A general overview and bibliometric analysis of seven ACM hypertext and web conferences

Swati Agarwal*
Department of Computer Science,
Indraprastha Institute of Information Technology,
Delhi (IIITD) New Delhi, India
Email: swatia@iiitd.ac.in
*Corresponding author

Nitish Mittal
Department of Computer Engineering,
Netaji Subhas Institute of Technology (NSIT),
New Delhi, India
Email: nitishmittal94@gmail.com

Ashish Sureka
Industrial Software Systems (ISS),
ABB Corporate Research Centre,
Bangalore, India
Email: ashish.sureka@in.abb.com

Abstract: Bibliometric analysis of published scientific papers is a widely used practice to conduct quantitative evaluations and assessments of conferences. In this study, we performed an in-depth bibliometric, scientometric and exploratory analysis of ACM SIGWEB sponsored conferences by visually analysing the DBLP database. We conducted a series of experiments and empirical analysis to answer several questions. Our results showed that the articles published in SIGWEB conferences stem from a variety of countries while the degree of cross-country collaboration is relatively low and that most co-authors of publications are by researchers who all reside in the same country. Collectively, SIGWEB conferences have a higher hosting rate and local community participation in places where the USA and Europe make the greatest conference contributions. Our results showed that the participation of female authors in SIGWEB conferences is increasing while in contrast, there is a huge gender imbalance in leadership and official conference positions.

Keywords: ACM SIGWEB; bibliometric analysis; DBLP; gender imbalance; computer science research; CSR; hypertext and hypermedia; conference assessment; publication analysis; topic evolution; academia-industry collaboration; scholarly output; web engineering.

Biographical notes: Swati Agarwal is a Senior PhD candidate in the Information Management and Data Analytics Group at the IIIT-Delhi. She graduated from the IIIT-Delhi with a Master’s degree in Computer Science with a specialisation in Data Engineering in July 2013. Her research interests include social computing, social informatics, text mining and analytics, natural language processing and information visualisation.

Nitish Mittal is a Software Development and Data Science Engineer at the Wadi.com, Dubai, UAE. He graduated in June 2016 from the NSIT, Delhi with a Bachelor of Engineering in the field of computer science. His research interests lie in the areas of data mining, big data analytics, computational social networks and information visualisation. He has also performed research in the software testing domain in the field of combinatorial interaction testing.

Ashish Sureka is currently serving as a Principal Scientist at the ABB Corporate Research Centre (India) and an adjunct faculty member at the IIIT-Delhi. His research interests are in the areas of software analytics, social media analytics, software repositories and process mining.

This paper is a revised and expanded version of a paper entitled ‘A glance at seven ACM SIGWEB series of conferences’ presented at ACM SIGWEB Newsletter, New York, NY, USA, June 2016.

1 Research motivation and aim

ACM special interest group on hypertext and the web (ACM SIGWEB)\(^1\) is a community of scientists, researchers, scholars and professionals in academia and industry who study various forms of Hypertext (HT), Document Engineering (DOCENG), social networks, Information and Knowledge Management (CIKM), digital libraries, hypermedia, Web Science (WEBSCI), data mining, web search and user modelling. ACM SIGWEB conducts several activities such as sponsoring and organising conferences, delivering awards, organising education and research development activities. One of the major activities of ACM SIGWEB is organising and sponsoring symposia, workshops and conferences aimed at providing platforms for the exchange of ideas and information. ACM SIGWEB (formerly known as SIGLINK) is now in its third decade and sponsors seven annual conferences spanning a wide range of topics. The seven conferences sponsored by ACM SIGWEB (as of 8 May 2016) are listed below:

1 Hypertext and Social Media (HT)
2 Joint Conference on Digital Libraries (JCDL)
3 Document Engineering (DOCENG)
4 Web Science (WEBSCI)
5 Information and Knowledge Management (CIKM)
6 Web Search and Data Mining (WSDM)
7 User Modelling, Adaptation and Personalisation (UMAP)
These seven ACM SIGWEB conferences are prestigious in the broad area of the web. They are long-running conferences that are convened across the world and, together, they provide an international forum for the presentation of research results and the exchange of ideas between academic and industry professionals. Assessment and evaluation of the quality, impact, status and evolution of the various SIGWEB conferences is important for the ACM SIGWEB community and for conference sponsors, steering committees, science and technology policy makers and government bodies. We believe that a scientometric and bibliometric analysis provides a scientific approach to systematically explore and evaluate the state of these conferences. The Digital Bibliography and Library Project (DBLP) computer science bibliography provides open bibliographic information on major computer science journals and proceedings and is a popular and widely used service. As of 8 May, 2016, the DBLP dataset contains 3,337,182 articles by 1,717,004 distinct authors published in various proceedings, including 4,731 conferences, 1,474 journals and other sources such as PhD theses and technical reports. DBLP provides an XML dump of the DBLP records and entries containing the metadata and attributes of indexed publications. The specific research goal of this study is to conduct a bibliometric analysis based on the DBLP data and proceedings metadata (metadata of articles, authors, affiliations and conferences) extracted from the ACM digital library. We parsed the ACM digital library pages of each article and analyse various aspects of the seven ACM SIGWEB-sponsored conferences. The research objective of this study is to present a reflection and a historical or general overview of the seven SIGWEB conferences. Furthermore, to accomplish this goal, our objective is to analyse the metadata for all papers published from the beginning of the conferences through 17 September 2015 that are available from the DBLP dataset, the ACM digital library, and from data extracted using other open source APIs. In particular, we aim to conduct an analysis on the following facets of the SIGWEB conferences:

1. paper selectivity and yearly trend in terms of the number of accepted and submitted papers
2. average number of authors per paper
3. prolific and most productive authors.
4. university and industry collaboration as indicated by joint authorship
5. scholarly output of various countries in the world that have contributed to SIGWEB conferences
6. cross-country collaboration (as indicated by co-authorship from researchers from different countries)
7. female participation as authors, and an investigation of gender imbalances or gaps
8. number and percentage of papers published by authors from the country hosting the conference
9. common research topics and themes as indicated by the author-generated keywords
10. topic evolution and trends across years
11. funding agencies that have sponsored research published in SIGWEB papers.
2 Related work and research contributions

In this section, we present related work on bibliometric analysis of scientific paper publication records in computer science conferences. Our literature survey reveals that analysing conference paper publications is an area that has attracted the attention of several researchers, and a number of interesting insights can be extracted from bibliometric analysis of conference proceedings. Table 1 display a summary of our literature survey and shows a list of bibliometric based studies across several conferences that span multiple topics: software engineering, information retrieval, databases, Computational Intelligence (CI), Computational Linguistics (CL), Natural Language Processing (NLP), HT, human-robot interaction, human-computer interaction, requirements engineering (RE), software engineering, and theory and information visualisation.

Table 1 A summary of closely related work to the bibliometric analysis of computer science conferences

<table>
<thead>
<tr>
<th>Study</th>
<th>Conference(s)</th>
<th>Empirical analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Bartneck and Hu (2009)</td>
<td>CHI</td>
<td>Scholarly output and contribution of countries.</td>
</tr>
<tr>
<td>3 Bartneck (2010)</td>
<td>HRI</td>
<td>Scholarly output and contribution of countries, organisations and authors. Conference growth, internalisation, collaboration and military funding.</td>
</tr>
<tr>
<td>5 Sharma et al. (2016)</td>
<td>RE</td>
<td>Interdisciplinarity, topic distribution, collaboration, authorship trends and scholarship output of countries.</td>
</tr>
<tr>
<td>6 Agarwal et al. (2016b)</td>
<td>Several Computer Science Conferences</td>
<td>Gender imbalance and low participation of women in computer science research.</td>
</tr>
<tr>
<td>7 Nascimento et al. (2003)</td>
<td>SIGMOD</td>
<td>Co-authorship graph and small world phenomenon.</td>
</tr>
<tr>
<td>8 Smeaton et al. (2002)</td>
<td>SIGIR</td>
<td>Co-authorship analysis and topical analysis.</td>
</tr>
<tr>
<td>9 Van Eck and Waltman (2007)</td>
<td>IJCNN, FUZZ-IEEE, CEC</td>
<td>Association between concepts in the field, study of field evolution and identification of emerging topics.</td>
</tr>
<tr>
<td>10 Hall et al. (2008)</td>
<td>ACL, EACL, NAACL, COLING HLT, TINLAP, IJCNLP</td>
<td>Development and diversity of ideas in field.</td>
</tr>
<tr>
<td>11 Jacovi et al. (2006)</td>
<td>CSCW</td>
<td>Core and sub-community, chasm papers</td>
</tr>
<tr>
<td>12 Lopes et al. (2016)</td>
<td>WEBIST</td>
<td>Community analysis, theme and topical analysis</td>
</tr>
<tr>
<td>13 Furukawa et al. (2015)</td>
<td>WWW</td>
<td>Chronological changes in research topics</td>
</tr>
</tbody>
</table>
Table 1 shows a summary of work that is closely related to the study presented in this paper. Bergamaschi et al. (2012) conducted a quantitative analysis of three SIGWEB sponsored conferences: World Wide Web (WWW), Hypertext and Hypermedia (HT) and JCDL. They computed the PageRank metric for all papers published in WWW, HT and JCDL over a ten-year period and plotted their distribution by publication year to study variations in the quality of these conferences over the years (Bergamaschi et al., 2012). Bartneck and Hu (2009) conducted a quantitative analysis on the countries and organisations that have contributed to the success of the human-computer interaction (CHI) conference. Their study revealed that just 7.8% of the participating countries are responsible for 80% of the papers in the CHI proceedings and, hence, that the scholarly output or productivity of countries contributing to CHI is skewed (Bartneck and Hu, 2009). Bartneck (2010) presented a historical overview of the International Conference on human robot interaction (HRI). They studied aspects such as conference growth, internationalisation, collaboration, rankings of countries in terms of scholarly output, ranking of organisations and authors in terms of publication productivity and extent of military funding for HRI papers (Bartneck, 2010). Henry et al. (2007) analysed four major human-computer interaction conferences: Computer-Human Interaction (CHI), User Interface Software and Technology (UIST), Advanced Visual Interfaces (AVI) and Symposium on Information Visualisation (InfoVis). They studied co-authorship and citation networks as well as conference impact. Their study revealed a power-law degree distribution with a few widely collaborating authors and highly cited articles (Henry et al., 2007). Sharma et al. (2016) presented insights from a bibliometric analysis of all the papers in the RE series of conferences over a period of 11 years, from 2005 to 2015. They studied several aspects such as the cross-disciplinary and interdisciplinary nature of RE research, topical analysis, university and industry collaboration, internal and external collaborations, authorship trends and scholarly output of various countries in the world (Sharma et al., 2016).

