# Measuring Conference Quality by Mining Program Committee Characteristics

Ziming Zhuang
zzhuang@ist.psu.edu

Ergin Elmacioglu
ergin@psu.edu

Dongwon Lee
dongwon@psu.edu

C. Lee Giles
giles@ist.psu.edu

The Pennsylvania State University
University Park, PA 16802, USA

## ABSTRACT

Bibliometrics are important measures for venue quality in digital libraries. Impacts of venues are usually the major consideration for subscription decision-making, and for ranking and recommending high-quality venues and documents. For digital libraries in the Computer Science literature domain, conferences play a major role as an important publication and dissemination outlet. However, with a recent profusion of conferences and rapidly expanding fields, it is increasingly challenging for researchers and librarians to assess the quality of conferences. We propose a set of novel heuristics to automatically discover prestigious (and low-quality) conferences by mining the characteristics of Program Committee members. We examine the proposed cues both in isolation and combination under a classification scheme. Evaluation on a collection of 2,979 conferences and 16,147 PC members shows that our heuristics, when combined, correctly classify about 92% of the conferences, with a low false positive rate of 0.035 and a recall of more than 73% for identifying reputable conferences. Furthermore, we demonstrate empirically that our heuristics can also effectively detect a set of low-quality conferences, with a false positive rate of merely 0.002. We also report our experience of detecting two previously unknown low-quality conferences. Finally, we apply the proposed techniques to the entire quality spectrum by ranking conferences in the collection.

## Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.4 [**Information Systems Applications**]: Miscellaneous; I.7.4 [**Document and Text processing**]: Electronic Publishing

## General Terms

Algorithms, Design, Experimentation, Measurement

*JCDL'07,* June 18–23, 2007, Vancouver, British Columbia, Canada.
Copyright 2007 ACM 978-1-59593-644-8/07/0006 ...$5.00.

## Keywords

Bibliometrics, program committee, social network analysis, data mining, ranking, impact factor, call for papers

## 1. INTRODUCTION

The potential of digital libraries is not only in making a large collection of knowledge easily accessible, but also in providing users with effective recommendation and filtering tools. When a user searches through literature or a librarian makes a subscription decision, bibliometrics are used to measure the quality and impact of venues and documents.

We tackle the scenario in which digital libraries in the Computer Science domain, such as the ACM Portal[1] and CiteSeer digital library[2], need to automatically measure the quality of academic conferences. This is a non-trivial problem for two reasons. First, the Computer Science discipline is unique in its publication practice: unlike almost every other field, peer-reviewed conferences play a role equally (if not more) important to that of the established journals. This is because the discipline's fast-moving pace of progress requires that new research findings to be distributed more quickly and on a broader scale. With competitive acceptance rates of 10-20% and often receiving more citations than journals [22], prestigious refereed conferences are one of the premium publishing venues for researchers in Computer Science.

Second, with the rapid growth of the Computer Science discipline, the number of conferences has also increased dramatically in recent years, which is evident in our collected data from DBWorld[3] (see Figure 1). Often confronted with an abundance of available venues, it is becoming more and more important for researchers and librarians to be discerning about the reputation (thus the quality) of the conferences. The problem is to automatically spot the reputable (or low-quality) ones among hundreds of *Call for Papers* (CFPs) announced each year. Formally, the problem can be defined as follows:

**Definition 1 (Conference Quality Measure)** *Given a set of conference CFPs $X$, where $x \in X$ contains multi-attribute information such as $\{$conference title, date, location, themes, topics, program committee, sponsors, $\ldots\}$, identify a set of*

[1] http://portal.acm.org/dl.cfm
[2] http://citeseer.ist.psu.edu/
[3] http://www.cs.wisc.edu/dbworld/

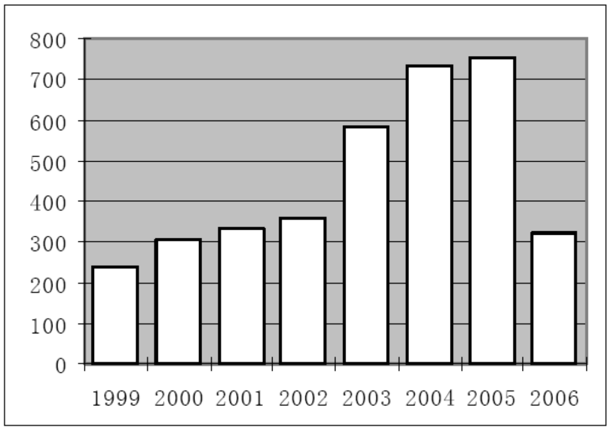


**Figure 1: Number of distinct conference Call for Papers announced on DBWorld in each year from 1999 until May 2006. The number of conferences is steadily increasing over time.**

*reputable or low-quality conferences $P$ ($\subseteq X$), such that $p$ ($\in P$) satisfies some constraints (to be given).* □

Existing techniques to measure the reputation and quality of venues are to associate conferences with certain bibliometrics, most notably citation-based metrics such as the Impact Factor [10]. That is, the quality of a venue is in direct proportion to its bibliometrics characteristics. For example, if the number of citations to the publications in a conference has passed a certain threshold, the conference is considered of good quality (to be elaborated in Section 2). However, in this paper, we claim that such techniques are inadequate to measure the quality of conferences in Computer Science:

1. For emergent or young conferences, historical citation statistics are not readily available, rendering citation-based metrics inapplicable.
2. Even for well-established conferences, citation statistics takes time to accumulate. A recent study of the major database conferences and journals between 1994 and 2003 shows that many of the citations reach back five and more years [22].
3. More generally, when a researcher looks at the CFP or browses the website of a conference, unless it is a well-known venue in his/her familiar domain, the researcher is not likely to be aware of its citation statistics, thus unable to judge its quality.

When we read a CFP and try to decide whether it is a reputable venue worth publishing in, most of us examine the list of Program Committee (PC) members and make our reasonable judgment. That is, we claim the following:

**Hypothesis 1.** *The quality of a conference is closely correlated with that of its PC members.* ■

This hypothesis, together with the three issues aforementioned, has inspired us to study the problem from a novel perspective – to evaluate the quality of conferences through analyzing the characteristics of their PC members.

In this paper, we explore an array of heuristics for mining the correlation between characteristics of the PC members and quality of the conferences. Given a large collection of conference CFPs, we first employ entity extraction techniques to recognize and extract the names of the PC members. By mining their characteristics (details to follow in Section 4), it allows the automatic identification of quality venues even when the publication citation statistics are scarce or unavailable, thus the aforementioned issues can be resolved. In particular, our paper makes the following contributions:

- We fill in the gap in current bibliometrics research, proposing a set of novel heuristics to measure the quality of conferences by mining the characteristics of the PC members.
- We discuss how to combine such heuristics under a classification scheme of which the performance is very promising. Using the same approach a handful of low-quality conferences are also effectively detected.
- Using real CFP data that we have gathered from the Web, our claims are validated empirically.
- We further demonstrate that our techniques can be applied for ranking and recommending conferences, and are able to discover emergent venues of good quality.

