Scientometrics (2010) 83:133–155
DOI 10.1007/s11192-009-0078-y

# Assessing the quality of scientific conferences based on bibliographic citations

**Waister Silva Martins · Marcos André Gonçalves · Alberto H. F. Laender · Nivio Ziviani**

Received: 9 December 2008 / Published online: 11 September 2009


**Abstract** Assessing the quality of scientific conferences is an important and useful service that can be provided by digital libraries and similar systems. This is specially true for fields such as Computer Science and Electric Engineering, where conference publications are crucial. However, the majority of the existing quality metrics, particularly those relying on bibliographic citations, has been proposed for measuring the quality of journals. In this article we conduct a study about the relative performance of existing journal metrics in assessing the quality of scientific conferences. More importantly, departing from a deep analysis of the deficiencies of these metrics, we propose a new set of quality metrics especially designed to capture intrinsic and important aspects related to conferences, such as longevity, popularity, prestige, and periodicity. To demonstrate the effectiveness of the proposed metrics, we have conducted two sets of experiments that contrast their results against a "gold standard" produced by a large group of specialists. Our metrics obtained gains of more than 12% when compared to the most consistent journal quality metric and up to 58% when compared to standard metrics such as Thomson's Impact Factor.

**Keywords** Bibliometrics · Citation analysis · Ranking · Classification

W. S. Martins · M. A. Gonçalves (✉) · A. H. F. Laender · N. Ziviani
Computer Science Department, Federal University of Minas Gerais, Belo Horizonte 31270-901, MG, Brazil
e-mail: mgoncalv@dcc.ufmg.br

W. S. Martins
e-mail: waistermortins@yahoo.com.br

A. H. F. Laender
e-mail: laender@dcc.ufmg.br

N. Ziviani
e-mail: nivio@dcc.ufmg.br

## Introduction

The Web contains large repositories of scientific literature, such as digital libraries and preprint archives, whose content is usually targeted to specific communities. A major issue, that is crucial for scientific information, is to determine the quality of this content. In many cases, the quality of a scientific paper can be derived directly from the quality of the venue in which it was published. Therefore, determining the quality of a publication venue is an essential information for researchers, since it can help in the search of works of higher quality that can influence an ongoing research or in the decision about to which venue to submit a work. The quality of a venue may also be used to help on decisions regarding promotions, awards, funding, and scholarships.

Two common strategies to determine the quality of publication venues are consultation with specialists and application of quality metrics. The first strategy involves collecting and analyzing the opinion of several specialists in a given field. This generally produces a very good assessment of the quality of the venues when the number of consulted specialists is significant, since the specialists can use all their knowledge and experience to provide precise opinions about them. Examples of projects that employ this strategy include VHB-Jourqual2,[1] Association of Business Schools - Academic Journal Quality Guide,[2] and ABDC Journal List.[3] However the cost and effort associated with consulting and collecting the opinion of a large number of specialists may be very high. Also, in very dynamic fields, in which new venues are created or cease to exist very frequently and in which venues may change quality often, this approach becomes infeasible.

In the second strategy, quality metrics designed by information specialists or specialized organizations (e.g., Thomson Reuters[4]) are applied to produce statistics trying to estimate the quality of the venues. Many of these metrics rely on bibliographic citations as their main source of information. Usually a citation graph is built in which venues are nodes and citations between them are edges. Then, statistics about the graph are generated, as the basis to estimate the quality of the venues.

Several metrics based on citation analysis have been proposed in the literature. Among them we can cite Thomson's Impact Factor (Amin and Mabe 2000), Y-Factor (Bollen et al. 2006) and *h*-index (Hirsch 2005). Most of these metrics have been designed for assessing journals, since, in most of the knowledge fields, these are the most important publication venues. However, in fields such as Computer Science and Electrical Engineering, where the knowledge advances in a very fast pace, scientific conferences[5] assume a crucial role to support rapid dissemination of research results (Patterson 2004). As show by Laender et al. (2008), the scientific production in Computer Science is strongly centered on conferences in a ratio of 2.49 conference papers to each journal article. Some of these conferences are very competitive, having acceptance rates close to 10%. Nevertheless, there are no consolidated criteria and metrics to qualify conferences. Moreover, not much is known about the performance of the existing journal metrics to predict the quality of conferences.

One difficulty to estimate the quality of conferences based on citation analysis is to obtain a significant amount of bibliometric data about conferences, which may involve

[1] http://pbwi2www.uni-paderborn.de/WWW/VHB/VHB-Online.nsf.

[2] http://www.the-abs.org.uk/?id=257.

[3] http://www.abdc.edu.au.

[4] http://scientific.thomson.com.

[5] We will use the term "conference" to also denote other types of scientific meetings such as symposia, workshops, etc.

crawling, integration, extraction, and mining of Web content. However, the expansion and popularization of large digital libraries and repositories of scientific literature, such as CiteSeer,[6] DBLP,[7]. Google Scholar,[8] and Libra,[9] have made this type of analysis possible. Some of these systems are very large and comprehensive, containing high quality metadata about scientific papers, covering a large number of conferences.

In this article, we conduct a study about the relative performance of existing journal metrics in assessing the quality of scientific conferences. The deficiencies and problems found in these metrics motivated us to propose a set of new quality metrics especially designed to capture intrinsic and important aspects related to conferences such as longevity, popularity, prestige, and periodicity.

In order to evaluate the effectiveness of the newly proposed metrics to assess the quality of conferences, we have conducted two sets of experiments. In the first one, our metrics were used to rank a set of Computer Science conferences and the results were contrasted against a "gold standard" produced by a large group of specialists. Then, we used our metrics to classify these conferences with respect to some pre-established quality levels, also according to the gold standard. Our metrics obtained gains up to 8.4% in ranking similarity and 7.8% in classification accuracy when compared to best journal quality metrics and of more than 58% when compared to standard metrics such as Thomson's Impact Factor.

The remaining of this article is organized as follows. Firstly, we describe related work. Then, we present an analysis of some of the most important existing citation-based metrics for assessing journals, followed by a detailed description of the citation-based metrics we propose for assessing scientific conferences. The datasets used in our experiments and the obtained results are described next. Finally, we present our conclusions.

## Related work

The most common form to evaluate the reputation and quality of a publication venue is through citation analysis. In citation analysis, the quality of a venue is directly related to the citations received by the papers published by that venue. There are several citation-based metrics proposed in the literature, for example, Thomson's Impact Factor (IF) (Amin and Mabe 2000), Y-Factor (Bollen et al. 2006), and *h*-index (Hirsch 2005). Among these metrics, IF is the most widely used one and has also been adopted by some digital libraries. However, since its conception, IF is largely criticized due to its sole dependency on citation counts (Saha et al. 2003). Seglen (1997), for instance, discusses a number of limitations regarding the validity and applicability of IF. In order to cope with the IF limitations, several other metrics have been proposed, some of which are discussed below and in the next section.

Clausen and Wormell (2001) present a deep analysis of one particular conference, namely, the International Online Information Meeting (IOLIM) conference. By means of statistical and bibliometric analysis they provide quantitative information about geographic distribution of members of organising/advisory committees, referees, panelists, authors, delegates and citations. Bibliographic citations are also used as a main indicator of the

[6] http://citeseer.ist.psu.edu.

[7] http://dblp.uni-trier.de

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

[9] http://libra.msra.cn.

intellectual impact of the IOLIM series of conferences. This is one the first works to discuss the problem and to make an effort to identify relevant sources of information for assessing the impact of conferences.