Agarwal et al. (2016b) conducted an exploratory data analysis on the DBLP bibliographical database to study the gender gap, employment imbalance and the low female participation in computer science research (CSR). Their study revealed that in CSR, only 21% of the authors (actively publishing in CSR conference proceedings) are female; the rest (79%) are male (Agarwal et al., 2016b). Nascimento et al. (2003) investigated the co-authorship graph obtained from all papers published at SIGMOD (the ACM Special Interest Group on Management of Data) between the year 1975 and 2002. Their study revealed that SIGMOD’s co-authorship graph exhibits a small-world phenomenon (Nascimento et al., 2003). Smeaton et al. (2002) conducted an analysis of 25 years of SIGIR (Special Interest Group on Information Retrieval) conference papers. They conducted an analysis of topics (the rise and fall of topics over the years) and a co-authorship analysis (Smeaton et al., 2002). Van Eck and Waltman (2007) conducted an analysis of three major conferences in the CI field: Joint Conference on Neural Networks (IJCNN), Conference on Fuzzy Systems (FUZZ-IEEE) and the IEEE Congress on Evolutionary Computation (CEC). They studied the associations between the main concepts in these fields for the periods 1996–2000 and 2001–2005 and the evolution of the fields and identified emerging areas within CI (Van Eck and Waltman, 2007). Hall et al. (2008) applied unsupervised topic modelling to papers published in several major CL and NLP conferences to study the development of ideas in these fields. Their study revealed convergence over time in the topic coverage of Association for Computational Linguistics (ACL), (COLING) International Conference on
Computational Linguistics and Empirical Methods on Natural Language Processing (EMNLP) as well an expansion of topic diversity. Jacovi et al. (2006) conducted an investigation of the structure of the ACM conference on Computer-Supported Cooperative Work (CSCW) and Social Computing conference using a structural analysis of the citation graph of the conference papers and related articles. They extracted CSCW’s core and most prominent clusters and also defined a metric to identify and analyse chasm-papers (scientific papers that are cited significantly more outside the conference than within) (Jacovi et al., 2006). Lopes et al. (2016) conducted a comprehensive analysis of the first ten editions of the Web Information Systems and Technologies (WEBIST) series of conferences. They conducted a community analysis of the conference by applying techniques based on social network analysis and also studied the prevalent themes and topics of publications from the conference series (Lopes et al., 2016). Furukawa et al. (2015) proposed a technique to mine the chronological changes in research topics as extracted from scientific papers published in conference proceedings and conference sessions using a text-analytics method. They studied the chronological networks of conference sessions from World Wide Web (WWW) conferences between the year 2002 and 2011 and presented an information visualisation of the scientific and technical streams formed by the session sequences (Furukawa et al., 2015). Agarwal et al. (2016a) presented a bibliometric analysis of nine ACM SIGWEB cooperating conferences. They conducted an analysis on ten years of publications records and research community of these conferences. The analysed the growth rate of the conference, evolution of various common and emerging research topics. They also conduct a quantity vs. impact analysis on the community of SIGWEB cooperating conferences. They analyse the most prolific authors in the community and the ranking of these conferences based on the number citations (h-index) and publication records.

2.1 Research contributions

In contrast to the existing work, the study presented in this paper makes several novel research contributions. Our work is an extension to the newsletter article by Agarwal et al. (2016d), who presented the first bibliometric and exploratory study on ACM SIGWEB-sponsored conferences. They analysed several years of publication records and calculated the overall participation rate of authors in various SIGWEB conferences and computed the number of unique affiliations of the authors who published papers in SIGWEB conference proceedings over the last 35 years. They also analysed the increment and decrement rate of the overall acceptance of papers in SIGWEB conferences. However, Agarwal et al. (2016d) presented an overview of SIGWEB conferences as an entire community. In contrast, the study presented in this paper is a much deeper and more comprehensive analysis of SIGWEB-sponsored conferences because all ACM SIGWEB conferences have their own distinguished community. We extended the research questions (RQs) proposed in Agarwal et al. (2016d) and conducted a timeline-based analysis of each conference individually. In contrast to Agarwal et al. (2016d), we present a timeline-based review of the average number of authors who have published papers in SIGWEB conference proceedings. We extend their proposed RQs and compute the correlation between the number of submissions received and the number of published papers for each conference individually. In this paper, we propose six additional research questions that cover dimensions not covered in previous literature.
Because the aim of this study is to present a comprehensive analysis of the SIGWEB community, we identify the most prolific profiles (those publishing the maximum number of papers) in SIGWEB conference proceedings. Agarwal et al. (2016d) presented the growth rate of conferences by analysing the number of distinct affiliations participating in these conferences. We extend that research question to analyse the scholarly output of each country because the author affiliations can change. Because SIGWEB focuses on a wide variety of topics that are of interest to both academic and industry researchers, we identified the type of each affiliation and analysed the prominence of industry-academia collaboration in SIGWEB conference proceedings. We further analyse the growth rate of cross-country collaboration in SIGWEB conferences by identifying the unique countries of co-authors. In contrast to Agarwal et al. (2016d), we analyse host-country participation in a SIGWEB conference by computing the percentage of papers published by authors from the country hosting the conference. We further study the evolution of common research topics, the themes of the conferences and their trends across the years. We also investigate the gender imbalance in the SIGWEB community by computing the participation rate of female researchers who are authors or hold leadership positions. We also analyse the participation and engagement of external funding resources that sponsor research published in SIGWEB conference proceedings.

3 Experimental setup and data collection

In the work presented in this paper, we conducted a study on the publication records of seven ACM SIGWEB-sponsored conferences. DBLP provides the largest open source bibliographical database of computer science journals and conference proceedings. We downloaded a snapshot of the DBLP database published on September 17, 2015 consisting of 4.7 million records, including 15.9K books, 1.7 million articles published in conference proceedings and 1.34 million journal articles and other publications. The focus of this paper is to conduct a study on ACM SIGWEB conferences; therefore, we parsed the DBLP database and extracted the publication records of seven conferences (as per the 2016 update on ACM SIGWEB website). We extracted a total of 162 conference proceedings for these seven conferences that are published in the ACM Digital Library. For all seven conferences, we were able to extract a total of 9,141 unique publication entries. The aim of this paper is to conduct a comprehensive analysis of these seven SIGWEB conferences across various facets and dimensions including articles, authors, topics, affiliations, countries and co-authorship analysis within the SIGWEB community. The DBLP database provides only the publication records of a conference or an article such as the conference name, authors’ names, year published and publisher (ACM, IEEE and Springer). However, these records are insufficient to conduct a comprehensive study on the ACM SIGWEB community. Therefore, we also extracted additional metadata for each SIGWEB conference by parsing the contents of the ACM Digital Library. Figure 1 illustrates the basic design and architecture of the experimental dataset collection process. The aim of ACM metadata collection is to enrich the information of four attributes of the bibliographical entries in the DBLP database. These are primarily named articles, authors, affiliations and conferences (A3C). We discuss each of these attributes in detail in the following subsections.
3.1 Conference metadata enrichment

We accessed the DOI link to ACM Digital Library of each conference by creating a unique conference link from the ‘ee’ field available in DBLP records. The ‘ee’ field contains the global (and sometimes local) URL that redirects users to the electronic version of an article available on the publisher’s server. Among 9,141 unique articles published in 162 proceedings records for the ACM SIGWEB conference, the DBLP database contains the DOI link for only 87% of the publications (7,996) co-authored by 13,610 unique authors. Due to the missing ‘ee’ links, we were not able to extract the ACM metadata for the remaining 13% (1,145) of the articles. Instead, we used the ‘ee’ link available for conference proceedings and extracted the following three fields to enrich the conference metadata:

1. **Dignitaries’ profiles:** the profiles of the dignitaries each conference (referred to as the authors of a conference proceeding in the ACM Digital Library) contains the list of editors, program chairs and general chairs of the proceedings for that particular year. We scraped the DOI link for each SIGWEB conference (available in DBLP records) and extracted the names from all the profiles available under the proceeding’s author list. During our manual inspection of ACM pages, we observed that each conference page provides the names of general chairs, but the availability of information for program chairs and editors varies among the different
conferences. Figures 2(a) and 2(b) show examples of two such proceedings that are available from the published ACM Digital Library. These snapshots reveal that the conference proceeding of CIKM 2011 contains only editor information while the proceedings of WSDM 2014 contain only information about general chairs and program chairs. For our experimental dataset, we were able to extract a total of 356 such profiles available from 135 conference proceedings.

2 Conference location: the DBLP database records the full name of conference proceedings along with the location and year of the conference. However, for older conferences such as CIKM and HT, there were several entries missing from our database snapshot. Therefore, we implemented a web page crawler to parse the ACM pages (accessed using ‘ee’ link of the conference) and fetch the complete name of the conference location from the appropriate place on the ACM pages. Figures 2(c) and 2(d) show snapshots of the complete names of conference proceedings and locations. Because the format of location names varies for different conferences [for example, USA is abbreviated in Figure 2(c) but the full name of the UK is written out in Figure 2(d)], we applied geocoding to country names to ensure consistency among the extracted locations. As shown in the designed architecture, we used the Google Geocoding API on the extracted locations to obtain the exact country code for the conference location. Using the ACM metadata and Google map API, we were able to extract the locations of 95% of the conferences (129 out of 135).

3 Acceptance rate: the ACM digital library provides a timeline of the overall acceptance rate for a conference computed over the last 10 to 15 years of conference proceedings. It also provides the statistics for the number of submissions received for the conference and the number of articles published that year. Despite the fact that each conference has a different acceptance rate for different types of submissions (i.e., posters, short papers, full papers and other articles), the ACM records only the overall acceptance rate and the total number of articles received by the conference for that particular year. This information is not public: it is provided to the digital library for the conference chairs only. We extracted this information for further analysis on SIGWEB conferences to enrich our experimental dataset, which was originally parsed from the DBLP records.

Figure 2 Snapshot examples of conference metadata available on the proceedings pages in the ACM digital library (a) profiles of dignitaries available from CIKM 2011 (b) profiles of dignitaries available from WSDM 2014 (c) conference location of WebSci 2014 (d) conference location of WebSci 2015 (see online version for colours)
3.2 Articles’ metadata enrichment

Similar to conference metadata extraction, we also used the ‘ee’ field available in the DBLP database to access the DOI link of each article from the digital library. We scraped the ACM page of each publication entry and extracted the following six additional attributes:

1. **ACM concepts and author tags:** each article published in the ACM digital library contains a list of certain key terms that define the related technical keywords and topic of the paper. In the ACM digital library, an article contains two types of key terms: author tags and ACM concepts. Author tags are user-defined keywords that appear under the keyword category of an ACM publication, while the ACM concepts of a paper are the keywords identified and sponsored by IBM Watson\(^6\). IBM Watson is a sponsored application used in the ACM digital library that analyses the text of papers and extracts a collection of semantic connections in the article. For example, for an article, ‘Penguins in sweaters, or serendipitous entity search on user-generated content’\(^7\) published in CIKM 2013, while the authors provided tags such as ‘entity networks’, ‘entity search’, ‘information systems applications’, ‘interestingness’, ‘metadata’, and ‘serendipity’, IBM Watson labels the concepts as ‘Yahoo! Answers’, ‘Wikipedia’, ‘question answering’, ‘metadata’, ‘web search engine’ and ‘user generated content’. To study the evolution of topics in various SIGWEB conferences, we extracted both the author tags and the concepts for each article. We parsed the ACM webpages of 7,996 articles (for which the DBLP contains the DOI link) and extracted the author tags for 99% of the articles (7,948) and ACM concepts for 79% of the articles (6,310). Based on the availability of author tags and concepts, we were able to extract a total of 5,743 distinct author tags and 14,203 unique ACM concepts.

2. **ACM references:** references are papers or reports that have been referred to or mentioned in a given article. The ACM records the complete bibliographical entry for each reference mentioned in the paper but it provides the paper id (a link, making an article clickable) of only those references that are also published under ACM digital library. For example, the article, “Understanding and controlling the filter bubble through interactive visualisation: a user study”\(^8\) by Nagulendra and Vassileva (2014) (published in the HyperText 2014 conference proceedings) has 36 references, among which only 20 reference articles are published by the ACM. We selected the IDs of these 20 references and referred to them as ‘ACM References’. For 7,996 SIGWEB articles, we were able to extract a total of 31,198 references, among which only three, 665 distinct papers are published by the ACM and, consequently, have an assigned ACM paper id. We extracted the list of these articles by scraping the ‘Reference’ section available on the ACM webpages.

