Malaysian Journal of Library & Information Science, Vol. 22, no. 1, April 2017, 69-82

Indexing of authors according to their domain of expertise

Tehmina Amjad[1] and Ali Daud[1,2]
[1]Department of Computer Science and Software Engineering,
International Islamic University, Islamabad, 44000, PAKISTAN
[2]Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, SAUDI ARABIA
e-mail: tehminaamjad @iiu.edu.pk (corresponding author);
ali.daud@iiu.edu.pk

ABSTRACT

Measuring the impact and productivity of an author is an important, yet a challenging task. Most of the existing methods for ranking or indexing of authors are based on simple parameters such as publication counts, citation counts and their combinations. These methods are topic independent, hence ignoring the intra-field differences. This study introduces a specific method for indexing of researchers to measure their productivity in a given field of interest, believing that an author can be interested in more than one fields and can have different level of expertise in all these fields. This paper proposes Domain Specific Index (DSI), a novel method for indexing of authors with respect to their fields of interest. Latent Dirichlet Allocation (LDA) is applied to capture the latent topics within text corpora. DSI calculates the standing of an author in all topics of his or her interest by considering topic based citations instead of using overall citations like traditional methods. The citations received by a multi-authored paper are divided among all its co-authors on the basis of their topic probability in that particular field. Results show that instead of giving credit of received citations equally to all co-authors of a paper, if a weight is given with respect to their level of interest in that field, more specific authors in that field will be ranked as top authors.

Keywords: Indexing; Domain specific modeling; Topic modeling; Topic based ranking; Citation analysis.

INTRODUCTION

Indexing for the quantification of an individual's research output informs us the authoritative researchers of a domain. Citations of papers published by the researchers in peer-reviewed journals are usually analyzed to index academic objects (journals, papers, and authors). An extensively used citation analysis method is the h-index (Hirsch 2005) which takes number of publications and their citations into account. It is robust in the sense that it is not sensitive to low cited or un-cited papers, but it is insensitive to sub-fields of an area, e.g. data mining, databases, information retrieval and image processing which are the sub-fields of computer science domain. This last trait can be considered as a disadvantage of the h-index and its following measures g-index (Egghe 2006), m-quotient (Burrell 2007a), a-index (Burrell 2007b), r-index and ar-index (Jin et al. 2007).

https://dx.doi.org/10.22452/mjlis.vol22no1.6

Domain based indexing is important to determine a person's expertise in a specific area of research. In real life, the researchers do not have any restrictions to work only in one field. They can be interested in more than one research field. However, it is not compulsory that they give equal amount of time to all fields of interest and attain equal level of expertise in all the fields they work. They can be an expert in one field while at the same time a middling in another. Finding the standing of an author in a particular field needs to be done by considering the weight of citations received only in that particular field, instead of the author's overall citations. The objective of this research is to find topic specific index of authors according to their level of interest and expertise in a precise domain of interest.

We explain this situation with the help of an example. A research institute is interested in hiring a person with expertise in information retrieval (IR) and the researcher's h-index is the main criteria they are following. Suppose that there are 3 people, A, B and C, who applied for a research position in the institute and they have a h-index of 10, 18, and 7, respectively. All three have IR as one of their research interests. From the h-index, it can be expected that the person B with h-index 18 has a higher chance to get selected. As the organization needs a person with expertise on IR, it would be good to know the h-index for both general and specific research expertise of the person. For example, A has a general h-index of 10 and IR h-index of 8; B has a general h-index of 18 and IR h-index of 5; and C has a general h-index of 7 and IR h-index of 3. In that case it is much easier to decide that A is more suitable for the research position. Although B has a higher h-index, but his expertise in IR is not as much as that of A, hence topic specific or domain specific index is more useful and applicable.

In order to deal with the limitations of h-index, g-index and other indexing measures whilst keeping the advantages of them, we propose Domain Specific Indexing (DSI) for domain/topic based indexing of authors. This proposed method has the ability to find not only distinguished authors of any field, but it also provides an author's standing in all the topics he or she has worked in with respect to citations received in this particular field. DSI quantifies citations with respect to the domain of citing and cited paper. The citations in which the domain of citing paper and cited paper is similar should be considered more important, than the citations that are received from more generic or different domains. The major contributions of our work described in this paper are (a) appraisal for the need of domain based indexing; (b) provision of a framework for domain specific indexing which can consider the citations received by authors in all topics they have worked in; and (c) experimental verification of the worth of the proposed h-index type topic based framework. The proposed DSI method gives most relevant results when we are interested in finding the authors by evaluating the quality of their work within a particular domain. Citations from papers in the same field depict quality of work in that particular field. If an author receives citations from papers that typically fall in the same domain, these should be given more weight, while the citations that came from papers not strictly from same domain or from generic papers must be given less weightage.