An approach to determine the impact of journals based on the centrality metric of social networks has been presented by Bollen et al. (2005). Bollen et al. (2008) and Bollen and de Sompel (2008) present a comparative study of usage and citation-based metrics. A large amount of usage data was collected and used to validate the application of usage-based metrics in measuring the impact of journals. Zhuang et al. (2007) propose a method for assessing the quality of scientific conferences. This method is based on program committee characteristics and makes the assumption that the quality of a conference is directly related to the quality of its program committee. In another method for assessing the quality of conferences, Yan and Lee (2007) consider that in each field there is a set of recognizable papers of good quality (the seeds). Thus, to determine the quality of a conference, one should look for those papers with a quality similar to the seeds. For this, three metrics that basically consider the set of authors a paper has in common with the seeds are proposed. The problem here is to find good seeds that are not too restrictive. Another work focused on database conferences is presented by Larsen and Ingweresen (2006). There, citations are used for ranking conferences within a digital library. A comprehensive citation analysis for two main database conferences (SIGMOD and VLDB) and three database journals (TODS, VLDB Journal, and SIGMOD Record) is presented by Rahm and Thor (2005). Souto et al. (2007) develop a classification model to support the (semi-)automatic evaluation of Computer Science conferences based on ontologies and inference rules.

Despite the existence of related work that explores other sources of evidence about the quality of conferences (Zhuang et al. 2007; Yan and Lee 2007; Souto et al. 2007), our work is focused on metrics based on citations, since this is the most widely used information about venue quality and is largely available in digital libraries and other similar systems on the Web. In comparison, information such as acceptance rate, program committee characteristics, special ontologies, etc, is much harder to obtain. This makes the application of our newly proposed metrics very practical.

## Analysis of existing citation-based metrics

In this section, we present some of the most popular citation-based metrics used to estimate the quality of journals. In the following discussion we analyze the characteristics of each of these metrics and show that none of them captures all aspects required by a good metric for assessing the quality of conferences.

### Citation Count

Citation Count (CC) is the simplest metric used to assess the quality of a journal and has also been used to estimate the impact of a paper as well as of a researcher or group of researchers. The CC of a journal X ($CC_X$) is calculated by counting the number of citations received by all papers published in X, then:

$$CC_X = |I_X|,$$

where $I_X$ is the set of citations from articles citing articles published in X. Despite being widely used, we can mention two negative points about this metric. The first one is that citations coming from a low quality journal have the same value as citations coming from a

high quality journal, i.e., CC does not weight citations according to the quality of the journal from where they come from. The second negative point is that older journals may be favored when compared to newer ones, simply because, by having articles published earlier, they may have received more citations over time than articles published in newer journals.

### Impact Factor

The Impact Factor (IF) is a metric proposed by the *Institute for Scientific Information*.[10] It measures the quality of a journal based on the average number of citations in a year given to those articles in a journal that were published during the two preceding years. The citation window is therefore one year, pointing back in time (synchronic) to the previous two years of publications in the journal. IF is the most popular metric to assess the quality of journals, being widely applied. A deep study on IF is presented by Amin and Mabe (et al. 2000). However, we have not been able to find any study on its application to conference assessment.

More specifically, the IF of a journal X in a given year Y is calculated by counting the number of citations received by articles published in X, in the two years previous to Y, from articles published in any journal in the year Y. This number is divided by the number of articles published in X in the two years previous to Y.

Like CC, IF does not also consider the prestige or quality of the journal the citation has come from. The main difference of IF when compared with CC is that CC promotes older journals that receive a large number of citations over time and IF measures the current popularity of articles published in a journal, thus new journals are not penalized.

### Weighted PageRank

A variation of the PageRank algorithm (Brin and Page 1998) adapted for citation analysis has been proposed by Bollen et al. (2006). The proposed algorithm, called Weighted PageRank (WPR), expresses the prestige of a certain journal, considering the prestige of articles that cite that journal. This prestige is captured by considering weighted edges that determine the amount of prestige transferred to each journal by each article. This weight is calculated according to the number of times a journal cites another one. The formal definition is given by the formulas:

$$\mathrm{WPR_X} = (1-d) + d \sum_{\forall \mathrm{Y} \in J_\mathrm{X}} (\mathrm{WPR_Y} \times w_{\mathrm{Y,X}})$$

and

$$w_{\mathrm{Y,X}} = \frac{W(\mathrm{Y},\mathrm{X})}{\sum_{\forall \mathrm{Z}} W(\mathrm{Y},\mathrm{Z})},$$

where $W$(Y,X) and $W$(Y,Z) represent the number of citations from journal Y to journal X and from journal Y to journal Z, respectively, $J_\mathrm{X}$ is the set of journals citing journal X, and $d$ is a constant that can vary between 0 and 1 and is responsible for alleviating the amount of prestige transferred from one journal to another. For the experiments reported later, we used the value of 0.85, which is the same one employed in the original work

[10] Now called Thomson Reuters.

(Bollen et al. 2006). In the first instance all conferences have the same value of prestige, i.e., $\mathrm{WPR_X} = 0$ for all $X$.

Y-factor

The Y-Factor, also proposed by Bollen et al. (2006), is a metric that combines IF and WPR. The Y-Factor of a journal X is given by:

$$\text{Y-Factor}_\mathrm{X} = \mathrm{WPR_X} \times IF_\mathrm{X}.$$

As IF involves the counting of citations that a journal received in this year, it is a metric that expresses popularity. WPR, on the other hand, is a metric that expresses the prestige of a journal. Thus, Y-Factor is a metric that combines the prestige and the popularity of a journal, i.e., a journal with a high value for Y-Factor is likely to be popular and have a high prestige.

*h*-index

A metric that captures the quality of the scientific production of an individual, known as *h*-index, has been proposed by Hirsch (2005). The success of the *h*-index has led to extensions aimed at using it for assessing journals (Braun et al. 2006). Similarly to its original definition, a journal has an *h*-index of *h* if there is at least *h* papers published in that journal that have received at least *h* citations each. *h*-index captures the number of published papers with more impact in a journal over time, therefore newer journals have disadvantages under this metric. Notice that we could shorten the period for calculating the *h*-index to help dealing with this problem, however this could be harmful to older/more traditional conferences.

Analysis of the metrics for assessing the quality of scientific conferences

In order to analyze the major deficiencies of journal metrics for measuring the quality of conferences, we first consider intrinsic aspects of the conferences and discuss some characteristics that should be addressed by a metric specifically designed for conference assessment, namely:

- *Longevity versus current popularity*. The longevity of a conference brings important information about its quality. An older and more traditional conference normally will have a higher absolute number of citations, having, for instance, a high CC value. However, a conference may lose importance over time or even cease to exist. Another negative aspect of considering only the conference longevity is that there may be new conferences with high quality. Thus, a good metric should also consider the importance and current popularity of a conference.
- *Prestige versus popularity*. Some journal metrics capture only popularity, thus valorizing those venues that have a large number of citations. Although popularity is a valuable information, metrics that consider prestige, i.e., valorize citations coming from most prestigious venues, may also capture important information about the quality of the venues being evaluated. Thus, a metric that combines these two features is likely to produce better results.
- *Conference size*. The size of a conference may be considered as one possible indicator of its quality because high quality conferences usually attract a large number of good

**Table 1** Characteristics addressed by each metric ("Yes" means that the metric addresses the characteristic, "No" otherwise)

| | Longevity | Current popularity | Prestige | Conference size | Periodicity | Data coverage and sparseness |
|---|---|---|---|---|---|---|
| CC | Yes | No | No | No | Yes | Yes |
| IF | No | Yes | No | No | No | No |
| WPR | Yes | No | Yes | No | Yes | Yes |
| Y-Factor | Yes | Yes | Yes | No | Yes | Yes |
| *h*-index | Yes | No | No | No | Yes | Yes |

submissions as well as a large number of attendees. Thus, this information should be captured by any good metric for conferences. Certainly, there are large conferences of low quality and small conferences of high quality, such as SIGCOMM and SIGMETRICS, therefore the size of a conference should not be used as the sole criterion for assessment.