3. **ACM citations:** the citations of a paper are the articles (reports, theses or scientific publications) that are referred to or mentioned in the given article. The ACM digital library computes the citation count of an article \(A\) for only those articles that are also published in the ACM digital library. For example, the article “Microblog-genre noise and impact on semantic annotation accuracy”\(^9\) by Derczynski et al. (2013) (published in the HT 2013 proceedings) has received 64 citations. However, ACM records the citation count as nine because only nine of the articles citing this work
are published in the ACM digital library. We extracted a list of these articles and refer to them as ACM Citations. ACM citations can be extracted by scraping the ‘cited by’ section available on ACM webpages. For all 7,996 articles, we were able to extract 3,044 ACM Citations, while the total count of ‘cited by’ articles is 32,627.

4 Unique authors: for each article, the DBLP records the distinct names of all the co-authors. For any two or more authors who have similar names, DBLP disambiguates their profiles by adding four padding bits as a suffix (0001, 0002, 0003, etc.) in their name. For example, three such similar names would be recorded as ‘Jie Wu’, ‘Jie Wu 0001’ and ‘Jie Wu 0002’. Because we enriched our dataset by extracting the ACM metadata for the DBLP records, we also extracted the ACM id for all authors to ensure consistency in our dataset. We scraped the ACM page of each article and extracted the unique ACM author id of each co-author in the paper. We then mapped the DBLP and ACM ids of authors to avoid inconsistency and noise in the dataset.

5 Publication type: the DBLP dataset provides information about the number of pages each article occupies. Based on the page length of each article, we divided them into one of three categories: full paper, short paper, demo paper or poster paper. To assign these labels, we manually checked the previous 15 years of SIGWEB conferences and checked the page limits for each category for each year. For the majority of SIGWEB conferences, the page counts of regular or full papers and short papers are eight and four pages, respectively. Demo or poster papers are two pages long. However, we observe that in the DBLP records, the page counts of SIGWEB articles vary from 1 to 60 for reasons unknown to us.

6 Funding source: funding resources are external sources that sponsor research. Funding resources are acknowledged during the copyright form submission of the article. ACM records the copyright form information for each article under the Publication tab of the digital library. If a research paper published in ACM conferences is supported by either a government organisation, industry or a research lab, ACM records that information and makes it publicly available as metadata of that paper, referred to as ‘funding resource’. We scraped the ACM pages of all publications available in our database and extracted the funding resource information for each article, if available. For example, the research published in the article “Modelling and Predicting Retweeting Dynamics on Microblogging Platforms” by Gao et al. (2015) has four external funding sources: “Doctoral Fund of Ministry of Education of China”, “Excellent Middle-Aged and Youth Scientists of Shandong Province”, “National Natural Science Foundation of China” and “Natural Science Foundation of Shandong Province”.

3.3 Authors’ metadata enrichment

To enrich the author metadata, we extracted the unique ACM numerical identifier for each author and created their ACM profile links by embedding their profile id into the lead author’s ACM page URL (http://dl.acm.org/author_page.cfm?id=author_id’). We used an HTML parser to scrape the author profile pages in the ACM digital library and extracted the following five attributes for each author available in our database:
1. **General information:** from the general information about the author, we extracted the author’s full name from the ACM page and their publication history (referred to as ‘bibliometrics’ in the ACM digital library). An author’s bibliometrics includes the number of articles published by the ACM digital library, the year of the author’s first and most recent ACM publication, the total number of citations this author has received from other ACM papers and the average number of ACM citations this author has received (refer to Section 3.2) to see the information available for each article.

2. **Colleagues and co-authors:** to mine the degree of collaboration in SIGWEB conferences, we created a network of authors in which a node represents an author and the link between two nodes represents co-authorship collaboration between those two authors in SIGWEB conference proceedings. The ACM records the distinct name of all co-authors (referred to as ‘colleagues’ of the given author). We parsed the ACM profile page of each author and extracted the ACM numerical id for each of that author’s colleagues. To create a network of all the co-authors who have published articles together in only SIGWEB conference proceedings, we created a network for the authors and co-authors available in our database. Our database reveals that in SIGWEB-sponsored conferences, the average number of colleagues (co-authors) per author is 27.3.

3. **Gender identification:** in addition to the ACM metadata for an author, we further extracted the gender of all distinct authors who have published articles in SIGWEB conference proceedings. We used the Genderise.io API to determine the gender of the authors in our database. As of 8 May, 2016, the Genderize.io database contains 216,286 distinct names from 79 countries and 89 languages. The API takes the person’s first name as input and classifies it into one of three classes: male, female and not applicable (NA). The API also returns a confidence score, referred to as a certainty factor or probability of gender. The confidence score varies on a scale from zero to one. If the gender of the given first name cannot be determined, the API returns NA. For a name such as Peter, the confidence factor is nearly one for the gender male, whereas for names such as Stevie, the confidence factor ranges between 0.6 and 0.7 (because several first names similar to Stevie have been used both for male and female names in the past). When the confidence score for the first-name gender match was below 60% (= 0.6), we classified it as NA. Note that several author names in our database did not contain the full author name: only initials were available for the first name. In such cases, we classified the gender as NA (for example, A. B. Smith). Among 13,610 unique profiles, we were able to extract the gender of 10,973 authors (7,956 were classified as male and 3,017 were classified as female), and 2,637 authors were classified as NA (including those authors with only initials available for the first name).

3.4 **Affiliation metadata extraction**

As discussed above, the DBLP database provides very limited bibliometric information about publication records, conference proceedings and authors’ profiles. Therefore, to conduct a comprehensive and detailed bibliometric analysis of the publications database, we enriched our data using ACM metadata and other third party APIs. In bibliometric
and publication information, author affiliation is an important attribute. Unlike other digital libraries such as IEEE and Springer, ACM records the affiliations of all the authors and co-authors who publish papers in ACM conference proceedings. Furthermore, it also keeps a track of the history of authors’ affiliations because they can change organisations. Therefore, the affiliation information may vary for different articles published by the same authors. To disambiguate the affiliation records of all the authors, we divided the affiliation metadata into two categories: article-specific affiliations and author-specific affiliations. We discuss both of these categories in the following subsections:

**Algorithm 1 Author Affiliation Disambiguation**

**Data:** Paper_id $P_id$, Author_id $A_id$

**Result:** Affiliations $AF$

**Algorithm** $ExtractAffiliations(P_id)$

1. Extract names of all co-authors
2. **for each** author $A_i \in A$
3. Extract the affiliation $AF_i$ of $A_i$
4. **add affiliation** $AF_i$ in the list $AF$

**end**

**Algorithm** $ExtractAffiliations(A_id)$

5. Extract the list of affiliation records $AF_i$ of given author $A_id$
6. create a list $AF$ from affiliations $AF_i$

1. **Article-specific affiliation:** article-specific affiliations are independent from any changes in the affiliation records of an author. In this category, we parsed the article webpage available in the ACM digital library and extracted the name of the affiliation of each author. Given an ACM paper id as a parameter, function $ExtractAffiliations(P_id)$ of Algorithm 1 returns the list of affiliations that map to the co-authors of the paper. Figure 3(a) shows an example of the affiliations specific to a single article that will be extracted for all co-authors.

2. **Author specific affiliation:** to enrich the authors’ metadata, we extracted the affiliation history of each author both from our database and from those available in the ACM digital library. Given an ACM author id as a parameter, function $ExtractAffiliations(A_id)$ of Algorithm 1 returns the list of all affiliations recorded in the ACM digital library. We parsed the ACM page for each author and built a crawler to extract the names of all the affiliations listed on that profile. Figure 3(b) shows an example of affiliation history records extracted from an author’s profile. For all the authors contributing to SIGWEB conferences and available in our database (13,610), we were able to collect a total of 8,734 distinct affiliations.

3. **Category of affiliation or organisation:** it was apparent that with the emergence of various technical and application oriented studies in the area of WEBSCI, researchers from various industries and academic institutions have been publishing articles in ACM SIGWEB conference proceedings. ACM records the name and
history of all the researchers who have published articles in the ACM Digital Library. However, it does not reveal the type of the affiliation (industry or academia). We used four different open source APIs arranged in a cascaded ensemble manner to identify the type of each affiliation. Figure 1 shows the complete design for this affiliation metadata extraction method. We first applied OpenStreetMap (OSM) API\textsuperscript{12} to 8,734 distinct affiliations (extracted in Section 3.3) to identify their type and geolocations on a world map. If OSM cannot determine the type or geolocation of any of the affiliations, we next applied Alchemy API\textsuperscript{13} to those entries. The Alchemy APIs predict the affiliation type with a reasonably high accuracy. After applying OSM and Alchemy on 8,734 affiliations, we were able to extract the types of 6,943 entries (~80\%) correctly; 5,587 affiliations were classified as academic college and universities, while 1,356 affiliations were classified as industries. However, Alchemy does not compute the geolocations of these affiliations accurately; consequently, we also applied the Google Map API\textsuperscript{14} and the Bing Geocoding API\textsuperscript{15} to pinpoint the locations of the remaining affiliations. Among 8,734 unique affiliations, we were able to extract the geolocations for 8,194 records, but were unable to identify the type or geolocations for ten records.

We created our experimental dataset by combining the four main attributes, Article, Author, Affiliation and Conference metadata (\textit{A3C}) extracted for the SIGWEB entries that appeared in the DBLP bibliographical records. The enriched database for \textit{A3C} is illustrated in Figure 1 and labelled as databases \textit{A}, \textit{B}, \textit{C} and \textit{D}. Our complete dataset is publicly available on Mendeley Data\textsuperscript{16} so that future researchers can replicate our research or use it for benchmarking, comparison and extension purposes (Agarwal et al., 2016c).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{A snapshot examples for author affiliation disambiguation in the ACM digital library (a) affiliation details of co-authors (b) affiliation history of one author (see online version for colours)}
\end{figure}

\begin{flushleft}
\textbf{On Integrating Network and Community Discovery}
\end{flushleft}

\textbf{Full Text:} [pdf]

\textbf{Authors:}
- Zhiyu Liu \quad University of Illinois at Urbana-Champaign, Urbana, IL, USA
- Charu Aggarwal \quad IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
- Jianwei Han \quad University of Illinois at Urbana-Champaign, Urbana, IL, USA

\textbf{Affiliation history}
- Northwestern University
- Simon Fraser University
- University of Illinois at Urbana-Champaign
- University of Illinois
- IEEE
4 Experimental results and analysis

In this section, we describe an exploratory bibliometric analysis performed on seven WEBSCI conferences sponsored by ACM SIGWEB community. As discussed in Section 3, all the UMAP proceedings are published in non-ACM digital libraries. Therefore, we conducted our analysis on only the remaining six conferences. We analysed the DBLP records and the ACM metadata of these conferences to study the various aspects of SIGWEB conferences and communities that can be useful for a variety of end users (the ACM and its sponsors, authors and policy makers). We proposed 11 RQs across various facets and dimensions of these SIGWEB conferences. To answer these RQs, we performed statistical and exploratory data analysis on the publications database enriched with article, author, affiliation and conference (A3C) attributes.

RQ1 What is the trend of submissions, publications and acceptance rate of each SIGWEB conference over the years?

Figure 4 variation in the number of submissions received, published articles and acceptance rate in each SIGWEB conference for each year (see online version for colours)

In this research question, we analysed the growth rate of ACM SIGWEB conferences based on the most basic and key dimension of bibliometric analysis by computing a timeline based variation of the overall publication records along with the overall acceptance rate of each SIGWEB conference. As discussed in Section 3.1, despite the fact that each conference category has a different manuscript acceptance rate, ACM records the overall number of submissions and acceptance rate of the conference for that particular year. We analysed the rate of increase or decrease in publications of a conference over a period of time. We extracted a list of the overall acceptance rate for all SIGWEB conferences available for the past 10 to 15 years and analysed the variation in
acceptance rate of SIGWEB conferences based on the maximum metadata available for each conference. Figure 4 contains six sub-plots (stacked column charts) showing the fractional variation in the number of submissions received, papers published and the acceptance rate of six SIGWEB conferences. Figure 4 reveals that for HT, JCDL and DocEng, we were able to extract a maximum of 12 years of acceptance-rate information, while for CIKM and WSDM, the ACM Digital Library contains only eight to nine years of data. Despite the fact that WebSci is a relatively new conference (seven years old), ACM provides only four years of metadata WebSci.

