige of a journal ($\nu = 1, f_n$ and $\mu = 1$) whereas the right ranking is generated by selecting the Topic-Sensitive Trend h-index for authors and the plain h-index for journals ($\nu = 3, f_n$ and $\mu = 1$).

Research Field	$S^{f_n}_{2,1,1}$
Non-Num. Algorithms-Problems	75,369
Network Architecture-Design	53,635
User Interfaces	47,393
Information Search-Retrieval	43,032
Design Methodology	42,164
Learning	40,067
Natural Language Processing	37,401
3-D Graphics and Realism	34,837
Systems	33,416
Graph Theory	32,445
Scene Analysis	31,846
Applications	31,832
Prob. Solving-Cont. Methods	31,175
Deduction-Theorem Proving	30,262
Comp. Geometry-Obj. Modeling	29,279

Research Field	$S^{f_n}_{2,3,1}$
Non-Num. Algorithms-Problems	78,626
Network Architecture-Design	52,719
User Interfaces	47,950
Information Search-Retrieval	43,187
Natural Language Processing	40,186
Design Methodology	39,907
Learning	38,695
Systems	37,427
3-D Graphics and Realism	35,907
Graph Theory	32,853
Language Classifications	32,480
Applications	31,178
Prob. Solving-Cont. Methods	30,742
Comp. Geometry-Obj. Modeling	30,491
Scene Analysis	30,475

Table 9 Attractive research fields for new scientists according to $S^{f_n}_{2,\nu,\mu}$ scores, for various author and journal evaluation metrics. Left: $\nu = 1$, $\mu = 1$. Right: $\nu = 3$, $\mu = 1$.

Research Field	$S^{f_n}_{2,1,f_n,1}$
Non-Num. Algorithms-Problems	151,373
Network Architecture-Design	89,276
Information Search-Retrieval	89,019
Graph Theory	84,130
Design Methodology	81,356
User Interfaces	79,736
Learning	76,624
Prob. Solving-Cont. Methods	71,169
Systems	64,558
Applications	64,424
Modes of Computation	60,279
Language Classifications	58,116
Systems and Software	58,009
User/Machine Systems	57,943
3-D Graphics and Realism	57,565

Research Field	$S^{f_n}_{2,3,f_n,1}$
Non-Num. Algorithms-Problems	155,311
Network Architecture-Design	87,913
Information Search-Retrieval	86,720
Graph Theory	84,959
User Interfaces	78,688
Design Methodology	77,521
Learning	75,659
Prob. Solving-Cont. Methods	70,929
Systems	70,487
Applications	64,971
Language Classifications	64,833
Modes of Computation	64,618
3-D Graphics and Realism	59,076
User/Machine Systems	58,560
Deduction-Theorem Proving	57,831

Table 10 Attractive research fields for new scientists according to $S^{f_n}_{2,\nu,\mu}$ scores, for various author and journal evaluation metrics. Left: $\nu = 1, f_n$, $\mu = 1$. Right: $\nu = 3, f_n$, $\mu = 1$.

According to the left ranking of Table 9, the area which is the most attractive for new scientists is *Non-numerical Algorithms and Problems*, followed by *Network Architecture and Design*. Recall from Figure 4 that the latter is most popular than the former, however, new scientists will not find it equally attractive. Surprisingly, the third most attractive research field for new scholars is *User Interfaces*, a topic which is ranked eighth in the corresponding popularity list. Another field of research which is attractive for new scientists but not so popular is *Information Search and Retrieval*.

Additionally, there are several popular research fields which are totally unappropriate for new scientists. The most representative example of such cases is

Research Field	$S^{f_n}_{3,1,1}$
Non-Num. Algorithms-Problems	88,847
Network Architecture-Design	68,286
Design Methodology	66,094
User Interfaces	58,970
Learning	55,308
Information Search-Retrieval	55,142
Scene Analysis	45,804
Applications	43,277
Deduction-Theorem Proving	42,339
Prob. Solving-Cont. Methods	41,713
Natural Language Processing	41,603
Graph Theory	39,058
Numerical Algorithms-Problems	34,354
3-D Graphics and Realism	34,133
Optimization	31,966

Research Field	$S^{f_n}_{3,3,1}$
Non-Num. Algorithms-Problems	87,722
Network Architecture-Design	64,577
Design Methodology	60,868
User Interfaces	57,081
Information Search-Retrieval	52,865
Learning	51,771
Scene Analysis	42,763
Natural Language Processing	41,847
Applications	40,960
Deduction-Theorem Proving	39,191
Prob. Solving-Cont. Methods	39,036
Graph Theory	37,704
3-D Graphics and Realism	33,769
Numerical Algorithms-Problems	31,812
Systems	31,353

Table 11 Attractive research fields for new scientists according to $S^{f_n}_{3,\nu,\mu}$ scores, for various author and journal evaluation metrics. Left: $\nu = 1$, $\mu = 1$. Right: $\nu = 3$, $\mu = 1$.