LITERATURE REVIEW

A lot of indexing measures are proposed for the indexing of individual's research output and to evaluate an individual's research work. H-index (Hirsch 2005) can be considered as the state-of-the-art and simplest among them, which takes into account both the papers and their number of citations. In recent years, it has become an increasingly important tool because it

considers both number of papers and citations into account for calculating author's productivity. The documents are arranged in descending order according to the citations received by them for a scientist. The h-index is then the document number N, equals to or less than the number of citations of respective paper and all the preceding documents have N or fewer citations. The h-index has been highly welcomed by the research community and used for indexing by many research indexing systems for example Microsoft Academic Search and ArnetMiner. However, it is insensitive to one or several outstandingly highly cited papers as it simply uses all citations of papers irrespective of more or less cited papers.

This aspect was criticized and g-index (Egghe 2006) was proposed as a counterpart. The publications of a scientist are arranged in descending order then g-index is the largest document number such that top g publications collectively received at least g^2 citations. Kosmulski (2006) proposed H(2)-index, and like the g-index, the calculation of the h(2)-index also gives more weight to highly cited articles. Burrell (2007a) discussed the limitation of h and g index of not considering career length and proposed a stochastic model based on h-index by considering the number of years of researcher's activity along with publication and citation rates.

Later, Burrell (2007b) discussed the h-index core and a-index by emphasizing the dynamic and time dependent nature of the publications and citations. H-index is used to identify the most prolific core of a researchers output. He said it can be expressed as average number of citations of published paper in h core i.e. total citation count is divided by h-index. A-index represents the average impact, as it is computed by the mean of citations, of h-core. Bornmann, Mutz and Daniel (2008) said that the process of determining a-index involves arithmetic mean which is influenced by extreme values. They proposed m-index, and instead of using arithmetic mean to measure the central tendency of citation distributions, it uses median of the number of citations received by the published papers in the h core to handle extreme values effect. Jin et al. (2007) discussed the process of determining a-index which involves the division by h-index which is unfair with the researcher with higher h-index. They proposed R-index, instead of dividing by h-index; it calculates the square root of citations of Hirsch's core publications. Along with r-index Jin et al. (2007) proposed AR-index which uses intensity of the citations of the published articles and life time of the publication as well. This makes indexing more sensitive as with the passage of time the researcher index not only increases but can decrease also.

Egghe and Rousseau (2008) proposed weighted h-index (h_w-index) based on the number of citations obtained by the published papers in h core and is sensitive to performance changes. Katsaros, Akritidis and Bozanis (2009) proposed f-index which takes co-terminal citations into account by considering them the generalization of self-citation and co-citations. They also showed how they can be used to capture the manipulation attempts in tempering scientometric indicators. Cabrerizo et al. (2010) merged the properties of both h and g indexes to create a hybrid index known as hg-index.

To compensate the limitations of single indicator, a few studies (Bornmann et al., 2008; Wildgaard, Schneider and Larsen 2014) recommended to combine the h-type indicators. This was a more user friendly approach as, it aims to categorize and merge pairs of indicators associated with the productive core. Cabrerizo et al. (2010) presented Q2-inex to relate two

diverse dimensions in a researcher's productive core, which are the number of papers and impact of papers. X-index (Claro and Costa 2010) was proposed as an indication of research level. It explains quantity and quality of the productive core and acknowledges the cross-disciplinary assessment with colleagues. V-index (Daud et al. 2013) also known as variation index was proposed to handle the issue of variation among the number of citations received by a researcher for his papers. S-index (Ko and Park 2013) was proposed which is an evaluation index based on the number of citations of each article in a particular journal and the rank of the article according to the number of citations. Along with the number of publications and citations (Amjad et al. 2015a; Amjad et al. 2015c), the impact of mutual influence of authors was considered when they work in collaboration for ranking of researchers.

However, the application of h-index and similar measures for quantification of an author gives incomplete picture, as these measures cannot distinguish between the citations from a paper of relevant topic and a paper of irrelevant topic. The limitation of all the above indices that multiple authors of a paper are given same credit, which is not fair is discussed in a few studies (Chai et al. 2008; Sekercioglu 2008; Wan, Hua and Rousseau 2007). They introduced weighted citation method for dividing the citations among the co-authors on the basis of their order in the paper and number of authors in each paper. Unfortunately, no investigation is made for indexing of researchers with respect to a specific topic of interest. It is believed that researchers can be interested in more than one research field. However, it is not compulsory that they have an equal level of expertise in all the fields they work in. They can be an expert in one field while at the same time a mediocre in another. Indexing them by using all the received citations generally, irrespective of the topic specificity is not sufficient to find their standing with respect to a particular topic. This motivates us to find topic specific index of authors according to their level of interest and expertise in a particular field.

OBJECTIVE AND METHOD

In this study the problem of topic specific indexing of authors is dealt as an information retrieval problem. Given $\boldsymbol{D} = \{D^t_i, \ldots\ldots, D^t_n\}$ be an $n \times t$ matrix representing t dimensional feature vectors of n objects, where n is the total numbers of papers ***(P)*** and $T_i \in R^t$, t is the number of topics of paper P_i. Each row of $n \times t$ matrix corresponds to one paper P_i and each column corresponds to its topic probability value corresponding to that topic.

