Who is Going to Win the Next AAAI Fellowship Award?

Evaluating Researchers by Mining Bibliographic Data

LiorRokach, Meir Kalech, Ido Blank, Rami Stern

Department of Information Systems Engineering,

Ben-Gurion University of the Negev, Israel

Abstract

Accurately evaluating a researcher and the quality of his work is an important task when decision-makers have to decide on such matters as promotions and awards. Publications and citations play a key role in this task and many previous studies have proposed using measurements based on them for evaluating researchers. Machine learning techniques as a way of enhancing the evaluating process have been relatively unexplored.We propose using amachine learning approach for evaluating researchers. In particular, the proposed method combines the outputs of three learning techniques (Logistics regression, Decision Trees and Artificial Neural Networks) to obtain a unified prediction with improved accuracy. We conducted several experiments to evaluate the model's ability to: (1) classify researchers in the field of artificial intelligence as AAAI fellows, and (2) predict the next AAAI fellowship winners. We show that both our classification and prediction methodsare better than previous measurement methods and reach a precision rate of 96% and a recall of 92%.

1. Introduction

Evaluating aresearcher is necessary for various decisions such as whether to hire, promote or grant him or her a competitive award. In most cases, the committee making the decision considers the candidate's list of publications. Since this factor can be deceiving, different

measurements have been developed that use citation information to evaluate and rank researchers.Unfortunately, the problem of how to utilize these measurements still remains and the question arises of how well these measurements indicate the quality of a researcher's work.

Previous studies have attempted to evaluate the accuracy of the measurements by using them to predict when researchers would be promoted (Jensen et al. 2009) or by checking their correlation with human assessments (Li et al. 2010). These studies examined only a small number of measurements and did not use machine learning techniques for combining multiple indices in the prediction process.

In this paper, we propose to use machine learning methods to evaluate and rank researchers based on their publications and citations. These methods use simple bibliographic measures about the researchers, such as the number of papers and citationsas well as advanced indices based on citation data such as the *h-index* (Hirsch, 2005; Bornmann and Daniel, 2007); the*g-index* (Egghe 2006) and various social indicators. Our process includes: (1) extracting bibliographic data from different data sources; (2) selecting features concerning simple measures and citation-based indices and (3) utilizingmachine learning methods to rank the researcher.

The significance of this study lies in using a committee machineapproach based on various bibliographic measurements for evaluating researchers. A committee machine assembles the outputs of various machine learning techniques to obtain a unified decision with improved accuracy. In particular, our paper examines two research questions: (1) How should multiple indices be combined using machine learning techniques? (2) Does social networking among researchers, implemented by co-authorship, improve the ranking of the researchers? In this paper, we evaluate bibliographic measurements empirically via various experiments on a large set of researchers.

In our case study, we focus on the AAAI Fellowship Award. This award recognizes a small percentage of the AAAI researchers who have made significant, sustained contributions to the field of artificial intelligence[1]. This award has become very selective since 1995. Between 1990-1994, 147 researchers won the award; from 1995 to 2009 only 92 researchers gained this coveted prize. We aim to classify researchers in the field of artificial intelligence as AAAI fellows and seek to predict who will win the next AAAI Fellowship Award. We believe the AAAI Fellowship Award is an interesting case study for evaluating the predictive performance of bibliographic measures for the following reasons:

1. **Award vs. Promotion:** Most of the previous studies on researcher evaluation focus on promotion or tenure-track tasks. We believe that a decision on promotion may involve factors other than research quality, such as the availability of positions. In this sense, predicting the possibility that a researcher may be a candidate for a highly prestigious AAAI fellowship, may reflect much better the quality of the researcher and his work.

2. **AI is a well-defined subdomain of computer science:**It is easier to comparescientists in the AI community than scientists from a broader domain such as "computer science" since each subdomain has a different citation pattern. For example, the citation patterns in AI and bioinformatics are very different, making it difficult to compare researchers from these two subdomains. This might explain why previous attempts to predict Turning Award winners were only partially successful.

3. **Data Availability:**There is ample bibliographic data about AI publications available and the AI community contains a sufficient number of AAAI Fellows to validate our methods. Furthermore, the bibliographic data includes different types of publications from journals, conferences, books and chapters over a period of many years.

[1]http://www.aaai.org/Awards/fellows.php

Utilizing a set of 292 researchers from the AI community, we evaluated our methods by implementing and testing three different tasks: (1) classifying a researcher as an AAAI fellow based on her bibliographic data; (2) predicting which researchers would win the competitive AAAI fellowship award; and (3) using an authorship network to measure the distance of a researcher from existing AAAI fellows. Our model,using simple bibliographic measures, citation-based indices and indicators associated with the authorship network of the researchers, provided promising results, with a false negative rate of 8% and a false positive rate of 2%. In addition, we found that our machine committee model was much better than a random model.

2. Scientific Background

This section includes two parts. The first part presents citation-based indices that were previously used for researcher evaluation. In the course of this paper we used these measurements for our machine learning methods. The second part presents studies that used such measurements for prediction.

The most common measurement in evaluating researchers was proposed by Hirsch to evaluate physicists (Hirsch, 2005). A scientist is said to have a Hirsch index (*h-index*) with size *h*, if *h* of his total papers have at least *h* citations each. Another primary measurement is Egghe's*g-index* (Egghe, 2006). This index is affected by the number of citations that the researcher has and their distribution among the researcher's various papers. *g-index* uses a decreasing order of the researcher's publications according to a key based on the number of citations she received. The *g-index* value is the highest integer (*g*) such that all the papers that were ranked in positions 1 to *g* have a combined number of citations of at leastg^2. The *g-index* aims to improve the *h-index* by giving more weight to frequently-cited articles.

The*h-index* measurement has several limitations.In particular, certain factors are ignored, such as the number of authors per paper or when the paper was first published. These limitations led to new variations and measurements of the *h-index*:

1. *Rational h-index distance*: This variation calculates the number of citations that are needed to increase the *h-index* by one point. Let *m* denote the additional citations needed, $hD=h+1-m/(2h+1)$ (Ruane and Tol, 2008).

2. *Rational h-index X*: A researcher has an *h-index* of *h* if *h* is the largest number of papers with at least *h* citations. However, a researcher may have more than *h*papers, say *n*, with at least *h* citations. Let us define $x=n-h$, $hX=h+x/(s-h)$ where *s* is the total number of publications (Ruane and Tol, 2008).

3. *e-index*: This index is based on the square root of the surplus of citations in the h-set beyond h^2, i.e., beyond the theoretical minimum required to obtain *h-index* of *h*. The aim of the*e-index* is to differentiate between scientists with similar h-indices but different citation patterns (Zhang, 2009; Zhang 2010).

4. *Individualh-index*: In order to reduce the effects of co-authorship, the individual *h-index* divides the standard *h-index* by the average number of authors in the papers that contribute to the *h-index* (Batista et al., 2006).

5. *Norm individual h-index*: This index first normalizes the number of citations for each paper by dividing the number of citations by the number of authors for that paper. Then the index is calculated as the *h-index* of the normalized citation counts. This approach is much more fine-grained than the former one; it accounts more accurately for any co-authorship effects that might be present (Harzling, 2010).

6. *Schreiber individual h-index*: Schreiber's method uses fractional paper counts (for example, one-third for three authors) instead of reduced citation counts, to account for shared authorship of papers. Then it determines the multi-authored h-index based

on the resulting effective rank of the papers using undiluted citation counts (Schreiber, 2008).

7. *Contemporary h-index*: This index adds an age-related weighting to each cited article; the older the article the less weight (Sidiropoulos et al., 2007).

8. *AR-index*: This is an age-weighted citation rate, where the number of citations for a given paper is divided by the age of that paper. The AR-index is the square root of the sum of all age-weighted citation counts over all papers that contribute to the *h-index* (Jin, 2007).

9. *AWCR*: This is the same as the AR-index but it sums over all papers (Harzling, 2010).

10. *AWCRpA*: This per-author age-weighted citation rate, although similar to AWCR, is normalized as to the number of authors for each paper (Harzling, 2010).

11. *pi-index*: This index is equal to one-hundredth of the number of citations obtained for the top square root of the total number of journal papers ('elite set of papers') ranked by the number of citations in a decreasing order (Vinkler, 2009).

