Journal of Informetrics 6 (2012) 586–599

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Development a case-based classifier for predicting highly cited papers

Mingyang Wang[a,b,*], Guang Yu[b], Jianzhong Xu[c], Huixin He[d], Daren Yu[d], Shuang An[e]

[a] College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, PR China
[b] School of Management, Harbin Institute of Technology, 150001, PR China
[c] School of Economics and Management, Harbin Engineering University, 150001, PR China
[d] School of Power Engineering, Harbin Institute of Technology, 150001, PR China
[e] School of Information and Computer Science, Northeastern University at Qinhuangdao, 066004, PR China

ARTICLE INFO

Article history:
Received 17 February 2012
Received in revised form 18 May 2012
Accepted 20 June 2012

Keywords:
Highly cited papers
Prediction
Case-based classifier

ABSTRACT

In this paper, we discussed the feasibility of early recognition of highly cited papers with citation prediction tools. Because there are some noises in papers' citation behaviors, the soft fuzzy rough set (SFRS), which is well robust to noises, is introduced in constructing the case-based classifier (CBC) for highly cited papers. After careful design that included: (a) feature reduction by SFRS; (b) case selection by the combination use of SFRS and the concept of case coverage; (c) reasoning by two classification techniques of case coverage based prediction and case score based prediction, this study demonstrates that the highly cited papers could be predicted by objectively assessed factors. It shows that features included the research capabilities of the first author, the papers' quality and the reputation of journal are the most relevant predictors for highly cited papers.

1. Introduction

The phenomenon of why some papers become highly cited while the vast majority of papers remain infrequently cited or uncited is a persistent question in the field of scientometrics. During the last decades, works have been done to explore the reason for this phenomenon: (i) The citation motivations of quoters were widely discussed (Bornmann & Daniel, 2008; Case & Higgins, 2000; Hewings, Lillis, & Vladimirou, 2010; Kim, 2004; Laband & Piette, 1994; Rong & Martin, 2008). But the quoters' citation motivations are subjective to a large extent, which makes it difficult to properly monitor the citation trend of papers. (ii) Some mathematical-statistical models were established to predict papers' future citation behaviors. Glänzel and Schubert (1995) presented a non-homogeneous birth-process model, and further discussed the statistical reliability of the model (Glänzel, 1997). Burrell presented a series of stochastic models in the presence of obsolescence to predict the future citation pattern of individual papers (Burrell, 2001, 2002a, 2002b, 2003). (iii) The bibliometric factors that have influences on papers' citation activities were widely investigated. Features associated with the authors (age, gender, social status, etc.), the papers (collaboration, document type, subject matter, etc.), and the journals (impact factor, etc.) were discussed (Bornmann & Leydesdorff, 2012; Danell, 2011; Fu & Aliferis, 2010; Levitt & Thelwall, 2008; Moed, 2010; Penas & Willett, 2006; Sagi & Yechiam, 2008; Xia, Myers, & Wihoite, 2011). However, these features are mainly the external bibliometric features of papers. The factors on papers' *quality*, which could be the kernel features dominate papers' citation activity, are left alone because of lacking an appropriate way to quantify it. Van Dalen and Henkens (2005) stated that the *quality* of paper could be approximated by the *impact* and *speed* with which knowledge is disseminated in the scientific community. The *impact* of

* Corresponding author at: College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, PR China.
E-mail address: wangmingyang@nefu.edu.cn (M. Wang).

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2012.06.002

one paper boils down to the number of citations registered by the Web of Science (of ISI). The *speed* with which one paper is disseminated in the scientific community is measured by the timing of the first citation.

Recently, we discussed the role of papers' quality on citation counts prediction based on papers published in the field of astronomy and astrophysics in 1980 (Wang, Yu, & Yu, 2011). The *quality* of papers is also measured by its impact and speed dimensions. The impact of one paper is expressed by its knowledge diffusion properties in our work. And the speed of one paper is measured by its first-cited properties, including its first-cited age and the citations obtained in its first cited year. By using the technique of multi-classifier system, we found that the papers' quality plays an important role on predicting papers' future citation activity.

The present experiments performed two additional analyses to extend our preliminary work. First, a new independent data was used. The papers used for our experiments were extracted from four journals in four different fields. It could be more generality than our previous investigation that confine the experiments to a single subject. Second, another modeling technique of case-based classifier (CBC) was introduced. It is well-known that there exist some noises in papers' citation activities. A robust mechanism is needed to make better performance for citation prediction. Recently, we found that the soft fuzzy rough set (SFRS) works well to deal with noisy samples (Hu, An, & Yu, 2010). Thus, the model of CBC in our experiments was constructed based on the hybrid use of case-based reasoning (CBR) and SFRS. And SFRS is used in the three kernel processes of CBR: feature reduction, case selection, and reasoning. If the predictors for highly cited papers are similar with our preliminary experiment when using different data set and modeling techniques, this result would indicate that the predictors are credible and they depend more on the choice of features rather than the choice of data and classifiers.

The remainder of this paper is organized as follows. Firstly, the related techniques about CBR and SFRS are given. Secondly, the CBC model integrating of CBR with SFRS is introduced. Finally, the experimental results are shown, and the typical features were extracted out for highly cited papers.