For digital libraries, there are several directions toward which this work can be applied. First of all, we provide a novel method of estimating the impact of conferences, and it can be fully automated. This method can filter through the CFPs of emerging conferences, spotting possibly prestigious ones to recommend to researchers and librarians. Second, the proposed heuristics can be applied to approximate the existing citation-based conference impact factors, especially when the citation records of this conference are scarce or inaccessible.

The remainder of our paper is organized as follows. In Section 2, we briefly survey the related work and put our paper into context. In Section 3, we describe our experimental framework and the real-world data sets which we have collected. In Section 4, we propose five heuristics to identify reputable conferences through PC characteristics analysis, and investigate each of them individually. In Section 5, we combine these heuristics in a classification scheme and examine how well they work in aggregation. We report the performance of our classification algorithm in detecting a set of low-quality conferences in Section 6. In Section 7, we extend the proposed heuristics for conference ranking. In Section 8, we discuss a number of implications based on the experimental results. In Section 9 we conclude our paper with directions for future work.

## 2. RELATED WORK

Measuring the quality of publication venues is an important task in bibliometrics. The most widely adopted method to this task is to use Garfield's Impact Factor (IF) [10]: the average number of times the published papers are cited up to two years after publication. Since the introduction of the IF, it has been heavily criticized primarily for its sole dependency on citation counts [23], and therefore many alternatives, e.g., H-index [11], PageRank-like measure [3], and download-based measures [4], have been proposed to rank computer and information science journals [13, 17]. Several citation-based metrics have been proposed for ranking documents retrieved from a digital library [15], and to measure the quality of a small set of conferences and journals in the database field [22]. A recent study [16] introduces topic modeling to further complement the citation-based bibliometric indicators, producing more fine-grained impact measures.

Recently, several studies [18, 1, 6] have analyzed the scientific collaboration networks in different disciplines. To the best of our knowledge, ours is the first study that utilizes several social network analysis metrics with a focus on the PC members of the Computer Science conferences.

What we describe in this paper is not an improvement or alternative of the IF. Rather, we present a data mining based technique that investigates the characteristics of the PC members listed on the CFP, which can quickly determine if a given conference is a reputable one or not. Unlike IF-like measures, our proposal does not require access to the citation records of the conference proceedings but instead analyzes information available on the CFP. By training and evaluating a classifier using real CFPs of Computer Science conferences, we demonstrate that the PC characteristics can be used as a quick indicator the quality of the conferences.

Our paper is inspired by the work [19] in which the authors built classifiers to detect spam web pages. However, our problem is arguably more difficult than theirs: spam web pages are relatively easier to judge, while reputable or questionable conferences are sometimes hard to be differentiated. Therefore, the main hypothesis of our paper, the quality of the PC members is correlated with that of the conference, plays a major role.

# 3. DATA COLLECTION

We used two datasets in the design and evaluation of our algorithm. First, we used the citation data from the ACM Guide covering a 54-year range from 1950 to 2004, which contained the metadata about 609,000 authors and 770,000 articles. The ACM Guide is a high-quality citation digital library that has a good coverage on the computing literature. We used the data to construct a collaboration graph [18], in which nodes represent authors and edges between any two nodes represent coauthorship (i.e., two authors have coauthored one or more papers). All edges in the graph are unweighted, that is, all have the equal importance and only signify whether a collaboration exists between two authors. Repeated collaborations between two authors can also be captured into the graph by weighting each edge with a value proportional to the number of publications two authors have coauthored. Such a graph with weighted edges may give more hints about the collaboration strengths among the authors, which we plan to investigate in future implementations. The ACM Guide has generated about 1.2 million edges in our collaboration graph. Note that ACM Guide itself does not have a notion of "unique key" such as Digital Object Identifier (DOI). Instead, it depends on the names of authors to distinguish them. Therefore, the classical name authority control problem [12] may arise (i.e., same author with various spellings or different authors with the same spelling). We carefully tried to minimize the impact of this problem and in experiments always used the full names whenever possible (e.g. "Dongwon Lee" instead of "D. Lee").

| Topic | Conferences |
|---|---|
| Database | SIGMOD, VLDB, ICDE, EDBT, ... |
| AI | AAAI, IJCAI, ICML, NIPS, KDD, ... |
| Application | WWW, ICCV, ACL, ... |
| System | HPCA, CCS, ASPLOS, DAC, ... |
| Theory | STOC, SIAM, FOCS, LICS, SCG, ... |

**Table 1: Examples of reputable conferences $R$.**

We collected 2,979 unique CFPs between February and May 2006 from DBWorld, a comprehensive and frequently-updated list of events in Computer Science (with the focus on database-related topics). Note that throughout the rest of the paper, we shall use the term "*conferences*" to represent conferences, workshops, and symposiums, as our proposal would work regardless of the type of event, as long as CFPs are given. We then extracted 16,147 what we believed to be distinct PC members from the CFPs with the same disambiguation techniques applied on the ACM dataset. Using one's first and last names as the mapping key, these PC members were matched with the ACM dataset. About 75.63% of the PC members had a 1:1 mapping to the ACM dataset.

Next, based on the conference impact factor ranking from CS Conference Ranking.org[4], we extracted the top 20 ranked conferences in each sub-field of Computer Science, and obtained their authoritative full names from DBLP[5]. Then we extracted from the DBWorld dataset 576 CFPs that approximately match the names of these top conferences. The resulting 576 CFPs were labeled as $R$, which formed a representative training set for the CPFs of the *reputable* conferences (see Table 1). The rest were labeled as $C$, which contained 2,403 CFPs and was disjoint from R. At the end, $R$ consisted of about 19.34% of all the CFP data collected, a reasonable sample of the top 20 ranked conferences.

# 4. IDENTIFY REPUTABLE CONFERENCES

In our paper [7], we examined a few techniques to analyze the conference Program Committee members for the purpose of detecting low-quality conferences. In this paper, we explore a much larger heuristic space and focus on a different end of the quality spectrum. The remainder of this section discusses the proposed heuristics in detail, and investigates the effectiveness of each heuristic as shown in evaluations performed on these two datasets $R$ and $C$.

## 4.1 Number of PC members.

As a hypothesis let us assume that a good conference reaches a large audience, receives a significant amount of submissions due to its popularity, and as a result requires a large program committee to facilitate the rigorous review process.