There are several works that present empirical experiments for evaluating researchers using the above measurements. Feitelson and Yovel(2004) compute the ranking of computer science researchers based on the total number of citations the researcher's papers received. They also created a theoretical model to predict the future number of citations. To evaluate their ranking model, they tried to predict the winners of the Turing Award. According to their results, the correlation between their model and the Turing Award winners was not sufficiently significant. Thus, their model could be used to supplement human judgment, but not to replace it. Unfortunately, they built their model based on data from CiteSeer[2] which is neither complete nor accurate.

[2]http://citeseerx.ist.psu.edu/

Jensen et al. (2009) used several measurement methods to predict which CNRS (French National Centre for Scientific Research) researchers would be promoted. They concluded that although there was a clear difference in the measurement values between the researchers that did get promotion and those that did not, their prediction model was successful for only half of the researchers. In this sense, predicting a competitive award, like AAAI fellowship, may reflect much better the quality of the researcher evaluation.

Another research, proposed by Li et al.(2010), tests the correlation between expert opinion on researcher quality and three known measurements (each measurement was tested individually). Although they found a significant correlation between the measurements and expert opinion, it was not enough to replace the human assessment of the researcher's quality.

Bornmann et al. (2008) compare nine different variants of the h-index using data from biomedicine and conclude that combining a pair of indices can provide a meaningful indicator for comparing scientists. They suggest that one of the indices should relate to the number of papers a researcher has published (as is the case with the h-index) while the second index will be related to the impact of the papers in a researcher's productive core (such as the a-index, which is the total number of citations divided by the h-index.) Similarly, Jin et al. (2007) propose combining the h index with the ar-index.

Social network analysis (SNA) has been previously used to examine the impact of individual researchers. For example, Kretschmer (2004) uses simple social distance indicators for analyzing co-authorship networks. Other more complicated network measures, such as betweenness centrality, are also appropriate for analyzing co-authorship networks. In particular, Liu et al. (2011) employ these metrics to evaluate the impact of individual

researchers on the recombination of knowledge and to show the effectiveness of these metrics.

The main contribution of our paper is that we propose a model that can combine many indices using machine learning techniques and evaluate it empirically. We show that by using machine learning, a low false rate is obtained in classifying researchers.

3. Methodology

To cope with the challenge of researcher evaluation, we implemented a supervised learning approach. Our process includes the following steps:

1. **Data Collection:**Collectingmetadata about the researcher's publications and citations.
2. **Feature Calculation:** Generating a training set with features composed of bibliographic data and different measurements such as h-index. The classes are determined according to the classification goal, such as winning an award.
3. **Feature selection:** Selecting the most indicative features.
4. **Model Training:** Building a classifier from the training set, using an induction algorithm.
5. **Evaluation:** Evaluating the predictive performance of the classifier.

Step 1: Data Extraction

In order to accomplish the first step we first extracted data from the DBLP[3]. The DBLP (Digital Bibliography and Library Project) is a bibliography database and website which indices more than 1.3 million papers on computer science. Since we are using DBLP as our primary source, the year range is determined by the bibliographic coverage of the DBLP

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

database. Although the DBLP has been indexing papers since 1936, coverage only became substantial (more than 1000 papers a year) from the early Seventies. Because we are trying to predict AAAI fellowships since 1995, we find the DBLP a good source for obtaining a candidate's publication list.

.

The database can be downloaded in a XML format. We first parsed the XML and loaded the data into a relational database. Then we queried for all researchers who have published at least five papers in AI journals or leading conferences. We set a threshold of five papers in an attempt to differentiate AI researchers from other types of computer science researchers.

The list of journals contains all journals in the sub-category "Computer Sciences – Artificial Intelligence" that is indexed by Thomson Reuters' Web of Knowledge[4]. In addition we compiled a list of the top five conferences in artificial intelligence after consulting several AAAI fellows who serve on the Fellows Selection Committee. It should be noted that the list is similar to other lists (see, for example, the top tier AI conferences that were included in the Alberta Computer Science Conference Rankings[5] or in theMicrosoft Academic Ranking[6].

The DBLP database contains 456,764 individual authors from among the entire computer science community. About 24,707 authors have written at least one AI paper and 2,140 persons have written at least five qualified AI papers (i.e., papers that were published in one of the AI journals or leading conferences described above). Moreover, all AAAI fellows have

[4]http://apps.isiknowledge.com/
[5]http://webdocs.cs.ualberta.ca/~zaiane/htmldocs/ConfRanking.html
[6]http://academic.research.microsoft.com/RankList?entitytype=3&topDomainID=2&subDomainID=5

more than five qualified AI papers. Thus the threshold of five papers, which approximately identifies the top 10% researchers in the AI field, can be used as an initial filter.

From the top 10% AI researchers, we selected a subset of 292 AI researchers. We then selected a set of 92 AAAI fellows consisting of all fellowship winners since 1995. As noted above, this award has become very selective since 1995. Between 1990-1994, 147 researchers won the award. From 1995 to 2009 only 92 researchers gained this coveted prize and this factexplains our selection. The remaining 200 researchers were randomly selected without replacement from the qualified list of AI researchers on the condition that they were not AAAI fellows (i.e., not even AAAI fellows that won prior to 1995). We have not used the entire qualified population (2,140 persons) because it would require more extensive resources to extract their citations. However, in our opinion, the sample we used was sufficiently large and similar to what other researchers in the field have regarded as adequate. It should be noted that we selected all AAAI fellows since 1995 and did not count on random selection. If we had done so, the resulting sample would have included only 13 fellows. Such a sample has too few instances for inducing reliable insights about the AAAI fellowships. This phenomenon is referred to in the literature as the class imbalance problem (Chawla et al.,2004). In particular, class imbalance usually occurs when, in a classification problem, there are many more examples of a certain class than another class. In such cases, standard machine learning techniques may be "overwhelmed" by the majority class and ignore the minority class. In fact, undersampling of the majority class (in our case, the non-fellows) is a well-known method in machine learning for overcoming the class imbalance problem.

For every researcher, we first queried the list of her papers in DBLP. The list includes all papers of the researcher as they appear in DBLP (i.e., all papers in the domain of computer science) and not only the papers that were published in one of the AI journals indicated

above. We took this approach since in the field of computer science it may not be sufficient to "rely on journal publications as the sole demonstration of scholarly achievement" (Patterson et al., 1999).

It should be noted that the abovementioned inclusion criterion of five qualified AI papers is used only for narrowing the list of candidates (from a total of 24,707 CS researchers in DBLP to only 2,140 researchers). Once a candidate satisfies the inclusion criterion, we explore all her papers (including non-AI qualified papers). We assume that a candidate can publish a high impact paper in another CS domain (such as the Journal of the ACM which targets a much broader audience than the AI community). Later on, we calculate the bibliographic indices of the candidate in two ways: a) using all her papers and b) using only the candidate's qualified AI papers. In the second instance we first filter out the non-qualified papers and only then calculate the index. Using machine learning techniques we can combine the various index variants in the same model.

For each paper we used a Web crawler to extract the details of the papers that cited the paper in question. We used Thomson Reuters Web of Knowledge (WoK) website and Google Scholar (GS) to obtain the citation information. Google Scholar and WoK are both used for obtaining the citations of the candidate's papers because they differ in their journal coverage and generally provide different citation records for the same target papers (Garcia-Perez, 2011). For example, WoK provides a limited coverage of non-English papers and almost no conference papers. On the other hand, the coverage of Google Scholar is uneven across disciplines and has very limited coverage of old papers (before 1996). As indicated by Meho and Yang (2007), GS "stands out in its coverage of conference proceedings" and the use of GS, in addition to WoK, "helps reveal a more accurate and comprehensive picture of the scholarly impact of authors". In fact, it has been shown that combining these different sources provides a more complete picture of the scholarly impact (Levine-Clark and Gil,

2009). Using the WoK database, we extracted the metadata details of almost 92,000 citing papers while the number of extracted citing papers from Google Scholar reached almost half a million.

Finally we used the DBLP database to generate the social network of the researchers. The nodes represent the CS researchers and the edges represent the co-authorship relations. We found DBLP to be an appropriate database because of its extended coverage of CS papers and because it can be fully downloaded and loaded into our database. We calculated social network-based features on the authorship distance between the researchers under examination and existing AAAI fellows. We describe this technique in detail in the next section.