Here, the goal is to order n objects (n papers) according to their topic probabilities and to find the $m \times t$ matrix representing t dimensional feature vectors of m objects, where m is the total numbers of authors ***(A)*** and $T_i \in R^t$, t is the number of topics of author A_i. . For each author we will get a t vectors representing his ranking in that specific topic.

This study proposes the domain specific indexing (DSI) method which calculates the standing of an author with respect to his or her domain of interest. DSI considers that an author can work in more than one field and hence, must have more than one index, one for each field he or she has worked in. DSI divides the citation received by an author with respect to the topics the author has worked in. It distinguishes between the received citations with respect to the topic probabilities in a given field, and the weight of the citations from the same domain are considered to be more central.

We argue that fetching the data relevant to the query, and ranking it is not sufficient for topic sensitive indexing. This does not involve the correlations present among the topics. The proposed method divides the citations received by an authors with respect to their topics and indexes the author in a topic sensitive way.

Dataset

An investigation is performed on a real world dataset from Arnetminer of about 1,572,277 and 2,084,019 citation relationships in the form of references. It entails all papers from Digital Bibliography and Library Project (DBLP) along with the abstract of papers (if available on the web), and the citation relationship between these papers in the form of references. Common text preprocessing procedures are applied by (a) removing stop-words, punctuations and numbers; (b) converting all words to lower case; and (c) removing words and authors having frequency less than 3 in the dataset. The German words also occur very frequently in this dataset, which are replaced by the simple English word. After preprocessing, the subset selected for experimentation contains 117,676 papers and 128,778 authors.

Graph Structure

We constructed a directed paper citation graph from the above mentioned dataset, for conducting the experimentation in this study. In paper citation networks the papers represent the vertices and their citation relationship represents the edges between them. Let graph *G=(V, E)*, where *V* represents the set of vertices and *E* represents the set of edges. Thus, the set of papers is represented by set *V={ v1, v2, … vn }*, where *n* is the total number of papers. For edges of paper citation graph the edges set can be seen as a set of any two paper's citation relationship, *E={(v1,v2), (v1, v3), … (vi, vj) },* in which *(vi, vj)* means vertex *vi* connects to vertex *vj*, i.e paper *vi* cites paper *vj*. Figure 1 shows a simple example of paper citation network with 4 vertices (publications) and five directed edges (citations).

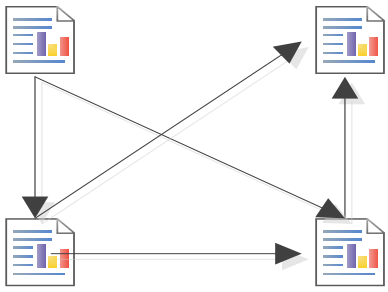

Figure 1: The Paper Citation Network

Selection of Queries

N-gram statistical package (Banerjee and Pedersen 2003) is used to find top frequent bigrams from the paper titles. A total of 100 from the top 260 frequent bigrams are selected as queries. Selection of hundred queries is made by just taking bigrams which represent commonly known research areas of computer science. The hundred queries are shown in Table 1.

Table 1: Selected 100 Bigram Queries

Selected queries				
Logic programming	Congestion control	Load balancing	Differential equations	Speech recognition
Hierarchical representation	Energy consumption	Moving object	Random walk	Sensitivity analysis
Systems modeling,	Computational complexity	Comparative studies	Intrusion detection	Fading channel
Business processes	Image retrieval	Organizing map	Relational databases	Virtual environment
Ubiquitous computing	Communication system,	Calculus theorem	Parallel programming	Hybrid systems
Feature selection	Machine translation	Sensitive visualization	Java programs	Intelligent tutoring
Data mining	Set programming	Video streaming	Discriminant analysis	Image segmentation
Service composition	Wavelet transform	Embedded system	Agent-based reasoning	Knowledge discovery
Collaborative filtering	User interface	Protein structure	Code generation	Multiple processors
Data reduction	Web search	Text classification	Geometric design	Peer-to-peer systems
Requirements engineering	Systems modeling	Surface reconstruction	Evolutionary algorithm	Fault tolerance
Open source	Combinatorial optimization	Fault diagnosis	Vehicle routing	Programming language
Markov models	Computer vision	Data stream	Mobile devices	Routing protocol
Information theory	Neural network	Efficient algorithm	Dimensionality reduction	Operating systems
Decision making	Reinforcement learning	Genetic algorithm	Data structures	Network monitoring
Graph cuts	Molecular dynamic	Trust management	Mixture model	Likelihood estimation
Storage system	Digital library	Desktop application	Performance analysis	Polynomial time
Signal processing	Pattern matching	Social network	Software development	Design methodology
UML model	Prototype implementation	Java program	Shared memory	Boolean functions
Problem solving	Security protocols	Resource allocation	Virtual reality	Quality service