Languages Classifications. This topic is the third most popular, however, it is not ranked among the 15 most attractive research fields. Apparently, the problems related to this research area are difficult to confront or even understand and they are not suitable for starters.

The data recorded in this Table leads to two important conclusions: At first, popularity does not coincide with attractiveness for new scientists. There are popular research fields which are not attractive and they can be characterized as "hostile" for starting scientists, such as *Language Classifications.* On the other hand, there are research fields which although unpopular, they provide excellent opportunities at the scientists in question. Examples of such cases are *User Interfaces* and *Information Search and Retrieval.*

The second ranking of Table 9 employs the trend h-index for evaluating the work of a researcher. Recall that this metric is sensitive to age of the incoming citations of an article. Compared to the previous case the top-4 entries are left unchanged, however, in the fifth position we encounter another interesting case. *Natural Language Processing* which is not among the twenty most popular research fields, is quite attractive for new scientists.

Regarding the rankings of Table 10, the Topic-Sensitive extensions of h-index and trend h-index are employed for authors. In these cases, to compute the value of $S^{f_n}_2$, we need to store for each author and each research area the value the corresponding metric. That is, an author does not perform equally at every scientific topic; this allows us to identify the individuals who are possibly very experienced, but they are considered as starters for a particular research area. The two most attractive research areas for new scientists are the same once again, whereas *Information Search and Retrieval* is found in the third position. The usage of TSh-index in the $S^{f_n}_2$ highlights *Graph Theory* and considers is as the fourth most suitable scientific toping for new scholars.

Tables 11 and 12 contain rankings of the most attractive fields of research according to the $S^{f_n}_{3,\nu,\mu}$ score. The left list of Table 11 is constructed by using the plain h-index metric for both authors and journals. Compared to the left list of Table 9 the ordering of the topics is slightly different. Therefore, *Non-numerical Algorithms and Problems* and *Network Architecture and Design* are again the most

Research Field	$S^{fn}_{3,1,f_n,1}$
Non-Num. Algorithms-Problems	174,918
Design Methodology	124,155
Information Search-Retrieval	111,011
Network Architecture-Design	108,845
Learning	103,935
User Interfaces	96,847
Graph Theory	93,240
Prob. Solving-Cont. Methods	87,801
Applications	86,004
Systems and Software	74,004
Scene Analysis	73,129
Deduction-Theorem Proving	72,581
Numerical Algorithms-Problems	70,121
Optimization	69,350
Models	68,961

Research Field	$S^{fn}_{3,3,f_n,1}$
Non-Num. Algorithms-Problems	174,565
Design Methodology	116,441
Information Search-Retrieval	105,393
Network Architecture-Design	103,921
Learning	100,244
User Interfaces	93,061
Graph Theory	92,073
Prob. Solving-Cont. Methods	85,573
Applications	85,142
Systems and Software	70,879
Deduction-Theorem Proving	70,229
Scene Analysis	69,260
Numerical Algorithms-Problems	67,792
Models	66,540
Optimization	65,760

Table 12 Attractive research fields for new scientists according to $S^{fn}_{3,\nu,\mu}$ scores, for various author and journal evaluation metrics. Left: $\nu = 1, f_n$, $\mu = 1$. Right: $\nu = 3, f_n$, $\mu = 1$.

appropriate research fields for new scientists, however in the third position *User Interfaces* is replaced by *Design Methodology*. The usage of this metric highlights two significant points: the eighth position of *Three Dimensional Graphics and Realism* and the ninth place of the *Systems* research fields. Both of them are not among the twenty most popular areas, however they can be considered at least promising for new scholars.

Now let us summarize the results we presented in this Subsection. In almost every ranking *Non-numerical Algorithms and Problems* and *Network Architecture and Design* are considered as the most attractive research fields for starting researchers. Other topics also include *User Interfaces*, *Information Search and Retrieval* and *Graph Theory*. The comparison of these results to the popularity ranking of Figure 4, leads to the conclusion that popularity and attractiveness do not coincide; there are popular research fields which are not suitable for starters (such as *Language Classifications*), whereas some others, not so popular, are ideal for them.

7 Conclusions

In this paper we studied the problem of identifying attractive research areas for new scientists. Since this is a new issue, we initially described the properties of the space where the problem is set and solved.

In the sequel, we identified the characteristics of the new scholars and the attributes of the attractive research areas. We distinguished popular research areas from attractive, and we stated that popularity does not render a topic of research attractive for new scientists. Therefore, to measure the attractiveness of a research field for a new scholar, we presented two scoring schemes which incorporate multiple different parameters such as the number and the recency of the published articles and their citations, the number and the reputation of the involved authors and the reputation of the publishing journals.