We avoided the need to address name ambiguity by relying on DBLP which has a disambiguation feature in place (Ley and P. Reuther , 2006). For example,there are 29 different authors named "*Wei. Wang*" in *DBLP*[7]. For each one of them, DBLP holds a separate publication list. Naturally this does not resolve all ambiguity problems. However, we believe that in our case it is less crucial since we are focusing only on AI researchers and the DBLP usually indexes the full name (and not only the last name and the initials of the first and middle name). Both factors reduce the possibility of ambiguity.

Another issue that needs to be addressed is the matter of errors in citation databases. Each citation dataset may have mistakes such as duplicate citations or phantom citations (García-Pérez, 2010). We removed duplicate citation by using the procedure presented in Kan and Tan (2008)[8]. In this paper we did not check for phantom citations, because it would have required us to go over the reference list of the citing paper and this list is not available in Google Scholar.

[7]http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/w/Wang:Wei.html
[8]It can be downloaded from: http://wing.comp.nus.edu.sg/~tanyeefa/downloads/recordmatching/

To summarize, the DBLP dataset was used for obtaining the publications list of the candidates. The DBLP dataset was also used to generate the co-authorship graph (the social network of the researchers). On the other hand, WoK and GS were used for extracting the metedata of the citation papers.

Step 2: Features Calculation

Three types of features were discerned. The first type of features were derived from what we regarded as simple bibliographic measures and included: total publications; total publications normalized by the number of authors; total citations; total citations normalized by the number of authors; citations per year; average number of citations per paper; average number of papers per year and seniority (number of years passed since the first publication). The second type of features, composed of citation-based indices, included all the 13 indices described in the Scientific Background section. The third type of features was derived from the co-authorship network.

As mentioned above, after obtaining from DBLP the publications list of a certain candidate, we went over the list and for each paper we queried the citation database (GS or WoK) and obtained all the citations for this paper. The citations are first parsed and their metadata are stored in the database with an indication as to which paper was cited. In order to calculate a certain index variant for a specific year, we first filtered out all non-relevant publications and citations, and then calculated the index based on the remaining papers and citations.

Each of the above features was calculated according to several variants:

1. Data Source: GS, WoK – For example, the h-index was calculated separately using the WoK citation and GS citation indices respectively.

2. Paper Type – This indicates the types of papers of the researcher in question that should be taken into consideration. We considered three types: all papers, journal paper only, AI only (based on the qualification list indicated above).

3. Citing Paper Type -- This indicates which citing papers were taken into consideration. As in the previous alternative, we differentiated between all, journal only and AI only.

4. Self-Citation Level: We differentiated between three different levels of self-citation: Level 0: all citations were taken into consideration; Level 1: we ignored citations, in which the researcher in question was one of the authors; Level 2: we ignored citations in which one of the original authors (not necessarily the researcher in question) was also one of the authors of the citing paper.

Based on the above parameters, we calculated up to 2 X 3 X 3 X 3 =54 variants for the same index. Each measure variant was calculated on a different subset of the documents. We used different variations of the same measure in order to evaluate diverse aspects of the researcher. For example, researcher A may have had a higher h-index than researcher B when all papers were taken into consideration (indicating a stronger impact of her papers among the general audience. At the same time, researcher A may have a lower h-index than her counterpart when only AI papers are taken into consideration (indicating that her impact in the AI community is lower). By exploiting the synergy among the variants we can make better predictions. In particular, we can analyze their correlation with the target class (the AAAI fellowship indicator), and induce what is the h-index variant mixture of a typical AAAI fellow. The sensor fusion perspective may also motivate the use of several variants of the sameindex (see, for example, Frolik et al., 2001). It has been shown that even if the sensor readings (the index's values in our case) are highly correlated, one can benefit by combining them. This can be explained by the fact that none of our indices are error-proof. By

combining different variants, where each one is calculated on a partially different set of papers, we can mitigate the faults of a subset of the indices.

The third type of features includes several social indicators. These indicators were calculated based on the co-authoring patterns of the researchers. Our hypothesis was that close research relationships among AAAI fellows increases the probability of winning. To examine this hypothesis, we modeled the relationships among AAAI fellows by a social network inspired by an Erdos number. An Erdos number describes the "collaborative distance" between a person and the mathematician Paul Erdos, as measured by authorship of mathematical papers (Newman, 2001). We used the DBLP to build the collaboration graph, where the nodes represent the researchers. An edge connects two researchers if they are co-authors.

The social indicators were calculated on a yearly basis in the following manner. For a given year, we took all papers published until that year (inclusive) and generated a social authorship network. Then we marked the nodes of all the researchers who won the AAAI fellowship up to that year. Finally we calculated the social network indicators for each candidate. Figure 1 illustrates the collaboration graph. To measure the collaboration distance of researcher *r* and the AAAI fellows, we measured three parameters: (1) the minimal path length between *r* and the closest AAAI fellow; (2) the average path length; and (3) the number of AAAI fellows whose distance to *r* was less than 5. These social distance indicators were chosen due to their simplicity and because variations of them were successfully used in the past for analyzing co-authorship networks (Kretschmer, 2004). Other more complicated network measures, such as betweenness centrality, are also appropriate for analyzing co-authorship networks (Liu et al. , 2011). However, we leave this for future research.

Each social indicator was calculated according to three different variants: a. using all papers; b. using only AI papers and c. using only journal papers. Thus we have 3x3=9 social-based

features. In addition there are 15 citation-based features[9]. Since each citation-based feature has 54 variants, we have 54 x 15=810 citation-based features. In addition, we have 4 simple bibliographic features[10]. Each simple bibliographic feature has 6 variants (2 different citation datasets x 3 types of papers). Thus we have 4 x 6 = 24 simple features. In total, we have 9+810+24=843 features.

The above features were calculated for each researcher on a yearly basis. Obviously the index for a certain year considers papers and citations up to that year. For example, when we calculated the h-index for a certain year, future papers and citations were not considered. The need to calculate the index for each year was one of the reasons why Step 1 above extracts the metadata of the citing papers (including years) in addition to the papers of the researchers.

As for the number of records, we analyzed AAAI fellows from 1995 up to 2009, i.e., a 15-year period. Of the 292 candidates that were selected, each candidate had one record per year (representing her status at the end of the year). Thus potentially we should have 15 x 292=4,380 records. However, if a candidate started her career a bit later (i.e., her first paper was published later than 1995) she has several empty records in the initial years. After removing these empty records, we had a total of 3,898 records.

Insert Figure 1 Here

Figure 1: Collaboration graph.

[9]h-index; rational h-index distance; rational h-index X; e-index; individual h-index; norm individual h-index; Schreiber individual h-index; contemporary h-index; AR-index; AWCR; AWCRpA; pi-index; total citations; total citations normalized by the number of authors; and average number of citations per paper

[10]total publications; total publications normalized by the number of authors; average number of papers per year and number of years passed since the first publication

Overall, the data set contains 3898 records and 843 input features. Each record represents a profile of a candidate in a particular year (end of the year). Each column represents a certain measure variant. In addition, we classified every record such that 'true' represents "AAAI fellow" and 'false' represents "not AAAI fellow".

Step 3: Feature Selection

As indicated in the previous section there were 843 input features in the dataset. The most important challenge was to select the features and to determine which had the most influence. The first step in coping with this challenge was to determine a method for coping with the dimensionality problem. It is well known that the required number of labeled instances for supervised learning increases as a function of dimensionality. The required number of training instances for a linear classifier is linearly related to the dimensionality and for a quadratic classifier to the square of the dimensionality. In terms of nonparametric classifiers, such as decision trees, the situation is even more severe. It has been estimated that as the number of dimensions increases, the training set size needs to increase exponentially in order to obtain an effective estimate of multivariate densities. This phenomenon is known as the "curse of dimensionality." Techniques that are efficient in low dimensions, such as decision trees inducers, fail to provide meaningful results when the number of dimensions increases beyond a 'modest' size.