In our first experiment, we investigated whether the number of PC members has a strong correlation with the quality of a conference. We plotted the distribution of the number of PC members in our dataset in Figure 2. This figure (and all the other figures in Section 4) consists of two sub-figures

[4] http://www.cs-conference-ranking.org/

[5] http://www.informatik.uni-trier.de/∼ley/db/

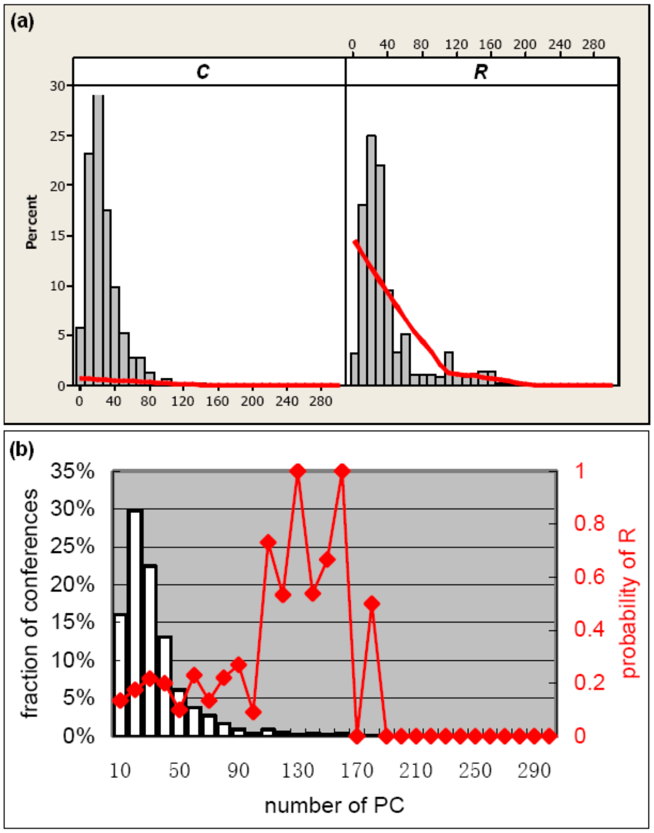


**Figure 2: (a) Conferences labeled as *R* tend to have more PC members, which coincides with (b), showing the prevalence of *R* is higher among conferences with more PC members.**

(style adopted from [19]). The sub-figure on top Figure 2 (a) is a histogram showing the differences in the frequency distribution of the two types of instances, $C$ and $R$. The sub-figure at the bottom Figure 2 (b) consists of a bar chart and a line chart. The X-axis represents a heuristic under investigation (currently showing *the number of PC members*). The Y-axis on the left, which applies to the bar chart, depicts the overall percentage of the 2,979 conferences that fall into a particular range of the current heuristic. The Y-axis on the right, which applies to the line chart, depicts the probability of the conferences within that particular range that are labeled as $R$ (i.e. judged as reputable conferences). Here the probability is calculated as the percentage of instances labeled as $R$.

Overall, Figure 2 shows that the number of PC members tends to be larger in $R$ than in $C$: the mean of $R$ is 37.28 compared with 26.83 in $C$. Although this heuristic exhibits a clear correlation, the chart becomes noisy and shows a number of spikes toward the right, most possibly due to the small number of data points within range.

## 4.2 Average number of publications by PC.

It is usually true that a reputable conference has a program committee of renowned researchers. Because one's publication record is an important indicator for the quality of his/her research, we studied in the second experiment whether the average number of publications of the PC members (data collected from the ACM dataset) is a good indicator for the quality of the conference. The results are shown in Figure 3. Although PC members of the conferences in $R$ appear to be more prolific, the difference is not highly significant between $R$ and $C$: mean of $R$ is 16.94 and $C$ is 13.44. The prevalence of $R$ actually drops as the average number of publications increases beyond about 40, most likely due to the scarcity of data.

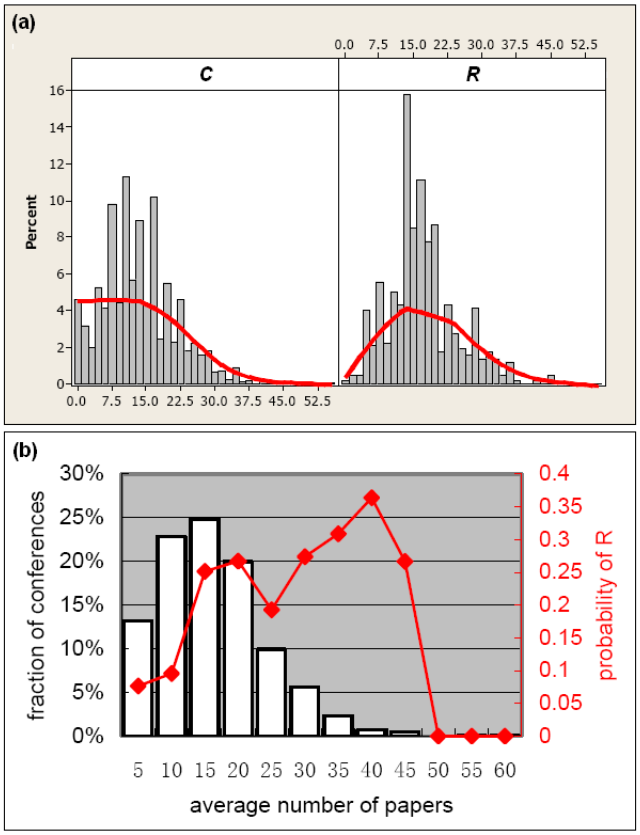


**Figure 3: PC members of conferences labeled as *R* tend to have more publications than their counterparts in *C*, however the difference is not very significant.**

## 4.3 Average number of coauthors of PC.

A related heuristic is to observe how frequently the PC collaborate with their peers in the field. The assumption is that renowned researchers tend to be more active in collaboration with their colleagues for publication. Thus we investigated whether the average number of coauthors of PC members has a positive correlation with the conference quality. Note that number of coauthors specifies the number of *distinct* collaborators on all publications of a given author. Figure 4 depicts that, generally speaking, a larger number of collaborators of the PC members implies a higher chance of the conference being a reputable one. The mean for $R$ is 16.34 and for $C$ is 13.09. However, using this heuristic alone may lead to false positives, as some prominent researchers mainly publish single-authored papers. But this is a diminishing trend in computer science as the average number of collaborators per author tends to increase steadily (Figure 5). Therefore we expect a low false positive rate.