<table>
<thead>
<tr>
<th>Conference</th>
<th>CIKM</th>
<th>HT</th>
<th>JCDL</th>
<th>DocEng</th>
<th>WebSci</th>
<th>WSDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>pub vs. sub</td>
<td>0.71</td>
<td>0.66</td>
<td>0.94</td>
<td>0.85</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>sub vs. acc</td>
<td>–0.08</td>
<td>–0.37</td>
<td>–0.59</td>
<td>–0.06</td>
<td>–0.96</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Figure 4 shows that for the majority of SIGWEB conferences (CIKM, JCDL and WSDM), the acceptance rate varies between 20% and 40%. In the HT and DocEng conferences, the acceptance rate fluctuates very frequently but with only minute changes (1% to 5%) in the number of submissions the conferences receive. The graphs in Figure 4 also show that, over the years, the acceptance rate for each conference has decreased by approximately 5% to 10%. We observe that rate of acceptance in these conference increases whenever the rate of received submissions falls below the rate of the previous year. For example, in 2013, the acceptance rate of the HT conference increased by 40% and the number of submissions received fell by 22% that year. Similarly, in 2004, the acceptance rates of DocEng conference increased by 15% while the number of submissions declined by 8%. As a new conference, WSDM has steadily decreased its acceptance rate, and recently, it has been publishing only approximately 15% of the submissions it receives.

For example, the last sentence in the following paragraph would read Table 2 shows that the WSDM conference has a strong positive correlation between submissions and publications and between submissions and acceptance rate.

We further computed Pearson’s correlations between the number of publications and the number of received submissions for each conference, as shown in Table 2. Table 2 reveals that in the CIKM, JCDL and DocEng conferences, the number of publications and submissions have a strong positive correlation. Hence, the number of publications has a tendency to increase along with an increment in the number of paper submissions. In contrast, the HT conference has a positive but only moderate correlation between the number of publications and submissions. Table 2 also shows that in the CIKM and DocEng conferences, the submissions and acceptance rates have either no correlation or a negligible negative correlation (–0.08 and –0.06 respectively). It is also evident from Figure 4 that, despite frequent fluctuations, the CIKM and DocEng conferences have acceptance rates between specific ranges. Table 2 shows that the WSDM conference has a strong positive correlation between publications and submissions and between acceptance rate and received submissions.

RQ2 What is the average number of authors per paper across years and conferences?
Newman (2001) investigated the structure of scientific collaboration networks by analysing data drawn from databases such as MEDLINE (biomedical research), the Los Alamos e-Print Archive (physics), and NCSTRL (computer science). Their findings revealed that the average paper had approximately three authors. However, they discovered notable exceptions in fields such as theoretical high-energy physics and computer science, in which smaller collaborations are the norm, resulting in an average of approximately two authors per paper, and their analysis of the SPIRES high-energy physics database showed an average of nine authors per paper. Guan and Ma (2004) analysed 9,632 computer science papers recorded in the INSPEC database between 1993 and 2002. Their study revealed that the average number of authors per paper varies between 1.5 and 2.5 across countries (USA, UK, Germany, India, Japan and China) and that the average number of authors for a scientific paper in Asian countries is slightly higher than for the number of authors in Occidental countries. Calculating statistics such as the average number of authors per scientific paper is useful in understanding collaboration networks, collaboration profiles and co-authorship indices. Our motivation was to understand whether researchers in the SIGWEB community work in small or large groups and what the distribution of articles is across single, paired, multi and mega (more than six or seven) authors per paper.

Figure 5 displays a line chart showing the average number of authors per paper for each individual SIGWEB conference and for all SIGWEB conferences together. We computed the average total number of authors in each paper published in a conference in a particular year and calculated their average. In Figure 5, the dots represent the availability of proceedings and co-author data. Our empirical analysis supports the results found by previous studies that investigated the average numbers of authors for scientific papers. Figure 5 reveals that overall average number of authors in all SIGWEB conferences is 2.83, but varies from a minimum average of 1.96 up to a maximum of 3.39 for the individual conference. We observe that the degree of collaboration in SIGWEB conferences has increased over the past decade. When we split our database into a timeline of three decades (<1994, 1994–2004 and 2005–2015), we found that, in recent years, the average number of authors for papers in each SIGWEB conference has
A general overview and bibliometric analysis of seven ACM hypertext conferences, 2022.

In older conferences such as HT, the minimum number of average authors is 1.96 (recorded in 1987), but between 1994 and 2015, the average number of authors for an article in HT rose to between 2.4 and 2.7, and that increased slightly up to 3.1 in 2003 and from 2011–2015. In contrast, the CIKM conference has experienced an increment in the average number of authors of up to 0.2 every year. The CIKM conference also recorded the maximum number of average authors (3.67) in 2014. However, we observed a different trend in the newer SIGWEB conferences (JCDL, WebSci and WSDM). Figure 5 reveals that the JCDL, WebSci and WSDM conferences always had an average of at least three authors for a paper. The number of average authors in JCDL in year 2002 are 3.05 while in WSDM in year 2011, the number of average authors are 3.87.

RQ3 Who are the most prolific authors in SIGWEB conference?

We believe that to understand a research community, it is important to know the most prolific authors and researchers who are publishing papers in that community. We defined prolific authors based on their publication records within the community and extracted the names of all the authors who have published papers in ACM SIGWEB conferences. Then, we identified the most prolific authors based on their contributions to these conferences. Figure 6 shows a timeline-based illustration of the top authors in all SIGWEB conferences. The X-axis of the graph shows the year of their first publication in a SIGWEB conference. For example, when a researcher is among most prolific authors for the HT conference, the X-axis for that author shows the year of his or her first publication for the HT conference only, while the Y-axis shows the number of publications in each conference (identified in the legend). Authors’ names separated with commas indicate different authors with equal numbers of articles and the same first publication year for that conference.

Figure 6 reveals that Edward A. Fox has the maximum number of publications (56) among all the prolific authors in SIGWEB conferences. Since 1987, Edward has served as a professor in the Dept. of Computer Science at Virginia Tech, and directs the Digital Library Research Laboratory there. In 2001, he also served as a general chair for the JCDL conference. Cheng Xiang Zhai, currently serving as a professor in the Dept. of Computer Science at Virginia Tech, and directs the Digital Library Research Laboratory there.
Computer Science at the University of Illinois at Urbana-Champaign, has the maximum number of publications (42) for the CIKM conference and his first publication in that conference was published in 2001. Cheng has also served as a program co-chair and senior program chair for the SIGIR, CIKM and NAACL-HLT conferences. F.M. Shipman (23), Wendy Hall (22) and D.E. Millard (17) are the top three prolific authors in HT; they have been a part of this community for over a decade. Wendy Hall is a professor of Computer Science at the University of Southampton, UK and the director of the WEBSCI Institute. She is among the very first computer scientists to perform research in multimedia, hypermedia and HT. Philip S. Yu (36), W. Lee (24) and J. Han (24) are among the four most prolific authors in the CIKM conference; they have been a part of the community since 1994. Philip is a professor in Department of Computer Science at the University of Illinois at the Chicago and a former manager of the Software Tools and Techniques group at the IBM Thomas J. Watson Research Centre. He also serves as editor-in-chief of the ACM TKDD (Journal of Knowledge Discovery from Data). The JCDL conference began in 1996; however, our database contains the records of the conference only since 2001. We found that all the top ten prolific authors in JCDL published their first paper for the conference between 2001 and 2004. Edward A. Fox (56), M.L. Nelson (39) and D. Bainbridge (30) are the top three researchers in the JCDL community.

Because WebSci and WSDM are the newest conferences, the most prolific authors for these conferences have similar numbers of articles and first publication years. Also, the number of publications of prolific authors is comparatively smaller than those of authors who have been a part of other SIGWEB conferences over the past decade (up to a maximum of six for WebSci and 11 for WSDM). Figure 6 also reveals that some of these prolific authors have been publishing papers for multiple SIGWEB conferences. For example, F.M. Shipman has been actively participating in both the HT (since 1989) and WebSci (since 2013) conferences. However, because not all the WebSci proceedings were available in our database, the year of his first publication in the conference could not be determined correctly.

RQ4 What is the degree of academia-industry collaboration and participation?

ACM SIGWEB conferences are some of the most popular and largest conferences in the field of WEBSCI. The six SIGWEB conferences studied in this paper cover a wide variety of topics such as HT, DOCENG, social networks, information and knowledge management, digital libraries, hypermedia, WEBSCI, data mining, web search and user modelling. Authors from research labs, academic institutions and industries have published their research in these conference proceedings. Agarwal et al. (2016d) presented a study on variations in the participation rate of distinct affiliations in ACM SIGWEB community and revealed that; overall, the ACM SIGWEB conferences have a relatively higher rate of author participation from distinct affiliations.

To answer this research question, we extended the study presented in Agarwal et al. (2016d) by analysing the rates at which different types of affiliations (industry and academic institutions) participate in SIGWEB conferences. As discussed in Section 3.4, we extracted the affiliation types using several open source APIs (OpenStreetMap,
A general overview and bibliometric analysis of seven ACM hypertext

Alchemy, Google Map API and the Bing Geocoding API). Based on the statistics collected in Section 3.4, we plotted the distributions of academic universities and industry (AUI) collaboration against each conference. We defined the degree or score of AUI collaboration \( D_{AUI} \) as the number of distinct publications in a conference in which one or more of the co-authors in a paper has different type of affiliation than the others. To compute this score, we fetched the affiliation of each author in the paper and assigned a Boolean number to the article based on the affiliation types of the authors. If all the authors of a paper were from academia or all from industry, we assigned \( D_{AUI} = 0 \) to that paper. Alternatively, if any of the authors had a different type of affiliation than the other co-authors (either industry or academia), we assigned \( D_{AUI} = 1 \) to that paper. Figure 7 shows examples of two such papers published in JCDL 2015. Figure 7(a) shows that all the authors of the paper are affiliated with academic universities; therefore, we assigned the degree of AUI collaboration as zero. In contrast, in Figure 7(b), two of the authors are affiliated with industry but one is affiliated with an academic institution; therefore, we assigned the degree of AUI collaboration as one.