2. Methods

2.1. Case-based reasoning (CBR)

CBR is an instance-based learning methodology for problem solving (Jorgenson, 2004). The basic assumption of CBR is that similar experiences can guide future reasoning, problem solving, and learning (Symth & Keane, 1998). It is the same way as human being's dealing with problems in daily life. There are three kernel steps in CBR:

(i) Feature reduction: The performance of a CBR classifier system can be significantly diminished by using too many input features. And the need to reduce the number of features for CBR is well discussed (Kupinski & Giger, 1999). It means that optimal features can help it produce better performance. So far, a number of algorithms have been developed for feature reduction (Hu, Yu, & Xie, 2006; Hu, Xie, & Yu, 2007; Jensen & Shen, 2009; Kwak & Choi, 1994). But it is well known that the data in real-world application are usually corrupted by noise, which brought great negative influences on the performance of CBR. In this regard, a robust reduction technique or algorithm is desirable in practice.

(ii) Case selection: When given a new case to be classified, CBR searches from case base for similar cases and composes predictive result on the basis of class labels of similar cases to the new case. Thus, the performance of CBR is also sensitive to the choice of case base. Variety methods for case selection have been developed, such as k-NN (Hart, 1968) and its improved algorithms (Gates, 1972; Tomek, 1976), case coverage and reachability (Cloete & Zyl, 2006; Li, Shiu, & Pal, 2006b), and so on (Jin, Liu, & Hou, 2010; Mitra, Murthy, & Pal, 2002). But a robust technique to identify and remove the redundant and noisy cases in real-world application is still needed.

(iii) Reasoning: Reasoning is to select a method to classify unseen cases with selected cases. The concept of case coverage is a widely used reasoning way for unseen cases (Li, Shiu, & Pal, 2006a; Li et al., 2006b). But it can only classify those unseen cases located in the range that the selected cases cover. An effective technique should be incorporated to evaluate the unseen cases out of the coverage in reasoning process.

Recently, we found that the soft fuzzy rough set works well to deal with noisy samples (Hu et al., 2010). The soft fuzzy rough set relies on the soft distance, which is distinguished with the statistical minimum distance in other rough set models. It makes SFRS to be well robust to outliers. In this paper, SFRS is introduced into CBR classification process to help better recognizing of suitable features and cases, and better classifying of unseen cases.

2.2. Soft fuzzy rough set (SFRS)

Here, some definitions for rough sets and fuzzy rough sets were given in order to better understand the mechanism of soft fuzzy rough set. More detailed discussion could be found in articles (Dubois & Prade, 1990; Pawlak, 1982), where the rough set and fuzzy rough set were first introduced.

Given a finite set of objects $U = \{x_1, x_2, \ldots, x_n\}$ described with a set of attributes $A = \{a_1, a_2, \ldots, a_m\}$, each object $x_i \in U$ can be formulated as a vector $x_i = \langle x_{i1}, x_{i2}, \ldots, x_{im}\rangle$, where x_{ij} is the j'th feature value of sample x_i. The set of $IS = \langle U, A\rangle$ is called an information system.

Pawlak's rough set model takes into consideration the indiscernibility between objects. The indiscernibility is typically characterized by an equivalence relation.

Definition 1. $\forall B \subseteq A$, an equivalence relation can be generated over the universe:

$$IND(B) = \{\{x, y\} \in U^2 | \forall a \in B, a(x) = a(y)\} \tag{1}$$

$IND(B)$ is called the B-indiscerniblility relation. If $\{x, y\} \in IND(B)$, then objects x and y are B-indiscernible. With $IND(B)$, U is partitioned into a family of equivalence classes of $U/IND(B)$ (or U/B). And the equivalence class of sample xi induced by B-indiscerniblility relation is denoted by $[xi]B : [x_i]_B = \{y \in U | \{x_i, y\} \in IND(B)\}$, which is a subset of samples having the same feature values as xi.

Given a classification task, the class labels of the objects are known in advance. Let X be a subset of objects belonging to the same class. The lower and upper approximations of X with respect to B can be defined as:

$$\begin{cases} \underline{B}X = \{x_i \in U | [x_i]_B \subseteq X\} \\ \overline{B}X = \{x_i \in U | [x_i]_B \cap X \neq \phi\} \end{cases} \tag{2}$$

The lower approximation of X consists of the samples whose equivalence classes consistently belong to X, while its upper approximation is the subset of samples whose equivalence classes have objects in X. If $\underline{B}X \neq \bar{B}X$, the approximating boundary of X is computed as $BND_B(X) = \bar{B}X - \underline{B}X$. It contains those objects that we cannot decisively classify into X on the equivalence classes of B.

Pawlak's rough set model could only deal with the features with discrete values. In practice, most of classification tasks are described with numerical features or fuzzy information. In this case, fuzzy similarity relations are used.

Definition 2. Given a nonempty universe U, R is a fuzzy binary relation on U if R satisfies:

(1) Reflexivity: $R(x, x) = 1$,
(2) Symmetry: $R(x, y) = R(y, x)$,
(3) Sup-min transitivity: $R(x, z) \geq T\{R(x, y), R(y, z)\}$.