Feature selection is a well-known approach for dealing with high dimensionality. The idea is to select a single subset of features upon which the inducer will run, while ignoring the rest. The selection of the subset can be done manually by drawing upon prior knowledge to identify irrelevant variables or by utilizing feature selection algorithms. In the last decade, many researchers have shown increased interest in feature selection and consequently many algorithms have been proposed, with some demonstrating remarkable improvements in

accuracy. Since the subject is too wide to survey here, the reader is referred to (Mengle and Goharian, 2009) for further reading.

In this paper we focus on ranking-based feature selection algorithms. These algorithms employ a certain criterion to score each feature and provide a ranking by measuring its value with respect to the binary class (either winning the AAAI fellowship or not). Given a feature ranking, a feature subset can be chosen by taking the top *k* features. In this paper we examined the following three criteria; all of them are implemented in the WEKA environment (Witten and Frank, 2005):

1. Chi-Square - chi-square was used to statistically ascertain the correlation between the target class (winning the AAAI fellowship) and the bibliometric indicators. We used the Chi2 algorithm (Setiono and Liu, 1995) which can be utilized for feature selection and discretization of the bibliometric indicators. For each bibliometric indicator, the algorithm tries to determine if adjacent intervals of the current indicator should be merged. For this purpose the chi-square statistical test is used to test the hypothesis that the target class value (winning or not winning) is independent of the two intervals. If the conclusion is that the class is independent, then the two adjacent intervals are merged. The merging process is repeated until there are no indicator values that can be merged. At the end of the procedure, the final chi-square result indicates the merit of the feature. Note that if an indicator is merged to only one value, it means that it has no merit and can be filtered out.

2. Gain Ratio- Gain ratio, originally presented by Quinlan in the context of Decision Trees (Mitchell, 1997), isdesigned to overcome a bias in the information gain (IG) measure. It measures the expected reduction of entropy caused by partitioning the examples according to a chosen feature. Given entropy E(S) as a measure of the

impurity in a collection of items, it is possible to quantify the effectiveness of a feature in classifying the training data.

3. Relief – This criterion estimates the quality of the features according to how well their values distinguish between instances that are near each other (Kira and Rendell, 1992). In each iteration, Relief randomly selects researcher x. It then searches the dataset for her two nearest neighbors from the same class (i.e., fellow or non-fellow as x), termed the "nearest hit H", and from the complementary class, referred to as "the nearest miss M". It updates the weights of the features that are initialized to zero in the beginning based on the simple idea that a feature is more relevant if it distinguishes between a researcher and her near miss and less relevant if it distinguishes between a researcher and her near hit. After completing the procedure, it ranks the features based on their final weight.

The criteria were examined in relation to the following highest ranked (top) features settings: 5, 10, 20, 30, 50, 100 and 200. Our preliminary results indicated that a gain ratio with the top 50 features provided the best predictive performance.

Step 4: Training the model

In this step we finally induce the classification model. The classifier aims to assess the probability that a particular researcher will become an AAAI fellow in a certain year. In this section, we examine various classification models for combining the different indices (and their variants). Since each model is based on a different assumption, the data fit is correspondingly different.

1. Logistics regression – This model assumes that the natural logs of the odds of a candidate becoming a fellow are a linear combination of the indices. It assigns a different weight for each one of the indices by fitting their values to the target class. The best fit aims to maximize the likelihood of the data given the fitted model. For

example, the following equation represents a fitted model. For the sake of simplicity, we used only two indices:

$$ln\left(\frac{p_i}{1-p_i}\right) = 0.71 + 0.00984 \cdot Number_{of\,Publications} + 0.10 \cdot h_{index}$$ 42

where p_i represents the probability of becoming a fellow. In this model, increasing the number of publications or the h-index of the candidate is associated with higher odds of becoming an AAAI fellow. In particular, according to this model, a researcher with 50 AI papers and an h-index of 20 has a 0.76 probability of becoming an AAAI fellow.

2. AdaBoost using decision tree – The decision tree combines the indices in a hierarchical fashion such that the most important index is located in the root of the tree. Each node in the tree examines a different index. Each candidate is assigned to one leaf that can be found by traversing the tree from the root to the leaf. A certain path is selected according to the values of the current candidate's indices. Decision trees assume that the space of the indices should be divided into axis-parallel rectangles such that each rectangle has a different fellowship probability. Figure 2 illustrates the classification of a researcher using a simple decision tree and its corresponding space partitioning. A different fellowship probability is assigned to each leaf. In particular, researchers with a total citation per author that is greater than 54 and with an h-index greater than 15 are associated with the top-right rectangle (the rightmost leaf) and have a probability of $p_{fellow}=0.17$ of becoming a fellow.

In this paper we built a decision forest, i.e., generating and combining several trees. This is a well-known approach for overcoming decision tree drawbacks (Breiman, 2001).

Insert Figure 2 Here

Figure 2: Illustration of decision tree

3. Multilayer Perceptron - This is a type of neural network in which the various measures are connectedbyan intricate network which consists of three node layers. Each node in the first layer represents a different measure. Each node in the first layer connects with a certain weight to every node in the following layer. The induction algorithm tries to find the best weights. Practically, a multilayer perceptron is nothingmore than a nonlinear regression in which the measures are combined using a sigmoid function. The logistics regression model described above is a single-layer artificial neural network.

Instead of simply using one of the above techniques, we applied a well-known practice in machine learning called committee machines (sometime associated with a more specific term such as ensemble learning or a mixture of experts) in which the outputs of several classifiers (experts) are combined. Each of the classifiers solves the same original task. Combining these classifiers usually results in a better composite global model, with more accurate and reliable estimates or decisions than can be obtained from using a single model. This idea imitates a common human characteristic -- the desire to obtain several opinions before making any crucial decision. We generally weigh the individual opinions that we receive and then combine them to reach a final decision (Polikar, 2006; Rokach 2010).

In this paper we combined three types of classifiers (decision trees, logistics regression and multilayer perceptron) by assigning the same weight for all classifiers. It is known that combining different types of classifiers can improve predictive performance, mainly due to the phenomenon that various types of classifiers have different “inductive biases” (Mitchell,

1997). In particular Ali and Pazzani (1996) and Rokach et al. (2006) show that combining diverse classifiers can be used to reduce the variance-error (i.e., error due to sampling variation) without increasing the bias-error (i.e., error due to an inadequate model). Additionally, many participants in prediction contests combine various models in order to achieve the best results (see, for example, Koren, 2009).

During the test phase, we sought to predict if a certain candidate would become an AAAI fellow in a certain year. We inputted the candidate's indicesfor that specific year into the induced classifiers. Each classifier outputted the probability of the candidate of becoming a fellow. We then combined the classifier outputs by averaging their estimated probabilities using the same weight. This combination method is known as a distribution summation and despite its simplicity; it is known to provide excellent results (Ali and Pazzani, 1996).

4. Experiments and Results

In the following sections we present three experiments focused on the following research questions:

1. Can we accurately classify researchers as winners/not winners? What features most affect the classification?
2. Can we predict the fellows for a given year?
3. Does the authorship network of the researchers improve the classification results?

4.1 Classifying Researchers

The goal of the first set of experiments was to examine the ability to classify researchers as AAAI fellows. Also, we wanted to examine what features influenced the classification model. In order to do this, we used a leave-one-researcher-out validation procedure. In every test iteration,the classifiers were trained on the records of all researchers except one. The

classifiers were then tested on the records (years) for the only researcher left out of the training data set. This validation process was repeated for all 292 subjects.

We used the following metrics to evaluate the classifier:

1. A false negative (FN) rate is defined as the proportion of researchers who are non-AAAI Fellows from all researchers who were predicted as AAAI Fellows.

2. A false positive (FP) rate is defined as the proportion of researchers who are AAAI Fellows from all researchers who were predicted as non-AAAI Fellows.

3. Precision is defined as the proportion of researchers who are AAAI Fellows from all researchers that were predicted as AAAI Fellows.

4. Recall is defined as the proportion of researchers who are predicted as AAAI fellows from all researchers who are AAAI fellows.