Figure 7 Examples of the degree of collaboration between academic universities and industries (\( D_{AUI} \)) in ACM SIGWEB conferences (a) \( D_{AUI} = 0 \) (b) \( D_{AUI} = 1 \) (see online version for colours)

We conducted a study on the affiliations of all 13,610 authors and analysed the degree of collaboration from various academic universities and industries. The bar chart in Figure 8(a) shows the collective extent of collaboration for each SIGWEB conference. The X-axis lists the conference names while the Y-axis shows the relative percentage of the degree of AUI collaboration. As Figure 8(a) shows, WSDM has the highest degree of AUI collaboration, i.e., 30.2%. Among 519 articles published in the conference proceedings and available in our database, 157 articles were co-authored by researchers from both academic institutions and industrial organisations. Figure 8(a) also reveals that, in terms of the number of publications, CIKM has the highest degree of AUI collaboration. However, the overall relative percentage of AUI collaboration in CIKM is only half (15.4%) that of the WSDM conference; among 4,264 articles, only 658 articles included authors from different affiliations types. Furthermore, although it is the oldest SIGWEB conference, the overall degree of AUI collaboration in the HT conference (9.6%) is relatively lower than that of the CIKM and WSDM conferences. As Figure 8(a) shows, among 1,027 articles, only 99 were co-authored by both academic and industry researchers.
We further performed a timeline-based analysis of the extent of AUI collaboration in all SIGWEB conferences. Figure 8(b) shows a line chart with the timeline along the X-axis while the Y-axis shows the degree of collaboration in terms of the number of publications whose $D_{AI} = 1$. As Figure 8(b) shows, for the majority of the time, the degree of AUI collaboration in SIGWEB conferences is below ten. Over the past 26 years, only the CIKM and WSDM conferences have had a substantial number of articles published by authors in collaboration from both academia and industry. Conferences such as HT and JCDL reach a level of ten such articles only once each, in 1991 and 2008, respectively. We observe that – similar to the number of publications – the acceptance rate and number of unique authors for the WSDM conference has a relatively higher rate of academic-industry collaboration and participation in the conference. The graph in Figure 8 shows that during the past 10 years, the extent of such collaboration in the CIKM conference has increased at reasonably high levels, reaching up to 102 in 2012 (approximately 20% of the total articles published in the conference that year). In contrast, at other SIGWEB conferences, the degree of AI collaboration in articles is below 25% of the total number of articles published in the conferences for that year.

RQ5 What is the scholarly output of various countries in SIGWEB conferences?

Bartneck and Hu (2009) presented a quantitative analysis of a publication database (created from historical proceedings) of the CHI conference and identified various countries that contributed to the conference. Their study revealed that less than 10% of the countries across the world have been participating and publishing articles in the CHI conference. Agarwal et al. (2016d) conducted a study on seven SIGWEB conferences and analysed the scholarly output of various countries in these conferences. In an extension to the work presented in Agarwal et al. (2016d), we analysed the scholarly output of a country based on the number of articles published from that country. In contrast Agarwal et al. (2016d) analysed the scholarly output of a country based on the number of co-authors affiliated with that country, an analysis that over-represents the relative...
percentage of scholarly output when there are large numbers of co-authors. For example, suppose that a given paper P1 is co-authored by three authors: A1, A2 and A3. Authors A1 and A2 are affiliated with the same university while A3 is affiliated with a different university in the same country, C1. Agarwal et al. (2016d) computes the scholarly output of country C1 as two whereas, in reality, the number of publications from C1 is one.

Figure 9 shows the choropleth map showing the collective distribution of the number of articles published from each country in SIGWEB conferences spanning 26 years. As Figure 9 shows, the USA has the maximum number of publications (3,907) – approximately 75% higher than the country with the 2nd highest scholarly output (990 articles) in SIGWEB conferences. Therefore, we set the maximum value of the colour bar to 1,600; that value shows the maximum colour intensity for USA. Figure 9 reveals that China (scholarly output of 990 articles), the UK (621), Germany (534), Canada (308) and France (300) are the other top-five countries contributing to SIGWEB conferences. Unlike the study presented in Bartneck and Hu (2009), we found that – in comparison with CHI-SIGWEB conferences have relatively high scholarly output from other advanced developed and developing countries such as Japan (238 articles), Australia (265), India (148), Korea (137) and Brazil (195). Our results revealed that many underdeveloped countries in the Americas and Africa have significantly lower participation in SIGWEB conferences. We further conducted a quantitative analysis on affiliation records of authors publishing in each conference. We computed the collective scholarly output of various countries in each conference over past 26 years. Table 3 shows the list of countries publishing the highest number of articles in each of the SIGWEB conferences. Due to space limitations, we present the results of only the top 10 and distinct countries with the highest scholarly output in their respective SIGWEB conferences. We also discuss the results only for HT conference which is the oldest SIGWEB conference. Table 3 shows the statistics of the top ten countries with maximum number of scholarly articles in each individual SIGWEB conference and all conferences combined.

Figure 9  Collective scholarly output of various countries in all ACM SIGWEB conferences (see online version for colours)
Table 3  List of countries publishing the highest number of articles for each SIGWEB conference

<table>
<thead>
<tr>
<th>Country</th>
<th>HT</th>
<th>WebSci</th>
<th>JCDL</th>
<th>WSDM</th>
<th>DocEng</th>
<th>CIKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>28</td>
<td>-</td>
<td>20</td>
<td>-</td>
<td>29</td>
<td>183</td>
</tr>
<tr>
<td>Austria</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brazil</td>
<td>-</td>
<td>5</td>
<td>36</td>
<td>11</td>
<td>73</td>
<td>-</td>
</tr>
<tr>
<td>Canada</td>
<td>39</td>
<td>-</td>
<td>32</td>
<td>17</td>
<td>31</td>
<td>188</td>
</tr>
<tr>
<td>China</td>
<td>-</td>
<td>7</td>
<td>58</td>
<td>90</td>
<td>-</td>
<td>799</td>
</tr>
<tr>
<td>Denmark</td>
<td>47</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>France</td>
<td>29</td>
<td>8</td>
<td>20</td>
<td>-</td>
<td>113</td>
<td>123</td>
</tr>
<tr>
<td>Germany</td>
<td>91</td>
<td>19</td>
<td>44</td>
<td>30</td>
<td>44</td>
<td>306</td>
</tr>
<tr>
<td>Ireland</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Israel</td>
<td>-</td>
<td>-</td>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Italy</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>21</td>
<td>139</td>
</tr>
<tr>
<td>Japan</td>
<td>-</td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td>141</td>
</tr>
<tr>
<td>Netherlands</td>
<td>44</td>
<td>16</td>
<td>-</td>
<td>14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>New Zealand</td>
<td>-</td>
<td>-</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Singapore</td>
<td>-</td>
<td>-</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>133</td>
</tr>
<tr>
<td>Spain</td>
<td>36</td>
<td>9</td>
<td>32</td>
<td>31</td>
<td>26</td>
<td>-</td>
</tr>
<tr>
<td>Switzerland</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UK</td>
<td>158</td>
<td>59</td>
<td>86</td>
<td>27</td>
<td>78</td>
<td>213</td>
</tr>
<tr>
<td>USA</td>
<td>407</td>
<td>65</td>
<td>963</td>
<td>340</td>
<td>162</td>
<td>1970</td>
</tr>
</tbody>
</table>

Figure 10 reveals that in the HT conference, the USA has the maximum number of publications (407), which is 62% higher than the country (UK) with the 2nd highest scholarly output (158). Therefore, similar to Figure 9, we set the maximum value for colour bar to 200 (the maximum colour intensity for USA) which does not normalise the values for countries with lower contributions. Following a similar pattern to the total number of publications and the AUI collaboration, the overall scholarly output of other countries for the HT conference is relatively lower than that of the USA and Europe. For example, while computing the collective scholarly output of countries for all SIGWEB conferences, China has the 2nd highest number of articles, but specifically for the HT conference, China holds the 18th position, with a total of 21 publications in the past 26 years: approximately 2% of its overall contribution to the SIGWEB community. We observe that other than the USA, only a few European countries such as Germany (91), Netherlands (44), UK (158) and Italy (54) made high contributions to the HT conference. Figures 9 and 10 reveal an interesting insight; Denmark holds the 23rd position in terms of collective publication rate for all SIGWEB conferences, but, for the HT conference, Denmark holds the 5th rank. It has 47 papers in the HT conference, which is 67% of its overall scholarly output in SIGWEB conferences (70 articles). Based on our empirical analysis, we find that the scholarly output of various countries in both old and new SIGWEB conferences is sparse and that authors from different regions of world have been publishing in these conferences. The USA, Europe and the UK have the highest participation rates in and contributions to the SIGWEB community.
A general overview and bibliometric analysis of seven ACM hypertext

Figure 10  Collective scholarly output of various countries in all ACM SIGWEB conferences (see online version for colours)

Figure 11  Examples of ACM SIGWEB articles with different degrees of cross-country collaboration ($DC^3$) varying from one to four (a) $DC^3 = 1$ (b) $DC^3 = 2$ (c) $DC^3 = 3$ (d) $DC^3 = 4$ (see online version for colours)

RQ6  What is the degree of cross-country collaboration across the world?

In accordance with the analysis conducted in previous RQs over the years, we found that researchers from various countries have been collaborating and co-publishing their research in SIGWEB conferences. In extensions of our previous RQs RQ4 and RQ5 on the degree of AUI collaboration and the overall scholarly output of countries in SIGWEB conferences, respectively, we further analysed the degree of cross-country collaboration ($DC^3$) within the SIGWEB community. We defined the degree of cross-country collaboration of an article as the number of distinct countries of the co-authors of a paper. Figure 11 shows examples of four different SIGWEB articles in which the number of distinct countries of co-authors vary from one to four. To compute the degree of cross-country collaboration in a conference, we calculated the total number of articles with different $DC^3$'s. We extracted the number of distinct countries of the co-authors of a paper using the affiliation metadata collected as described in Section 3.4 and used for RQ5.
Figure 12  Distribution of the degree of cross-country collaboration in various SIGWEB conferences (see online version for colours)

![Graphs showing distribution of degree of cross-country collaboration in various SIGWEB conferences](image)

Figure 12 is a graph of the timeline for each conference showing the variation in the degree of cross-country collaboration. As per the statistics of the affiliation metadata (see Section 3.4), an article is co-authored by researchers from at most six countries. Figure 12 reveals that in all SIGWEB conferences, the maximum number of articles have a $DC^3$ of one, indicating that all the co-authors are affiliated with the same country. Figure 12 also reveals that number of articles published by authors in collaboration between two countries is significantly higher than the degree of collaboration among articles with authors from more than two countries. The graphs in Figure 12 show that for all available proceedings of SIGWEB conferences, the number of articles with a $DC^3$ of two fluctuate very frequently. Furthermore, these articles are relatively rare in comparison to the total number of publications in that year. For example, in DocEng 2012, the number of articles with a $DC^3$ of two is only 15% of the total articles published in DocEng that year (six out of 41 articles). Despite having the greatest number of articles with a $DC^3$ of two (557 articles over the years), the overall cross-country collaboration in CIKM is relatively low in comparison to the total number of publications that year. However, the pattern of increases and decreases in the number of articles with a $DC^3$ of two is similar to the number of articles with no cross-country collaboration (i.e., $DC^3 = 1$). Figure 12 also reveals that even in older conferences such as HT, the number of articles with a $DC^3$ of three is significantly lower than in new conferences such as WSDM. In the HT conference, the distribution of such papers varies between one and four while in WSDM it can reach a maximum of eight papers (in 2014). The maximum number of articles with a $DC^3$ of four is even smaller for ACM SIGWEB conferences. The statistics based on the extracted metadata reveal that there have been only three papers co-authored from five distinct countries, published in 2008 (CIKM) and 2013 (CIKM and DocEng). Further, except for one article in WebSci 2011 with a $DC^3$ of six,
the number of articles having more than five country affiliations is zero for all SIGWEB conferences.