- *Periodicity*. Conferences may be annual, biennial or even triennial. A metric that does not consider a time window compatible with this specific time aspect may harm a conference or a particular group of conferences.
- *Data coverage and sparseness*. Contrary to what occurs with journals that have much more organized collections of metadata about them, there are few digital libraries that include a large spectrum of citation data about conferences. This means that the data collection process for assessing conferences may involve data originated from various sources. Thus, very often, a conference that is indexed by a digital library may not necessarily have all its editions or occurrences covered by that digital library. This may lead to a situation in which the citation data available for a conference does not cover a continuous time window. Similar to what occurs for periodicity, a metric that considers a time window that is not flexible enough to alleviate data coverage and sparseness problems due to possible missing editions may harm specific conferences.

Table 1 shows which of the above characteristics are addressed by the previously described journal metrics. As we can see, none of the journal metrics addresses all characteristics. Although one may argue that these characteristics could also be considered when assessing journals, it is clear that some of them are key for providing a more accurate assessment of conferences (Martins et al. 2009). For instance, a metric like IF, which considers a fixed time coverage, may harm those conferences for which citation data is not available for that specific time interval. Also, none of the journal metrics provides a means to take into consideration the size of a conference as an assessment feature. All these observations indicate that new metrics specific for assessing conferences are required.

## Proposed citation-based metrics for conferences

In this section, we propose new specific metrics for assessing the quality of conferences, which have been designed based on the analysis of existing citation-based metrics presented in the previous section.

### Conference Impact Factor

IF is the most popular metric for assessing the quality of journals, but its use for conferences should be investigated more deeply due to two important aspects not considered by this metric: the periodicity of the conferences and the coverage and sparseness of the data source. Using IF as originally proposed may be harmful to conferences that are not covered by the digital library in the specific time window used by IF, which considers the citations of papers published in a year to papers of the previous two years. For example, for a biennial or triennial conference, at most one edition of the conference would influence the calculation of IF, in terms of in- and out-citations. Also, as discussed earlier, data from some editions of a conference may not be covered by the digital library. Thus, even annual conferences may be subject to the same problems mentioned for biennial or triennial conferences. To address these issues, we propose a redefinition of IF, called Conference Impact Factor (CIF), which employs a larger time window, increasing the probability of obtaining data for the conference being assessed.

The CIF of a conference X in a given year Y is calculated by counting the number of citations that are made by papers of conferences published between the years Y and Y-3 to papers published in X in the three years previous to Y-3 and dividing that value by the total number of papers published in X in the three years previous to Y-3. Thus, we increase the time window for a period of six years. In this way, even triennial conferences (worst case) will have guaranteed two editions in that period. With a larger time window, the probability of a conference being impaired is lower, even if there are coverage problems. On the other hand, by considering a larger time window, CIF loses a little of "recentness". This time window could be further extended increasing the chances of obtaining more data. However, there exists a tradeoff between the recentness and the coverage, thus we chose to use the smallest time window that could help dealing with the worst case scenario.

Notice that the CIF definition and those of the subsequent metrics include only citations among conferences, since our dataset include only conference papers, as will be described in section Datasets. Extensions of these metrics to include other types of document such as books and journal articles are trivial.

It should be stressed that CIF is just one of new proposed metrics, one that builds upon some ideas of previously existing journal metrics, but that adapts them to better work for the context of conferences. As discussed in the previous section, there are several other intrinsic aspects of conferences that must be captured. This is done in each of the new metrics discussed next.

### Conference Citation Impact

CIF is the ratio between the total number of citations and the total number of papers published in a conference in a certain period of time. Using a similar idea applied to CIF, we exploit a time window of six years and propose the Conference Citation Impact (CCI). CCI measures the ratio between the total number of citations received by a conference in a period of time and the total number of citations received by all conferences in that period. CCI has the same advantages and disadvantages of CIF, but it tends to promote conferences that have a larger number of papers. Despite using some ideas of CIF, e.g., a larger time window, the modification proposed by CCI of considering the relative proportion of the number of citations to a given conference with regard to the total amount of all

citations, produces a new metric very suitable for conferences, significantly improving results when compared to CIF, as shown by our experimental results.

More specifically, the CCI of a conference X in a given year Y is calculated by counting the number of citations that are received by papers of X published between the years Y-3 and Y, by papers published in the three years previous to Y-3, and then dividing that value by the total number of citations made by papers published between the years Y-3 and Y to papers published in all conferences (from a same domain) in the three years previous to Y-3. CCI is multiplied by the number of conferences in order to produce a value with the same order of magnitude of CIF. This multiplication does not affect the measure generated by CCI, however it allows its combination with other metrics such as CIF with the same importance or weight in the final result. We notice that, the domain restriction is necessary since different domains may have very different citation patterns. This can be even further elaborated for sub-domains within a larger one, although this may be too fine-grained. We touch in some aspects of this in the next section but we leave a more complete elaboration of this issue for future work. In the definition of the next metrics, it should be implicit that all computations performed over all conferences or the total number of conferences refer to conferences in the same domain.

### Combined Conference Factor

Two important issues that are not explicitly addressed by CIF or CCI are the conference longevity and the CS. CCI tends to prefer conferences with more papers, however conferences may have few papers with many citations, therefore producing a high value for CCI. In order to try to explicitly capture the size of a conference, we propose the Conference Size (CS) metric, which estimates the size of a conference X by analyzing the number of papers published in X, since usually only the most important and largest conferences have an infrastructure to receive and publish a large quantity of papers. Again, we should stress that there are examples of large conferences of low quality and this factor should not be the only one taken into consideration. However, there is a tendency for low quality conferences to be smaller while conferences of high quality tend to be larger. This tendency is confirmed by the good results obtained by this metric in our experiments, as shown in section Results.

We define the CS of a conference X as the quantity of papers published in X in all years ($NA_X$) divided by the total number of published papers in all conferences (TNP). The result is multiplied by the number of conferences (NC), so that this value has the same order of magnitude of CIF. Formally, the CS of a conference X is defined as:

$$CS_X = \frac{NC \times NA_X}{TNP}.$$

To measure the importance of a conference over time, i.e., the conference longevity, we propose another simple metric called Conference Longevity (CL). The CL of a conference X consists of dividing $CC_X$ by the total number of citations that all conferences received (TNC) over time. This result is multiplied by NC, as was done for CS, so that its value has the same order of magnitude of CIF. Formally, the CL of a conference X is defined as:

$$CL_X = \frac{NC \times CC_X}{TNC}.$$

By combining CIF and CCI, and the two metrics defined above, we propose a new metric, called Combined Conference Factor (CCF), defined as:

$$\mathrm{CCF_X} = \mathrm{CIF_X} + \mathrm{CCI_X} + \mathrm{CS_X} + \mathrm{CL_X}.$$

CCF is a new and original metric that addresses most of the intrinsic characteristics of conferences, not explicitly covered by traditional journal metrics. It promotes traditional, large, and popular (longevity and current popularity) conferences, i.e, conferences with high quality indicators. Its suitability for conference assessment is confirmed in our experiments, as we shall see later when discussing our experimental results. It should be stressed, though, that a conference does not necessarily need to meet all the criteria included in CCF to be considered a conference of good quality. For example, a smaller conference may be have a enough longevity and citation record to be considered a high quality conference. CCF addresses most of the issues previously discussed, being prestige the only aspect not explicitly covered by CCF. This motivated us to develop the next metric, the Conference Factor.