5. The F-measure indicates the harmonic mean of the last two metrics.

Table 1 summarizes the results. The rows represent the various classifiers. In the first two rows we can see the anchor results for two simple naive classifiers which either classify all researchers as false or all researchers as negative. Such naive classifiers have, of course, a false positive rate of 0%, but a false negative rate of 100% and vice versa. Note that for the first case, the precision value is not defined. While these two classifiers perform badly, they can be used to put our results in the properperspective. The next row shows the best result obtained with our machine committee system using the top 50 features.As can be seen, our classifier significantly improves the false negative of the naive classifier but with a low false positive rate. In order to conclude which classifier performs best, we first used the adjusted Friedman test on the F-measure in order to reject the null hypothesis and then the Bonferroni-Dunn test to examine whether the best classifier performs significantly better than the other classifiers (García et al., 2010). Specifically, in Table 1, the null-hypothesis, that all classifiers perform the same and the observed differences are merely random, was rejected using the

adjusted Friedman test. We proceeded with the Bonferroni-Dunn test and found that the classifier trained using the top50 features statistically outperform all others with a 95% confidence level.

Table 1: False negative and positive rates of different classifiers.

Insert Table 1 Here

The results of the machine committee system were very encouraging from the predictive performance point of view but at the same time the classifiers were incomprehensible. Thus, we used the Ripper algorithm (Cohen, 1995) which can generate rules to determine under what conditions a researcher will receive the AAAI Award. The performance of the Ripper algorithm is presented in Table 2. The predictive performance is lower than the machine committee system but the obtained list of rules is comprehensible. For instance, this next rule is a result of this classifier: *IF (the number of publications > 9) AND (e-index > 12.071) AND (the average number of citations per paper > 4.618)* ➔ *Fellow=TRUE (11.0/1.0)*. The meaning of the right-hand of the rule is that there are *11+1=12* cases which satisfy the conditions from which *11* cases also satisfy the consequent (i.e., Fellow=TRUE). The null-hypothesis, that the two classifiers perform the same, can be rejected using the Wilcoxon test with a confidence level of 95%. Thus, we conclude that from the predictive performance perspective, the machine committee should be preferred .

Table 2: Comparing Machine Committee with Ripper Classification Rules

Insert Table 2 Here

In Table 3 we analyze how social indicators affect general predictive performance. We can see that relying only on social features provides false negatives; this is much worse than all the other classifiers which do not consider authorship network. We further experimented with the impact of using social features with the simple bibliographic measures and the citation-based indices. Surprisingly, we found that such a combination improves the results as shown in the second row. These results are even better, both in terms of false negative as well as false positive rates, than the results of our classification model which does not use social features (row 3). These results are very impressive and show that authorship distance features offer a promising direction in evaluating researchers. The null-hypothesis, that all classifiers perform the same and that the observed differences are merely random, was rejected using the adjusted Friedman test. We proceeded with the Bonferroni-Dunn test and found that the approach involving the use of all features (including social features) outperforms all others with a 95% confidence level.

Table 3: Analysis of Social Features

Insert Table 3 Here

Table 4 presents experiments that examine the most influential features on the success of the classification. For this task we ran the same experiments as before but considered only a few subsets of features:

1. All features but social (All features except for the social indicators)

2. The simple bibliographic measures that are associated only with raw bibliographic data (such as the number of publications)

3. The features that are associated with only citation-based index measures (such as the h-index).

4. Only h-index variant features (54 input features in total)

5. Only the number of publication variant features (6 features in total)

6. The best single feature. Among all features we found that the highest F-Measure was provided by the g-index, calculated over the WoK data source, using only AI authored papers and all journal citing papers, including self-citations (Level 0).

It is interesting to see that the false positive of each one of the individual features (rows 4,5 and 6) is much worse than the combination of all the features as presented in the second column. This means, for instance, that the number of publications and the *h-index,* which are usually considered as influential factors for researcher evaluation, in fact, fail when they are regarded as the sole evaluation tool. The combination of the simple bibliographic measures (row 2) and the citation-based indices (row 3) presents results that are very close to the combination of all the features. The null-hypothesis, that all classifiers perform the same, was rejected using the Friedman test with a confidence level of 95%. The Bonferroni-Dunn test indicates that the hypotheses that "All Features but Social Features", "Simple bibliographic measures Only" and "Citation based Indices Only" perform the same at confidence levels of 95% and 90%, respectively, and cannot be rejected. However, the same test indicates that "All Features but Social Features" significantly outperforms "All h-index Variants", "Number of publications Variants" and g-index at a confidence level of 95%.

Table 4: Comparing various subsets of features.

Insert Table 4 Here

Table 5 presents the top ten features selected using the feature selection procedure. It should be noted that the same feature can be selected more than once (for example, the g-index in Table 5), but each time there is a different variant (i.e., it is calculated based on a different set of papers). Forty-five features of the top 50 features are citation-based indices; two of them are social indicators and the rest are simple bibliographic-based measures. Thus, the citation-based indices dominate the top 50 list. It should be noted, however, that there are many more citation-based features to begin with.

Table 5: Top ten features.

Insert Table 5 Here

Table 6 presents the performance of the top single feature in each category and the performance of the top 5 features in each category. The penultimate row indicates the performance obtained by combining the top 5 features of all categories (a total of 15 features, 5 features from each category). The last row presents the performance obtained by the top 15 features selected from all features (and not from each category separately). The results indicate that combining features from all categories is better than taking features from only one category. Moreover, in terms of predictive performance, the last procedure (i.e., selecting the features from all categories) is slightly better than the penultimate procedure (joining the top features in each category). Nevertheless the penultimate procedure balances the various aspects of the researcher and does not rely mainly on citation features.

Table 6: Comparing feature selection methods.

Insert Table 6 Here

We tested if combining several variants of the same index can improve the predictive performance of the AAAI task. Table 7 presents the results obtained by using: a. a single WoK-based h-index using all papers (i.e., data source=WoK; paper type=All, citing paper type=All, self-citation level=0); b. a single GS- based index using all papers; c. a combination of all WoK-based h-index variants; d. a combination of all GS-based h-index variants; e. a combination of all h-index variants. The results indicate that the F-measure is improved by more than 5% when the variants of the same index are combined.

Insert Table 7 Here

Table 7: Illustrating how combining variants of the same index can improve predictive performance

We examined if the combination of the three types of classifiers actually improves the predictive performance. Table 8 presents the predictive performance obtained by each model separately and by combining them into one model. As can be seen, combining the models improved the performance of the F-Measure by 4%.

Insert Table 8 Here

Table 8: Comparing the performance of various models

Note that we classify each candidate as a fellow or non-fellow in each year separately, from 1995 to 2009. For each year, the candidate has a different profile snapshot and thus the classifier may assign her a different fellowship probability. Among the candidates, we also examined the actual fellows. The earliest year in which the model assigns a fellowship probability that is greater than 0.5 to a fellow is considered to be the predicted year. This can be smaller or greater than the actual year. We measured the difference between the first year the model classified a researcher as a winner and the year that she actually won. This measurement indicates the deviation of our classifier from the optimum.

Figure 3 presents the time lag in years in comparison to the actual time of winning. The x-axis represents the time lag in years. Negative values represent an earlier winning declaration and positive values represent a delay. We can see that 31% of the researchers were classified as winners too early, and 55% were classified too late. 13% of the researchers were classified for the same year they actually won the award. However, the classification of most of the researchers, 76%, was characterized by a lag of four years. In the next section we investigate this point in greater depth.

Insert Figure 3 Here

Figure 3: Time lag in years.

To summarize, our best classifier offers a clear improvement over other models. The combination of all the features presents the best results, much better than the accepted measurements. About 92% of the researchers who won the AAAI award were classified as such by the model. Moreover, only 2% of the non-fellow researchers were classified as fellows (false positive rate). A more profound analysis indicates that 56% of the researchers that won the award but were never classified as winners (false negative), actually won in the

last 3 years (2007 to 2009). The time lag explains the reason that the model did notclassify them as winners. In addition, as shown in Figure 3, most of the errors had a lag of only a few years, and in fact, the peak of the graph obtains the value 0 (where the model is exactly right).

4.2 Predicting the Next Winners

In the previous experiment, we explored the question of whether researcher *r* deserved to receive the award. In the second set of experiments, we attempted to determine who was going to win the award next year. When predicting a ranking for the year *y*, the training set included the data on all the researchers until that year (but not including it); the testing set included the data for the year *y*. For example, when computing the ranking for 2003, all the data from the years 1995-2002 was used for the training set and the data of 2003 was used for the testing set. By dividing the data in this way, we simulated real scenarios because when trying to determine the winners for 2003, we could only know what happened until 2002. This experiment was performed for the last 10 years (2000-2009).