We say R is a fuzzy similarity relation. Here, T is the abbreviation of triangular norm, which is a binary operation on interval [0, 1] satisfying the following conditions: (i) Commutativity: $T(x, y) = T(y, x)$; (ii) Associativity: $T(x, T(y, z)) = T(T(x, y), z)$; (iii) Monotonicity: $y \leq z \Rightarrow T(x, y) \leq T(x, z)$; (iv) The number 1 acts as the identity element: $T(x, 1) = x$. Fuzzy similarity relations are used to measure the similarity of the objects characterized with continuous features. The fuzzy equivalence class $[x]_R$ associated with x and R is a fuzzy set on U, where $[x]_R(y) = R(x, y)$ for all $y \in U$. Based on fuzzy equivalence relations, the fuzzy lower and upper approximations were defined.

Definition 3. Let U be a nonempty universe, R be a fuzzy similarity relation on U and $F(U)$ be the fuzzy power set of U. Given a fuzzy set $A \in F(U)$, the lower and upper approximations are defined as:

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y \in U} \max\{1 - R(x, y), A(y)\} \\ \bar{R}A(x) = \sup\limits_{y \in U} \min\{R(x, y), A(y)\} \end{cases} \tag{3}$$

In our experiments, A stands for the set representing the papers' three levels {HCPs, MCPs, LCPs}, which is a crisp set:

$$A(y) = \begin{cases} 1, & y \in A \\ 0, & y \notin A \end{cases}$$

Accordingly, the fuzzy lower and upper approximations in Eq. (3) becomes:

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y \in U - A} \{1 - R(x, y)\} \\ \bar{R}A(x) = \sup\limits_{y \in A} R(x, y) \end{cases} \tag{4}$$

For each sample $x \in U$, its fuzzy lower approximation to A is the dissimilarity between x and the nearest sample $y \notin A$. And the fuzzy upper approximation to A is the similarity between x and the nearest sample $y \in A$. If the Gaussian function is introduced to compute the similarity $R(x, y)$:

$$R(x, y) = \exp\left(\frac{-\|x - y\|^2}{\delta}\right) \tag{5}$$

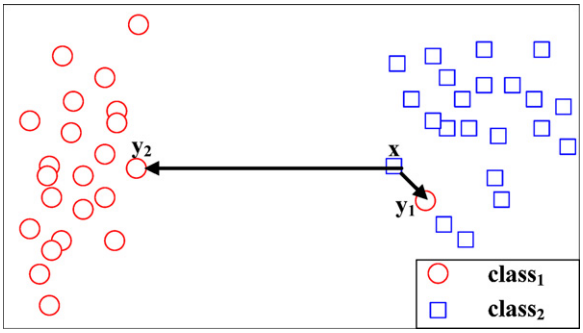

Fig. 1. The influence of noise on the membership of x to the fuzzy lower approximation of the class.

Then $1 - R(x, y)$ could be taken as a general distance function $d(x, y)$ between x and y. And the lower and upper approximation of fuzzy rough set becomes:

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y \in U-A} \{d(x, y)\} \\ \bar{R}A(x) = \sup\limits_{y \in A} \{1 - d(x, y)\} = 1 - \inf\limits_{y \in A} \{d(x, y)\} \end{cases} \tag{6}$$

The lower approximation of fuzzy rough set $\underline{R}A(x)$ is the distance between x and the nearest sample $y \notin A$. And the upper approximation of fuzzy rough set $\bar{R}A(x)$ is the similarity between x and the nearest sample $y \in A$.

However, the lower approximation of fuzzy rough set is not robust to outliers. Taking Fig. 1 as an example, the membership of object x to the fuzzy lower approximation of $class_2$ is the distance between x and y_1, that is $d(x, y_1)$. But Fig. 1 shows that y_1 is a noisy sample. Assuming that y_1 does not exist, the fuzzy lower approximation of x to $class_2$ will be the distance between x and y_2. Obviously, the distance of $d(x, y_2)$ can better and correctly represent the membership of fuzzy lower approximation. Thus, the existence of noisy samples alters the lower approximation of a class, and then deteriorates the classification results.

It is well-known that there exist some noises in papers' citation activities. For instance, some works suggested that the highly cited papers tend to be published in high impact journals (Van Dalen & Henkens, 2001, 2005), but there are also papers present in poorly impact journals. The noisy samples would disturb the citation trend analysis. The SFRS could help to eliminate the noises. It relies on the soft distance, which is distinguished with the statistical minimum distance in fuzzy rough set.

Definition 4. Given an object x and a set of objects Y, the soft distance between x and Y is defined as:

$$SD(x, Y) = \underset{d(x,y) y \in Y}{arg\ \sup} \{d(x, y) - \beta m_Y\} \tag{7}$$

The main idea of soft distance is to enlarge the distance by neglecting noisy samples. Taking Fig. 1 as an example, the distance between object x to $class_2$ should be $d(x, y_1)$ in fuzzy rough set as discussed above. Though y_1 is an outlier, the fuzzy rough set cannot neglect it. But for soft distance, y_1 can be ignored and the soft distance would be $d(x, y_2)$.