### Conference Factor

Similarly to what was done for the Y-Factor, the Conference Factor (C-Factor) is a metric that combines prestige and popularity. However, while the Y-Factor is calculated multiplying WPR by IF, C-Factor replaces IF by CCF, since we believe CCF is more suitable for conference assessment. The C-Factor of a conference X is defined as:

$$\text{C-Factor}_\mathrm{X} = \mathrm{WPR_X} \times \mathrm{CCF_X}.$$

## Experimental evaluation

### Datasets

In this section we describe the two datasets used in our experiments. The first one is a ranking of Computer Science conferences that we produced as the result of an electronic poll conducted with specialists as part of a project in Brazil, called Perfil-CC, aimed at assessing the production quality of the top Brazilian Computer Science graduate programs (Laender et al. 2000). This ranking was used in our experiments as a "gold standard" in order to evaluate the effectiveness of our proposed metrics. The second dataset is a collection of citation metadata crawled from the Libra Academic Search.

#### *The Perfil-CC Ranking*

Here we explain the methodology used to build the Perfil-CC Ranking. Initially, several conference lists obtained from various sources were combined. After, we performed a cleaning process to remove duplicates and those conferences with the following characteristics: (1) conferences not yet consolidated (less than four editions), (2) conferences with submission by abstract (not full paper), and (3) regional conferences or conferences restricted to a specific country. To facilitate the voting process, the conferences remaining after the cleaning step were divided into 27 groups, representing a possible division of the Computer Science field (Laender et al. 2000).

To validate and improve the quality and coverage of our list of conferences, we invited all Brazilian Computer Science researchers that hold an individual grant from CNPq (The Brazilian National Research Council, similar to NSF),[11] and faculty members of all Computer Science graduate programs in the country, to assess the quality of the conferences in the list. All participants could suggest additions and removals of conferences from the list, as well as possible group changes. At the end of this process, we reached a list with just over a thousand conferences.

Next, we conducted an electronic poll about the quality of each conference in the list. In this poll, each researcher could vote only once in each group, but was allowed to vote in all groups, if desired. Each conference could be classified into one of three categories (A, B or C) according to its quality, being A the highest quality category. To avoid conferences being over-evaluated, for each group there was a limit of no more than 40% of conferences classified as A, having the sum of conferences A or B not exceeding 80%. Besides these, there were two other categories: NE ("not evaluated"), when a researcher chose not to vote for a particular conference, for considering not being able to do it (because of lack of knowledge about the conference, for example), and NC ("not considered"), to disqualify a conference whose quality was considered too low to even be listed. This step involved 312 subjects, 147 of which were CNPq researchers, making a total of 875 votes in all groups, which means that on average each subject voted in about three groups.

The final step was turning the votes into a number that would capture the quality of the conference. We investigated several scenarios considering different criteria to count and weight the votes. The differences among these scenarios were not significant, thus we choose to use the scenario which some consulted specialists considered the best. Accordingly, the weights of 3, 2 and 1 were assigned to votes A, B and C, respectively, and votes in NE and NC received zero weight. To normalize the conference scores to a number between 0 and 1, the score obtained with the weights was divided by three times the number of valid votes received by each conference. We considered as valid votes, all votes different from NE and NC. Thus, votes NE and NC were not counted and did not influence the final result. Conferences that had a number of valid votes lower that 40% were removed from the final result. This percentage was chosen experimentally to avoid cases of conferences highly classified but with very few valid votes. After this step, the final list of evaluated conferences reached an exact number of one thousand conferences. The final generated ranking can be accessed at http://www.latin.dcc.ufmg.br/perfilccranking/.

To further validate our ranking, we compare it against the CORE Ranking of ICT Conferences.[12] This ranking was generated as a result of a process similar to the one used in the Perfil-CC project, i.e., CORE, the Computing Research Association of Australasia, conducted a ranking exercise through a series of meetings attended by Australian academics, involving approximately 1,500 Computer Science conferences. There is a large intersection between the CORE and the Perfil-CC conference lists, the main exceptions being regional and local conferences (e.g., Asian-pacific conferences) which are included in the CORE list. An important difference between the projects is the number of categories and the distribution of conferences in each category. CORE divides conferences into five categories: A+ or highest quality; A or high quality; B or medium quality; C or low quality; and L or regional conferences of low quality. The distribution of conferences in the CORE list is shown in Table 2.

[11] These researchers receive this grant based on the quality of productivity and are considered as top researchers or leaders in their respective fields.

[12] http://www.core.edu.au.

**Table 2** Percentage of conferences in each category of the CORE ranking

| Category | A+ | A | B | C | L |
|---|---|---|---|---|---|
| Size (%) | 6 | 27 | 31 | 29 | 6 |

**Table 3** Confusion matrix contrasting the Perfil-CC and CORE categories

| | | CORE | | |
|---|---|---|---|---|
| | | A | B | C |
| Perfil-CC | A | 55 | 16 | 4 |
| | B | 19 | 15 | 7 |
| | C | 3 | 5 | 9 |

To compare the Perfil-CC and CORE rankings, we consider a set of 194 conferences extracted from the Perfil-CC list and used in our experiments (an explanation about the choice of these conferences is given next when we describe the Libra dataset). From these 194 conferences, 133 are common to both lists. To compare both rankings, we map conferences of categories A+ and A to a same unique category labeled A, containing only high quality conferences. We also do not consider category L, since none of regional conferences were in our list. As a result, both lists have three categories after the modifications. Table 3 shows the confusion matrix contrasting the categories of Perfil-CC against the categories of the CORE ranking. As we can see, 79 (59.4%) of the 133 conferences were classified in the same category in the two rankings. This can be thought as a high value, if we consider the differences in distribution in each category and in cultural factors of the consulted communities that can influence the opinion about some conferences. Notice, for instance, that the number of conferences misclassified in the extremes (A as C and C as A) is very small. Therefore, we believe that the produced rankings are consistent and reflect an accurate classification of the quality of a considerable number of conferences.

The large number of conferences handled by the Perfil-CC and CORE projects reinforces the importance of an automatic process for conference assessment, since the execution of a polling process like these is very costly. Another important consideration is the large number of recently created conferences, which makes it difficult to keep an updated list of assessed conferences.

*The Libra Dataset*

Libra Academic Search or simply Libra is a digital library focused on the Computer Science field. Libra allows free search of bibliographic data helping users to find scientific papers of interest. It currently has more than 3 million documents organized according to the publication venue. Venues are divided into 23 groups, similar to Perfil-CC groups.