When testing a researcher *r* in year *y*, the classification model returned the probability of every researcher winning the award. We ranked the researchers by sorting them in decreasing order according to their probabilities.

To verify the accuracy of the ranking for a specific year *y*, we checked the position of the actual winners in year *y* in the ranking. Assuming there are *m* winners in year *y*, a perfect accuracy is given in case all the winners are located in the first *m* positions of the ranking scale. For the accuracy metric, we defined the variable *CurrentWinners* which is associated with every position in the ranking.

CurrentWinners: indicates the number of researchers who won the award in year *j* and are ranked in the positions 1 to *i*.

The higher the value of *CurrentWinners (i,j)*, the better the accuracy of the ranking. Figure 4 presents the value of *CurrentWinners* for the years 2003, 2006 and 2009, correspondingly. The x-axis represents the number of positions and the y-axis is the number of winners. The upper curve represents the values of *CurrentWinners* in an optimal ranking while the middle curve represents the values of our ranking. We compared these values to a baseline random ranking presented as the diagonal curve. In our model we selected the top candidates that have the best odds of becoming a fellow. The winning probability estimation was provided by the trained model. On the other hand, in the random model we simply assumed that all candidates have the same probability of winning. Thus the top candidates are randomly selected without replacement, as if in a lottery. This random model simulateda situation in which we have no bibliographic knowledge about the candidates. Obviously, a random curve grows linearly since the positions of the winners are uniformly distributed.

Insert Figure 4 Here

Figure 4: CurrentWinners for year 2003, 2006 and 2009.

We can see that our prediction model is much better than the random model. To examine the accuracy of our ranking, we calculated the AUC (area under curve) of each graph, using the trapezoidal rule, normalizing it by the AUC of the optimal ranking. We compared it to the

normalized AUC of a random ranking. The results of the normalized AUC measurement are presented in Figure 5. The x-axis represents the years and the y-axis represents the normalized AUC. As can be seen, our prediction model is always much more accurate than the random model and is close to the optimum. On average, it is 86% of the optimum, while the random model is only 50%.

Insert Figure 5 Here

Figure 5: Normalized AUC of the CurrentWinners, comparing our prediction model and a random model.

Unfortunately, predicting the AAAI Fellows accurately is not realistic, since, in reality, two researchers with exactly the same bibliographic data, will win the award in approximately the same year, but with a deviation of one or two years. Since we want to consider the correct classification despite a mistake in the correct year, we defined also a similar variable *FutureWinners* that gives a score to correct classifications in future years rather than only to the current year:

FutureWinners (i,j) indicates the number of researchers who won the award in a year greater than *j*, and are ranked in the positions 1 to *i*.

Again, the higher the value of *FutureWinners(i,j)*, the better accuracy of the ranking. Figure 6 presents the results of the normalized AUC measurement for the *FutureWinners*, compared to a random model. We can see that in the course of the years, our model is always much better than a random model. From 2004, it is even 1.5 times more accurate than the random model. Moreover, it is interesting to see that the accuracy of predicting the *FutureWinners* increases in the course of the years since our prediction model is improved by learning from more data.

Insert Figure 6

Figure 6: Normalized AUC of the *FutureWinners*, comparing our prediction model and a random model.

5. Summary and Future Work

In this paper we adopt existing bibliometric indices and off-the-shelf machine learning techniques to identify outstanding AI researchers. Our main contribution focuses on putting the right pieces together, including the idea of combining social network data with bibliometric indices and empirically demonstrating the potential usefulness of the proposed configuration. In particular, we show that combining various bibliometric index variants by generating a machine learning committee can improve the predictive performance. We empirically evaluated our approach via three sets of experiments on researchers from the AI field. In the first experiment we trained a classifier to classify researchers as AAAI fellows. We showed that a classifier which uses both simple bibliographic measures and citation-based indices reduces the false negative rate the most. We examined the improvement of the classifier by using authorship graph parameters. We found that a classifier that uses solely authorship graph parameters produces a high false negative rate. However, adding such parameters to the classifier that we presented in the first experiment significantly improves the classifier. In the second experiment we tried to predict the next AAAI winner. We showed that our prediction model is much better than a random model.

In the future we plan to investigate in greater depth the influence of the authorship network on the evaluation of researchers. In addition to the number of citations, we would like to consider the ranking of the researcher who has been cited and his authorship graph. In addition, we plan to examine the influence of the publication types on researcher evaluation. In many cases

only journal papers are considered; we would like to address the impact of journal papers on the evaluation of researchers and whether we should consider conference papers too.

References

1. Ali, K.M. and Pazzani, M.J. (1996), Error reduction through learning multiple descriptions, Machine Learning.
2. Batista, P.; Campiteli, M.; and Kinouchi, O. (2006),Is it possible to compare researchers with different scientific interests? *Scientometrics*68(1):179–189.
3. Bornmann, L., & Daniel, H. D. (2007),What do we know about the h-index? Journal of the American Society for Information Science and Technology, 58(9), 1381-1385.
4. Bornmann, L., Mutz, R., & Daniel, H. D. (2008),Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. Journal of the American Society for Information Science and Technology, 59(5), 830-837.
5. Breiman, L., (2001), Random forests,Machine learning, 45(1):5-32.
6. Chawla, N.V. and Japkowicz, N. and Kotcz, A. (2004), Editorial: special issue on learning from imbalanced data sets, ACM SIGKDD Explorations Newsletter, 6(1):1-6.
7. Cohen W. W. (1995), Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123.
8. Egghe, L. (2006), Theory and practice of the g-index. *Scientometrics*69:131–152.
9. Fayyad U. M.,Irani K. B. (1993), Multi-interval discretization of continuousvalued attributes for classification learning. In: Thirteenth International Joint Conference on Articial Intelligence, 1022-1027.
10. Feitelson, D. G., and Yovel, U. (2004), Predictive ranking of computer scientists using citeseer data. *Journal of Documentation* 60:44–61.
11. Freeman, L. C. (2004),*The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press.

12. Frolik, J. and Abdelrahman, M. and Kandasamy, P., (2001), A confidence-based approach to the self-validation, fusion and reconstruction of quasi-redundant sensor data, IEEE Transactions on Instrumentation and Measurement, 50(6):1761-1769.
13. García S., Fernández A., Luengo J., and Herrera, F. (2010), Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Information Sciences, 180(10):2044-2064.
14. García-Pérez, M. A. (2010), Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. Journal of the American Society for Information Science and Technology, Volume 61, Issue 10, Article first published online: 2 JUN 2010
15. García-Pérez, M. A. (2011), Strange attractors in the Web of Science database, Journal of Informetrics, Volume 5, Issue 1, January 2011, Pages 214-218
16. Harzing, A.-W. (2010),*The Publish or Perish Book*. Tarma Software Research Pty Ltd.
17. Hirsch, J. E. (2005),An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46):16569–16572.
18. Jensen, P.; Rouquier, J.-B.; and Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*78(3):467–479.
19. Jin, B. (2007),Thear-index: complementing the h-index. ISSI Newsletter 3:6.
20. Jin, B., Liang, L., Rousseau, R., &Egghe, L. (2007), The R- and ARindices: Complementing the h-index. Chinese Science Bulletin, 52(6), 855–863.
21. Kan M.Y. and Tan Y. F., (2008), Record Matching in Digital Library Metadata. In *Communications of the ACM (CACM)*, Volume 51, Issue 2, pages 91-94, February 2008

22. Kira, K. and Rendell, L.A., (1992), The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of Ninth National Conference on Artificial Intelligence, 129–134.

23. Koren Y, (2009), The BellKor solution to the Netflix Grand Prize. Available at http//:www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf;

24. Kretschmer, H., (2004), Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. Scientometrics, 60(3), 409–420.

25. Levine-Clark M. and Gil E.L. (2009), A comparative citation analysis of Web of Science, Scopus, and Google Scholar, *Journal of Business and Finance Librarianship*14 (1): 32–46

26. Ley M. and Reuther P., (2006), Maintaining an online bibliographical database: The problem of data quality. In EGC'2006, 2006, 2 Volumes, pages 5-10, 2006.