Labeling for Topic based Clustering and Query Relevance

The aim of this study is domain based indexing of authors according to their field of interest. This involves considering the citations of a paper for indexing of an author with respect to his or her topic probability for that specific field. Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) is used to identify the topical features, and to organize the dataset into 100 topic based clusters. LDA is an unsupervised generative model which considers that each document is a mixture of some topics and each word's creation is associated with one of the document's topics. It generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers the discrete distributions of topics

per document. It must be noticed here that LDA cannot automatically label the modeled topics. We used Vector Space Model (VSM) (Salton, Wong and Yang 1975), for labeling of data modeled by LDA. We organized dataset into 100 clusters and then assigned these clusters, the most suitable titles by using vector space model (VSM). For assigning the titles we used 100 bigram queries shown in Table 1. We calculated the relevance of each cluster with each query using VSM and assigned the most relevant query as title to that cluster.

Calculation of Ad hoc h-index

For the purpose of comparison and validation of proposed method we introduced the Ad hoc h-index. Ad hoc h-index is based upon the general h-index of an author along with the query relevance of that author with a given topic. After assigning titles to the queries we calculated Ad hoc h-index values for all authors in the dataset. For this purpose we followed the algorithm below:

Algorithm: Calculate Ad hoc h-index

Abbreviations: Author *A*, Query Relevance *Q Rel*, Topic *T*, Vector Space model *VSM*
Required Input: *A*

1: $\forall\ A_i$: $Find\ h\_index(A)_i$
2: $\forall\ A_i$: $Find\ T(A)_i$
3: $\forall\ A_i\ find\ Q\ Rel_i\ applying\ VSM\ for\ each\ T_i$
4: $For\ A_i\ i \coloneqq 1\ to\ n\ do$
$Ad\ hoc\ h\_index(A)_i \coloneqq Q\ Rel_i * h\_index(A)_i$
end for

Calculation of Domain Specific Index (DSI)

Finally, we calculated the DSI values for indexing of all authors using the following algorithm:

Algorithm: Calculate DSI

Abbreviations: Paper *P*, Author *A*, Citations *C*, Query *Q*, Topic *T*, Topic Probability *T_Prob*, Domain Specific Citation *DSC*, Domain Specific Index *DSI,* proportionality constant *a*
Required Input: *P*

1: $\forall\ P_i$: $Find\ A_i, and\ C_i$
2: $for\ Q\ 1 \coloneqq 100\ do$
$\forall\ P_i\ find\ T_{Prob_i}\ of\ Pi\ using\ LDA$
end for
3: $for\ P_i\ i \coloneqq 1\ to\ n$
$for\ T_j j \coloneqq 1\ to\ 100$
$DSC_{i,j} = T\_Prob_{i,j} * C_i$
end for
end for
4: $For\ A_i\ i \coloneqq 1\ to\ n$
$For\ T_j j \coloneqq 1\ to\ 100$
$DSI_i\ (A_i, T_j) = \sqrt{\frac{\sum DSC_{i,j}}{a}}$
end for
end for

The DSI provides the standing of an author in each topic to portray a very clear picture of his or her interest in a given research area. Hence, the process of selection of topic specific authors becomes reliable.

RESULTS AND DISCUSSION

To evaluate the results of DSI, we conducted a series of experiments. Firstly, we calculated the general h-index for all authors; secondly, we identified the fields in which an author has worked in. Thirdly, we calculated the query relevance score of each author for each topic, and named the resultant as Ad-hoc h-index. We used this index for the purpose of evaluation of results of DSI. We applied the similarity measures, OSim and KSim for finding how well the results are calculated by using DSI rather than using generic h-index and Ad hoc h-index.

OSim is a similarity measure used for comparison of rankings. *OSIM(t1, t2)* indicates the degree of overlap between the top n results of two rankings, $\tau 1$ and $\tau 2$. We define the overlap of two sets *R1* and *R2* (each of size K) as follows:

$$OSim(\tau 1, \tau 2) = \frac{|R1 \cap R2|}{k}$$

Where *R1* and *R2* are the sets of rankings of top-k authors with respect to topics contained in $\tau 1$ and $\tau 2$, respectively. We performed our experiments for different values of K, i.e., k=25, 50, 75 and 100 for both methods i.e. the Ad-hoc method that uses general h-index, and the proposed DSI method. Figure 2 shows OSim for a number of pair of queries that have overlapping authors in DSI and Ad hoc h-index. We can see that DSI has separated the results more efficiently and there is less overlapping authors when we use DSI as compared to Ad hoc h-index. For example for top 25 authors, using ad-hoc h-index there were 1134 queries that have overlapping authors related to them, while for DSI there were 805 queries with overlapping authors. Thus, DSI reduces this overlap showing that it can distinguish topic specific authors in an effective way.