The choice of Libra as a source of metadata about citations has two main reasons. The first reason is the volume of information in Libra. The number of conferences is considerable and papers of these conferences have a large amount of metadata including citations, fundamental for this work. DBLP, one of the largest Computer Science digital libraries, would be a good alternative because it covers a lot of conferences and the quality of their metadata is very high; however very few DBLP entries have citation information. The second reason for choosing Libra is its coverage with respect to the Perfil-CC conference

**Table 4** Libra coverage of conferences for six selected Perfil-CC groups

| Group | Libra coverage (%) |
|---|---|
| Machine Learning | 84.2 |
| Databases, Information Retrieval, Digital Libraries, Data Mining | 88.9 |
| Computational Biology | 83.3 |
| Human–Computer Interaction, Collaborative Systems | 88.9 |
| Networks, Distributed Systems, P2P Systems | 68.3 |
| Web, Multimedia and Hypermedia Systems | 79.1 |

**Table 5** Description of data crawled from Libra

| | |
|---|---|
| Crawled conferences | 194 |
| Crawled papers | 109,969 |
| Total of references made/citations received by papers in the dataset (only internal) | 145,282 |
| Total of papers making references to other papers in the dataset (only internal) | 27,759 |
| Total of papers receiving at least one citation | 43,749 |

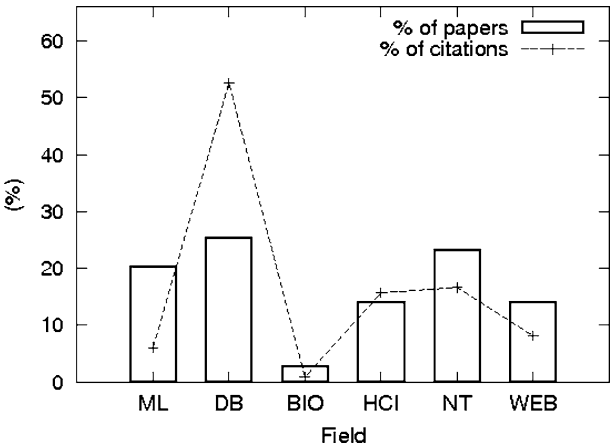


**Fig. 1** Percentage of papers and citations for each group from all crawled papers and captured citations

list, which for some groups reaches more than 80%, as can be seen in Table 4. CiteSeer, a third option, was the first digital library in the Computer Science field that indexed and connected citations automatically. However, CiteSeer provides no organization of papers by venue, which makes it difficult a crawling focused on conferences of interest as well as to find out the coverage of conferences with respect to the Perfil-CC Ranking. Moreover, the number of documents in CiteSeer, just over 760 thousands, is much smaller than in Libra.

To perform the experiments, we crawled from Libra the information about papers and citations of conferences from six groups of the Perfil-CC list, namely: Machine Learning (ML); Databases, Information Retrieval, Digital Libraries, Data Mining (DB); Computational Biology (BIO); Human–Computer Interaction, Collaborative Systems (HCI);

Networking, Distributed Systems, P2P Systems (NT); and Web, Multimedia and Hypermedia Systems (WEB). These groups were chosen prioritizing those that have greater coverage with respect to the Perfil-CC conference list. The results of this crawling is summarized in Table 5. An analysis based on each group was also performed. Figure 1 shows the percentage of papers and citations from each group. The DB group has the highest number of papers and most citations. This is the actual distribution gathered from Libra, and reflects the fact the we put together, under this label, several large Computer Science subfields related to information systems, such as Databases, Information Retrieval and Data Mining, which also have a good coverage in that Digital Library, making the group very large by nature. The difference to other groups is significant and may favor conferences of this specific group in our experiments.

## Results

### *Ranking comparison*

For each of the metrics presented (the newly proposed and the existing ones) a ranking was generated and compared with the Perfil-CC Ranking according to several evaluation metrics. The first of these metrics, called Top 30, is based on comparing the first 30 elements of each ranking with the first 30 elements taken from the Perfil-CC Ranking, i.e., how many conferences they have in common. This metric is effective only for conferences of higher quality as it compares only conferences of the Top 30 in the ranking.

Another way to compare the similarity between two rankings is to calculate the distance between each corresponding element in the rankings. Simple Distance, the second evaluation metric used, is defined as:

$$D(m_1, m_2) = \frac{1}{N} \sum_{\forall i} \frac{|P_{m1}(i) - P_{m2}(i)|}{N},$$

where $N$ is the number of elements in the ranking, and $P_{m1}(i)$ and $P_{m2}(i)$ are the elements in the $i$th position in the ranking built with the metrics $m_1$ and $m_2$, respectively. The function $D(m_1, m_2)$ produces as output a number between 0 (when the rankings are equal) and 0.5 (when the rankings are in reverse order).

When comparing two rankings, differences in higher positions should be weighted more importantly than in lower positions. This means that it is more important to improve the ranking by trying to put the best conferences on the top positions. The Simple Distance is a metric that does not take this factor into account, because the distances between all elements have the same weight. For this reason, we also used another evaluation metric, Weighted Distance, defined by Sidiropoulos and Manolopoulos (2000) as:

$$D_{\mathrm{w}}(m_1, m_2) = \frac{1}{N \sum_{\forall i \in V} w(m_1, m_2, i)} \sum_{\forall i \in V} d_{\mathrm{w}}(m_1, m_2, i),$$

$$d_{\mathrm{w}}(m_1, m_2, i) = |P_{m1}(i) - P_{m2}(i)| \times w(m_1, m_2, i),$$

$$w(m_1, m_2, i) = \frac{1}{min(P_{m1}(i), P_{m2}(i))}.$$

Finally, we also used a well known metric to compute similarity between two rankings, the Kendall Tau Coefficient ($\tau$) (Kendall 1938), defined as:

**Table 6** Results of the rankings comparison using Top 30, Simple Distance, Weighted Distance and Kendall Tau Coefficient

| | Top 30 | $D$ | $D_w$ | $\tau$ |
|---|---|---|---|---|
| CC | 19 | 0.2109 | 0.1191 | 0.3840 |
| IF | 13 | 0.2490 | 0.3233 | 0.2626 |
| WPR | 18 | 0.2159 | 0.1224 | 0.3652 |
| Y-Factor | 16 | 0.2359 | 0.1298 | 0.3053 |
| *h*-index | 18 | 0.2167 | 0.1257 | 0.3640 |
| CIF | 12 | 0.2352 | 0.2586 | 0.2973 |
| CCI | 19 | 0.2077 | 0.1210 | 0.3718 |
| CCF | **21** | 0.2031 | 0.1166 | 0.3915 |
| C-Factor | 19 | **0.1951** | **0.1132** | **0.4161** |

Best results are in boldface style

$$\tau = \frac{2P}{\frac{1}{2}N(N-1)},$$

where $N$ is the number of elements in the rankings and $P$ is the number of pairs that are in the same order in the two rankings subtracting the elements that are out of order, i.e., each pair of elements is analyzed and for each pair that appears in the second ranking in the same order as in the first, we add 1 to P, and if the pair does not appear in the same order, we subtract 1. The value of $\tau$ is $+1$ when the rankings are equal and $-1$ when the rankings are exactly reverse.