It is a problem that how many samples should be taken as noises and neglected? In the definition of soft distance, the penalty factor β is used to control the number of overlooked samples. If we overlook one sample, $d(x, y)$ will minus β. Parameter $m_Y = |\{y_i | d(x, y_i) < d(x, y)\}|$ shows the number of overlooked samples. If $d(x, y') - \beta m_Y (y' \in Y)$ is the largest of $\{d(x, y) - \beta m_Y (\forall y \in Y)\}$, the distance $d(x, y')$ would be taken as the soft distance between x and Y. We also discussed the value domain of β, and found that $\beta = 0.1$ is a good choice (Hu et al., 2010). It means that if the soft distance increases 0.1, there's one sample at most is taken as outlier and neglected. In the present experiments, β is also assigned as 0.1.

Based on the soft distance, the soft fuzzy rough set is defined as follows.

Definition 5. Let U be a nonempty universe, R be a fuzzy similarity relation on U and $F(U)$ be the fuzzy power set of U. The soft fuzzy lower and upper approximations of $A \in F(U)$ are defined as:

$$\begin{cases} \underline{R^S}(A)(x) = 1 - R\left(x, \underset{y\ A(y) \le A(y_L)}{arg\ \sup} \{1 - R(x, y) - \beta m_{Y_L}\}\right) \\ \overline{R^S}(A)(x) = R\left(x, \underset{y\ A(y) \ge A(y_U)}{arg\ \inf} \{R(x, y) + \beta n_{Y_U}\}\right) \end{cases} \tag{8}$$

where

$$\begin{cases} Y_L = \{y | A(y) \le A(y_L), y \in U\}, y_L = \underset{y}{arg} \inf\limits_{y \in U} \max\{1 - R(x, y), A(y)\} \\ Y_U = \{y | A(y) \ge A(y_U), y \in U\}, y_U = \underset{y}{arg} \sup\limits_{y \in U} \min\{R(x, y), A(y)\} \end{cases}$$

m_{Y_L} is the number of the samples overlooked in computing the soft fuzzy lower approximation $\underline{R}^S A(x)$, n_{Y_U} is the number of the samples overlooked in computing the soft fuzzy upper approximation $\overline{R^S}A(x)$.

If A is a crisp set, the membership of x to $\underline{R}^S A(x)$ is:

$$\underline{R}^S A(x) = 1 - R(x, y_{AL}) \tag{9}$$

where

$$y_{AL} = \underset{y \ A(y)=0}{arg\ \sup} \{1 - R(x, y) - \beta m_{Y_L}\} = \underset{y \ A(y)=0}{arg\ \sup} \{d(x, y) - \beta m_{Y_L}\} = \underset{y}{arg} SD(x, U - A)$$

Obviously, $\underline{R}^S A(x)$ equals to the soft distance from x to $U - A$.

Similarly, the membership of x to $\overline{R^S}A(x)$ is:

$$\overline{R^S}A(x) = R(x, y_{AU}) \tag{10}$$

where

$$y_{AU} = \underset{y \ A(y)=1}{arg\ \inf} \{R(x, y) + \beta n_{Y_U}\} = \underset{y \ A(y)=1}{arg\ \sup} \{1 - R(x, y) - \beta n_{Y_U}\} = \underset{y \ A(y)=1}{arg\ \sup} \{d(x, y) - \beta n_{Y_U}\} = \underset{y}{arg} SD(x, A)$$

$\overline{R^S}A(x)$ equals to the soft distance between x and the sample that is used to compute the soft distance from x to A.

As discussed in Fig. 1, the soft distance is more robust than the statistical minimum distance. It makes the soft fuzzy rough sets be more robust to noises than the fuzzy one.

In classification learning, it is natural to desire that the membership of each sample belonging to its decision is as large as possible.

Definition 6. Given a decision table $DS = \langle U, C \cup D \rangle$, U is a nonempty universe, C is the set of attributes and D is the decision attribute. For $\forall B \subseteq C$, the membership of an object $x \in U$ belonging to the soft positive region of D on B is defined as:

$$POS_B^S(D)(x) = \sup_{x \in U/D} \underline{B^S}(X)(x) \tag{11}$$

The soft fuzzy dependency of decision D on feature subset B is defined as:

$$\gamma_B^S(D) = \frac{\sum_{x \in U} POS_B^S(D)(x)}{|U|} \tag{12}$$

Dependency is the ratio of the samples in the lower approximation over the universe, which is widely used to measure the classification performance of attributes. A larger dependency of feature subset means that it has better capability to distinguish different classes.

Based on CBR and SFRS, we designed a case-based classifier (CBC) model for prediction highly cited papers. In the model, the technique of SFRS would be utilized in the three kernel steps of CBR to lessen the negative influences of noises.

2.3. A case-based classifier (CBC) based on CBR and SFRS

The proposed CBC model is also composed of three stages of feature reduction, case selection and reasoning.