**Figure 13** Collective degree of cross-country collaboration in SIGWEB conferences (see online version for colours)

Figure 12 reveals that there is a lack of a higher degree of cross-country collaboration in SIGWEB conferences. Therefore, to analyse the rate of collaboration among various countries, we created a network of all the countries that have a publication in a SIGWEB conference and a record in our experimental dataset. Figure 13 shows a collective network graph of all the SIGWEB conferences in which the nodes represent the countries and the edges are the number of distinct articles co-authored by two countries. The size of each node shows the total number of articles published by a country in collaboration with other countries (node degree). Figure 13 reveals that USA and China have the maximum number of articles (218) published in collaboration – a value that is approximately 53% higher than the second-highest rate of collaboration between the USA and the UK (104 articles). Figure 13 shows that the USA has collaborated with the majority of countries across the world (a node degree of 957). For example, the USA has published 55, 49, 44, 40, 54 and 34 articles in collaboration with Brazil, Germany, Italy, India, Canada and Australia, respectively. Countries such as Denmark, Netherlands, Spain, Austria, Greece, Singapore, Qatar and Japan have substantially higher numbers of SIGWEB articles published in collaboration with other countries. The degree (the number of articles published in collaboration) of these countries in the network varies from 30 to 60. In contrast, various underdeveloped regions in South Africa, North America and South Asia have both low participation and low cross-country collaboration in SIGWEB conferences. For example, countries such as Viet Nam, Senegal, United Arab Emirates, Pakistan and Romania have a maximum degree of two papers published in collaboration.
with other countries. The network graph shown in Figure 13 also supports the statistics presented in RQ5, namely, that although SIGWEB conferences attract participants from all over the world, the rates of participation and collaboration by countries in South America, Asia, South Africa and a few in Europe are relatively low.

RQ7 What is the participation rate of female authors in SIGWEB conferences?

Agarwal et al. (2016b) presented a study on the gender gap, employment imbalance and the poor rate of female participation in CSR measured in terms of authorship in scholarly publications. They conducted their analysis on publication records from 81 conferences across various domains of CSR (i.e., computer science, data engineering, software engineering and theory). They performed a collective analysis on the gender of authors publishing in various conferences of these domains. In contrast to their work, to answer this research question, we studied the gender imbalance in each individual SIGWEB conference and in the SIGWEB conferences as a whole. Furthermore, our results do not replicate the results from their study because other than CIKM, none of the SIGWEB conferences are among those 81 CSR conferences. We analysed the gender gap in the SIGWEB research community across the following facets:

1. Number of distinct female authors publishing articles in SIGWEB conferences.
2. Number of female authors among most prolific author and dignitary profiles in the SIGWEB community.
3. Number of female authors holding a leadership position in SIGWEB conferences.

4.1 Distinct female authors

Based on the statistics of the authors’ gender information extracted as described in Section 3.3, we conducted an analysis on the authors and studied the gender imbalance in SIGWEB conferences. Figure 14 shows the distribution of distinct female authors who have published articles in SIGWEB conferences since they began until September 15, 2015. Figure 14(a) shows a boxplot diagram of the collective distribution of female authors in each conference. We computed the number of articles published in each SIGWEB conference as well as those co-authored by at least one female author. We created an array of the number of such articles for every year that the conference proceedings were available in our database and arranged them into three sub-plots to avoid the normalisation of lower values. Figure 14(a) reveals that the majority of the SIGWEB conferences have a maximum of 60 distinct female authors who have published a paper in these conference proceedings since 1987 and that the minimum number of female authors in a SIGWEB community is ten. Following a pattern similar to the total number of authors in a conference, the rate of female participation in CIKM (maximum of ~ 350) and JCDL (maximum of ~ 150) is significantly higher than female participation in other SIGWEB conferences. Further, the median rate of female participation in CIKM is the 2nd highest (46) when compared to other SIGWEB conferences, and JCDL has the highest rate of female participation, with a median of 91 articles.
Figure 14  Distribution of the rate of female authors publishing in SIGWEB conferences (a) collective distribution of female authors in each SIGWEB conference (b) variation in female author participation rate over the years (see online version for colours)

Figure 14(b) shows a timeline-based review and the trend of female participation in SIGWEB conferences over the years. Figure 14(b) reveals that, despite being the oldest and most popular SIGWEB conference, the rate of female participation in HT is always below 30, which is only 15% of the total authors participating in the HT conference. The rate of female research participation in HT increases only twice, in 2007 and 2009, reaching maximum values of 34 and 50, respectively. Figure 14(b) shows that in comparison to other SIGWEB conferences, JCDL has had a relatively higher rate of female author participation since the beginning of the conference. As shown in Figure 14(a), JCDL has only two outliers, with minimum values of 44 and 51 recorded in 2015 and 2010 respectively. Figures 14(a) and 14(b) also reveal that not only is the collective female participation in the DocEng conference minimal but also that it has not improved over the years. DocEng has an even lower rate of female authors than the newer SIGWEB conferences. While comparing the results of RQ2 and RQ7, our empirical analysis revealed that there is a huge gap in the number of female authors participating in intra-SIGWEB conferences. We did not find a major gender imbalance in ACM SIGWEB conferences. In comparison to other CSR conferences (presented in Agarwal et al. 2016b), female authors participate in ACM SIGWEB conferences at rates that are up to 50% higher.
4.2 Most prolific female authors

To analyse the specific contributions of female authors to the SIGWEB community, we further analysed the profiles of the most prolific authors and the dignitaries for each SIGWEB conference based on author/dignitary gender. Table 4 shows the scatter map of female authors who are among top ten and most prolific authors in SIGWEB conferences (analysed in RQ3). The annotations for these authors in Table 4 shows that among the 60 most prolific authors (identified in RQ3), only six female authors.

Table 4 Most prolific female authors in the SIGWEB community

<table>
<thead>
<tr>
<th>Author</th>
<th>#Articles</th>
<th>First article</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>W. Hall</td>
<td>22</td>
<td>1993</td>
<td>HT</td>
</tr>
<tr>
<td>C.C. Marshall</td>
<td>16</td>
<td>1987</td>
<td>HT</td>
</tr>
<tr>
<td>M.G. Pimentel</td>
<td>16</td>
<td>2001</td>
<td>DocEng</td>
</tr>
<tr>
<td>H.Y. Balinsky</td>
<td>13</td>
<td>2005</td>
<td>DocEng</td>
</tr>
<tr>
<td>Lisa Harris</td>
<td>3</td>
<td>2013</td>
<td>WebSci</td>
</tr>
<tr>
<td>S.T. Dumais</td>
<td>11</td>
<td>2009</td>
<td>WSDM</td>
</tr>
</tbody>
</table>

Table 5 Female authors holding leading positions (general chair) in SIGWEB community

<table>
<thead>
<tr>
<th>Year</th>
<th>CIKM</th>
<th>DocEng</th>
<th>HT</th>
<th>JCDL</th>
<th>WSDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>N. Pissinou</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1997</td>
<td>N. Pissinou and K. Makki</td>
<td>-</td>
<td>W. Hall</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1998</td>
<td>N. Pissinou</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2000</td>
<td>S. Gauch and E. Rundensteiner</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2001</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>C. Borgman</td>
<td>-</td>
</tr>
<tr>
<td>2003</td>
<td>-</td>
<td>-</td>
<td>H. Ashman</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2005</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>M. Marlino</td>
<td>-</td>
</tr>
<tr>
<td>2007</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>E. Rasmussen</td>
<td>-</td>
</tr>
<tr>
<td>2008</td>
<td>N. Frayling and G. Kazai</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2010</td>
<td>G. Kazai</td>
<td>-</td>
<td>-</td>
<td>J. Hunter</td>
<td>-</td>
</tr>
<tr>
<td>2011</td>
<td>G. Kazai</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2013</td>
<td>-</td>
<td>S. Marinai</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2015</td>
<td>-</td>
<td>C. Vanoirbeek</td>
<td>-</td>
<td>S. Allard L. Hang</td>
<td>-</td>
</tr>
</tbody>
</table>

Cathy Marshall, one of the most productive authors in the HT conference, is currently a Principle Scientist at Microsoft Research’s Silicon Valley Lab and also serves as an adjunct Professor in the Department of Computer Science and Engineering at the Centre for the Study of Digital Libraries, Texas A&M University. Because data for only four years of the proceedings of WebSci are available in the ACM Digital Library, the number of articles published by authors in the conference is already lower than those from authors in other conferences. Table 4 also reveals that in one of the oldest and most popular SIGWEB conferences (CIKM) no female authors appear among the top ten profiles of
the conferences. Despite being a new conference, only one female author is listed among top ten profiles of the WebSci conference, Lisa Harris, who is currently serving as an Associate Professor in Marketing, Deputy Director of the WEBSCI Institute and Program Leader of the Digital Marketing at the Southampton Business School at the University of Southampton. Our analysis reveals that despite the relatively high rate of female participation in SIGWEB conferences, the individual participation rates of these authors is relatively lower than those of male authors.

4.3 Female authors holding a leadership position in the community

Agarwal et al. (2016b) presented a gender-based analysis of various authors holding a leadership position in CSR conferences. Their study revealed that there is a huge gender imbalance in terms of the number of female authors holding dignitary positions in CSR conferences. We performed a gender-based analysis of the proceedings authors (editors and general chairs) for the SIGWEB conferences and identified the distribution of male and female authors who have made substantial contributions to the conference and community. Table 5 shows the scatter map of all the female authors holding various leadership positions in SIGWEB conferences such as editors and general chairs. The X-axis shows the timeline of the SIGWEB conferences. Table 5 reveals that except for CIKM, none of the SIGWEB conferences have had a female researcher serve as general chair since their beginnings. In contrast, during the first few years of the CIKM conference, five female researchers served in dignitary positions. Yelena Yesha is a tenured Professor at the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County and also directs the National Science Foundation (NSF), a government organisation for funding Hybrid Multicore Productivity Research in the USA. Niki Pissinou is a founder of the ACM GIS professional forum and currently serves as a Professor at the School of Computing and Information Sciences, Florida International University, Miami. Table 5 also reveals that during the first few years of the CIKM conference, female general chairs were also a part of the editor team for the proceedings. However, over the past 13 years only two female authors (G. Kazai three times and N. Frayling only once) have held editor positions and no female researchers have been the general chair of the conference. Gabriella Kazai is a former research consultant at Microsoft Bing and at Microsoft Research and currently serves as a Head of Data Science at Lumi, Semion Ltd., a start-up company begun by the founders of last.fm. She has also served as a Senior Program Committee member for various CSR conferences including SIGIR, CIKM, CHI, HComp, ECIR and WI. Despite being the oldest and most popular WEBSCI conference, HT has only recently had two female researchers serve as Editors (S. Marina i and C. Vanorbeek). Similarly, spanning 20 years, JCDL has had female researchers serve as editors only five times. Our analysis of the six SIGWEB conferences supports the study presented in Agarwal et al. (2016b), which showed that – similar to numerous other CSR conferences – ACM SIGWEB has a huge gender imbalance within the community.

RQ8 What is the degree of author’s participation from host country?

As the most popular and reputable conferences in the domain of WEBSCI and data analytics, we believe that the ACM SIGWEB conferences represent an opportunity for participating researchers. Furthermore, over the past three decades, these conferences
have been organised all over the world and authors from various countries have participated and availed themselves of these opportunities. Therefore, as an extension to RQ5, in this research question, we analysed the degree of local community participation of researchers in SIGWEB conferences when the conference is hosted in their own countries. We extracted the location of each SIGWEB conference for every year of the proceedings available in the ACM digital library and computed the distinct number of articles published by authors from the host country. Due to space limitations, we do not present results for each conference; instead, we selected a one older and one newer conference and compared the host country participation in each of those conferences.

Figure 15 shows a choropleth map of host country participation in the HT and WSDM conferences. The colour intensity in the choropleth map shows the maximum rate of participation from a country (if the conference was hosted in that country more than once). Figure 15(a) reveals that during past 26 years, the HT conference has been hosted mostly in the USA and in several European countries. HT has been hosted 8 times in the USA, while our empirical analysis revealed that the highest number of articles (32) in any year from the USA occurred in 1991. We observe that despite a decrease in the total number of publications, the rate of local community participation from USA is still higher in comparison to that of other countries. For example, in 2012, the number of articles published from the USA was 15, which is 50% of the total number of articles published in the conference that year. Unlike the USA, South America and Europe have significantly lower participation rates in the HT conference when it is hosted in those regions. For example, in 2014, articles by researchers from Chile constituted only 2% of the articles in the conference.