The results of the ranking comparisons generated by all citation metrics and using all evaluation metrics defined above can be seen in Table 6. For the Simple Distance ($D$) and Weighted Distance ($D_w$), lower values represent better results. As can be seen, CC obtained the best results among the existing metrics, followed closely by the WPR and *h*-index metrics. C-Factor, a new metric, obtained the ranking most similar to the Perfil-CC Ranking under all evaluation metrics but Top 30 (CCF obtained the best result under this metric), with gains, considering the best result of the existing metrics (i.e, those of CC), of up to 8.4% in the case of $\tau$. Against the second best metric (WPR), our gains were of about 14% in $\tau$, for example. Notice that some of these best results were produced against recently proposed metrics. If we consider most traditional and largely used metrics, such as Impact Factor, our gains are of more than 58% (in case of $\tau$). These results, when compared with those of the existing metrics, show the importance of having considered the intrinsic aspects of conferences when designing these two metrics.

Analyzing Table 6 in more detail, it can be seen that CIF has produced a more similar ranking to the Perfil-CC Ranking than IF. This should be expected due to the increased time window which decreases the probability that some conferences are harmed by the lack of data. CC and *h*-index, which explore the conference longevity and suffer less with the problems of coverage and periodicity, also obtained good results. The reason for this is that cases in which a conference looses importance over the years are rare. The normal tendency is that a conference will become consolidated as time goes by and its importance will increase. Therefore, the conference longevity may be enough in some cases to judge its quality. The good results of WPR demonstrate the importance of prestige for assessing conferences. This fact can also be observed in the small improvements of C-Factor over CCF and in the superiority of the Y-Factor results when compared with IF.

Looking more deeply at each of the generated rankings, we found some coverage problems that may explain some of the results. Of the 194 conferences considered, 14 (7.22%) of them did not have any citation. For IF, that considers a time window of only

**Table 7** Top 10 conferences in the Perfil-CC Ranking and in the IF, CCF and C-Factor rankings

| Rank | Perfil-CC Ranking | IF | CCF | C-Factor |
|---|---|---|---|---|
| 1 | INFOCOM | CIDR | SIGMOD | SIGMOD |
| 2 | SIGCOMM | SIGCOMM | VLDB | VLDB |
| 3 | CSCW | SIGMOD | CHI | CHI |
| 4 | CHI | VLDB | SIGCOMM | SIGCOMM |
| 5 | KDD | PODS | INFOCOM | INFOCOM |
| 6 | WWW | NSDI | WWW | ICDE |
| 7 | ICML | IMC | ICDE | SIGIR |
| 8 | ACM-MM | KDD | PODS | PODS |
| 9 | SIGMOD | EDBT | SIGIR | ICML |
| 10 | ICC | UIST | KDD | WWW |

3 years, the number of conferences not having any citations is 74 (38.14%); using the time window of 6 years, as proposed for CIF, this number decreases to 47 (24.23%). Consequently, these conferences without any citations will be ranked in the last positions by practically all metrics. This analysis shows the importance of considering the conference longevity and is a justification for the good results of the metrics that consider this aspect. The metrics incorporating CS and that, therefore, consider the number of papers published by the conference can also better assess the quality of conferences without citations. The lack of citation data will negatively impact the assessment of these conferences, but they will be no longer obligatorily ranked in the last positions, since there is some information to evaluate them. This analysis is also valid for conferences that have a bad citation coverage.

To have a better idea about the rankings generated by our metrics, Table 7 shows the Top 10 conferences in the Perfil-CC Ranking and in the rankings generated by IF, CCF, and C-Factor. As can be seen, the CCF and C-Factor rankings are very similar in the Top 10 positions, having nine conferences in common. Despite being a bit different from the Perfil-CC Top 10 ranking, all conferences that appear in the Top 10 positions in the CCF and C-Factor rankings are definitively high quality conferences. These two metrics favor conferences from the DB group, i.e., the Top 10 include six DB conferences in CCF and five in C-Factor. This is due to the fact the DB group has the best coverage in Libra. On the other hand, the Top 10 in the IF ranking are very different from those in the Perfil-CC Ranking, including three conferences classified below the 50th position in that ranking (EDBT 56th, NSDI 100th, and CIDR 183th). This shows that the new proposed metrics are more consistent and produce a great improvement when compared with the IF metric, the most traditional metric for journals.

Table 8 shows the importance of the metric CS as a quality indicator. The table presents the Top 10 and Bottom 10 conferences ranked by CS. As we can see, Top 10 includes two conferences ranked below the 100th position in the Perfil-CC Ranking. Moreover, CHI, INFOCOM, ICDE, SIGMOD, and VLDB are good examples of large conferences of very high quality. In addition, in the Bottom 10 there are only conferences ranked between the 100th and 193th positions in the Perfil-CC Ranking. This indicates that CS is a valuable indicator, however, it should not be used in isolation.

One interesting result that required a deeper analysis was the Top 30. The conferences in our sample ranked in the Top 30 according to the Perfil-CC Ranking are probably very important and popular conferences that should have a greater coverage in Libra and, therefore, would be correctly ranked by our metrics. However, the best result for the Top 30 was obtained with CCF, with 21 conferences correctly classified in the Top 30, i.e., the

**Table 8** Top 10 and Bottom 10 conferences in the Perfil-CC Ranking and, their respective position according to CS

| Top 10 | | | Bottom 10 | | |
|---|---|---|---|---|---|
| Rank | Conf. | Pos. | Rank | Conf. | Pos. |
| 1 | CHI | 4 | 185 | SAPIR | 193 |
| 2 | INFOCOM | 1 | 186 | DIWeb | 178 |
| 3 | ICPR | 52 | 187 | SWWS | 151 |
| 4 | NIPS | 24 | 188 | NSDI | 100 |
| 5 | ICDE | 13 | 189 | W4A | 174 |
| 6 | SIGMOD | 9 | 190 | EuroITV | 142 |
| 7 | VLDB | 12 | 191 | PDC | 175 |
| 8 | WebNet | 176 | 192 | WCW | 188 |
| 9 | ICME | 23 | 193 | GIR | 167 |
| 10 | KES | 110 | 194 | Web3D | 158 |

best result ranks nine conferences outside of the Top 30. An analysis of this showed that seven of these top conferences were not ranked in the Top 30 by none of the metrics. We investigate two possibilities: (i) Libra does not have enough coverage for these seven conferences hindering their ranking in the Top 30 and (ii) citation data only is not sufficient to rank these conferences correctly and more information is needed. These seven conferences are: International Conference on Fuzzy Systems (FUZZ), International Conference on Communications (ICC), International Joint Conference on Neural Networks (IJCNN), European Conference on Machine Learning (ECML), International Conference on Web Services (ICSW), IEEE/WIC/ACM International Conference on Web Intelligence (WI), and ACM/IFIP/USENIX International Middleware Conference (MIDDLEWARE).

From the seven conferences that failed to rank in the Top 30 by practically all metrics, five (FUZZ, ICC, IJCNN, ICSW, WI) have serious coverage problems. The problem of coverage can be divided into two. First, Libra has not enough data on the conference and, second, the crawled dataset is not large and comprehensive enough. For example, conferences on Machine Learning may be impacted because of the absence of conferences on Artificial Intelligence in this dataset, since there is probably a strong connection between these two groups of conferences. The two remaining conferences, ECML and MIDDLEWARE, have a considerable amount of information in Libra, but still are not ranked in the Top 30 by any of the metrics. In these two cases additional information (different from citations) is possibly necessary for achieving a correct assessment.