OSim only measures the degree of overlap of two rankings. Therefore, to indicate the degree to which the relative ordering of the top *n* results of two rankings are in agreement, we also included KSim which is a variant of the Kendall's τ distance measure. We present the KSim definition as follows:

Consider two ordered lists of rankings τ_1 and τ_2, each of length *n*. Let U be the union of the URLs in τ_1 and τ_2. If δ_1 is U -τ1, then let τ'_1 be the extension of τ_1, where τ'_1 contains δ_1 appearing after all the URLs in τ_1. Similarly, we can produce τ'_2 by extending τ_2. The KSim is defined as follows:

$$KSim(\tau 1, \tau 2) = \frac{|(u, v): \tau 1 \text{ and } \tau 2 \text{ agree on order of } (u, v), u \neq v|}{|U||U-1|/2}$$

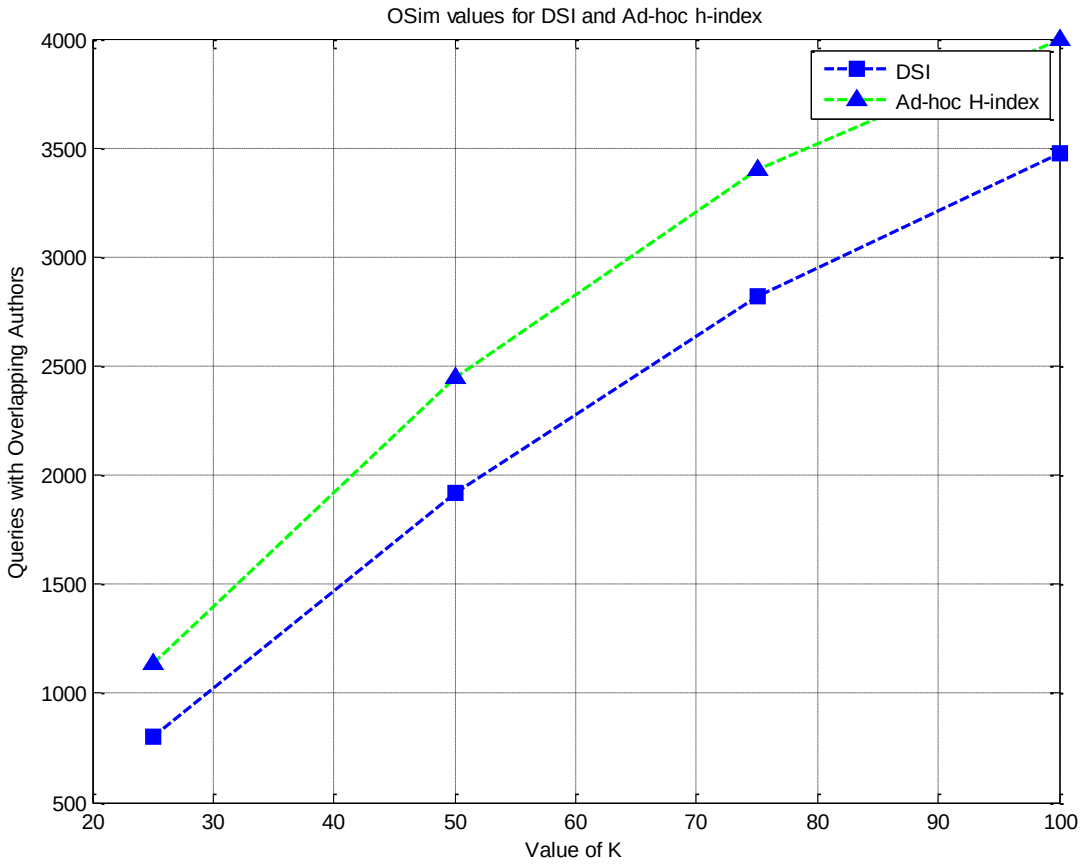

Figure 2: OSim Values for DSI and Ad-hoc h-index for Different Values of k

Figure 3 shows the numbers of pairs of queries that have different order of DSI and Ad hoc h-index values for authors in their respective author's listings for different values of top k authors (25, 50, 75 and 100). We notice that with smaller k, i.e. for top 25 and top 50 authors, DSI has lesser number of pairs that have different order and when the size of k increases (top 75 and top 100 authors), the results of DSI and Ad hoc h-index have become very close to each other. Most of the times we are interested in finding the top most authors of a domain so in such cases the DSI method can be used reliably, as it shows significantly better results for smaller values of k.

Table 2 shows the top 5 authors using DSI method and Ad hoc h-index for some selected queries. We have selected these queries on the assumption that authors belonging to these domains will have less overlap with other domains. In analyzing the citations received by these authors, we came to know that the citing sources are relevant to the topic of the cited source. For the purpose of evaluation we have calculated the average and standard deviation of the publication count and citation count of top 5 authors of all topics (number of citations and publications are subject to the selected dataset). We observed smaller values for standard deviation for authors retrieved by DSI as compared to Ad hoc h-index, showing that results of DSI are more stable. These authors appear at the top in the field because of the quality of their

work, as we believe that citations are a measure of quality of work, if source and destination of a citation are from same domain (Amjad et al. 2015b).