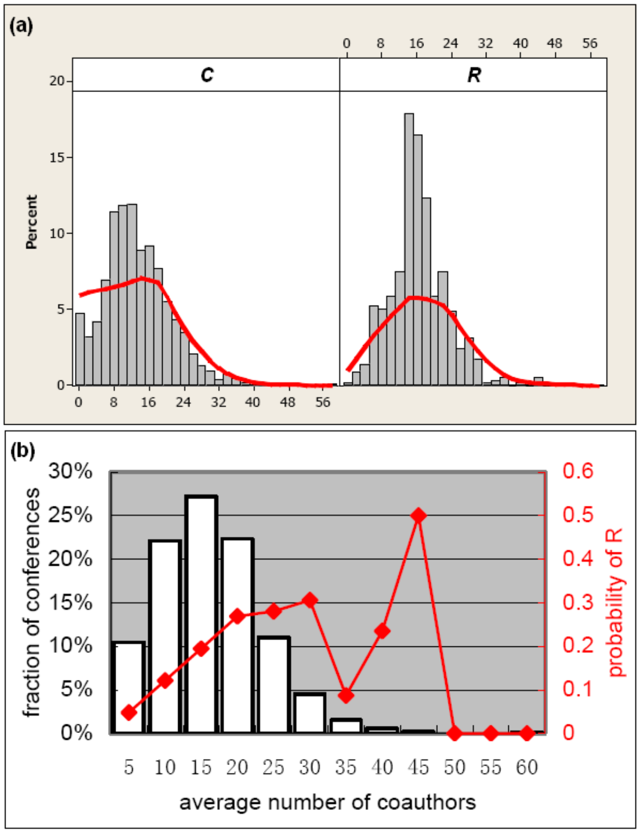


**Figure 4: The distribution is reasonably similar to that of 4.2. PC members of conferences labeled as *R* tend to have more collaborators.**

## 4.4 Closeness centrality of PC.

The *closeness centrality* measure is one of the methods in Social Network Analysis (SNA) to quantify an individual's location in a community [25]. Most prominent ones are often located in the strategic locations in the social network of the community. The *closeness centrality* measure can be defined as how close an author is on average to all other authors. Then, authors with high *closeness* values could be viewed as those who can access new information quicker than others, and similarly, information originating from those authors can be disseminated to others quicker [18]. Formally, the closeness of a node $v$ in a connected graph $G$ is defined as follows:

$$CC(v) = \frac{n-1}{\sum_{w \in G} d(v,w)}$$

where $d(v,w)$ is the pair-wise geodesic (i.e., shortest distance) and $n$ is the number of all nodes reachable from $v$ in $G$.

In our fourth experiment, we studied whether the average *closeness* of the PC members has a correlation with the quality of the conference. Here we calculated the *closeness* value for every PC member based on the collaboration graph constructed using the ACM dataset. The assumption is that a high quality conference has a program committee composed of a group of renowned researchers, who are prominently located in the social network of the community.

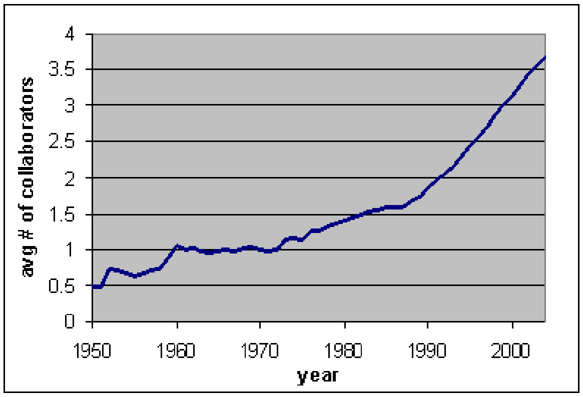


**Figure 5: The average number of distinct collaborators per author in the ACM dataset is steadily increasing over time.**

As can be observed in Figure 6, the prevalence of reputable conferences is generally on the rise as the average *closeness* of the PC members increases, with a spike toward the right when the average *closeness* reaches 0.1. The mean of $C$ is 0.056 compared with 0.062 of $R$.

## 4.5 Betweenness centrality of PC.

While the *closeness centrality* measure depicts how visible an individual is in the community, the *betweenness centrality* measure shows how influential an individual is over the information flows in the social network. Sometimes the interactions between any two indirectly connected authors (i.e., they never collaborated with each other before) might depend on the other authors who connect them through their shortest path(s). These authors potentially play an important role in the network by controlling the flow of interactions. Hence the authors who lie on most of the shortest paths between pairs of authors can be viewed as the *hubs of collaboration* in the community. This notion, known as the *betweenness* of a node $v$, $B(v)$, measures the number of shortest paths between pairs of nodes passing through $v$, and formally defined as follows [8]:

$$BC(v) = \sum_{w,x \in G} \frac{d(w,x;v)}{d(w,x)}$$

where $d(w,x)$ is the shortest path between $w$ and $x$, and $d(w,x;v)$ is the shortest path between $w$ and $x$ passing through $v$. The equation can also be interpreted as the sum of all probabilities that a shortest path between each pair of nodes $w$ and $x$ passes through node $v$.

Figure 7 depicts the correlation between the average *betweenness* of the PC members and the quality of the conference. We again calculated the *betweenness* value for each PC member based on the collaboration graph constructed using the ACM dataset. To our surprise, it appears that there is no strong correlation between the two. A spike exists when the average *betweenness* approaches 0.0003, however there are only five instances within that range.

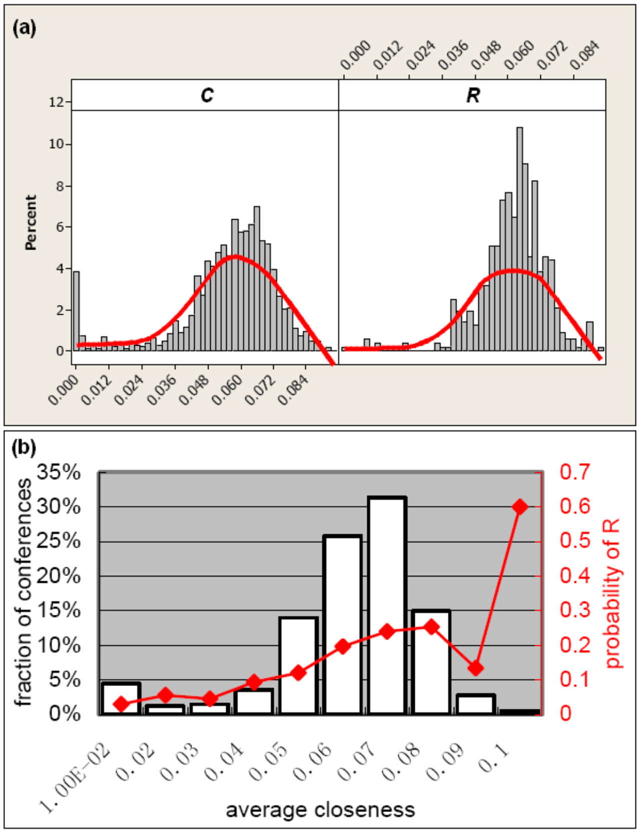


**Figure 6: (a) The distribution between *C* and *R* shows some differences, and (b) it exhibits a positive correlation between the average *closeness* and the prevalence of *R*.**