Table 9 shows the number of conferences mistakenly ranked outside the Top 30. To determine how many positions above the 30th position the conferences are being ranked, we calculated the Simple Distance between the ranked position and the position 30. We

**Table 9** Top 30 results

| | CC | IF | WPR | Y-Factor | *h*-index | CIF | CCI | CCF | C-Factor |
|---|---|---|---|---|---|---|---|---|---|
| NEr | 11 | 17 | 12 | 14 | 12 | 18 | 11 | 9 | 11 |
| DMTop30 | 59.09 | 68.41 | 49.33 | 47.43 | 65.75 | 76.61 | 60.45 | 52.78 | 47.45 |

NEr is the number of conferences mistakenly ranked out of the Top 30 and DMTop30 is the average of the simple distance to the 30th position for these conferences

can see that, on average, Y-Factor and C-Factor were the metrics that made the best prediction for the mistakenly ranked conferences, ranking them approximately 48 positions away from the 30th position, i.e., at the 78th position. This indicates that there is missing information preventing the metrics to correctly rank these conferences in the Top 30. This missing information may be due to coverage problems of Libra or the exclusive dependency on citation data of our metrics. Thus, this should be further investigated in the future.

*Classifying conferences according to their quality*

The Perfil-CC project, besides providing a ranking for a thousand of conferences, also presents a division of these conferences into three categories (A, B and C), as explained before. Using this information, the conference distribution of our crawled sample in each category is known, being, respectively, for categories A, B and C the following: 88 (45.36%), 71 (36.60%) and 35 (18.04%).

We compared how the metrics being studied (the new and the existing ones) performed on the task of classifying the conferences of our sample into these three categories. For this, we used the position that the conference was ranked by each metric and the conference distribution into the categories. Thus, conferences classified by each metric until the 88th position were considered as belonging to category A, those between the 89th and 160th positions were considered as belonging to category B and those above the 160th positions were considered as belonging to category C.

To compare the performance of each metric in this task, we employed information retrieval measures commonly used for classification: accuracy, precision, and recall. Accuracy simply measures the sum of hits in each category. Precision is the ratio of correctly classified instances from a set of elements assigned to a given category while recall is the number of conferences correctly classified in a category divided by the correct number of elements of that category. We used precision and recall values averaged over all categories (Macro-Precision and Macro-Recall). To illustrate, we will define these measures using only two categories (A and B) as follows:

| | | Prediction | |
|---|---|---|---|
| | | A | B |
| True label | A | *a* | *b* |
| | B | *c* | *d* |

$$\text{Accuracy} = \frac{a+d}{a+b+c+d}$$

$$\text{Recall}_\text{A} = \frac{a}{a+b} \quad - \quad \text{Recall}_\text{B} = \frac{\text{d}}{\text{c}+\text{d}}$$

$$\text{Macro-Recall} = \frac{\text{Recall}_\text{A} + \text{Recall}_\text{B}}{2}$$

$$\text{Precision}_\text{A} = \frac{a}{a+c} \quad - \quad \text{Precision}_\text{B} = \frac{\text{d}}{\text{b}+\text{d}}$$

$$\text{Macro-Precision} = \frac{\text{Precision}_\text{A} + \text{Precision}_\text{B}}{2}$$

Here $a$ is the number of conferences correctly classified in category A, $b$ is the number of conferences mistakenly classified in category B, $c$ is the number of conferences

mistakenly classified in category A, and $d$ is the number of conferences correctly classified in category B.

In the experiments, the values of recall are not presented because they are the same of precision for all metrics. This was expected because of the method employed to transform the position of a conference in the ranking generated by each metric in one of three categories (A, B or C), which follows the distribution presented earlier. Thus, since the number of conferences classified in each category is fixed and a conference mistakenly classified necessarily appears in another category, then we have $b = c$. For example, the number of conferences from category A classified as belonging to category B is equal to the number of conferences from category B classified as belonging to category A and, therefore, precision is equal to recall.

Accuracy, precision and recall capture only the performance of the metrics in classifying conferences in the correct category, not considering the "gravity" of the error. For example, a conference A erroneously classified as C by a metric represents a much more serious error than classifying the same conference as B. To compare metrics and check which ones provide classifications with fewer more serious errors, we proposed the Category Point metric. For this we defined three different scenarios with different scores. Since the results of the three scenarios were very similar, we present only one of them. The Category Point is calculated as follows: for each conference ranked in the correct category, we sum up two points, for each conference A or C classified as B we sum up 1 point, for each conference B classified as A or C, we sum up 1 point, for each conference A classified as C or C classified as A no points are added.

Table 10 summarizes the results for all metrics. As before, the worst result was obtained by IF. CCF and C-Factor were again the best performing metrics. CCF classified 57.22% of the conferences in the correct category against 53.09 and 51.03% of WPR and CC, respectively, presenting a gain of 7.8% over WPR, the best of the existing metrics, of 12.13% over CC, the second best of the existing metrics in this task, and of 28% over IF. C-Factor, on the other hand, obtained the highest values of precision and Category Point, presenting gains of 5.8% in accuracy, 5.9% in precision and 3.4% for Category Point (10 points), over the best results of the existing metrics. When compared to IF the C-Factor gains are around 33% in precision and 11% in Category Point.

CIF results far exceeded IF, which shows that IF should not be used for assessing conferences in the way it was originally defined. CCI results were slightly better than the ones for CIF using the same time window, probably due to the promotion of conferences with larger number of published papers. $h$-index had intermediate results. A more complete

**Table 10** Classification results

| | Accuracy (%) | Precision (%) | Category Point |
|---|---|---|---|
| CC | 51.03 | 47.77 | 278 |
| IF | 44.33 | 39.70 | 258 |
| WPR | 53.09 | 49.65 | 278 |
| Y-Factor | 47.94 | 42.63 | 266 |
| $h$-index | 47.94 | 45.22 | 272 |
| CIF | 51.03 | 46.23 | 274 |
| CCI | 53.09 | 47.93 | 278 |
| CCF | **57.22** | 52.47 | 286 |
| C-Factor | 56.19 | **52.58** | **288** |

Best results are in boldface style

**Table 11** Confusion matrix generated by C-Factor

| | Prediction | | |
|---|---|---|---|
| | A | B | C |
| True label | | | |
| A | 59 | 22 | 7 |
| B | 21 | 36 | 14 |
| C | 8 | 13 | 14 |

analysis of *h*-index is needed to try to extract some features that may be incorporated to other metrics designed for conferences. CC is a very simple metric, but again obtained good results, reinforcing the importance of the conference longevity. Of the existing metrics, WPR, the metric that captures the conference prestige, was the one that obtained the best result overall. Y-Factor had bad results, most probably due to its use of IF. CCF and C-Factor were the best metrics for conference classification among the metrics studied since they were especially designed to cover specific aspects related to the evaluation of the quality of conferences.

Table 11 shows the confusion matrix generated by C-Factor, the metric that obtained the best value for Category Point. We can see that seven conferences from category A were classified as category C and eight conferences from category C were classified as category A. Notice that Category Point penalizes these cases where none of the others measures do. These results might be explained by the following facts. All seven conferences from category A classified as C have less than 10 citations, being three of them new conferences for which there is not yet much information about. The other ones are consolidated conferences which possibly suffer from coverage problems. Of the eight conferences from category C classified as A, three of them are conferences from the DB group and one is a conference from the HCI group, the two groups with the largest number of papers and citations in our dataset, which might explain the fact that they were well classified. Among the other four conferences, two have a large number of papers and one a large number of citations, which, again might explain their classification. The last C conference classified as A has a medium number of citations, but all of them are very recent (high current popularity), which yields a high value for CCI and CIF, two metrics that highly influence C-Factor.