27. Li, J.; Sanderson, M.; Willett, P.; Norris, M.; and Oppenheim, C. (2010), Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. *J. Informetrics*4(4):554–563.

28. Liu, X. and Kaza, S. and Zhang, P. and Chen, H., (2011), Determining inventor status and its effect on knowledge diffusion: A study on nanotechnology literature from China, Russia, and India, Journal of the American Society for Information Science and Technology, Volume 62, Issue 6, pages 1166–1176.

29. Meho, L. I., & Yang, K. (2007), Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. Journal of the American Society for Information Science and Technology, 58(13), 2105-2125.

30. Mengle, S.S.R. and Goharian, N. (2009), Ambiguity measure feature-selection algorithm, Journal of the American Society for Information Science and Technology, 60(5): 1037-1050.

31. Mitchell T. (1997), Machine Learning, McGraw-Hill.

32. Newman, M. E. J. (2001),The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America* 98(2):404–409.

33. Oppenheim, C. (2007),Using the h-index to rank influential British researchers in information science and librarianship. Journal of the American Society for Information Science and Technology, 58(2), 297-301.

34. Patterson D., Snyder L., Ullman J. (1999), Computing Research Association, "Best Practices Memo: Evaluating Computer Scientists and Engineers for Promotion and Tenure," Computing Research News, vol. 11, no. 4, Sep. 1999, pp. A-B.

35. Polikar R., (2006), Ensemble based systems in decision making, IEEE Circuits and Systems Magazine 6 (3): 21–45.

36. Ruane, F., and Tol, R. (2008), Rational (successive) h -indices: An application to economics in the Republic of Ireland. *Scientometrics*.

37. Rokach L. (2010), Ensemble-based classifiers, Artificial Intelligence Review, 33(1):1-39

38. Rokach L., Maimon O.,Arbel R. (2006), Selective voting - Getting more for less in sensor fusion, *International Journal of Pattern Recognition and Artificial Intelligence*, 20(3):329-350

39. Schreiber, M. (2008),To share the fame in a fair way, h m modifies h for multi-authored manuscripts. *New Journal of Physics* 10(4):040201.

40. Setiono R. and Liu H., (1995), Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence.

41. Sidiropoulos, A.; Katsaros, D.; and A., B. Y. M. (2007), Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*72:253–280.

42. Vinkler, P. (2009),The g-index: a new indicator for assessing scientific impact. *J. Inf. Sci.* 35:602–612.

43. Zhang, C.-T. (2009), The e-Index, Complementing the h-Index for Excess Citations. *PLoS ONE* 4(5): 1-4.

44. Zhang, C.-T. (2010), Relationship of the h-index, g-index, and e-index. Journal of the American Society for Information Science and Technology, 61:625–628.

Table 1: False negative and false positive of different classifiers.

Features	FP Rate	FN Rate	Precision	Recall	F-Measure
None (All Negative)	0%	100%	NA	0%	NA
None (All Positive)	100%	0%	32%	100%	48%
Top 50	2%	8%	96%	92%	94%
Top 100	4%	13%	91%	87%	89%
Top 200	4%	17%	90%	83%	86%
All Features	5%	14%	89%	86%	87%

Table 2: Comparing Machine Committee Model with Ripper Classification Rules

Features	FP Rate	FN Rate	Precision	Recall	F-Measure
Top 50 Using Multiple-Classifiers	2%	8%	96%	92%	94%
Top 50 Using Decision Rules	6%	16%	87%	84%	85%

Table 3: Analysis of Social Features

Features	FP Rate	FN Rate	Precision	Recall	F-Measure
Social Features Only	7%	52%	77%	48%	59%
All Features (Including Social)	5%	14%	89%	86%	87%
All Features but Social Features	8%	21%	82%	79%	81%

Table 4: Comparing various subsets of features.

Features	FP Rate	FN Rate	Precision	Recall	F-Measure
All Features but Social Features	8%	21%	82%	79%	81%
Simple Bibliographic Measures Only	10%	20%	80%	80%	80%
Citation-based indices Only	6%	24%	85%	76%	80%
All h-index Variants	6%	47%	80%	53%	64%
Number of Publications Variants	27%	45%	49%	55%	52%
g-index (WoK, Paper=AI,Citing=Journal,Self-Citation=Level 0)	5%	42%	84%	58%	68%

Table 5: Top ten features.

Feature Category	Feature	Source	Manuscript Type	Citing Manuscript Type	Self-Citation Level
Citation Based	g-Index	WoK	AI	Journal	0
Social Based	# Fellows whose distance < 5	DBLP	All	---	---
Citation Based	g-Index	Google	All	AI	2
Citation Based	Norm Individual h-index	WoK	AI	Journal	0
Simple Bibliographic Based	# Individual Publications	Google	AI	---	---
Citation Based	Norm Individual h-index	WoK	AI	Journal	2
Citation Based	Schreiber Individual h-index	WoK	AI	AI	1
Social Based	The average path length	DB	All	---	---
Citation Based	Norm Individual h-index	WoK	Journal	Journal	2
Citation Based	Rational H Index	Google	Journal	Journal	2

Table 6: Comparing feature selection methods.

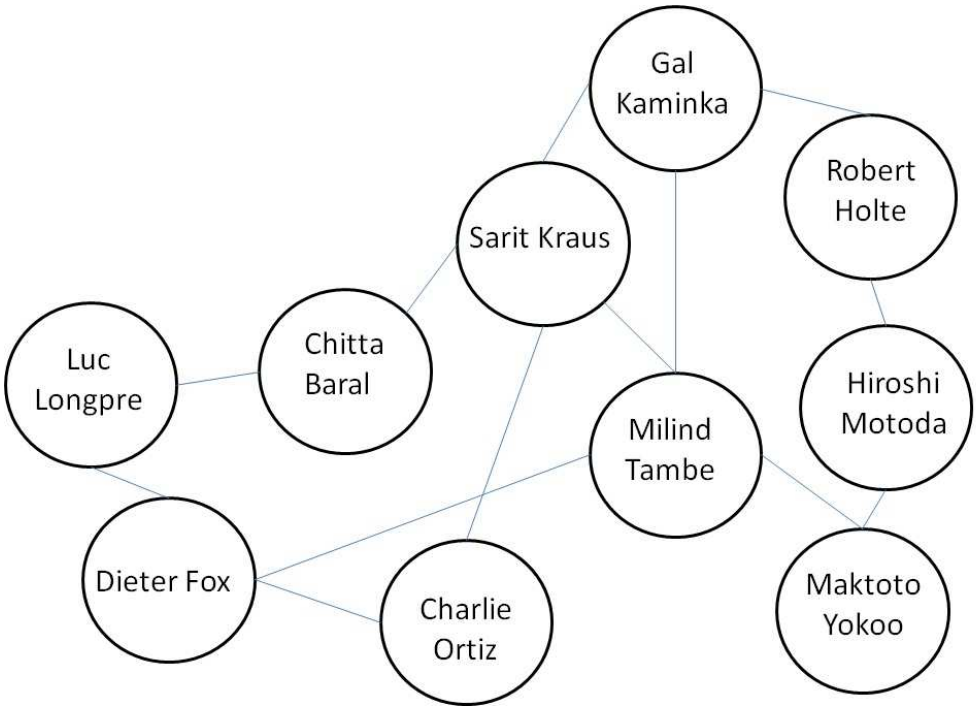
Features	FP Rate	FN Rate	Precision	Recall	F-Measure
Top one citation index	9%	23%	81%	77%	79%
Top one bibliographic measure	12%	24%	75%	76%	76%
Top one social indicator	10%	55%	67%	45%	54%
Top 5 citation indices	8%	21%	83%	79%	81%
Top 5 bibliographic measures	10%	18%	79%	82%	80%
Top 5 social indicator	7%	50%	77%	50%	61%
Joining the 5 top of all categories	5%	11%	89%	89%	89%
Top 15 features selected from all categories	4%	10%	91%	90%	91%

Table 7: Illustrating how combining variants of the same index can improve predictive performance