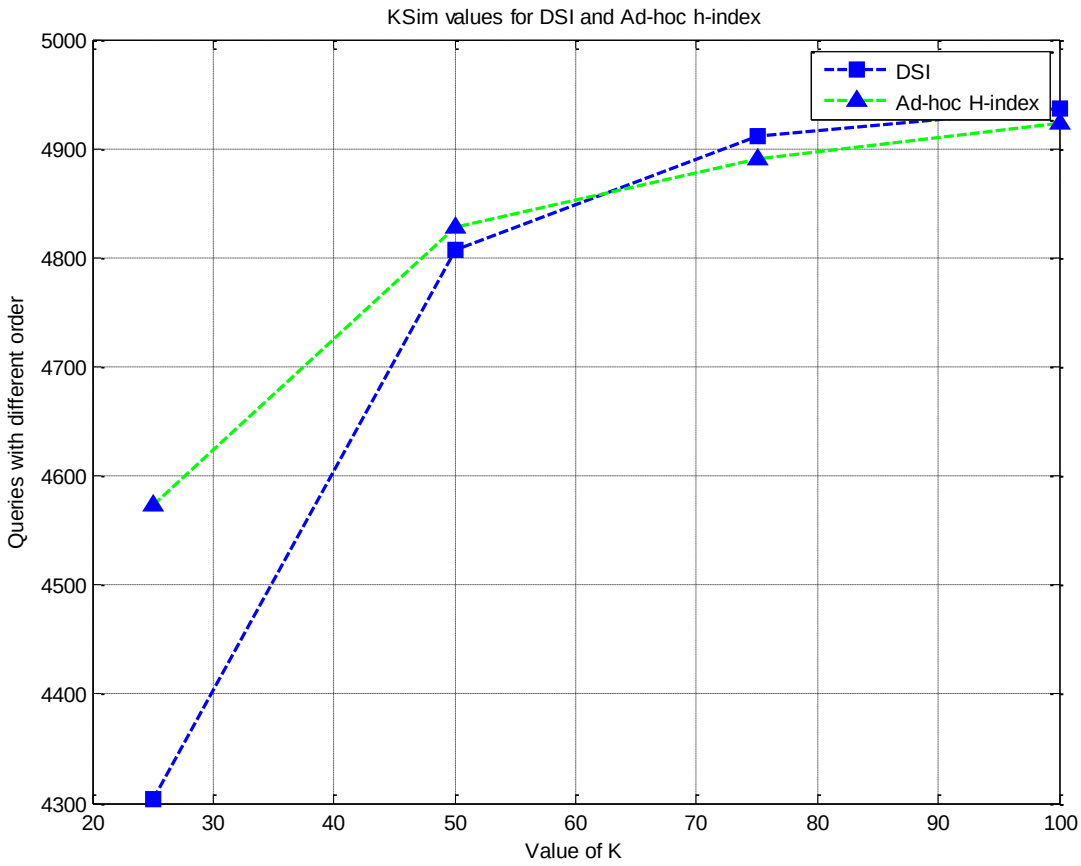

Figure 3: KSim Values for DSI and Ad-hoc h-index for Different Values of k

Now we present a brief qualitative analysis of top authors in these queries. Narendra Ahuja retrieved by DSI is a renowned name in the area of computer vision and robotics. David G. Lowe retrieved by Ad hoc h-index is another famous name in the area of computer vision, object recognition, and computational models of human vision. In our dataset David G. Lowe having 8 publications has received more citations but his citations are from more general topics, hence he cannot receive high rank in DSI, whereas, Narendra Ahuja appears on top position with citations from sources that are more relevant to the topic. Barbara Hammer is a professor in theoretical computer science for cognitive systems with neural networks as one of her research interests. Thomas Eiter from Vienna University of Technology, ranked on top by DSI for query logic programming has research interest in the following areas – knowledge representation and reasoning, computational logic, algorithms and complexity in AI, declarative problem solving, non-monotonic formalisms and databases. These areas are highly related to the term logic programming. Whereas, Ad hoc h-index ranked Francesca Rossi on top for this query and her profile shows artificial intelligence, constraint reasoning, preference modelling and aggregation, computational social choice which are less relevant to logic programming as compared to research interests of Thomas Eiter. Hermann Ney and Philipp

Koehn, ranked on top by DSI and Ad hoc h-index respectively both have machine translation as one of their research interests, and Table 2 shows a higher productivity and received citations for Hermann Ney as compared to Philipp Koehn.