2.3.1. Feature reduction

Here, a SFRS-based feature reduction algorithm was established. The algorithm employs the soft fuzzy dependency as the feature evaluation function and the sequential forward selection as the search strategy. The output of the algorithm is a feature ranking set $F' = \{f'_1, f'_2, \ldots, f'_{|F''|}\}$. Given the set F'_{k-1} with $k-1$ features selected, the k'th feature is determined by $\max_{f \in F - F'_{K-1}} \{SFD_{\{F'_{K-1} \cup \{f\}\}}(D)\}$. Thus, the feature with the maximum soft fuzzy dependency is extracted out in every circulation. Finally, a sequential of feature subsets of $F'_1 = \{f'_1\}, F'_2 = \{f'_1, f'_2\}, \ldots, F'_{F'} = \{f'_1, f'_2, \ldots, f'_{|F''|}\}$ is generated.

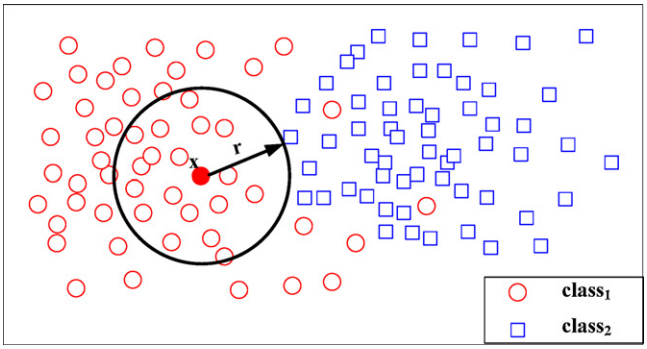

Fig. 2. Sketch for SFRSC.

Algorithm 1 *(Feature reduction algorithm).*

Input: X—the sample set;
F—the original feature set.
Output: F'—the reduction set of F.

Step 1 Initialize $F' = \phi$;
Step 2 Compute the reduction set
While (F is not empty)
1. For each feature $f \in F$
   Compute the soft fuzzy dependency $SFD_{\{F' \cup \{f\}\}}(D)$ of f;
2. Add the feature f' with the maximum soft fuzzy dependency to F':
   $F' = F' \cup \{f'\}$;
   $F = F - \{f'\}$;
3. Return F' and stop.

2.3.2. Case selection

Here, a case selection method based on SFRS was proposed. The concept of case coverage is used. It describes the completeness of a CBR system, which is the range of problems that the system can solve (Li et al., 2006a, 2006b). It helps to identify which cases should be removed and which preserved, with the main goal being to reduce the number of cases while maintaining the classification accuracy.

Definition 7. Given a decision table $DS = \langle U, C \cup D \rangle$, U is a nonempty universe, C is the set of attributes and D is the decision attribute. For any case $\forall x \in U$, the SFRS-based coverage (SFRSC) of x on $B \subseteq C$ is defined as:

$$Cover_B^R(x) = \left\{ x' | S(N(R(x, x'))) \leq \sup_{x \in U/D} \underline{R^S}(X)(x) \right\} \tag{13}$$

Fig. 2 shows the sketch for SFRSC. For case $x \in class_1$, assuming that its soft fuzzy lower approximation is:

$$\left\{ \sup_{x \in U/D} \underline{R^S}(X)(x) \right\} = \underline{R_S} class_1(x)$$

Then the SFRSC of x should cover all the samples located in the sphere centered at x with a radius of $\underline{R_S} class_1(x)$.

Thus, the coverage of a case is the range of problems (cases) which can be solved using this case. And the case with larger coverage should make a larger contribution to the completeness of CBR. Those case(s) whose coverage set is the largest should be selected out first, and this process of case selection continuous until all the cases in the case base are solved using the selected cases.

Algorithm 2 *(Case selection algorithm).*

Input: X—the original case base.
Output: X'—the selected case base.

Step 1 Initialize $X' = \phi$;
Step 2 Compute the SFRSC
While (X is not empty)
1. For each case $x \in X$
   Compute the SFRSC of x;
2. Add the case x' with the maximum SFRSC to X':
   $X' = X' \cup \{x'\}$;
   $X = X - \{x'\}$;
3. Return X' and stop.

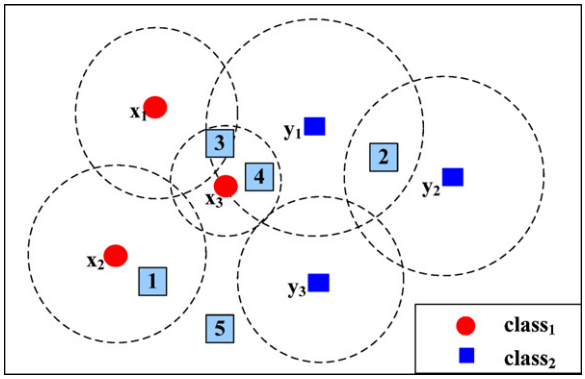

Fig. 3. Sketch for case coverage based prediction.

The feature reduction and case selection process generates a case base with fewer features and cases. Using the constructed case base and the reasoning technique of CBR, the highly cited papers could be predicted.

2.3.3. Reasoning

Here, a combined reasoning method based on the concepts of case coverage and case score is proposed. The main goal of this method is to classify both cases located in and out of the SFRSC.