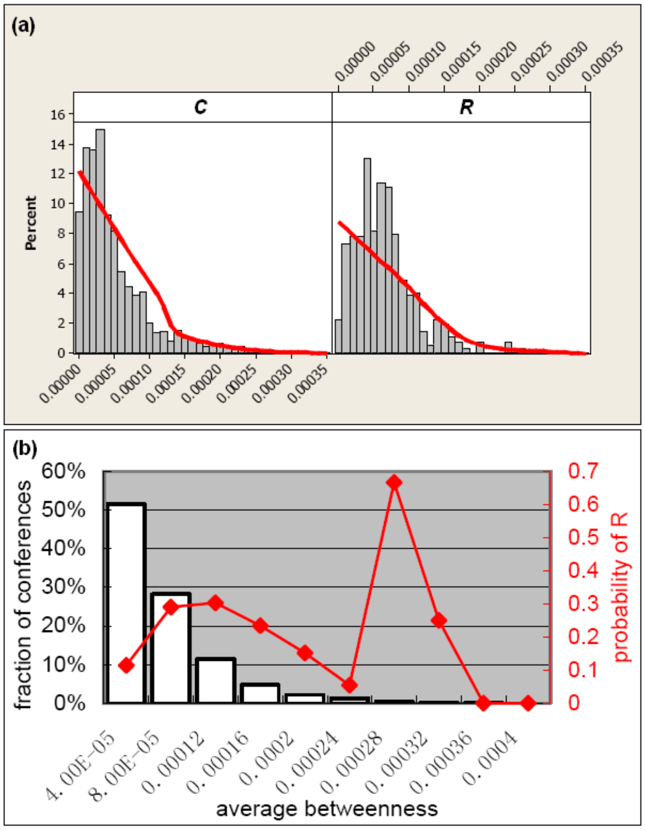


**Figure 7: (a) The average *betweenness* of PC members in *R* is higher than *C*, however (b) shows no clear correlation between this factor and the quality of the conference.**

## 5. COMBINING HEURISTICS FOR CLASSIFICATION

In the previous discussion, we proposed a number of heuristics to identify reputable conferences through statistical analysis of the characteristics of the associated PC members. Although most of the characteristics exhibited various degrees of correlation with the quality of the conference, few of them has enough distinguishing power when used individually. For example, using the "number of PC members" as heuristic, the prevalence of R was considerably high within the range of 110 - 170, however only less than 5% of the conferences in our dataset fell within that range. Therefore in this section, we study how to combine these heuristics to determine high-quality conferences more efficiently and accurately.

### 5.1 Naive classification.

Without loss of generality, our problem in Definition 1 can also be cast to a *binary classification* problem, formally described as follows:

**Definition 2 (Binary Classification)** *Given a set of conference CFPs $X$, classify $x$ ($\in X$) into one of the two classes: (1) a class of reputable conferences $P$ ($\subseteq X$), and (2) a class of low-quality conferences $P'$ ($P'$=$X$-$P$).* □

To solve the above problem, a number of classification schemes were used in our experimentation, including the Bayesian scheme, the decision tree based scheme, the rule-based scheme, and Support Vector Machines [24]. All classification schemes yielded consistent results. However, due to space constraints, we only report the results from the C4.5 decision-tree classification scheme [20], which performed slightly better than the other evaluated classification schemes. We used all five heuristics described in Section 4 as the feature space.

We employed the ten-fold stratified cross validation technique [14] to evaluate the classification accuracy. This technique randomly divided the judged data into 10 partitions of equal size, and performed 10 training/testing phases in which nine partitions were used for training and the remaining tenth partition was used for testing. Therefore, in each training/testing iteration, 2,681 instances were used in training, leaving out 298 instances for testing.

First, we classified the instances using each of the five heuristics individually. The number of PC members heuristic performed the best, correctly identifying 2,436 (81.77%) of the instances, and only misjudged 22 (0.91%) of the $C$ instances to be $R$. The *average closeness* heuristic came in second, with the ability to correctly classify 2,403 (80.66%) of the instances.

Then, we combined all the aforementioned heuristics under the C4.5 classification scheme. After the ten-fold cross validation, the results were promising: 2,570 (86.27%) of the

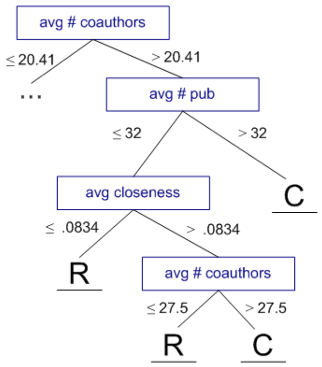


**Figure 8: A portion of the induced C4.5 decision tree for the combined heuristics.**

| class | precision | recall |
|---|---|---|
| C | 0.877 | 0.965 |
| R | 0.751 | 0.434 |

**Table 2: Precision and recall for naive classifier.**

judged instances were classified correctly while 409 (13.73%) were classified incorrectly. The false positive rate for class $R$ was low at 0.035. Table 2 shows a precision-recall matrix, in which the precision measure indicates the percentage of correctly classified instances in each class, and the recall measure indicates the fraction of instances that have been correctly classified. A portion of the resulting decision tree from the classification scheme is shown in Figure 8. In this tree, for example, a conference with an average number of coauthors of PC smaller than 20.41, number of PC smaller than 8, and an average number of publications of PC smaller than 20 is classified as $C$.

## 5.2 Boosting and bagging.

Boosting [9] and bagging [21] are two of the most popular techniques for improving the accuracy of a given classification algorithm [21], and have been proven effective in detecting spam web pages [19]. The rationale of both techniques is to produce an accurate classification by combining the power of multiple less accurate classifiers that are trained iteratively. However, the prediction of each classifier is weighted in boosting, while no weight is assigned and the majority prediction wins in bagging.

In this section, we report the results from adopting these techniques to improve the performance of our classifier. A ten-fold cross validation was used in all the experiments. Table 3 shows the precision and recall for the two classes $C$ and $R$, after applying bagging to the C4.5 classifier combining all the heuristics. After 10 iterations, both precision and recall were improved dramatically, especially for $R$. Overall, the number of correctly classified instances was 2,663 (89.33%), increased from 2,570 (86.27%) before bagging.

