Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science

Waister Silva Martins, Marcos André Gonçalves,
Alberto H. F. Laender, Gisele L. Pappa

Computer Science Department
Federal University of Minas Gerais
31270-901 Belo Horizonte, Brazil
{waister,mgoncalv,laender,glpappa}@dcc.ufmg.br

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 approaches for assessing the quality of publication venues has been proposed for journals. In this paper, we characterize a large number of features that can be used as criteria to assess the quality of scientific conferences and study how these several features can be automatically combined by means of machine learning techniques to effectively perform this task. Within the features studied are citations, submission and acceptance rates, tradition of the conference, and reputation of the program committee members. Among our several findings, we can cite that: (1) separating high quality conferences from medium and low quality ones can be performed quite effectively, but separating the last two types is a much harder task; and (2) citation features followed by those associated with the tradition of the conference are the most important ones for the task.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Measurement

Keywords

Machine Learning, Classification, Digital Library, Conference Assessment

1. INTRODUCTION

An important service that can be provided by digital libraries of scientific works and similar repositories is to inform the *quality* of the publication venue in which an article was published. The quality of a publication venue is an important information for many researchers; it can help in the search for good material that can influence their ongoing research or in the decision to which venue to submit a piece of finished research. The quality of publication venues is also useful to help on decisions about promotions and awards as well for deciding about grants and scholarships funded by research agencies and institutions.

Scientific journals are the focus of most of the existent studies in the literature about the quality of publication venues, given that, in most of the knowledge and scientific fields, this type of venue is the most important one. However, in some fields such as Computer Science and Electric Engineering, conferences[1] are important channels for the fast dissemination of research results [11]. As shown in [8], the scientific production in Computer Science is strongly centered on conferences in a ratio of 2.49 conference papers for each journal article.

Most of the existent quality metrics use citation analysis to infer the quality of publication venues. Citation analysis consists of analyzing citation patterns and the frequency of citations among journals and books. For journals, Thomson's Impact Factor [1] is the most popular and accepted citation analysis method. Some works, however, argue against its sole use as a quality measure, since it captures only one aspect of the quality spectrum [14, 15]. In the case of scientific conferences, the situation is even worse, since there are no consolidated metrics or criteria.

Another popular approach to assessing the quality of publication venues is to consult and analyze the opinion of a large number of specialists in a given knowledge field. This approach normally produces a good classification, when the number of specialists is significant enough, given that these specialists use their experience and knowledge to provide precise opinions about the quality of the venues. Examples of projects that employ this strategy include VHB-Jourqual2[2], ABDC Journal List[3] and CORE Ranking of ICT Conferences[4]. However, the cost for consulting and collecting the opinion of a large number of specialists may

JCDL'09, June 15–19, 2009, Austin, Texas, USA.

[1]We use the term 'conference' to also refer to other types of scientific meeting such as workshops, symposia, etc.
[2]http://pbwi2www.uni-paderborn.de/WWW/VHB/VHB-Online.nsf
[3]http://www.abdc.edu.au
[4]http://www.core.edu.au

be extremely high, even impractical, particularly in very dynamic fields, in which venues are created, cease to exist, and change quality very often. Moreover, this method may cause some distortions among subfields of a large knowledge field when there are a disproportionate number of people and venues within these subfields, depending on the way the consultation is conducted, as we shall see in our experimental results.

Among the innumerous criteria used by the specialists to give an opinion about the quality of a conference, we can cite: tradition, visibility, acceptance rate, reputation of the program committee members, among others. In this paper, we study the discriminative power of several of these criteria, as well as of their combinations, in being able to distinguish high, from medium, from low quality conferences.

For conducting our study, we collected a large amount of data about these criteria for 194 Computer Science conferences from digital libraries, conference sites, and other Web-based sources. By using automatic classifiers, we combined these criteria in order to find the best mixture of evidence for the task at hand. Thus, in sum, the contributions of this paper are: (1) the characterization of a large number of features that can be used as criteria to assess the quality of scientific conferences; (2) a study of how these several features can be combined by means of machine learning techniques to automatically and effectively classify conferences; and (3) a deep analysis and discussion about the relative difficulty of the problem.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes the datasets and the framework used in our experiments. Section 4 presents a characterization of the conference features analyzed. Our experiments and results are presented and discussed in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

As already mentioned, most of the existing work on assessing the quality of publication venues is focused on journals. This happens because this type of venue is the most important one in the majority of the knowledge and scientific fields. However, in fields that require a fast dissemination of knowledge, such as Computer Science and Electric Engineering, conferences assume an important role since the publication process in this type of venue is normally shorter than for journals. However, there are no consolidated metrics for assessing conferences.

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) [1], Y-Factor [4], and H-index [7]. Among these metrics, IF is the most widely used. However, since its conception, IF is largely criticized due to its sole dependency on citation accounts [14]. In [15], for instance, a number of limitations regarding the validity and applicability of IF are discussed. Thus, in order to cope with the IF limitations, several other metrics have been proposed in the literature for assessing journals. For instance, Bollen et al. [3] present an approach to determine the impact of journals based on the centrality metric of social networks. In [2], the authors present a comparative study of usage and citation-based metrics. In their work, they collected a large amount of usage data and used them to validate the application of usage-based metrics in measuring the impact of journals.

Although most existing methods aim at assessing the quality of journals, a number of methods specifically designed for scientific conferences have been recently proposed. In [20], Zhuang et al. propose a method that is based on program committee characteristics and make the assumption that the quality of a conference is directly related to the reputation of its program committee members. Thus, given a set of conferences, by mining their program committee characteristics, their method is able to automatically determine the prestigious ones. In another work [18], Yan and Lee consider that in each knowledge 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 doing this, they propose three metrics that basically consider the set of authors a paper has in common with the seeds. The problem here is to find good seeds that are not too restrictive. Finally, in [16], Souto, Warpechowskil, and Oliveira have developed a classification model to support the (semi-)automatic assessment of Computer Science conferences based on ontologies and inference rules.