Table 2: Top 5 Authors for Selected Queries using DSI and Ah-hoc h-index

DSI			**Ad hoc h-index**		
Query: Speech recognition					
	pubs	**Cit**		**pubs**	**Cit**
Narendra Ahuja	27	420	David G. Lowe	8	1356
Jie Yang	44	280	David Zhang	40	521
Thomas S. Huang	48	508	Anil K. Jain	54	1285
Philip R. Cohen	24	317	Paul J. Besl	3	643
Venu Govindaraju	21	128	David J. Kriegman	13	473
Avg	***32.8***	***330.6***		***23.6***	***855.6***
Stdev	***12.3***	***144.3***		***22.2***	***429.6***
Query: Neural network					
	pubs	**Cit**		**pubs**	**Cit**
Barbara Hammer	25	163	Amin Vahdat	51	1182
Edgar Korner	8	31	Donald F. Towsley	109	1576
Shun-ichi Amari	33	224	M. Frans Kaashoek	60	2867
Hiroyuki Nakahara	10	70	Michael N. Nelson	5	402
Heiko Wersing	9	70	Brent B. Welch	6	360
Avg	***17***	***111.6***		***46.2***	***1277.4***
Stdev	***11.3***	***79.4***		***43.2***	***1028.7***
Query: Logic programming					
	pubs	**Cit**		**pubs**	**Cit**
Thomas Eiter	61	571	Francesca Rossi	24	298
Diego Calvanese	35	414	Ken Kennedy	64	1320
Thomas Lukasiewicz	30	229	Georg Gottlob	58	950
Nicola Leone	35	418	Lotfi A. Zadeh	9	276
Torsten Schaub	23	86	Raghu Ramakrishnan	78	1608
Avg	***36.8***	***343.6***		***46.6***	***890.4***
Stdev	***14.4***	***188.2***		***28.9***	***598.2***
Query: Machine translation					
	pubs	**Cit**		**pubs**	**cit**
Hermann Ney	38	733	Philipp Koehn	9	378
Kevin Knight	33	467	Kevin Knight	33	467
Hwee Tou Ng	21	306	Daniel Marcu	41	929
Bonnie J. Dorr	26	137	Robert L. Mercer	6	206
Rada Mihalcea	18	155	Eiichiro Sumita	16	296
Avg	***27.2***	***359.6***		***21***	***455.2***
Stdev	***8.3***	***247.6***		***15.3***	***281.9***

* pubs, cits, Avg and Stdev represent publications, citations, average and standard deviation respectively

Figure 4 shows the standard deviation of citations received by the top 10 authors ranked by DSI and Ad hoc h-index. From DSI fit line and Ad hoc h-index fit line we can see that DSI has

relatively smaller standard deviation as compared to Ad hoc h-index showing the strength of the proposed method.

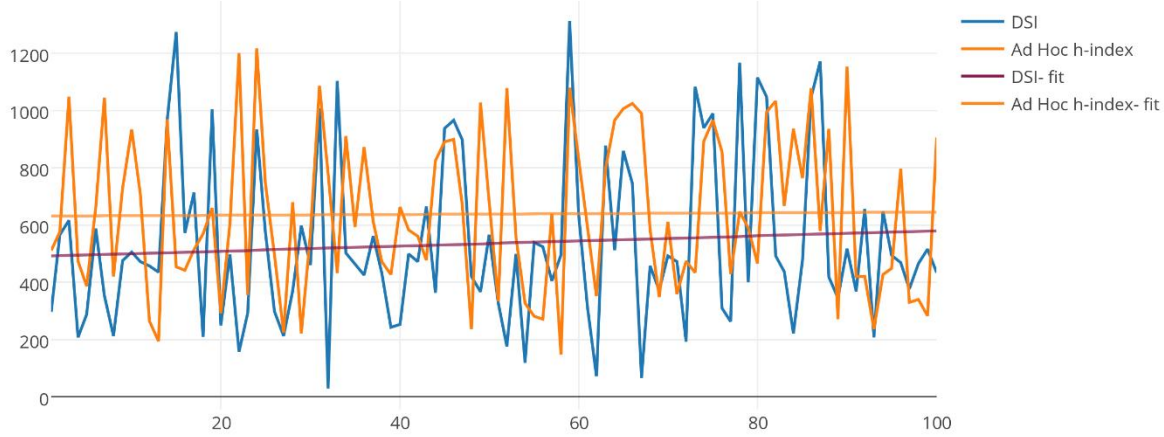

Figure 4: Standard Deviation for Citations of Top 10 Authors Ranked by DSI and Ad hoc h-index

CONCLUSIONS AND FUTURE WORK

In this study, we aim to find the standing of an author with respect to his or her level of expertise in a given topic, rather than a generic ranking. The proposed method, DSI, calculates topic specific index of an author, for all the topics an author has worked in. We applied LDA to find the probability of an author's association with a given topic. LDA is capable of generating soft clusters, ensuring that an author can belong to more than one cluster, which means that an author can work in more than one field. By using LDA, one paper of a researcher can be assigned to multiple topics with high probability score for the topic it is more related to, and low probability score for the topic with which it is related less. This is usually the case with research papers, as they can be relevant to more than one topic at the same time. We identified the fields of interest of all authors in the dataset, and calculated their respective Ad-hoc h-index values for each field. We also calculated the domain specific h-index (DSI) values for all authors, in all the fields they worked in. Results show that by using DSI, we can find the productivity indexes of all authors with respect to their fields of interest. DSI gives more realistic picture of an author's interest rather than the general h-index. DSI has ability to find not only the distinguished authors of any field, but also shows an author's level of expertise in all the topics he or she has worked in. The results show that DSI is a reasonable solution to find productivity and indexing of authors with respect to their fields of interest.