The boosting technique further improved the performance of our classifier. After another 10 iterations, we were able to correctly classify 2,739 (91.94%) instances, misjudging only 240 (8.06%). Table 4 shows the precision and recall after applying the boosting technique. Notice that the recall for $R$ was again dramatically improved beyond bagging. 422 instances in $R$ were correctly identified, increased from 311 with bagging.

| class | precision | recall |
|---|---|---|
| C | 0.899 | 0.979 |
| R | 0.859 | 0.54 |

**Table 3: Precision and recall after bagging.**

| class | precision | recall |
|---|---|---|
| C | 0.938 | 0.964 |
| R | 0.831 | 0.733 |

**Table 4: Precision and recall after boosting.**

# 6. DETECT LOW-QUALITY CONFERENCES

In April 2005, a group of MIT students pulled a prank[6] on the conference – "World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI)" – known for sending unsolicited invitation emails to people in academia. The MIT students used software to generate bogus research papers, complete with context-free grammar, and submitted two of them to the conference. To their surprise, one of the gibberish papers was accepted without any reviews. The event received much attention, being covered in various media and has became an amusing topic for debate among scientists. Inspired by this happening, we continued our evaluation by investigating whether the proposed heuristics could be used to detect conferences on the other side of the reputation spectrum - the so-called *low-quality conferences*. We collected the CFPs of 18 low-quality conferences[7] by consulting colleagues and reading the online comments (or complaints) about certain conferences[8]. These conferences were labeled as $LQC$.

We then tested each of the five heuristics on the two datasets $C$ and $LQC$. Overall, the differences were consistently obvious. In the interest of space, we report two most distinguishing heuristics in this section. It is apparent that on average the PC members in $C$ are much more prolific than those in $LQC$: the mean for $C$ is 13.44 compared with merely 1.54 for $LQC$ (see Figure 9).

On the other hand, the prevalence of $LQC$ appeared to have a very strong correlation with the average closeness of the PC members (see Figure 10). The spike toward the left-hand side of the figure indicates that the lower the average closeness value, the more likely a conference belongs to the $LQC$ class.

We also trained and tested a naive classifier to differentiate $C$ and $LQC$ using each and then all the aforementioned heuristics. When combined, the classifier correctly judged 2,406 (99.38%) of all the 2,421 instances. The precision and recall values for filtering out $LQC$ conferences were 0.996 and 0.998, and the false positive rate for $LQC$ was very low at merely 0.002.

At the end, to verify our judgment on $LQC$, we emulated what the MIT students did as follows:

[6] Details about the MIT students' prank and the SCIgen tool to generate random research papers can be found at http://pdos.csail.mit.edu/scigen/

[7] Identities of these conferences can be provided upon request.

[8] See http://www.inesc-id.pt/∼aml/trash.html and http://del.icio.us/tag/fakeconference

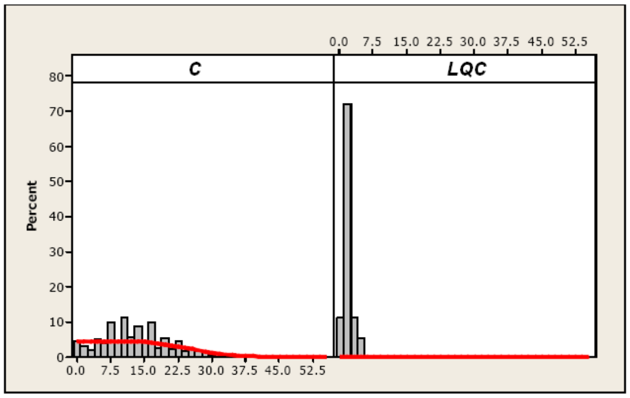


**Figure 9: Overall the number of publications by PC members in LQC is significantly lower than that in C; the highest average number of papers by PC members in LQC is only 3.8.**

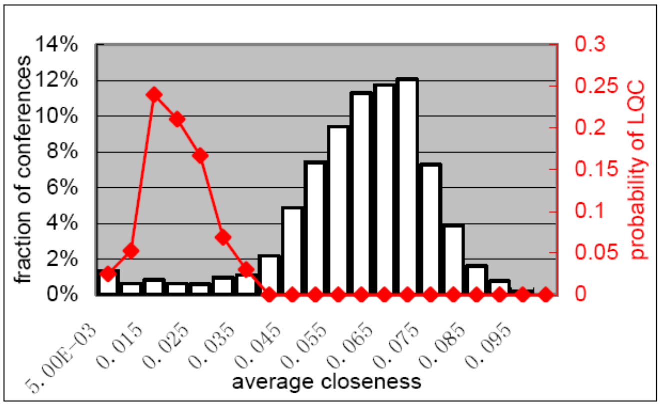


**Figure 10: Conferences labeled as LQC dominated the portion of conferences that had lower average closeness values.**

1. (April 8, 2006) We selected two conferences: *Conference A* and *Conference B*[9]. Both well exhibited the characteristics of *LQC*, appeared to be organized by the same group of people, and had a nearby submission deadline.

2. (April 10, 2006) We made up three bogus papers, *P1*, *P2*, and *P3* using the SCIgen software which the MIT students originally used in the prank. An example is shown in Table 5. We then measured how *authentic* those three papers were by using the *Inauthentic Paper Detector*[10] developed by a group of researchers at the Indiana University [5]. This software estimates the probability that a given article is likely to be written by human. Both *P1* and *P3* were judged as "*inauthentic*", with probabilities (that they were written by humans) of 28.9% and 38%, respectively. The other paper, *P2*, was determined as "*authentic*" with the probability of 61.5%. In order words, the software determined that both *P1* and *P3* were likely to be written by machines but *P2* by human (incorrectly). Nevertheless, all three papers were full of nonsense gibberish such that any serious reading by human reviewers can detect that they were bogus.

| |
|---|
| Abstract: Secure theory and IPv4 have garnered profound interest from both mathematicians and experts in the last several years. Given the current status of random modalities, scholars obviously desire the visualization of Internet QoS. In this position paper, we use knowledge-based methodologies to prove that vacuum tubes can be made introspective, optimal, and probabilistic... |

**Table 5: Snippet of a bogus paper *P1*.**

3. (April 24 - May 1, 2006) We submitted the three papers to the conferences: (1) *P1* to *Conference A* on April 24, (2) *P2* to *Conference B* on April 26, and (3) *P3* to *Conference A* on May 1.

4. (May 15, 2006) To our surprise, we got notifications that both *P1* and *P2* were accepted without any review. For some reasons that we could not know, *P3* was rejected. Despite our subsequent request for reviews or any rationale for the acceptance of the papers, we have not received any response so far.