Our work is inspired in [20]. However, we address a much harder problem since we do not only differentiate between prestigious and non-prestigious conferences but we are able to distinguish high, from medium, from low quality conferences. Moreover, our method does not rely only on aspects of the program committee, but explores a large number of features as criteria to assess the quality of conferences, which are combined by means of machine learning techniques to automatically and effectively classify them. We also provide a deeper discussion on the difficulty of the problem based on a number of different experiments.

3. DATASETS AND EXPERIMENTAL FRAMEWORK

3.1 Datasets

In this section, we describe the three datasets used in our experiments. The first one is a ranking of Computer Science conferences that was built 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 [8]. This ranking is used in our experiments as a "gold standard" in order to evaluate the effectiveness of several feature combinations for the task of assessing the quality of scientific conferences. The second dataset is a collection of citation data crawled from the Libra Academic Search[5]. Finally, the third dataset consists of a bulk of data describing specific information about conferences (e.g., number of editions, list of program committee members, acceptance rate, etc.) manually collected from the Web.

3.1.1 The Perfil-CC Ranking

The Perfil-CC ranking comprises a list of 1,000 conferences divided in three categories (A, B, and C) according to their quality, being A the highest quality category. This list was obtained from several sources after a cleaning pro-

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

cedure was performed to remove duplicates and conferences with less than four editions, with submissions based only on abstracts, and with a regional or national audience. Participated in the poll for assessing these conferences, 312 specialists, all of them faculty members of Computer Science graduate programs in Brazil. Of these 312 specialists, 147 were qualified researchers that hold an individual grant from CNPq (The Brazilian National Research Council, similar to NSF)[6]. To facilitate the assessment process, the conferences in the list were divided into 27 groups, representing a possible division of the Computer Science field [8].

During the poll, each specialist 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). To avoid conferences being overevaluated, 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 specialist 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.

After finishing the poll, a final step took place to turn the votes into a number that would capture the quality of each conference. Several scenarios, considering different criteria to count and weigh the votes, were considered. The differences among these scenarios were not significant, and the one considered the best by some consulted specialists was chosen. 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 3 times the number of valid votes received by each conference. All votes different from NE and NC were considered valid votes. Thus, votes NE and NC were not counted and did not influence the final result. Conferences that had a number of valid votes lower than 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. The final generated ranking can be accessed at `http://www.latin.dcc.ufmg.br/perfilccranking`.

It is worth noticing here that the large number of conferences handled by the Perfil-CC project reinforces the importance of an automatic process for conference assessment as we discuss in this paper, since the execution of a polling process like this is very costly. The whole process that included data collection and processing (standardization and cleaning), creation of the software infrastructure, conference grouping, discussion, the polling itself, and compilation of results, took about six months, with the help of many volunteers. Another important consideration is the large number of recently created conferences, which makes it difficult to keep an updated list of assessed conferences.

We finally note that, in order to further validate the Perfil-CC ranking, we contrasted it with the Australian CORE Ranking of ICT Conferences and confirmed that there is a large agreement between the two [10].

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

Group	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 1: Libra coverage of conferences existing in the selected Perfil-CC groups.

Crawled Conferences	194
Crawled Papers	109,969
Total number of citation (only internal)	145,282
Papers which have citation data (only internal)	27,759
Papers which have any citation	43,749

Table 2: Statistics about the Libra crawled data.

3.1.2 The Libra Dataset

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

The choice of Libra as a source of citation data 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 list, which for some groups reaches more than 80%, as can be seen in Table 1. 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.

For our experiments, we crawled from Libra bibliographic and citation data of papers appearing in 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 had greater coverage with respect to the Perfil-CC conference list. The results of this crawling is summarized in Table 2.

3.1.3 Additional Data

In addition to the bibliographic and citation data taken from Libra, we manually collected from digital libraries, conference sites, and other Web sources a large amount of data about the conferences included in the six Perfil-CC groups listed in Table 2. This data comprises, for each conference,

its periodicity (annual, biennial or triennial), its designation (conference, workshop, symposium, etc.), its number of editions up to 2007, and its submission rate, acceptance rate, and program committee member list for the 2005-2007 triennium editions. This data, together with the citation data taken from Libra, allowed us to exploit several feature groups and analyze their ability to classify conferences according to the three Perfil-CC categories. These feature groups are discussed in Section 4.

3.2 Experimental Framework

Our approach to automatically assess the quality of conferences combines several features in order to effectively classify them. However, the number of possible feature combinations is very large, which makes the search for an optimum solution untreatable even automatically. Thus, in order to address this problem, we apply several machine learning methods. These methods provide effective heuristics to search over the solution space, which allow us to efficiently find near-optimum solutions. In our experiments, we considered the following machine learning methods: Bayesian, decision trees, rule-based, and Support Vector Machines. However, we only present the results obtained using a random forest classifier [6], since this was the method that produced the best results. In addition, we used a technique know as bagging [12] to help improve the results.

To compare the performance of our classifiers, we employed information retrieval metrics [13] that are commonly used in classification problems, which are accuracy, precision, recall, and F-measure (F1). 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 instances correctly classified in a category divided by the correct number of elements of that category. F1 summarizes both recall and precision. We used precision, recall, and F1 values averaged over all categories (Macro-Precision, Macro-Recall, and Macro-F1). We show in Figure 1 a confusion matrix that defines these metrics using only two categories (A and B) to simplify.

		Prediction A	Prediction B
True Label	A	a	b
	B	c	d

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

$$Recall_A = \frac{a}{a+b} \quad — \quad Recall_B = \frac{d}{c+d}$$

$$Precision_A = \frac{a}{a+c} \quad — \quad Precision_B = \frac{d}{b+d}$$

$$F1_X = \frac{2 * (Precision_X * Recall_X)}{Precision_X + Recall_X}$$

Figure 1: Evaluation Metrics.