2.3.3.1. Case coverage based prediction. For the cases located in the SFRSC of selected cases, the class labels to them could be directly determined. And:

(a) If unseen case q is only located in the SFRSC of case p, then q would be appointed to the same class label as p.
(b) If unseen case q is located in several SFRSC from different selected cases, then q would be appointed to the maximum class that most cases belong.
(c) If unseen case q is not located in SFRSC of any selected cases, or q is located in the SFRSC of equal cases from different classes, then the case coverage based prediction could not work.

Fig. 3 shows the sketch for case coverage based prediction. q is an unseen case. x_1, x_2 and x_3 are cases coming from $class_1$. y_1, y_2 and y_3 are cases coming from $class_2$. If q is located at the range labeled as "1", that is, q only belong to the SFRSC of x_2, then q would be classified as $class_1$. If q is located at the range labeled as "2", that is, q belong to the SFRSC of y_1 and y_2, it would be classified as $class_2$. If q is located at the range labeled as "3", that is, q belong to the SFRSC of x_1, x_3 and y_1, it would be classified as $class_1$. But if q is located at the range labeled as "4" or "5", it could not be classified if only considering the cases' coverage. In this situation, the case score based prediction would work better.

2.3.3.2. Case score based prediction. For the cases located outside the SFRSC of any selected cases, or those located in the SFRSC of equal cases from different classes, their classification labels could be determined by case score based prediction algorithm. Every unseen case would be given a score by each of selected cases. The score is the product of two factors:

(a) Similarity (s): It shows the distance between the unseen case and the selected case. Here, Euclidean measure, which is a widely used distance measure, is used to determine the similarity between cases.
(b) Classification certainty (m): It shows the classification capability of the selected cases. Here, it is determined by the membership of the selected case to the soft fuzzy lower approximation of its own class.

Finally, the unseen case would be classified into the class giving highest score.

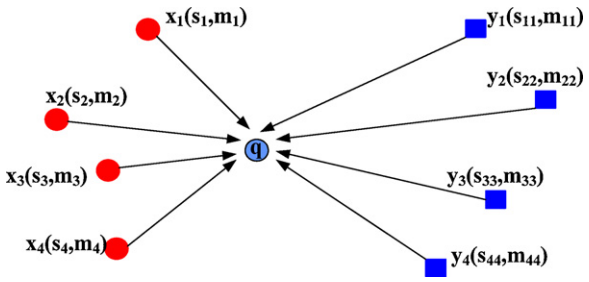

Fig. 4. Sketch for case score based prediction.

Algorithm 3 *(Case score based prediction algorithm).*

Input: *CS*—the selected case set;
x—the unseen case.

Output: *score*—the score gained from each class;
label—the class label of *x*.

Step 1 Computer the number, *C*, of different classes in *CS*.

Step 2 Compute the score from each class
Initialize $score = 0$;
1. For c: 1 to C

$$score_{class_c} = \frac{\sum_{i=1}^{N_c} s_i \times m_i}{\sum_{i=1}^{N_c} m_i}$$

end;
2. $label(x) \leftarrow \arg\max_c \{score_{class_c}\}$;
3. Return $label(x)$.

Fig. 4 shows the sketch for case score based prediction. q is an unseen case. $x_1(s_1, m_1), x_2(s_2, m_2), x_3(s_3, m_3)$, and $x_4(s_4, m_4)$ are cases coming from $class_1$. Where s_i ($i = 1, 2, 3, 4$) is the similarity between x_i and q, and m_i ($i = 1, 2, 3, 4$) is the classification certainty of x_i to $class_1$. $y_1(s_{11}, m_{11}), y_2(s_{22}, m_{22}), y_3(s_{33}, m_{33})$ and $y_4(s_{44}, m_{44})$ are cases coming from $class_2$. Where s_{ii} ($i = 1, 2, 3, 4$) is the similarity between y_i and q, and m_{ii} ($i = 1, 2, 3, 4$) is the classification certainty of y_i to $class_2$. For the unseen case q, every selected case will generate a score for q, that is $s_i \times m_i$ (or $s_{ii} \times m_{ii}$). And, the total score given by $class_1$ should be:

$$score_{class_1} = s_1 \times m_1 + s_2 \times m_2 + s_3 \times m_3 + s_4 \times m_4$$

The total score given by $class_2$ should be:

$$score_{class_2} = s_{11} \times m_{11} + s_{22} \times m_{22} + s_{33} \times m_{33} + s_{44} \times m_{44}$$

Finally, the class label to unseen case q is determined by:

$$label(q) = \arg\max_j \{score_{class_1}, score_{class_2}\}$$

where $j \in \{1, 2\}$. Sometimes, the number of cases from different class may be not equivalent. Then the weighted score would be used:

$$score_{class_j} = \frac{\sum_{i=1}^{N_j} s_i \times m_i}{N_j}$$

where N_j is the number of cases from $class_j$.

Thus, for multi-class problems, the label to the unseen cases q should be determined by:

$$label(q) = \arg\max_j \{score_{class_1}, score_{class_2}, \ldots, score_{class_C}\}$$

where C is the number of classes.