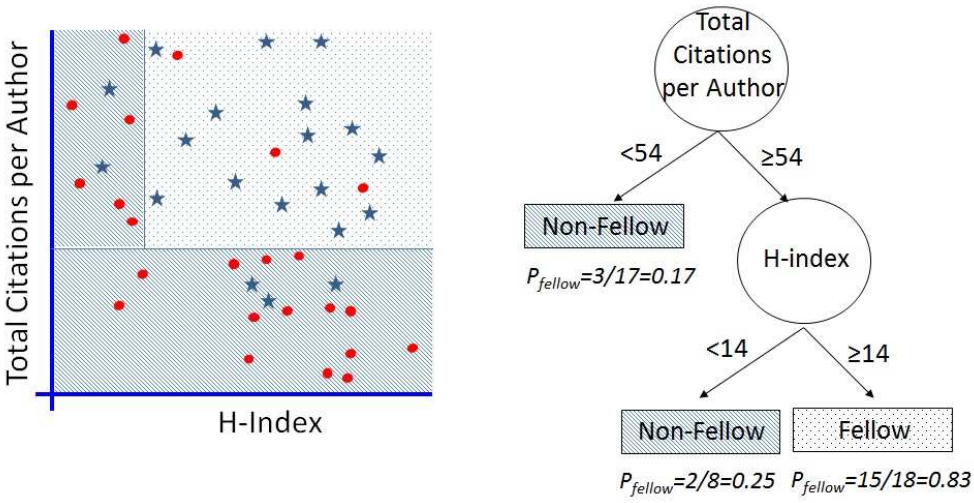
Features	FP Rate	FN Rate	Precision	Recall	F-Measure
A single WoK-based h-index variant using all papers	8%	51%	74%	49%	59%
A single GS-based h-index variant using all papers	10%	59%	66%	41%	51%
All WoK-based h-index Variants (27 variants)	8%	51%	74%	49%	59%
All GS-based h-index Variants (27 variants)	7%	49%	77%	51%	61%
All h-index Variants (54 variants)	6%	47%	80%	53%	64%

Table 8: Comparing the performance of various models

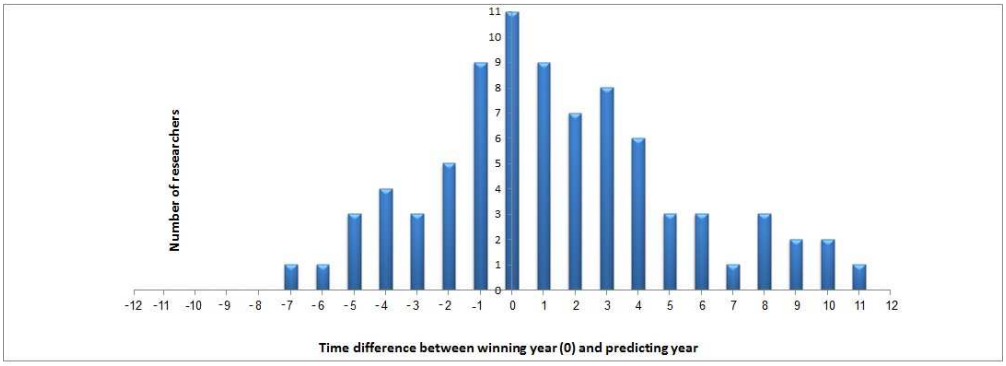
Model	FP Rate	FN Rate	Precision	Recall	F-Measure
AdaBoost	4%	12%	92%	88%	90%
Logistics Regression	6%	14%	88%	86%	87%
Multilayer Perceptron	7%	17%	85%	83%	84%
Combined	2%	8%	96%	92%	94%



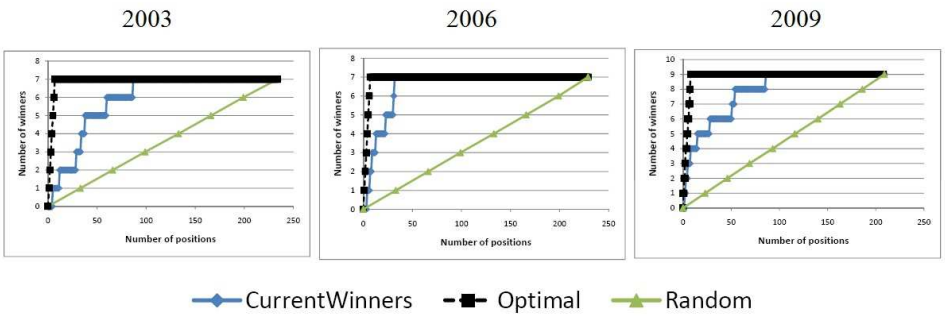
256x179mm (96 x 96 DPI)



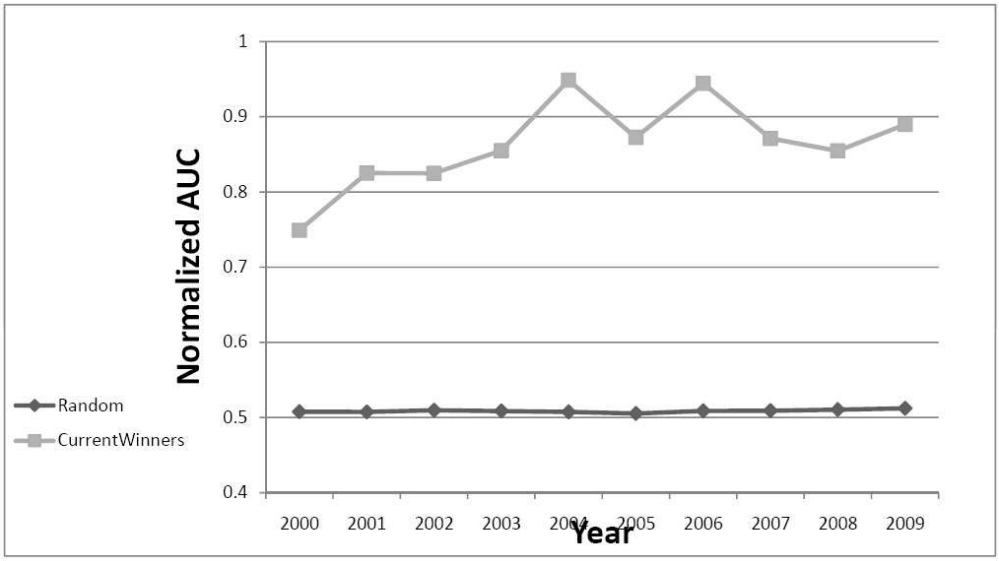
256x131mm (96 x 96 DPI)



287x131mm (96 x 96 DPI)



332x172mm (96 x 96 DPI)



269x153mm (96 x 96 DPI)

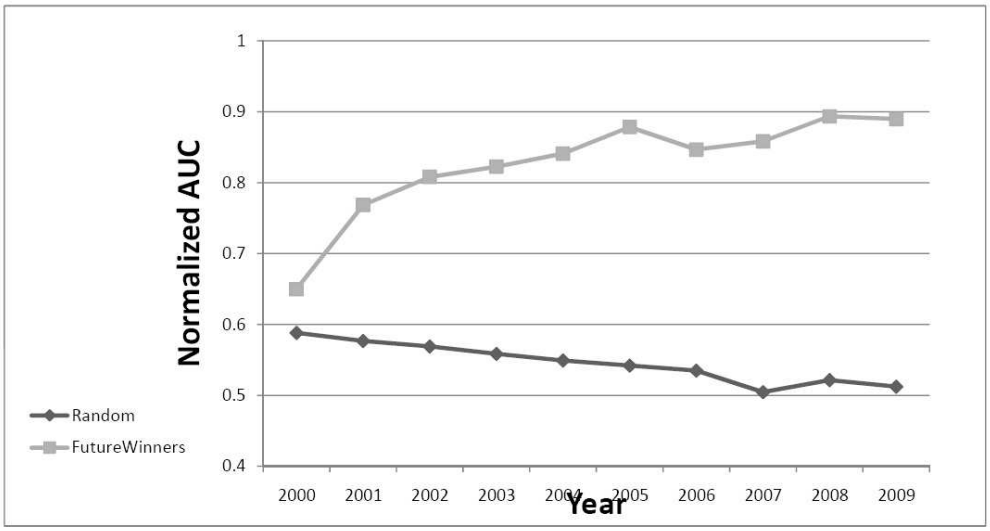


Figure 6
285x153mm (96 x 96 DPI)