In our experiments, we adopted a three-fold cross-validation experimental procedure instead of the usual 10-fold one. The reason for that was the small size of our final dataset (194 conferences). In addition, all results are averaged over 33 executions. In some of our experiments the number of conferences considered was even smaller, thus, in these cases, we randomly chose 50% of the conferences for test and 50% for training. The experiments in which we used this division will be explicitly indicated. In all the result tables, we show for the mean value of all the 33 runs and the 95% confidence interval (CI) for the mean.

4. CONFERENCE FEATURES

The final dataset used in our study comprises 21 features that can be divided into four groups, according to the conference aspects they are associated with: citation counting, submission and acceptance rates, tradition, and program committee characteristics. Other more general features, such as the conference periodicity, were also indirectly used. This section analyzes and discusses these different feature groups and their ability to classify conferences in three categories, namely A, B and C (from highest to lowest quality level). We also present cumulative distribution graphs that illustrate the discriminative power of some of these features in our dataset.

4.1 Citation Features

Features based on bibliography citations are the most commonly used to assess the quality of publication venues. Hence, there is a variety of metrics based on citations proposed in the literature. In this paper, we exploit a large subset of them, including the most popular metrics used for journal quality assessment, and some other metrics specially designed for conference quality assessment.

Among the metrics exploited for journal assessment are H-index [5], Citation Count (CC), Average Citation Count (AVGCC), Weighted PageRank (WPR) [4], Y-Factor [4], and Impact Factor (IF) [1]. Each of these metrics emphasize different quality aspects of publication venues. CC, for instance, measures popularity. WPR takes into account citations in high quality venues, valuing prestige, whilst IF measures current popularity.

Regarding the metrics for conference assessment, we proposed four new metrics: Conference Impact Factor (CIF), Conference Citation Impact (CCI), Conference Combined Factor (CCF), and Conference Factor (C-Factor). We briefly explain these metrics here and a more complete description of them can be found in [10].

CIF is a redefinition of IF that uses a temporal window of six years instead of three, increasing the probability of obtaining available data for the conferences being evaluated. CCI represents the ratio of the total number of citations a conference received by the total number of citations that all conferences received in a given period of time. CCI uses the same temporal window as CIF. The CCF of a conference X is defined as $CCF_X = CIF_X + CCI_X + CS_X + CL_X$, where CS (Conference Size) estimates the size of X based on the number of papers presented at it and CL (Conference Longevity) estimates the longevity of X by dividing CC_X by the total number of citations that all conferences received over time (the older it is, the more references it has). As observed, CCF combines different metrics aiming at promoting larger, traditional, and popular conferences, i.e., conferences with high quality indicators. At last, C-Factor takes into account how prestigious a conference is, and is formally defined as $C\text{-}Factor_X = WPR_X \times CCF_X$.

In order to give some insights on how citation-based metrics behave, Figure 2(a) presents the cumulative distribution of the total number of citations (Citation Count - CC) per category. Notice that in the graphs of Figure 2, given a value a in the x-axis, the curves show the fraction of conferences that have some value (e.g., PC Size) not greater than

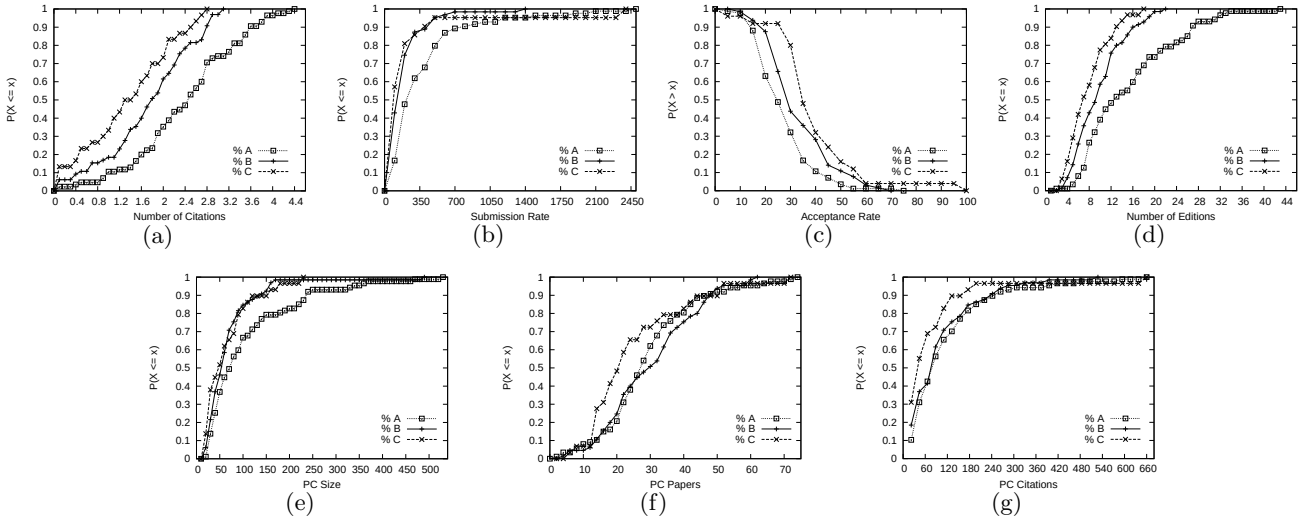

Figure 2: Cumulative distribution of the number of citations (CC) in log_{10} scale (a), submission rate (b), acceptance rate (c), number of editions (d), size of the program committee (e), average number of papers published by PC members (f) and average number of citations received by PC members (g).

a. The exceptions are Figures 2(a) and 2(c); the former is in log_{10} scale and in the latter the interpretation is the opposite: $P(X > x)$. As we can see, the graph in Figure 2(a) shows that conferences classified as A usually receive more citations than conferences classified as B, which in turn receive more citations than conferences classified as C. Also, the curves are very distinct, with no intersections. From this, we conclude that CC is a good discriminator for the three conference categories.

From all the citation-based metrics studied, we chose to show only the cumulative distribution for CC because it is the simplest one. It is also directly or indirectly used to define other citation metrics.