Figure 15 Local community participation by host countries in ACM SIGWEB conferences
(a) HyperText (b) WSDM (see online version for colours)

Similar to HT, the host country participation of the US in the WSDM is relatively higher than the host country participation rate of other countries. WSDM has been hosted in USA 3 times, and our database revealed that each of those years, 78% of the articles published in the conference proceedings were from the US. In contrast, in 2013, the number of articles from Italy constituted only 10%. Similar to the US, China also had a relatively large number of publications in WSDM in 2015 (i.e., 36% of the total articles
A general overview and bibliometric analysis of seven ACM hypertech

published in the conference that year). In contrast, during the first few years of the conference, DocEng has had lower participation from host countries (a minimum of 5%) but the values have gradually grown over the years. Our analysis revealed that in 2014 (when DocEng was hosted in the US), 41% of articles were published by authors from the local community.

Figure 16  Word-cloud presentation of the most popular topics in SIGWEB conferences
(a) CIKM (b) HyperText (c) DocEng (d) JCDL (e) WebSci (f) WSDM

![Word-cloud presentation of the most popular topics in SIGWEB conferences](image-url)
Figure 17  Distribution of ACM concepts for more than 50 articles published in seven ACM SIGWEB conferences (see online version for colours)

We further performed a relative comparison of collective participation from a country when a conference was hosted or not hosted in that country. Table 6 shows the statistics of the collective participation of authors in SIGWEB conferences from various countries. A value of $Y$ denotes the relative percentage of a country’s participation when the conference is hosted in that country, while $N$ denotes the relative percentage of a country’s participation when the conference is not hosted in that country. Blank values signify that the conference has never been hosted by a country. Given a country, $C$, and a timeline, $T$, of the years when the conference is hosted in $C$, the community participation in the country is calculated as follows:

$$Y = \frac{NumP_T}{NumP} \quad \text{and} \quad N = \frac{NumP_{T_C}}{NumP - NumP_T}$$

where $NumP$ is the total number of publications in a conference, $NumP_T$ is the total number of publications in a conference for a given timeline $T$, and $NumP_{T_C}$ is the total number of publications in a conference for a given timeline and country $C$. Table 6 reveals that all the countries have significantly higher participation in a conference when it hosted in that same country compared to those years in which the conference was held in another country.

RQ9 What are the most common topics of research that are frequently published in SIGWEB conferences?

As discussed in the previous sections, ACM SIGWEB covers a wide range of topics in the domain of WEBSCI. Some of these topics remain the same in all the conferences sponsored by ACM SIGWEB community while other topics vary according to the research theme of the conference. Understanding the core research theme, focus and
A general overview and bibliometric analysis of seven ACM hypertext

Scope of a conference is an important parameter for authors who would like to publish articles in conference proceedings. These themes and topics are labelled based on multiple attributes such as the application of the research, the methodologies and techniques proposed in the paper, evaluation metrics and experimental datasets. Our empirical analysis revealed that the most common and popular topics in the SIGWEB community are specific to each conference: the conferences share only core and broadly defined topics. To analyse the most common topics of each SIGWEB conference, we extracted both the author-generated keywords and the ACM concepts available in the ACM Digital Library. We believe that these author tags and the concepts annotated by IBM Watson are much more detailed and more closely related to the topic of the paper than are the terms in the ACM computing classification system (CCS). To avoid the noise in tags (repetitive keywords with different names or abbreviations), we used the ACM concepts to study the topics and themes of SIGWEB conferences. Figure 16 shows a word-cloud presentation of the most popular topics in each SIGWEB conference based on the number of publications annotated with these concepts. For all SIGWEB conferences, we were able to extract a total of 5,743 unique ACM concepts annotated by IBM Watson. As the word cloud shows, each of these conferences covered a wide variety of topics. For older conferences such as CIKM and HT, we were able to extract a total of 3,624 and 1,382 unique topics, respectively, while for newer conferences such as WebSci and WSDM, we extracted 596 and 938 unique concepts, respectively. Figure 16(a) shows that the majority of articles published in the CIKM conference proceedings are based on core themes and cover WEBSCI topics on a broader level, for example, information retrieval was annotated in 519 articles, while web search engine was annotated in 286 articles. In contrast, the publications associated with tags such as social media (66), semantic web (44), or other domain-specific keywords such as tf-idf (23) appear relatively less often in the CIKM conference. Similar to CIKM, the concepts found in the articles of the HT and JCDL conferences are specific to the core themes and major topics of the research performed within those communities. For example, hypermedia, and World Wide Web appear often in HT, while digital library appears often in the JCDL conference [refer to Figures 16(b) and 16(d)]. The HT conference also includes a large number of publications from the application domains of hypermedia and www. Therefore, we can find several other common topics in the word cloud that are not a part of the conference’s core themes. For example, social media websites are a major platform for conducting hypermedia and hyperlink studies. Furthermore, with the emergence of Twitter in recent years, we can find social networks and Twitter (32) as a part of the most common and popular topics in the HT conference. Social media websites, blogs and forums contain tags associated with each post that make the content easily searchable. Tags and metadata of a post on social media are important features for information extraction and hyperlinking. Therefore, we find tags (metadata) among one of the most common topics in HT; they are annotated in a total number of 50 papers. Unlike CIKM or HT, WebSci is more of an application-oriented conference. Therefore, we find that the most popular and common topics in the conferences are application specific and not related to the core theme of WEBSCI conferences. For example, social media (33), social network (32), Twitter (29), Facebook (15), user generated content (8), data analysis (7) and microblogging (6) appear relatively often. Given that we had data for only four years of the WebSci proceedings in our database, we believe that these numbers are relatively higher and constitute the most popular and common topics in the conference.
Table 6
Percentage of conference participation from various countries when hosting and not hosting a SIGWEB conference

<table>
<thead>
<tr>
<th>Country</th>
<th>CIKM</th>
<th>HT</th>
<th>JCDL</th>
<th>DocEng</th>
<th>WebSci</th>
<th>WSDM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y N</td>
<td>Y N</td>
<td>Y N</td>
<td>Y N</td>
<td>Y N</td>
<td>Y N</td>
</tr>
<tr>
<td>UK</td>
<td>7.50 4.68</td>
<td>30.23 13.61</td>
<td>- -</td>
<td>17.48 11.81</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>China</td>
<td>25.59 14.83</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>21.88 13.05</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>2.51 1.66</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>1.10 1.24</td>
</tr>
<tr>
<td>Germany</td>
<td>7.14 7.18</td>
<td>13.51 8.88</td>
<td>- -</td>
<td>13.95 6.69</td>
<td>21.43 7.74</td>
<td>- -</td>
</tr>
<tr>
<td>US</td>
<td>52.51 38.57</td>
<td>52.33 27.32</td>
<td>71.93 58.95</td>
<td>34.84 23.68</td>
<td>41.44 22.35</td>
<td>68.25 62.41</td>
</tr>
<tr>
<td>Canada</td>
<td>6.93 4.17</td>
<td>13.24 3.31</td>
<td>4.11 1.52</td>
<td>12.20 4.56</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>Portugal</td>
<td>1.79 0.46</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>Austria</td>
<td>- -</td>
<td>5.26 2.45</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>Italy</td>
<td>- -</td>
<td>24.32 3.92</td>
<td>- -</td>
<td>14.55 2.34</td>
<td>- -</td>
<td>8.42 3.52</td>
</tr>
<tr>
<td>Denmark</td>
<td>- -</td>
<td>9.33 4.52</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>Netherlands</td>
<td>- -</td>
<td>10.53 4.45</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>France</td>
<td>- -</td>
<td>4.76 3.09</td>
<td>- -</td>
<td>30.26 16.82</td>
<td>8.77 2.16</td>
<td>- -</td>
</tr>
<tr>
<td>Australia</td>
<td>- -</td>
<td>- -</td>
<td>6.06 1.30</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>Brazil</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>24.49 10.85</td>
<td>- -</td>
</tr>
<tr>
<td>Switzerland</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>4.44 3.18</td>
<td>- -</td>
</tr>
<tr>
<td>Spain</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>- -</td>
<td>6.25 6.07</td>
</tr>
</tbody>
</table>
Based on the statistics collected in Section 3.2, we find that the collective distribution of SIGWEB articles for each tag varies between one and 681. Our database reveals that among 5,743 unique concepts, a total of 5,100 concepts appear as labels for fewer than ten articles, and 2,875 tags are associated with only one article. Most of these outlier tags are for non-technical research-specific articles, and are not related to the usual topics seen in the conference. For example, several such tags are ‘spectral theory of ordinary differential equations’, ‘personal knowledge management’, and ‘computerised physician order entry’. Figure 17 shows the overall distribution of various ACM concepts that appear as labels for 50 to <100 articles (left) and more than 100 articles (right). Figure 17 reveals that the majority of the concepts with fewer than 100 articles are either application-specific concepts or proposed metrics and solution-approach-specific key terms. Some examples include ‘query optimisation’ (64), ‘business process’ (57), ‘blog’ (69), ‘support vector machine’ (73), ‘text mining’ (76) and ‘user-generated content’ (51). In contrast, concepts that are generalised and are common to all SIGWEB conferences are associated with large numbers of articles. Some examples include ‘information retrieval’ (681), ‘world wide web’ (565), ‘digital library’ (475), ‘web search engine’ (449), ‘social network’ (309), ‘database’ (295), ‘semantics’ (287), ‘metadata’ (248), ‘cluster analysis’ (241), ‘machine learning’ (154), ‘data analysis’ (133), ‘data mining’ (130) and ‘dataset’ (106).

**Figure 18** Most frequent pairs of concepts and topics in SIGWEB conferences (see online version for colours)
To analyse the correlation between most popular topics or concepts we calculated correlations between common pairs of these topics that occur together most frequently. Figure 18 shows a network graph of the most popular concepts that co-occur most frequently in SIGWEB articles. The colour intensity and the size of each node shows the relative degree (the number of articles annotated with that topic or concept). Figure 18 reveals that in SIGWEB articles, information retrieval is the most popular concept that co-occurs with other popular and commonly used concepts. It also shows that in the web domain, pairs such as ‘Information retrieval, Web search engine’, ‘Web search engine, WWW’, ‘Information retrieval, WWW’, and ‘Web page, WWW’ co-occur more frequently than pairs such as ‘Ranking, Ontology’. Similarly, in the domain of social networking, pairs such as ‘Twitter, Microblogging’ and ‘Twitter, Social Media’ co-occur more frequently than pairs such as ‘Microblogging, Database’. The network graph reveals that the frequency of one concept in significant pair impacts the frequency of another concept. We believe that the sparsity in these concepts and topics is also made possible as conferences grow and new emerging advanced topics are introduced in that domain. Therefore, we conducted a timeline-based review of several numbers of new topics introduced in each SIGWEB conference. We studied the evolution of these topics in the following research question.

RQ10 What is the trend of the evolution of topics in SIGWEB conferences?

Research shows that certain themes and topics in a conference can later evolve into more advanced and important research topics as new topics emerge (Coulter et al., 1998). Some of these topics arise over years, including the topics related to various inter-disciplinary domains. Coulter et al. (1998) presented a study based on 13 years of publication records from Software Engineering conferences (1982–1994). They performed a co-word analysis on the index terms of these publications extracted from ACM-CCS. Their study revealed that the software engineering research field is growing rapidly; several new topics have been introduced over the years such as ‘object oriented themes’, ‘user interfaces’, and ‘software reuse’. They also identified the core themes of these conferences and the topics that remained constant over a span of 13 years such as ‘requirement and specification’, ‘quality assurance’, ‘software development’ and ‘verification and validation’. In contrast to the existing study, and as an extension to the previous research question (RQ9), we analysed the evolution of various topics and research themes from ACM SIGWEB conferences over the years.