## 7. RANKING CONFERENCES

Many applications exist for digital libraries to automatically rank conferences based on their intrinsic quality. In inter- and multi-disciplinary academic units or in emerging topics such as bioinformatics, people with various backgrounds work together. However, there are numerous scenarios when measuring the quality of venues becomes important (e.g. making subscription decisions, recommending the most appropriate venue to submit work, promotion and tenure process, etc). What we have studied in previous sections is a supervised binary classification problem: to classify a given conference into either the reputable or the low-quality class.

In this section, we re-cast the problem into a ranking problem, applying our techniques to cover the whole quality spectrum:

**Definition 3 (Conference Ranking)** *Rank a set of conference $X$ by a scoring function $f(x) = \nu$, where $x \in X$ and $\nu \in [0, 1]$.* □

Furthermore, we make the following observations:

**Observation 1** *Using a set of reputable conferences $R$, we can use the probability function $P$ as the scoring function: $f(x) \equiv P(x \in R)$.* □

**Observation 2** *If one ranks $X$ according to the descending order of $\nu$, then the top-$k$ and the bottom-$k$ are equivalent to reputable and low-quality conferences, respectively.* □

These two observations suggest a general ranking scheme to sort the conferences based on the likelihood of being reputable. We have run a series of ranking experiments on various combinations of the heuristics. Due to space

[9] Identities of both conferences are withheld to avoid potential legal issues, but can be provided upon request.

[10] http://www.inauthentic.org

FOCS 2004, ER 2004, ICDT 2005, MobiSys 2004, DEXA 2004, VLDB 2005, WWW 2004, ICDM 2004, SIGIR 2003, SIGMOD 2004, ACM SAC 2004, ICDE 2006 ...

**Table 6: A sample of the overall top-ranked conferences, which is highly overlapping with those in $R$.**

constraints, we report two experiments in which we use a combination of all the aforementioned heuristics. In both experiments, we rank a set of conferences with unknown quality based on their probability of belonging to the set of hand-labeled reputable conferences $R$ (previously described in Section 3 and used in Sections 4 and 5). Such probability is estimated by the C4.5 classifier [20] as the confidence of its prediction.

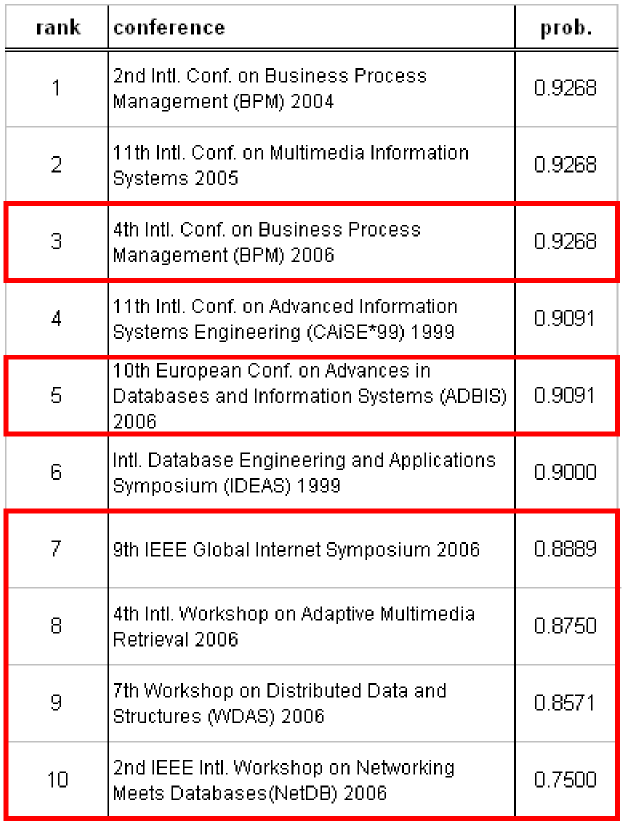
An overall ranking is reported by the first experiment. Not surprisingly, we see in Table 6 that most of the top-ranked conferences are overlapping with those that are present in $R$.

More interesting results are reported in the second experiment, in which we only consider conferences that do ***not*** belong to $R$. A motivation of this experiment is to investigate "how well the conferences in the middle region of the quality spectrum are ranked." Table 7 shows the top-10 conferences (that have not been labeled as $R$) sorted by the prediction confidence. We closely examine these conferences by reading the conference websites and a random sample of the accepted papers, and consulting with our colleagues. Some are workshops co-located with prestigious conferences (e.g, IEEE INFOCOM '06, ICDE '06) and some are conferences that are ranked decently by the same source[11] from which we have collected our sample of reputable conferences (e.g., IDEAS is ranked 29th and ADBIS is 31st).

Another interesting observation of this ranking is that six out of ten are fairly recent venues in 2006. This argues for the advantage of our ranking techniques – since these recent conferences most likely do not have citation statistics available, existing citation-based metrics such as the Impact Factor [10] become inapplicable. By exploiting the proposed heuristics, however, we are able to discover not only the *long-established* venues, but also the *emerging high-quality* ones.

## 8. DISCUSSION

There are a number of ways in which we can improve the performance of our methods. For the task of detecting low-quality conference, a false positive judgment could be extremely detrimental, since we do not want to mistakenly label a legitimate and reasonably good conference as a low-quality one. Thus the goal is to have a high precision. Although the proposed heuristics scored well in terms of precision in our experiments, more representative training samples, especially of the low-quality conferences, should be collected in the next phase of evaluation. On the other hand, for the task of identifying reputable conferences, there are definitely rooms for improvements on the current recall value. In order to achieve better classification accuracy, several other heuristics can be utilized, e.g., both the affiliations and the number of accumulated citations of PC members can be good indicators for the general quality of the program committee.



| rank | conference | prob. |
|---|---|---|
| 1 | 2nd Intl. Conf. on Business Process Management (BPM) 2004 | 0.9268 |
| 2 | 11th Intl. Conf. on Multimedia Information Systems 2005 | 0.9268 |
| 3 | 4th Intl. Conf. on Business Process Management (BPM) 2006 | 0.9268 |
| 4 | 11th Intl. Conf. on Advanced Information Systems Engineering (CAiSE*99) 1999 | 0.9091 |
| 5 | 10th European Conf. on Advances in Databases and Information Systems (ADBIS) 2006 | 0.9091 |
| 6 | Intl. Database Engineering and Applications Symposium (IDEAS) 1999 | 0.9000 |
| 7 | 9th IEEE Global Internet Symposium 2006 | 0.8889 |
| 8 | 4th Intl. Workshop on Adaptive Multimedia Retrieval 2006 | 0.8750 |
| 9 | 7th Workshop on Distributed Data and Structures (WDAS) 2006 | 0.8571 |
| 10 | 2nd IEEE Intl. Workshop on Networking Meets Databases(NetDB) 2006 | 0.7500 |

**Table 7: Top 10 conferences that do *not* belong to R. This shows how the ranking algorithm works in the middle region of the quality spectrum. Several recent venues (highlighted in rectangles) make their way into the list, which would not be possible in solely citation-based ranking scheme due to the lack of citation statistics.**

The proposed heuristics rely on the completeness and correctness of the list of the PC members extracted from CFPs. One potential issue is that a small portion of CFPs do not have a complete list of PC members, e.g., only showing the committee chairs and organizers. In such cases, it requires further action to harvest the list of PC members (for example, by crawling the conference web sites) before the proposed heuristics can be applied.