Although citation-based metrics allow us to measure very distinct and important aspects of conference quality, they have some drawbacks: problems such as self-citations, biased citations, and non-citations [9] are still far from being solved, and may greatly impact features based on them.

4.2 Submission/Acceptance Features

The paper submission rate of a conference reflects its popularity and size. High quality conferences are usually larger and attract a great number of submissions (and of participants too). On the other hand, these conferences usually also have a very low acceptance rate. The acceptance rate reflects the difficulty of having a paper accepted for presentation at a conference. In our study, we use the conferences' submission and acceptance rates, averaged over the 2005-2007 triennium, as features for conference quality assessment.

Figure 2(b) shows the cumulative distribution of the submission rate. As observed, the submission rate for conferences from categories B or C is biased towards lower values, whereas for conferences from category A it is biased towards higher values, as expected. However, the distribution curves for conferences from categories as B and C are very similar and intersect in some points, not allowing a clear distinction between them. In addition, it seems that the distinction between these two categories of conferences may be affected by the lack of available data. For instance, in our collected dataset, for 11.3% of the conferences classified as B and 40.0% of the conferences classified as C we have not been able to find this information as opposed to only 4.5% in the case of conferences classified as A. Regardless of that, the submission rate is a good feature to distinguish between categories A and others, but not between B and C.

Figure 2(c) shows the cumulative distribution for the acceptance rate. Note that conferences from category A are usually more biased towards lower values of acceptance rate than conferences from category B, which in turn are also more biased towards lower values of acceptance rate than conferences from category C. Also notice that only 16.6% of the conferences belonging to category A have an acceptance rate greater than 40%, while from categories B and C these figures raise to 35% and 48%, respectively. Furthermore, almost all conferences from category C (92%) have an acceptance rate greater than 30%. These figures show that the acceptance rate might confuse the classifier in certain situations, such as those in which conferences from categories B and C have very low acceptance rates. Apart from that, the acceptance rate can be considered a good candidate feature to help classifying conferences.

4.3 Tradition Features

The number of editions of a conference is directly related to its tradition. Older and more traditional conferences have usually higher quality. However, there are also young conferences of very high quality. Thus, although the number of editions clearly reflects the tradition of a conference, it may not be by itself a good conference quality predictor.

Figure 2(d) shows the cumulative distribution of the number of editions per conference category. In this case, conferences from category A are biased towards a larger number of editions, while conferences from categories B and C are biased towards smaller numbers of editions. Around 40% of the conferences from category A have more than 16 editions.

For categories B and C, only 10% and 16% of the conferences reached this number, respectively. Also note that there is not a big distance between curves B and C, indicating that the number of editions for conferences in these categories is not that different.

4.4 Program Committee Features

Several program committee (PC) characteristics can also be used as features to assess the quality of conferences, assuming that the quality of a conference is directly related to the reputation of its program committee. In [20], for instance, the authors evaluate the quality of conferences based on the size of their program committees, the average number of papers published by the program committee members, and the number of coauthors of each these members. They also used two measures based on network centrality: the closeness and the betweenness centrality [17]. These last two measures are calculated using a coauthor network. Observe that none of the features cited above are based on the citations received by the program committee members, even though this is a common source of information when evaluating researchers.

In this paper, we extend the set of features proposed in [20] and also consider the number of citations, the average number of citations per work, and the H-index [7] of each program committee member. Note that all the features considered in this section, except the PC size, refer to the scientific production of the PC members.

Figure 2(e) shows the cumulative distribution of the size of the program committee. The curve representing conferences A is biased towards bigger program committees than the curves for categories B and C. Here, again, curves for categories B and C are very similar, showing that this feature may not effectively distinguish conferences from these two categories. The average number of papers published by the program committee members is also a weak feature for classifying conferences, as showed in Figure 2(f). Note the intersections between the three curves.

There is also a great similarity between the average number of citations received by the PC members of conferences classified as A and B, as observed in Figure 2(g). In turn, the curve for conferences from category C is biased towards smaller values, showing that the PC members of these conferences are less cited.

4.5 Other Features

Besides the features presented in the previous sections, we use in our study three others that cannot be directly considered quality indicators, but we believe will help in the classification process. The first of these features is the way a conference is designated. We used this information because it usually indicates how large a conference is. In general, workshops are smaller than symposia, that in turn are smaller than actual conferences. Nevertheless, we can easily find some exceptions to this rule. The second feature is the conference periodicity. We want to analyze whether annual, biennial, and triennial conferences present different characteristics. At last, we use information about the group the conference belongs to, i.e., ML, DB, BIO, HCI, NT, and WEB, as considered in the Perfil-CC project (see Table 1). This feature is important because different groups might use different citation patterns and authorial criteria, as well as have a different number of conferences and researchers.

	Top 5	Top 10	Top 15	Top 20	# of Features
Citation	4	8	9	10	10
Submition	0	1	2	2	2
Edition	1	1	1	1	1
Committee	0	0	3	7	8

Table 3: Number of features in the top rank positions according to the information gain.

Group	Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)
Citation	56.03(±1.77)	65.06(±1.76)	67.62(±3.00)	67.62(±3.00)
Sub/Accep	42.31(±1.62)	47.57(±1.44)	56.24(±2.75)	56.24(±2.75)
Tradition	47.52(±1.79)	58.03(±2.45)	62.40(±3.77)	62.40(±2.37)
Prog Comm	47.84(±1.78)	51.27(±1.75)	64.07(±2.98)	56.78(±2.01)
All	56.49(±1.68)	67.12(±1.92)	69.35(±3.23)	67.74(±1.83)

Table 4: Results obtained for each of the feature groups when classifying conferences according to the Perfil-CC Criteria. CI is the 95% confidence interval for the mean.

4.6 Feature Evaluation

Considering all the features described in the previous sections, we created a conference dataset with 21 features. We then applied a popular feature selection method, namely the information gain [19], to evaluate the individual discriminative power of each one of them.