Figure 19 shows a line graph of the distribution of topics in each SIGWEB conference over the past 2.5 decades. As Figure 19 shows, compared to all the other SIGWEB conferences, CIKM has the maximum variation in the number of unique topics associated with the articles published in the conference in every year. Despite being the oldest and most popular SIGWEB conference, HT has had relatively fewer new topics introduced over the years; the majority of publications in the HT conference are consistent with the common and core topics of the conference. For example, hypermedia and hyperlinks have been the core research topics in HT since 1989. During the first few years of the conference, HT included articles labelled with approximately 100 unique topics. Unlike other SIGWEB conferences, the number of topics in HT decreased continually for 13 years. However, as social media, blogs and forums emerged; the number research topics for HT also increased – at a rate of 75% (refer to Figure 19). For example, blog, social network and multimedia topics were introduced in HT in 2002. Furthermore, with the emergence of microblogging websites, topics such as ‘Twitter’,
‘user generated content’, and ‘microblogging’ were introduced in 2010–2011. Figure 19 reveals that DocEng has had a relatively constant rate for the number of distinct topics in conference each year. We found that, despite including new topics in the conference every year, the number of unique topics per year remains approximately the same. For example, in 2009, published articles were labelled with 143 topics that were different and unique from 2008, but the total number of unique topics in 2008 and 2009 was the same (i.e., 179). Figure 19 reveals that there has been a rapid change in the number of unique topics and concepts in the WSDM conference over the years. The distribution of unique topics in WSDM varies from 98 to 302 per year. These topics cover a wide range of concepts: web mining, web crawlers, user-generated data, social media, information retrieval and ranking are some examples of the concepts that have been core and major topics of research in WSDM since 2008, while some topics that were either research-specific or that have risen over time occur significantly less often for WSDM articles. For example, concepts such as Wordnet and Yago are associated with only very few research papers in WSDM. Similarly, semantic web, ontology, Facebook, big data and Twitter are concepts that have been introduced in the conference since the emergence of Web 3.0 and microblogging portals.

Figure 19 Evolution of the number of topics in SIGWEB conferences over the years (see online version for colours)

RQ11 What are the major funding agencies that sponsor SIGWEB papers?
A majority of university research all over the world is funded by extramural grants from governments and industry. Academic research funded by government and industry leads to scholarly publications in conference proceedings and journals. Several publishers, including ACM, provide a way (normally at the point of submitting the camera-ready copy and copyright form) for authors to report the name of the funding agency supporting the research. The name of the funding source is then made publicly available online as part of the paper’s metadata. We believe that analysing the funding data for ACM SIGWEB conference papers will help the SIGWEB research community understand trends of funding expenditures, the relationship between funding and research outcomes (such as the number and quality of articles published as an outcome of a given research
funding) and answer questions such as which funding agencies are funding what types of research and where those research outcomes are being published.

**Figure 20** A word cloud of all funding agencies supporting various studies published in ACM SIGWEB conferences (see online version for colours)

Our dataset revealed that among 7,996 articles, only 17% (1,386) were funded by third-party organisations. There are a total of 104 unique funding agencies in our database. We computed the number of papers supported by all these organisations. Figure 20 displays a word-cloud of all the funding resources extracted from SIGWEB articles published in the ACM Digital Library. The sizes of the text in the word-cloud indicate the relative ratio of the number of articles funded by each agency. The minimum number of articles funded by an agency is one but reaches a maximum of 33. In total, there are 72 unique organisations that have funded at most one article published in SIGWEB conferences; however, on average, each agency has funded 2.28 articles. Figure 20 shows the top five funding agencies in terms of the number of studies supported, which are: Seventh Framework Program (33), NSF (17), National Research Foundation of Korea (11), Science Foundation of Ireland (9), and the Division of Information and Intelligent Systems (7). These top five funding sources are outliers; the remaining 99 funding agencies have sponsored at most 6 articles. The ‘Seventh Framework Program’, also known as ‘FP7’, is a major research funding organisation founded in 2007 in Europe. It continued until 2013, funding seven years of programs (research collaborations with government and societal projects). The NSF is a government funding organisation in the USA founded in 1950. The NSF has funded a variety of projects in computer science, physics, mathematics, and social science research. As of June 2016, NSF had funded approximately 1,000 research studies in the computer science domain. Figure 20 reveals that the organisations providing support for WEBSCI and data analytics research (published in SIGWEB conferences) are mostly agencies based in the USA, China, Singapore and Europe.
5 Threats to validity

In this paper, we conducted our study on only those SIGWEB proceedings that are published in the ACM Digital Library. Because the majority of SIGWEB proceedings are published in the ACM Digital Library, we discarded the proceedings published under IEEE, Springer, IEEE Computer Society and other publishers. We excluded these entries from our database due to inconsistencies in the ACM metadata available from each publisher. For example, ACM provides the names of all co-authors for a paper along with a hyperlink to each profile. In contrast, IEEE creates a hyperlink for only the lead author of the paper, which makes it difficult to obtain the metadata for all the authors. As discussed in Section 3, our proposed framework for attribute collection has dependencies on various third party and open source APIs. As per May 8, 2016, the Genderize.io API contains 2,162,866 distinct names associated with 79 countries and 89 languages. For our database, the average probability of obtaining a correct gender classification from this API is 94%. However, we observed that the API does not predict the gender of Chinese authors’ names with such a high accuracy, which generates false alarms. Further, creating the ground truth manually for all the authors and computing the accuracy of gender classification is highly impractical. Our dataset contains 13,610 unique authors. To annotate the gender requires searching for the name of each author on the web and checking their available profiles for verification. When no picture is available, the gender must be gleaned from the authors’ bios.

Among the other third-party APIs used in this study, the OpenStreetMap API (Nominatim) yields low-accuracy results for noisy text (spelling errors and incomplete addresses) and it cannot predict the type of a given affiliation. While computing the number of ’cited by’ features of SIGWEB articles within the SIGWEB community, we were able to extract only the valid ACM ids of papers because ACM does not list articles published by non-ACM publishers. We also found that some authors have multiple accounts on ACM. For example, Miyi Chung, a colleague of Kevin B. Shaw (ACM id: 81100396643) was mentioned three times in the list of his colleagues. Each of Chung’s profiles (81545460456, 81100292812 and 81545460656) has different year information and different publications. In this paper, we performed a correlation analysis of the number of publications and submissions in a conference. The acceptance rate of conferences varies for different categories of papers such as posters, short papers and full papers. However, the ACM digital library publishes only the overall acceptance rate of the conference (considering all types of submissions together). Conferences that have a large number of posters and short papers increase the relative numbers of publications and author participation in those conferences. Furthermore, it is not possible to filter for poster papers or perform analyses on only short or full papers because the paper type is not recorded in all the proceedings in the DBLP or the ACM digital library.
6 Conclusions and future works

ACM SIGWEB is a community of researchers and organisations supporting seven conferences in the web science area. ACM SIGWEB is a rapidly growing area that focuses primarily on a variety of topics and research themes around HT and hypermedia, information retrieval, social networks, knowledge management, DOCENG and digital libraries. In this paper, we conducted a bibliometric and exploratory analysis of the SIGWEB conferences and their communities. To conduct our study, we extracted the publication records for SIGWEB conferences from the DBLP bibliographical database and enriched the resulting records by extracting ACM metadata for each article, author, affiliation and conference from the ACM digital library. We formulated 11 RQs that cover various facets and dimensions and analysed several aspects of the SIGWEB community and the SIGWEB conferences. Our results revealed that the overall acceptance rate of each conference has been decreasing over the years and varies between 20% to 40%. Each year that the number of submissions decreases for a conference, we see an increasing acceptance trend. Our findings reveal that the average number of co-authors per article in both older and newer SIGWEB conferences varies up to maximum values of 2.7 and 3.87, respectively. Over the past decade there has been an upward trend in the degree of author collaborations. The scholarly output of each author revealed that the researchers publishing the most in SIGWEB conference proceedings are those authors who have been involved in the conference since the beginning. Researchers publishing from various industries and academic institutions have also been active participants in the conference. We conclude that the degree of collective collaboration among various industries and universities is relatively lower in the SIGWEB conferences than in other CSR conferences. CIKM and WSDM are the two conferences that have recently had a large number of publications co-authored by researchers from different types of organisations. The SIGWEB group is a widespread community that stretches across the globe and includes publications from various countries of the world. However, we conclude that the USA, UK, China and a few European countries contribute the most to the community.

Our empirical analysis also revealed that despite an upward trend in the degree of co-authorship, the overall collaboration of authors from distinct countries is relatively low; only those countries with the highest scholarly output in SIGWEB conferences have been collaborating with each other. We performed a gender analysis on the SIGWEB-published authors and conclude that unlike other CSR conferences the participation rate of female authors is significantly high in SIGWEB conferences (~ 50% of the total number of authors participating in the conference); however, there is a huge imbalance among researchers in leadership positions (general chairs and editors). Our analysis revealed that female authors who have served as editors or as general chairs in the conference have been part of the conference since the beginning. Similar to the trends for scholarly output of a country, only a few countries such as the USA, some European countries and China have higher participation rates in SIGWEB conferences from local communities when they are the host country of a conference. Our empirical analysis on articles’ metadata revealed that certain topics and themes of the conference remain the same over the years and some evolve into major and advanced topics (core themes) as new topics emerge in the community. We conclude that SIGWEB conferences have been growing rapidly as new topics and research themes have arisen. With the emergence of new studies in the domain of WEBSCI, many third party organisations (both government
A general overview and bibliometric analysis of seven ACM hypertext

and industry) have provided funding resources for the published studies. SIGWEB conferences metadata reveals that most of the organisations supporting WEBSCI and data analytics studies are based in the USA, UK, China and Europe. In this paper, we conclude that SIGWEB is a rapidly emerging and expanding community from many aspects, including articles, conferences, and number of authors. However, researchers affiliated with some of the more advanced developed and developing countries participate in the conference at low rates.

6.1 Future directions

We plan to create a web-based dashboard showing the graphical visualisations and statistics presented in this paper. The dashboard will consist of interactive charts dynamically populated with data in real time. The back-end data about the conference will be updated as and when more entries are recorded in the DBLP. The motivation is to provide a continuous display of the most recent data and insights rather than just a one-time study on a limited dataset. In this paper, we explored only eleven RQs due to the limited space available; however, in future work, we plan to investigate additional RQs such as identifying chasm papers, identifying the top 10 most-cited papers for each conference, determining the correlation between best-paper awards and the citation impact of the respective papers, core and sub-community analyses of authors, the small-world phenomenon, analysis of the largest connected sub-components in the co-authorship graph, a comparison of SIGWEB sponsored conferences with IEEE sponsored conferences in the area of HT and web, characteristic path length and clustering coefficients, identification of duos, trios and pivotal authors, out-degree distribution and minimum average path length over time, and statistics on internal and external collaborations.

References


Notes

1 http://www.sigweb.org/
2 http://dblp.uni-trier.de/
3 http://www.sigweb.org/conference/sigweb-conference/20-conferences/cooperating-conferences
4 http://www.sigweb.org
5 https://developers.google.com/maps/
7 http://dl.acm.org/citation.cfm?id=2505680&CFID=670930778&CFTOKEN=87104195
8 http://dl.acm.org/citation.cfm?id=2631811
9 http://dl.acm.org/citation.cfm?id=2481495
10 http://dl.acm.org/citation.cfm?id=2685303
11 https://genderize.io/
12 http://wiki.openstreetmap.org/wiki/
13 http://www.alchemyapi.com/products/demo/alchemylanguage
14 https://developers.google.com/maps/documentation/geocoding/intro
16 https://data.mendeley.com/datasets/dn5d8fbkb9/2