In the following section, the proposed CBC model is used to predict the future highly cited papers.

3. Data

Four representative journals were chosen from four different fields. All the papers published in these four journals in year 1985 were collected basing on the web version of the Science Citation Index (SCI) produced by ISI. Our goal is to develop a three labels' prediction model rather than a continuous one because error metrics for continuous loss functions are difficult to interpret. The model predicts whether a paper would grow up into highly cited papers (HCPs), medium-cited papers

(MCPs) or low-cited papers (LCPs) within 15 years of publication. A paper set with different kinds of citation impact should be established firstly.

3.1. Selection criteria

Aksnes suggested that there are two different approaches for defining HCPs, involving absolute and relative thresholds (Aksnes, 2003). For the absolute one, a fixed threshold is used as a definition. For example, articles cited more than 500 times are defined as HCPs. The limitation of using a fixed threshold is that the highly cited fields generate a predominance of HCPs. Therefore, the relative standard is often adopted to identify the most highly cited papers within each field. A paper was considered as highly cited if it has received more than a certain multiple of the citations of the average paper within the scientific subfield. Such a method of selection has similarities to the method applied by Glänzel, Rinia, and Brocken (1995) in a study of HCPs in physics. In addition, the sample identified should be manageable from a practical point of view, meaning that the number of papers should not be too large (or too small) for carrying out the different surveys.

In this paper, we used a relative standard of 10, which means that a publication has been considered as highly cited if the number of citations received during the time period is at least 10 times the mean citation rate in the particular subfield. The particular threshold of selection is somewhat arbitrary. Another definition or set of criteria would give a different sample. Still, the identified HCPs represent the very top papers in their fields.

For MCPs and LCPs, a relative standard is also used. A publication is been considered as low-cited if the number of citations received is less than the mean citation rate of the particular subfield. And the rest publications in this particular subfield are considered as medium-cited.

3.2. Identifying the sample data

It is a time-consuming work to gather information from all the papers for establishing papers' feature space. We established a sample set for our experiments. Using the score value of 10 as a selection criterion, about 0.5% of papers in each journal were identified as HCPs (The *Journal of Mathematical Physics* is an exception, where the ratio of HCPs is about 0.1%). All the HCPs identified in each journal were contained in the sample set. But, there are large amount of papers were identified as MCPs and LCPs using the selection criteria. Here, only 10% of MCPs and LCPs were randomly selected out as the sample data for each journal. The distributions of papers in the three levels for the four journals are shown in Table 1.

3.3. Establishing the feature space

Both the external and the quality information about these papers were considered in constructing the feature space. The external features mainly come from three aspects: the authors, the journals, and the external features of the paper itself. Sixteen external features were extracted out. The quality features come from papers' knowledge diffusion and first citation properties in the scientific community. Nine features, including papers' citing diffusion properties in the period of five years after their publication, and the information associated with papers' first citation were picked up. Table 2 shows the feature space for these papers.

Here, a five-year interval was used to extract papers' quality features. The reason is that it is often used in bibliometric analyses and is intermediate with respect to a short and a long-term citation window. Since the variability of citedness is expected to increase with the size of the citation window, a five-year interval is sufficient long term for a distinct polarization pattern to occur (Aksnes, 2003).

There are larger differences in each feature among journals. Each feature in the feature space is normalized first in each journal according to Max–Min normalization method. And every one was transformed into the feature with value located in [0, 1]. Then, the normalized data in the four journals are combined into the final sample data for our experiment.

4. Experimental analyses

In this section, we present and analyze the experimental results of the proposed CBC model. The feature reduction process is based on the feature reduction algorithm (Algorithm 1). The case selection process is based on the case selection algorithm (Algorithm 2). And the reasoning results are based on two algorithms: case coverage prediction algorithm and case score prediction algorithm (Algorithm 3). The main criterion for evaluating the performance of the proposed CBC model and

Table 1
Distributions of papers in each journal.

Journals	Abv.	HCPs	MCPs	LCPs	Total	Initial accuracy
IEEE Transactions on Automatic Control	ITAC	1	8	17	26	0.677
Journal of Applied Physics	JAP	10	48	133	191	0.680
Journal of Experimental Medicine	JEM	1	9	20	30	0.702
Journal of Mathematical Physics	JMP	5	13	33	51	0.633

Table 2
Feature space.

Features	Definitions	Sources
x_1	Number of authors	External
x_2	Whether there is international cooperation	External
x_3	Whether any author is American	External
x_4	The *h* index of the first author before publication of this paper	External
x_5	The number of papers published by the first author before this paper	External
x_6	The total citations to the papers published by the first author before this paper	External
x_7	The average citations to the paper published by the first author before this paper	External
x_8	The maximum number of papers published by the authors before this paper	External
x_9	The maximum total citations to the papers published by the authors before this paper	External
x_{10}	The maximum *h* index of the authors before publication of this paper	External
x_{11}	The impact factor of the journal publishing this paper	External
x_{12}	The number of papers published in the journal in year 1985	External
x_{13}	The length of this paper	External
x_{14}	The document type of this paper	External
x_{15}	The language of this paper	External
x_{16}	The number of references listed in this paper	External
x_{17}	The first-cited age of this paper	Quality
x_{18}	The first-citations of this paper	Quality
x_{19}	The total citations to this paper in its first five years after publication	Quality
x_{20}	The number of countries citing this paper in its first five years after publication	Quality
x_{21}	The number of document types of papers citing this paper in its first five years after publication	Quality
x_{22}	The number of institutions citing this paper in its first five years after publication	Quality
x_{23}	The number of languages citing this paper in its first five years after publication	Quality
x_{24}	The number of journals citing this paper in its first five years after publication	Quality
x_{25}	The number of subjects citing this paper in its first five years after publication	Quality