Table 3 shows the most discriminative features according to their information gain. Note that, among the 10 most important features, eight are based on citations. Also notice that the features based on program committee characteristics do not appear between the top 10. These features are the most difficult to obtain, since they require not only the list of program committee members, but also a high quality bibliography data source. The top five features, according to the information gain ranking, are: CCF, C-Factor, CCI, CC, and Number of Editions.

5. EXPERIMENTS

This section presents the results of experiments we performed to (1) assess the quality of the proposed groups of features in discriminating conferences in different categories, (2) identify the main issues that make this problem hard for automatic classification, and (3) study alternatives to lessen the effects of these issues.

5.1 Classifying Conferences According to the Perfil-CC Criteria

Recall that in Section 4 we analyzed four groups of features (citation, submission/acceptance, tradition, and program committee) that might be used for classifying conferences according to their quality. In this section, we study the impact that each of these groups has in the automatic conference quality classification process. Table 4 shows the results obtained when classifying the 194 conferences of our dataset into three categories (A, B and C), according to the Perfil-CC criteria. As observed, the best results were found when using the features based on citation counting, which obtained an accuracy of 56.03% and F1 of 67.62%. However, in general, the results presented in Table 4 may be considered not satisfactory. This is because only the classifier using the citation features presented an accuracy above 50%. This could be a result of using each feature group independently.

Hence, we combined the four groups of features (plus the

Citation		Prediction A	Prediction B	Prediction C
True Label	A	**62.50%**	32.95%	4.55%
	B	40.85%	**49.30%**	9.85%
	C	28.57%	48.57%	**22.86%**

Prog Comm		Prediction A	Prediction B	Prediction C
True Label	A	**62.50%**	35.23%	2.27%
	B	47.89%	**38.03%**	14.08%
	C	48.57%	40.00%	**11.43%**

Tradition		Prediction A	Prediction B	Prediction C
True Label	A	**62.50%**	32.96%	4.54%
	B	46.48%	**45.07%**	8.45%
	C	45.71%	45.71%	**8.58%**

Sub/Accep		Prediction A	Prediction B	Prediction C
True Label	A	**51.14%**	29.54%	19.32%
	B	49.30%	**42.25%**	8.45%
	C	51.43%	31.42%	**17.15%**

All		Prediction A	Prediction B	Prediction C
True Label	A	**60.23%**	36.36%	3.41%
	B	33.80%	**56.34%**	9.86%
	C	25.71%	51.43%	**22.86%**

Figure 3: Confusion matrices obtained by each of the feature groups when classifying conferences according to the Perfil-CC.

way the conference is designated, its periodicity, and the group it belongs to), and the results are reported in the last line of Table 4. The accuracy of 56.49% and F1 of 67.74% are statistically the same as the ones obtained when using only the citation features.

Figure 3 shows the confusion matrix for all the groups of features. Note that the majority of the classifiers correctly classifies conferences in category A. These are usually traditional conferences, with feature values very different from the ones in categories B and C (as showed in Figure 2). In contrast, many conferences from category C are classified as B, showing that there is some similarity among them.

A significative number of conferences from category B are predicted as being from category A. These are the conferences that are very close to the threshold that divides the conferences in the Perfil-CC ranking into the three categories. This problem will be discussed in detail in Section 5.1.2.

The fact that combining all the feature groups did not lead to statistically better results (when compared to the results obtained by the citation features) was intriguing. Hence, we performed a detailed analysis of the dataset, and identified three problems that justify the results obtained and show the difficulties in classifying the conferences in categories A, B and C. The first of these problems is the amount of missing data, since we have not been able to find all required data for some conferences. This problem is studied in more detail in Section 5.1.1.

The other two problems are related to our "gold standard", the Perfil-CC ranking. The first of them is related to the subjective aspects of the Perfil-CC polling process. For instance, when classifying conferences, specialists may have been biased towards the venues they have published in or cited frequently. Moreover, people from different fields may use different criteria to classify a conference as belonging to category A, B or C. We will study the impact of the latter by running different classifiers for conferences from different groups in Section 5.1.3.

The last problem refers to the thresholds used to divide the conferences into three categories. In order to divide the Perfil-CC ranking into three categories, thresholds were set. As a consequence, when comparing conferences at the "top" positions of each category, we can certainly say they belong to different categories. However, as we go down the ranking and get closer to the next category (i.e., closer to the threshold line), the features of certain conferences are equivalent

	Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)
194 Conf	56.49(±1.68)	67.12(±1.92)	69.35(±3.23)	67.74(±1.83)
156 Conf No Miss	62.88(±1.89)	69.55(±1.90)	77.53(±2.71)	73.08(±1.78)
143 Conf No Out	67.90(±1.41)	74.96(±1.92)	83.53(±2.03)	78.81(±1.50)

Table 5: Results obtained by the classifier with the original dataset (194 Conf), with no missing data (156 Conf No Miss) and with no missing data or outliers (146 Conf No Out).

to those at the top positions of this category. These are the conferences that mislead the classifier.

We illustrate this problem with an example, the *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. According to the Perfil-CC ranking, this conference belongs to category A (but it is at a bottom position of this category, close to the top positions of category B). Looking at its features, only the tradition of the conference (20 editions) is comparable to those of conferences of high quality. The other features correspond to those of conferences from category B. Hence, for the classification model, this conference will always be classified as B, although its category is A according to the Perfil-CC ranking, probably because, when assessing this conference, the specialists considered it important for their community, since it is an IEEE sponsored conference. Notice that our classifier does not consider this kind of information. We analyze the impact of the category thresholds in Section 5.1.2.

5.1.1 Removing Conferences with Missing Data and Outliers

Missing data was previously pointed out by our analysis as one of difficulties found by the classifier, leading to lower accuracy rates. For the experiments discussed in this section, we removed from the initial dataset conferences with missing data for at least one group of features. The new dataset, with 156 conferences, was used to train the classifier *with all the features*. Table 5 shows the new results obtained, where accuracy and F1 reached 62.88% and 73.08%, respectively. Figure 4 shows the confusion matrix for this experiment. Note that, in this case, none of the conferences belonging to category A were wrongly classified as C. This is evidence that, in the previous experiment, missing data was deceiving the classifier to wrongly classify examples from category A to C. At the same time, now only 5.89% of the examples in category C are correctly classified. This is due to the great similarity between examples in categories B and C.