## 9. CONCLUSION AND FUTURE WORK

In this paper, we proposed a number of heuristics to identify reputable conferences by mining the characteristics of the Program Committee members. When combined under a classification scheme, these heuristics performed promisingly, achieving a satisfying accuracy in differentiating conferences with greater impacts from the rest of the crowd. Evaluation results also showed that our heuristics were effective in detecting some extremely low-quality conferences. The same heuristics were also applied to rank and recommend conferences, which produced reasonably good results and was able to discover emergent venues of good quality.

[11] CS Conference Ranking.org; see Section 3

We believe this study filled in a gap in the current bibliometrics research by introducing some novel quality measures of publication venues. The findings of this work have a number of implications. They shed light on the patterns of PC members in reputable conferences as well as low-quality ones. As conferences become increasingly prevalent as the major outlet for publication in Computer Science, the impact of such quality indicators will be even more significant. The outcome of this study can be directly applied in existing digital libraries to complement and enhance the existing bibliometrics for ranking and recommending reputable publishing venues.

Directions of future work lie in many possibilities. We plan to investigate a number of additional heuristics, including affiliations and origin countries of the PC members, the number of citations to their publications, the number of topics and tracks of the conference as indicated in the CFPs, the organizers and the sponsors of the conferences, etc. The underlying correlations between the quality of the PC and the impact factors of the conference will be studied. We are also interested in studying whether the same heuristics can be applied to conferences in other research domains, and to other types of publication venues, most notably research journals. We will continue our work to apply the current and future findings to rank conferences, and eventually to recommend certain conferences based on the social distance and similarity in interests between the user and the PC members.

## 10. DEMO SYSTEM

A prototype of the proposed work is available at:

`http://pike.psu.edu/confranking/`

We will also provide the implementation and the datasets used in this paper for download in the near future.

## 11. ACKNOWLEDGMENT

The authors gratefully acknowledge partial support from NSF, IBM corporation, and Microsoft Research, and thank the anonymous reviewers for their comments.

## 12. REFERENCES

[1] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, *Evolution of the social network of scientific collaborations*, Physica A, 311(3-4):590-614, 2002.

[2] P. A. Bernstein, D. DeWitt, A. Heuer, Z. Ives, C. S. Jensen, H. Meyer, M. T. Ozsu, R. T. Snodgrass, K. Y. Whang, and J. Widom, *Database Publication Practices*, Proceedings of the 31st Very Large Data Bases Conference (VLDB), 2005.

[3] J. Bollen, M. A. Rodriguez, and H. Van de Sompel, *Journal Status*, retrieved at http://www.arxiv.org/abs/cs.GL/0601030, 2006.

[4] J. Bollen, H. Van de Sompel, J. Smith, and R. Luce, *Toward alternative metrics of journal impact: A comparison of download and citation data*, Information Processing and Management, 41(6): 1419-1440, 2005.

[5] M. M. Dalkilic, W. T. Clark, J. C. Costello, and P. Radivojac, *Using Compression to Identify Classes of Inauthentic Texts*, Proceedings of the 2006 SIAM Conference on Data Mining, April 2006.

[6] E. Elmacioglu and D. Lee, *On Six Degrees of Separation on DBLP-DB and More*, ACM SIGMOD Record, Vol. 34, No. 2, page 33-40, 2005.

[7] E. Elmacioglu and D. Lee, *Oracle, Where Shall I Submit My Papers?*, to appear in the Communications of the ACM (CACM).

[8] L. C. Freeman, *A Set of Measures of Centrality Based on Betweenness*, Sociometry, Vol. 40, pp. 35-41, 1977.

[9] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences, Vol. 55(1): 119-139, 1997.

[10] E. Garfield, *Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas*, Science, Vol:122, No:3159, p. 108-111, 1955.

[11] J. E. Hirsch, *An index to quantify an individual's scientific research output*, Proceedings of the National Academy of Sciences, Vol. 102(46):16569-16572, 2005.

[12] Y. Hong, B.-W. On, and D. Lee, *System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach*, In 8th European Conf. on Digital Libraries (ECDL), page 134-144, Bath, UK, September 2004.

[13] P. Katerattanakul, B. Han, and S. Hong, *Objective Quality Rankings of Computing Journals*, Communications of the ACM, Vol.45, No.10, Oct. 2003.

[14] R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Proceedings of the International Joint Conference on Artificial Intelligence, pages 1137-1145, 1995.

[15] B. Larsen and P. Ingwersen, *Using citations for ranking in digital libraries*, In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06), pages 370-371, Chapel Hill, NC, USA, 2006.

[16] G. Mann, D. Mimno, and A. McCallum, *Bibliometric Impact Measures Leveraging Topic Analysis*, In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06), pages 65-74, Chapel Hill, NC, USA, 2006.

[17] S. Nerur, R. Sikora, G. Mangalaraj, and V. Balijepally, *Assessing the Relative Influence of Journals in a Citation Network*, Communications of the ACM, Vol.48, No.11, 2005.

[18] M. Newman, *Who is the best connected scientist? A study of scientific coauthorship networks*, Physical Review, Vol. 64, 2001.

[19] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, *Detecting Spam Web Pages through Content Analysis*, Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, 2006.

[20] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan-Kaufman, New York, USA, 1993.

[21] J. R. Quinlan, *Bagging, Boosting, and C4.5*, Proceedings of the 13th National Conference on Artificial Intelligence Conference (AAAI), Vol.1:725-730, 1996.

[22] E. Rahm and A. Thor, *Citation analysis of database publications*, SIGMOD Record, Vol. 34, No. 4, 2005.

[23] S. Saha, S. Saint, and D. A. Christakis, *Impact factor: a valid measure of journal quality?*, Journal of the Medical Library Association, Vol. 91(1): 42-46, 2003.

[24] B. Scholkopf, C.J.C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning.* MIT Press, Cambridge, MA, USA, 1999.

[25] S. Wasserman and K. Faust, *Social Networks Analysis: Methods and Application*, Cambridge University Press, United Kingdom, 1994.