Table 12 shows the accuracy of the metrics for each group of conferences. The results in bold configure the best result in each group. As can be seen, none of the metrics is the best in more than two groups. A deeper study about the characteristics of each group is necessary to better understand this phenomenon and to be able to suggest which metric should be used in each situation or group. In the absence of such understanding, we suggest to use CCF and C-Factor as the best metrics in general. In terms of the existing metrics, WPR is the most consistent one, working better when compared to other existing metrics, for example, CC, mainly when the coverage increases. However, the new metrics again had the best results, obtaining the best accuracy values in all groups but one, the DB group. In this group, the one with the largest coverage of papers and bibliographic citations, the WPR obtained the best result, followed by C-Factor. In this case, the network of citations is very dense and prestige became a very important feature. For the BIO and WEB groups, in which most of the conferences were created a few years ago (approximately seven years on average), the best result was obtained by CCI, the metric that captures the current popularity.

**Table 12** Accuracy by group

| | ML | DB | BIO | IHC | NT | WEB |
|---|---|---|---|---|---|---|
| CC | 37.50 | 46.30 | 40.00 | 68.23 | 54.55 | 45.45 |
| IF | 40.63 | 40.00 | 30.00 | 66.67 | 38.18 | 51.52 |
| WPR | 40.63 | **60.00** | 50.00 | 75.00 | 54.55 | 39.39 |
| Y-Factor | 43.75 | 45.00 | 40.00 | 70.83 | 36.51 | 54.55 |
| *h*-index | 37.50 | 47.50 | 40.00 | 62.50 | 49.09 | 48.48 |
| CIF | 46.87 | 45.00 | 50.00 | 62.50 | 50.90 | 54.55 |
| CCI | 50.00 | 47.50 | **60.00** | 66.67 | 47.27 | **60.61** |
| CCF | **56.25** | 52.50 | 40.00 | 75.00 | 56.36 | 57.58 |
| C-Factor | 40.63 | 55.00 | 40.00 | **83.33** | **58.18** | 54.55 |

Best results for each group are in boldface style

**Table 13** Classification results using the CORE ranking

| | Accuracy(%) | Precision(%) | Category Point |
|---|---|---|---|
| CC | 57.14 | 47.26 | 202 |
| IF | 54.89 | 44.04 | 198 |
| WPR | 54.14 | 43.23 | 198 |
| Y-Factor | 54.14 | 43.23 | 198 |
| *h*-index | 58.65 | 48.51 | 202 |
| CIF | 60.15 | 50.14 | 204 |
| CCI | **62.41** | **51.84** | **208** |
| CCF | 59.40 | 47.81 | 204 |
| C-Factor | 54.13 | 44.74 | 198 |

Best results are in boldface style

We also verify the effectiveness of all metrics with regard to the CORE "gold standard". As previously done, we did not consider category L and labeled as category A, conferences of categories A+ and A in the CORE ranking. Table 13 shows the results. As before our metrics obtained the best results. CCF obtained good results, however CCI was the best metric. Among existing metrics, *h*-index obtained the best results. A better understanding of these results involves a deeper analysis of the CORE ranking and is currently outside of the scope of this article. The main point of these experiment was to contrast our metrics with regard to another "gold standard" generated by a different community, thus helping to further validate our results, by showing that it is not biased towards one ranking only.

## Conclusions

In this article, we have presented a study about the relative performance of existing metrics designed for journal evaluation in the context of scientific conferences. Based on the deficiencies found on these metrics regarding characteristics such as longevity, popularity, prestige, and periodicity that are especially important for conferences, we have proposed a set of new citation-based metrics to assess the quality of conferences. Our experiments in

two tasks—conference ranking and conference classification—show that the new metrics significantly outperform the journal-based ones in almost all evaluation metrics when compared to a "gold standard" generated by specialists.

Despite the good results, there are still a lot of room for improvements, as shown by the considerable differences in the rankings and classifications generated by our metrics and the "gold standard". As future work, we intend to better analyze the impact of each incorporated characteristic in our metrics and how each of them influences the results. For example, our analysis of the results by group has shown that there is no clear winner in all groups and we want to explore this further in order to better understand when to apply each metric based on features of the groups. Improvements in the metrics (e.g., turning some of them in diachronic) are always worth trying.

Another obvious extension is to consider other important characteristics that go beyond citations, such as program committee information and acceptance rate. Finally, we intend to exploit machine learning techniques such as decision trees and genetic programming, whose results are easily interpreted, in order to design better techniques for conference quality assessment.

**Acknowledgements** This research is partially funded by the Brazilian National Institute of Science and Technology for the Web (MCT/CNPq Grant Number 573871/2008-6), by the InfoWeb project (grant number 55.0874/2007-0), and by the authors's individual research grants from CNPq.

## References

Amin, M., & Mabe, M. (2000). Impact factors: Use and abuse. *Perspectives in Publishing, 1*, 1–6.

Bollen, J., & de Sompel, H. V. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *JASIST, 59*, 136–149.

Bollen, J., de Sompel, H. V., & Rodriguez, M. A. (2008). Towards usage-based impact metrics: first results from the MEASUR project. In *Proceedings of the 8th ACM/IEEE joint conference on digital libraries*, ACM, New York, USA, pp. 231–240.

Bollen, J., de Sompel, H. V., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing and Management, 41*, 1419–1440.

Bollen, J., Rodriguez, M. A., & de Sompel, H. V. (2006). Journal status. *Scientometrics, 69*, 669–687.

Braun, T., Glänzel, W., & Schubert, A. (2006). A hirsch-type index for journals. *Scientometrics, 69*, 169–173.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*, 107–117.

Clausen, H., & Wormell, I. (2001). A bibliometric analysis of IOLIM conferences 1977–1999. *Journal of Information Science, 27*, 157–169.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*, 16569–16572.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*, 81–93.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., Silva, E. S., & Ziviani, N. (2008). Assessing the research and education quality of the top Brazilian computer graduate programs. *ACM SIGCSE Bulletin, 40*, 135–145.

Larsen, B., & Ingwersen, P. (2006). Using citations for ranking in digital libraries. In *Proceedings of the 6th ACM/IEEE joint conference on digital libraries*, Chapel Hill, NC, p. 370.

Martins, W. S., Gonçalves, M. A., Laender, A. H. F., & Pappa, G. L. (2009). Learning to assess the quality of scientific conferences: A case study in computer science. In *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries*, Austin, TX, pp. 193–202.

Patterson, D. A. (2004). The Health of Research Conferences and the Dearth of Big Idea Papers. *Communications of the ACM, 47*, 23–24.

Rahm, E., & Thor, A. (2005). Citation analysis of database publications. *SIGMOD Record, 34*, 48–53.

Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: A valid measure of journal quality?. *Journal of the Medical Library Association, 1*, 42–46.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314*, 498–502.

Sidiropoulos, A., & Manolopoulos, Y. (2006). Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software, 79*, 1679–1700.

Souto, M. A. M., Warpechowski, M., & de Oliveira, J. P. M. (2007). An ontological approach for the quality assessment of computer science conferences. *Proceedings of the 2007 workshop on quality of information systems, 4802*, 202–212.

Yan, S., & Lee D. (2007). Toward alternative measures for ranking venues: a case of database research community. In *Proceedings of the 7th ACM/IEEE joint conference on digital libraries*, ACM, New York, USA, pp. 235–244.

Zhuang, Z., Elmacioglu, E., Lee, D., & Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE joint conference on digital libraries*, ACM, New York, USA, pp. 225–234.