In a second experiment, besides the conferences with missing data, we also removed from the dataset conferences we considered as outliers. Outliers are conferences that did not receive appropriated evaluations from the specialists. We identify these because their characteristics are very different, in average, from those of other conferences in the same group. There are many reasons for that, such as having low impact to the communities the specialists belonged to or addressing very specific subjects. Examples of such conferences are the *International Conference on Genetic Algorithms (ICGA)* and the *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. In total, 13 conferences were manually identified as outliers and removed from the dataset. The new dataset, with 143 conferences, presented statistically better results than the original one, with an accuracy of 67.90% and F1 of 78.81% (see Table 5). These results show that the 13

All		Prediction A	B	C
True	A	**79.27%**	20.73%	0.00%
Label	B	36.84%	**59.65%**	3.51%
	C	35.29%	58.82%	**5.89%**

Figure 4: Confusion matrix for experiments with no missing data.

		Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)
1	A and B	69.13(±1.68)	73.30(±1.92)	70.34(±2.61)	71.49(±1.68)
2	A and B (threshold A)	75.30(±1.37)	78.30(±0.87)	92.05(±1.68)	84.54(±0.93)
3	A and B (threshold B)	76.98(±1.50)	78.63(±2.79)	58.59(±3.93)	66.32(±2.85)
4	B and C	65.44(±1.95)	69.92(±1.27)	85.34(±3.08)	76.62(±1.59)
5	B and C (threshold B)	80.77(±0.71)	81.93(±0.42)	98.26(±0.85)	89.33(±0.44)
6	B and C (threshold C)	59.52(±2.25)	63.03(±1.78)	69.14(±3.73)	65.61(±2.31)
7	A and C	74.43(±1.83)	79.06(±1.47)	87.94(±2.49)	83.03(±1.32)

Table 6: Results obtained when classifying the conferences into two categories with the original and modified category thresholds.

removed conferences were certainly leading the classifier to learn wrong patterns about the data, and then wrongly classifying new conferences.

5.1.2 Analyzing the Category Thresholds

Another problem that presented difficulties to the classifier was related to the thresholds used to divide the conferences in the Perfil-CC ranking into three categories. In order to better understand this problem, this section describes three experiments involving the original dataset with 194 conferences. We used the original dataset in order to isolate the factors we want to analyze. Experiments considering the identified problems simultaneously are left for future work. First, the original classification problem is subdivided into smaller problems, and we test the classifier when working with only two of the three original categories: A and B, B and C, and A and C. After, we vary the original threshold and evaluate the impact of this change in the results. At last, we turn back to the 3-category problem, and evaluate the effects of changing the original thresholds that divide categories A and B, and B and C, simultaneously.

Table 6 summarizes the results. As expected, when discriminating between two categories with the original threshold (lines 1, 4 and 7), the worst results were obtained by the classifier when separating categories B and C. This emphasizes the similarities of the conferences in these categories. The better results are the ones obtained when classifying conferences into categories A and C. This confirms the results presented in Section 4, where the studied features showed different characteristics for these two categories.

In the second experiment, we vary the values of the original thresholds chosen to separate categories A from B, and B from C. First, a "gray zone" was created. It included all the conferences that had scores in the Perfil-CC ranking 10% greater or smaller than the original score of the first conference in the lower quality category. Then, two configurations were tested. In the first one, all conferences in the gray zone are considered as conferences belonging to the higher quality category (lines 2 and 5). In the second, they are included in the lower quality category (lines 3 and 6). The results obtained are presented in Table 6.

Regarding the experiments varying the thresholds for categories A and B, the values of accuracy and precision in both experiments are statistically better than the ones obtained with the original threshold. However, the values of recall and, consequently, F1, are significantly higher when the conferences in the "gray zone" are considered as being from category B. These numbers gives us an indication that conferences in the top raking positions of category B are really similar to those from category A.

	Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)
Threshold A	62.27(±1.46)	69.24(±1.34)	82.91(±2.25)	75.28(±1.25)
Threshold B	70.20(±1.34)	76.09(±3.49)	56.95(±4.04)	64.13(±2.91)

Table 7: Varying the thresholds for conferences from categories A, B and C simultaneously.

For the conferences in the "gray zone" of categories B and C, when we considered all of them as belonging to category B (line 5), we found results that are statistically better than the ones obtained with the original threshold. This shows that the reduced set of examples in category C has now features that can be easily used to distinguish them from examples in category B. In contrast, when considering the conferences in the "gray zone" as being from category C (line 6), the classification problem becomes more difficult, and worse results are obtained. These experiments stress once again the similarities between conferences from categories B and C.

At last, we vary both thresholds between categories A and B, and B and C, simultaneously. In the case of the B/C threshold, the previous experiments show that it is better to increase the threshold in 10%, including old examples from category C in category B. Nevertheless, for the thresholds between categories A and B, we believe a more detailed analysis is required. Hence, in the next experiments we kept the new B/C threshold fixed as discussed above and propose to evaluate the A/B threshold following the same approach as before. Table 7 reports the results obtained when considering the conferences in the "gray zone" as belonging to category A, and then to category B. Note the results obtained for accuracy and F1. When we increase the number of conferences in category A, we have a more balanced classification, with a higher F1 and lower accuracy. In turn, when we increase the number of conferences in category B, we have the opposite situation, a higher value of accuracy and lower values of recall and F1. This emphasizes the need for a better individual analysis for conferences in the "gray zone" of categories A and B.

5.1.3 Classifying the Conferences by Group

Another difficulty found by the classifier when separating conferences into different quality levels is related to the specialists' subjectivity in their manual classifications. For instance, researchers working on different groups have very different profiles. Older and larger research fields may have different citation patterns, authorial criteria, number of conferences, etc. These differences may have a big impact upon the Perfil-CC criteria. Hence, in this section we describe some experiments we conducted to evaluate separately conferences from different groups. Although this information was present in the dataset, it might not have been properly exploited by the machine learning algorithms.