In our work, it was also necessary to determine a method for evaluating the work of a researcher. There are several widespread metrics for this task, however, we introduced a set of topic-sensitive extensions which can make the aforementioned metrics sensitive to the research field we examine each time. With these extensions we are able to determine the value of a scientist's work for a particular research field.

Our methods have been attested experimentally by employing a large set of self-crawled research articles. The experiments provided some significant conclusions: The first is that there are exist some research fields which despite their popularity, they are not attractive for scholars who are now starting their career. On the other hand, some research fields are unpopular however, they provide excellent opportunities at these scientists.

References

1. Introducing the Impact Factor. http://thomsonreuters.com/ products_services/science/academic/impact_factor/
2. Banks, M.: An Extension of the Hirsch Index: Indexing Scientific Topics and Compounds. Scientometrics **69**(1), 161–168 (2006)
3. Bharati, P., Tarasewich, P.: Global Perceptions of Journals Publishing e-commerce Research. Communications of the ACM **45**(5), 21–26 (2002)
4. Bornmann, L., Daniel, H.: Does the h-index for Ranking of Scientists Really Work? Scientometrics **65**(3), 391–392 (2005)
5. Braun, T., Glänzel, W., Schubert, A.: A Hirsch-type Index for Journals. Scientometrics **69**(1), 169–173 (2006)
6. Ding, Y., Chowdhury, G., Foo, S.: Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management **37**(6), 817–842 (2001)
7. Egghe, L.: Theory and Practise of the g-index. Scientometrics **69**(1), 131–152 (2006)
8. Egghe, L.: Dynamic h-index: the Hirsch Index in Function of Time. Journal of the American Society for Information Science and Technology **58**(3), 452–454 (2007)
9. Garfield, E.: Citation Analysis as a Tool in Journal Evaluation. Science **178**(4060), 471–479 (1972)
10. Getoor, L.: Link-Based Classification. Advanced Methods for Knowledge Discovery from Complex Data pp. 189–207 (2005)
11. Hirsch, J.: An Index to Quantify an Individual's Scientific Research Output. Proceedings of the National Academy of Sciences **102**(46), 16,569 (2005)
12. Katerattanakul, P., Han, B., Hong, S.: Objective Quality Ranking of Computing Journals. Communications of the ACM **46**(10), 111–114 (2003)
13. Katsaros, D., Akritidis, L., Bozanis, P.: The f index: Quantifying the Impact of Coterminal Citations on Scientists' Ranking. Journal of the American Society for Information Science and Technology **60**(5), 1051–1056 (2009)
14. Lee, W.: How to identify emerging research fields using scientometrics: An example in the field of Information Security. Scientometrics **76**(3), 503–525 (2008)
15. Lowry, P., Romans, D., Curtis, A., PricewaterhouseCoopers, L.: Global Journal Prestige and Supporting Disciplines: A Scientometric Study of Information Systems Journals. Journal of the Association for Information Systems (JAIS) **5**(2), 29–80 (2004)
16. Noyons, E., Moed, H., Van Raan, A.: Integrating research performance analysis and science mapping. Scientometrics **46**(3), 591–604 (1999)
17. Ohniwa, R., Hibino, A., Takeyasu, K.: Trends in research foci in life science fields over the last 30 years monitored by emerging topics. Scientometrics pp. 1–17 (2010)
18. Rainer Jr, R., Miller, M.: Examining Differences Across Journal Rankings. Communications of the ACM **48**(2), 94 (2005)
19. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Generalized Hirsch h-index for Disclosing Latent Facts in Citation Networks. Scientometrics **72**(2), 253–280 (2007)
20. Sidiropoulos, A., Manolopoulos, Y.: A Citation-Based System to Assist Prize Awarding. ACM SIGMOD Record **34**(4), 60 (2005)
21. Sidiropoulos, A., Manolopoulos, Y.: A New Perspective to Automatically Rank Scientific Conferences Using Digital Libraries. Information Processing & Management **41**(2), 289–312 (2005)

22. Sidiropoulos, A., Manolopoulos, Y.: Generalized Comparison of Graph-Based Ranking Algorithms for Publications and Authors. Journal of Systems and Software **79**(12), 1679–1700 (2006)
23. Small, H.: Tracking and predicting growth areas in science. Scientometrics **68**(3), 595–610 (2006)
24. Tseng, Y., Lin, Y., Lee, Y., Hung, W., Lee, C.: A comparison of methods for detecting hot topics. Scientometrics **81**(1), 73–90 (2009)
25. Upham, S., Small, H.: Emerging research fronts in science and technology: patterns of new knowledge development. Scientometrics **83**(1), 15–38 (2010)