In future we are interested in involving the temporal dimension for indexing and productivity analysis along with topic sensitivity. Authors can switch their fields of interest with time, thus, considering the time dimension can be significant. Along with the topic sensitivity we can also further enhance the granularity by adding the contextual features as weighted vectors. This can help us in finding, for example, the most productive author of a field by adding number of publications as weighted vector, adding author's academic genealogy as the weighted vector

can contextualize the method to find the authors academic background. This particular contextualization can portray the drift of interest of authors from one topic to another with time.

ACKNOWLEDGEMENT

The work is supported by the Indigenous Ph.D. Fellowship Program of Higher Education Commission (HEC) Pakistan.

REFERENCES

Amjad, T., Daud, A., Che, D. and Akram, A. 2015a. MuICE: Mutual Influence and Citation Exclusivity Author Rank. *Information Processing and Management*, Vol.52, no.3: 374-386.

Amjad, T., Ding, Y., Daud, A., Xu, J., and Malic, V. 2015b. Topic-based heterogeneous rank. *Scientometrics*, Vol.104, no.1: 313–334.

Amjad, T., Daud, A. and Akram, A. 2015c. Mutual Influence based Ranking of Authors. *Mehran University Research Journal of Engineering & Technology*, Vol. 34, no. S1: 103-112

Banerjee, S. and Pedersen, T. 2003. The design, implementation, and use of the ngram statistics package, *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidenberg, pp. 370–381.

Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, 993–1022.

Bornmann, L., Mutz, R. and Daniel, H.-D. 2008. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine, *Journal of the American Society of Information Science & Technology,* Vol. 59, no.5: 830–837.

Burrell, Q.L. 2007a. Hirsch's h-index: A stochastic model. *Journal of Informetrics,* Vol. 1, no.1: 16–25.

Burrell, Q.L., 2007b. On the h-index, the size of the Hirsch core and Jin's A-index. *Journal of Informetrics,* Vol. 1, no. 2: 170–177.

Cabrerizo, F.J., Alonso, S., Herrera-Viedma, E. and Herrera, F. 2010. q2-Index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, Vol. 4, no.1: 23–28.

Chai, J.C., Hua, P.H., Rousseau, R. and Wan, J.K. 2008. The adapted pure h-index. In: H. Kretschmer and F. Havemann (eds), *Proceedings of WIS 2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics and Ninth COLLNET Meeting.* Humboldt-Universität zu Berlin.

Claro, J. and Costa, C.A. 2010. A made-to-measure indicator for cross-disciplinary bibliometric ranking of researchers performance. *Scientometrics,* Vol. 86, no.1: 113–123.

Daud, A., Saleem Yasir, S.M. and Muhammad, F. 2013. V-index an index based on consistent researcher productivity, Paper presented at Multi Topic Conference (INMIC), 2013 16th International. IEEE, pp. 61–65.

Egghe, L. 2006. An improvement of the H-index: the G-index. *ISSI Newsl*etter, Vol. 2, no.1: 8–9.

Egghe, L. and Rousseau, R. 2008. An h-index weighted by citation impact. *Information Processing & Management,* Vol. 44, no.2: 770–780.

Hirsch, J.E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 102, no.46: 16569–16572.

Jin, B., Liang, L., Rousseau, R. and Egghe, L. 2007. The R-and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, Vol. 52, no.6: 855–863.

Katsaros, D., Akritidis, L. and Bozanis, P. 2009. The f index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science and Technology*, Vol. 60, no. 5: 1051-1056.

Ko, Y.M. and Park, J.Y. 2013. An index for evaluating journals in a small domestic citation index database whose citation rate is generally very low: A test based on the Korea Citation Index (KCI) database. *Journal of Informetrics*, Vol. 7, no.2: 404–411.

Kosmulski, M. 2006. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, Vol. 2, no. 3: 4–6.

Salton, G., Wong, A. and Yang, C.-S. 1975. A vector space model for automatic indexing. *Communications of the ACM,* Vol. 18, no.11: 613–620.

Sekercioglu, C.H. 2008. Quantifying coauthor contributions. *Science,* Vol. 322, no.5900: 371.

Wan, J.-K., Hua, P.-H. and Rousseau, R. 2007. The pure h-index: calculating an author'sh-index by taking co-authors into account. *COLLNET Journal of Scientometrics and Information Management*, Vol.1, no. 2: 1-5.

Wildgaard, L., Schneider, J.W. and Larsen, B. 2014. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics,* Vol. 101, no 1: 125–158.