The results are presented in Table 8. As the number of conferences in some groups is very small, we used 50% of the examples for training and 50% for test in all experiments with groups. Notice that the best results are the ones for the HCI and DB groups, with accuracies of 67.31% and 65.74%, and F1 of 77.22% and 70.93%, respectively. These

	Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)	# Conf
BD	65.74(±2.89)	79.65(±5.30)	67.64(±5.86)	70.93(±3.93)	40
BIO	43.64(±4.96)	53.84(±7.19)	67.68(±10.44)	57.58(±7.23)	10
HCI	67.31(±3.37)	82.36(±5.07)	76.46(±5.44)	77.22(±3.20)	24
ML	43.52(±3.30)	54.36(±3.96)	62.61(±5.47)	56.97(±3.35)	32
NT	51.78(±2.05)	57.38(±1.81)	81.12(±4.03)	66.72(±1.79)	55
WEB	49.21(±3.29)	58.40(±5.38)	59.31(±6.91)	56.28(±4.67)	33
OFFG	55.55(±1.61)	68.78(±2.19)	67.91(±3.02)	67.91(±1.90)	-
ONG	56.49(±1.68)	67.12(±1.92)	69.35(±3.23)	67.74(±1.83)	-

Table 8: Results obtained when classifying conferences according to their groups. ONG and OFFG are results for all conferences including and excluding the group feature from the dataset.

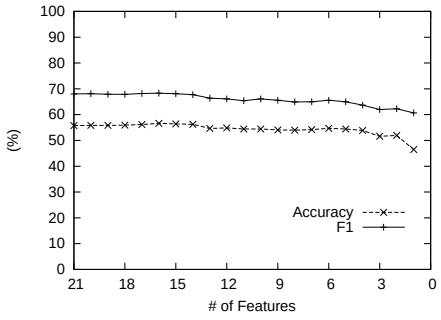

Figure 5: Accuracy and F1 obtained when reducing the number of features.

two groups are very homogeneous and the conference categories well defined. Moreover, these groups are the ones with more complete information about citations in Libra. As previously showed, citation features are among the most important ones for classification.

On the other hand, the worst results are the ones obtained by the ML and BIO groups, with accuracies of 43.52% and 43.64%, and F1 of 57.58% and 56.97, respectively. Notice that the BIO group has only 10 conferences and a small number of specialists from that community participated in the poll. Hence, they might not have been accurately evaluated. Concerning the ML group, a manual analysis showed that a significative number of high quality conferences, e.g., the *IEEE Congress on Evolutionary Computation (CEC)* and the *International Conference on Hybrid Intelligent Systems (HIS)*, does not have enough distinct feature values to accurately separate them between medium and low quality conferences, making the classification process harder. Notice that these results are similar to those obtained when we did not separate the conferences in groups, except for the HCI and DB groups for which the results are better.

At last, we executed one final experiment in which the group information was excluded from the original dataset. As showed in Table 8, the results (OFFG) are equivalent to those obtained when the group is considered (ONG).

In conclusion, all the results reported in this section support the idea that the group information is not essential for the classification process.

5.1.4 Reducing the Number of Features

The analyses carried out in Sections 4 and 5.1 show the challenges of classifying conferences in three quality levels. In special, we show that using all feature groups or only citation information in the classification process leads to similar results. This section analyzes the impact of removing features from the original dataset. Recall that in Section 4.6 we ranked the features according to their information gain. Here we reuse this information to incrementally remove the less discriminative features. Figure 5 presents the results of accuracy and F1. Note that reducing the number of features up to 6 does not have any big impact on the results. Only after that the values of accuracy and F1 degrade.

	Accuracy (CI)	Precision (CI)	Recall (CI)	F1 (CI)
A and B+C	72.30(±1.43)	76.73(±2.71)	57.11(±3.02)	64.89(±2.18)
A and Thr. A+B+C	79.88(±1.33)	80.50(±3.30)	51.23(±3.91)	61.72(±3.20)
Top 20	94.55(±0.64)	87.72(±4.55)	56.04(±5.24)	66.98(±4.61)
Top 30	90.83(±0.84)	87.31(±4.42)	48.48(±5.05)	60.91(±4.62)

Table 9: Results obtained when classifying conferences in categories HQ and OT.

A and B+C		Prediction A	Prediction B
True Label	A	**54.55%**	45.45%
	B	13.21%	**86.79%**

Top 20		Prediction A	Prediction B
True Label	A	**45.00%**	55.00%
	B	1.15%	**98.85%**

A and Thr. A+B+C		Prediction A	Prediction B
True Label	A	**58.73%**	41.27%
	B	6.11%	**93.89%**

Top 30		Prediction A	Prediction B
True Label	A	**50.00%**	50.00%
	B	1.22%	**98.78%**

Figure 6: Confusion Matrices for experiments with only two categories.

5.2 Classifying Conferences in Two Categories

Given the difficulties and challenges found when classifying conferences according to the three Perfil-CC categories, this section aims to simplify the problem. It transforms the previous 3-category problem into a 2-category problem: high quality (HQ) and others (OT). We considered different scenarios to relabel the three categories into two. First, conferences from categories B and C were merged to form category OT, category A becoming the HQ category. In a second approach, conferences at the bottom positions of category A (previously included in the "gray zone") were put together with the conferences from categories B and C in the category OT, the remaining conferences from category A becoming category HQ. At last, only conferences at the Top 20 and Top 30 positions of the Perfil-CC ranking were included in category HQ, all others being included in category OT.

The results of the different scenarios proposed above are summarized in Table 9. Observe that all accuracies are statistically better than the ones obtained with three categories, but in this case the accuracy is not the best measure to evaluate the performance of the classifier (since the category distribution for the examples is very unbalanced). The values of recall, in turn, are statistically lower. The confusion matrices for these experiments are showed in Figure 6. Note that the difficulty now is to classify conferences in category HQ, as it has fewer examples. Also, in some cases, conferences of the old category A are now in category OT and can create some confusion.

The first attempt to improve these results was to create a cost matrix, where wrongly classified examples from category HQ are associated with a higher cost. This approach forces the classifier to perform better in categories with fewer examples, obtaining a more balanced classification. In this approach, conferences correctly classified were assigned cost 0, and conferences from category OT incorrectly classified were assigned cost 1. For conferences from category HQ in-

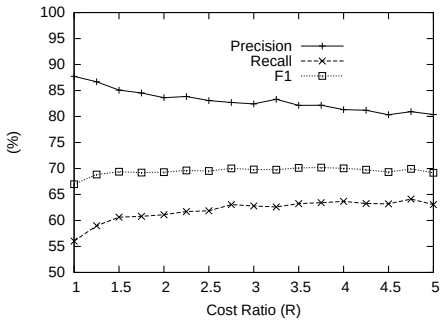

Figure 7: Precision, Recall, and F1 obtained varying *R*.

correctly classified as OT, a cost ratio (R) varying from 1 to 5 (0% to 500% greater than the classification error of conferences in category OT) was tested. Note that when $R = 1$, no cost matrix is used.

Figure 7 shows the results of using the cost matrix when only the top 20 conferences in the Perfil-CC ranking are considered in category HQ. Notice that when R is set to 3.75 the F1 is equal to 70.23%. When R is 1, the value of F1 is 66.98%. Nevertheless, the most interesting results are observed analyzing the precision and recall curves. As we increase the values of R, the values of recall also increase, as the values of precision decrease. Higher values of recall mean more balanced classification of examples in both categories. This is reflected in the value of F1, which measures precisely the relation between precision and recall.

6. CONCLUSIONS

In this paper, we tackled the problem of classifying conferences into pre-defined levels of quality by using a significant number of features about conferences collected from several Web sources. These features try to cover diverse aspects that are taken into consideration by specialists when manually judging the quality of a conference. When combining these by means of machine learning techniques, we obtain relative success on separating high quality conferences from medium and low quality ones. However taking these last two categories of conferences apart proved to be a much harder task. Classification with only two categories (high and low quality) was also shown to be more accurate.

Our approach is different from most of the related work in the sense that it does not focus in one specific type of feature but instead in their effective combination, thus capturing a number of important aspects that are in fact used in practice for this type of evaluation. In doing so, we also performed a characterization of the discriminative power of several types of feature concluding that the most effective ones are those based on citations followed by information about the tradition of the conference.

As future work, we first intend to invest in ways for automatically extracting the most important features for classification, as the lack of information proved to be an important aspect influencing in the performance of the methods. Also, we intend to investigate different ways of combining the features, e.g., through regression, if we consider this problem as one of ranking conferences. Besides, we also intend to increase the number and the diversity of the features used, considering, for example, sponsors and the quality of the papers presented at the conferences.

7. ACKNOWLEDGEMENTS

We thank members of our research group for helping us collect the initial raw data on which part of our study is based. 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 (MCT/CNPq Grant Number 55.0874/2007-0), and by the authors' individual grants from CNPq and FAPEMIG.

8. REFERENCES

[1] M. Amin and M. Mabe. Impact factors: Use and abuse. *Perspectives in Publishing*, 1:1–6, 2000.

[2] J. Bollen, H. V. de Sompel, and M. A. Rodriguez. Towards usage-based impact metrics: first results from the MESUR project. In *Proc. JCDL*, pages 231–240, 2008.

[3] J. Bollen, H. V. de Sompel, J. A. Smith, and R. Luce. Toward alternative metrics of journal impact: a comparison of download and citation data. *IP&M*, 41(6):1419–1440, 2005.

[4] J. Bollen, M. A. Rodriguez, and H. V. de Sompel. Journal Status. *Scientometrics*, 69(3):669–687, 2006.

[5] T. Braun, W. Glänzel, and A. Schubert. A hirsch-type index for journals. *Scientometrics*, 69(1):169–173, 2006.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. of the National Academy of Sciences*, 102:16569–16572, 2005.

[8] A. H. F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. S. Silva, and N. Ziviani. Assessing the Research and Education Quality of the Top Brazilian Computer Graduate Programs. *ACM SIGCSE Bulletin*, 40(2), 2008.

[9] M. H. MacRoberts and B. R. MacRoberts. Problems of citation analyses: Critical Review. *JASIST*, 40(5):342–349, 1989.

[10] W. S. Martins, M. A. Gonçalves, A. H. F. Laender, and N. Ziviani. Assessing the Quality of Scientific Conferences Based on Bibliographic Citations. *Scientometrics*, (to be published), 2009.

[11] D. A. Patterson. The Health of Research Conferences and the Dearth of Big Idea Papers. *CACM*, 47(1):23–24, 2004.

[12] J. R. Quinlan. Bagging, Boosting, and C4.5. In *Proc. AAAI*, volume 1, pages 725–730, 1996.

[13] B. Ribeiro-Neto and R. Baeza-Yates. *Modern Information Retrieval*. Addison Weslay Longman Publishing Co., 1999.

[14] S. Saha, S. Saint, and D. A. Christakis. Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 1(62):42–46, 2003.

[15] P. O. Seglen. Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079):498–502, 1997.

[16] M. A. M. Souto, M. Warpechowski, and J. P. M. de Oliveira. An Ontological Approach for the Quality Assessment of Computer Science Conferences. *Proc. IQIS Workshop*, LNCC 4802:202–212, 2007.

[17] S. Wasserman and K. Faust. Social Networks Analysis: Methods and Application. 1994.

[18] S. Yan and D. Lee. Toward alternative measures for ranking venues: a case of database research community. In *Proc. JCDL*, pages 235–244, 2007.

[19] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. ICML*, pages 412–420, 1997.

[20] Z. Zhuang, E. Elmacioglu, D. Lee, and C. L. Giles. Measuring conference quality by mining program committee characteristics. In *Proc. JCDL*, pages 225–234, 2007.