the achieved feature subsets is their classification accuracies. And the classification accuracy A_i depends on the number of samples correctly classified and is evaluated by the formula $A_i = (t/n) \cdot 100$, where t is the number of sample cases correctly classified, and n is the total number of sample cases.

In order to enhance the reliability of reasoning, a three-fold cross validation is used. It divides the whole data set into multiple pair of training and test sets (i.e. Fold-1, Fold-2 and Fold-3). Each training set uses 2/3 of the entire data records; with the rest 1/3 data records use for testing set. The cases in training set serve as the selected cases, the class labels of whom are known. And the cases in test set serve as the unseen cases. The class labels of these unseen cases are determined by the selected cases using the case coverage prediction algorithm and case score prediction algorithm. The class labels determined would be compared with the original class labels of these unseen cases. Then the classification accuracy is calculated. We repeated the procedure three times so that every case has been used exactly once for testing. And the average classification accuracy (i.e. the mean of Fold-1, Fold-2 and Fold-3) is taken as the final classification accuracy.

For convenience, some notations were introduced which would be used throughout the experiment:

$$\text{Reduced feature} = \frac{|\text{Reduced feature set}|}{|\text{Original feature set}|}$$

$$\text{Reduced case} = \frac{|\text{Reduced case set}|}{|\text{Original case set}|}$$

$$\Delta\text{Accuracy} = \text{classification accuracy after reduction} - \text{initial classification accuracy}$$

The initial classification accuracy is defined as the classification accuracy when using the original datasets, which have not been reduced either for the number of their features or cases. The last column of Tab.1 has shown the initial classification accuracy for the four journals.

Here, the classification performances in three different mechanisms are considered:

4.1. Classification performance only after feature reduction

Basing on the feature reduction algorithm (Algorithm 1), only the features are reduced for each of the four journals. Since a sequential of feature subsets of $F'_1 = \{f'_1\}, F'_2 = \{f'_1, f'_2\}, \ldots, F'_{F'} = \{f'_1, f'_2, \ldots, f'_{|F''|}\}$, is generated with Algorithm 1, the subset which could achieve the highest classification accuracy is selected out as the final reduced feature subset in this section.

Table 3 shows the results only after feature reduction for each journal. Here, ΔAccuracy is the difference between the classification accuracy after feature reduction and the initial classification accuracy. It shows that higher accuracy have been achieved, though the number of features is reduced dramatically (about 69% features were reduced). Using the reduced feature space, the average problem-solving accuracy is 0.851, and the average enhancement is 0.178.

Table 3
Accuracies after feature reduction.

Journals	Reduced feature subset	Reduced feature	Accuracy	ΔAccuracy
ITAC	$\{x_4, x_{24}, x_{25}, x_{17}, x_{20}, x_{19}, x_{22}, x_{18}\}$	0.68	0.822	0.145
JAP	$\{x_4, x_{17}, x_{25}, x_{24}, x_{20}, x_{19}, x_{22}, x_{14}\}$	0.68	0.892	0.212
JEM	$\{x_{25}, x_4, x_{24}, x_{17}, x_{20}, x_{19}, x_{22}\}$	0.72	0.86	0.158
JMP	$\{x_4, x_{24}, x_{17}, x_{20}, x_{25}, x_{19}, x_{22}, x_{10}\}$	0.68	0.83	0.197
Average	–	0.69	0.851	0.178

Table 4
Accuracies after case reduction.

Journals	Reduced case	Accuracy	ΔAccuracy
ITAC	0.12	0.733	0.056
JAP	0.16	0.812	0.132
JEM	0.11	0.788	0.086
JMP	0.19	0.765	0.132
Average	0.145	0.775	0.101

4.2. Classification performance only after case selection

Basing on the case selection algorithm (Algorithm 2), only the cases are reduced for each of the four journals. Table 4 shows the results for the accuracy of the case selection for each journal. Here, ΔAccuracy is the difference between the classification accuracy after case selection and the initial classification accuracy. It shows that about 14.5% cases were reduced. But the classification performance for each of the four journals is enhanced. Using the reduced case space, the average problem-solving accuracy is 0.775. The average enhancement is about 0.101 compared with the initial accuracy.

4.3. Classification performance after feature and case reduction

Here, both the features and cases are reduced using the feature reduction and case selection algorithm. Also, the feature subset achieving the highest classification accuracy is chosen out as the final feature subset. Table 5 shows the results for the classification accuracies for each of the four journals. Here, ΔAccuracy is the difference between the classification accuracy after feature and case reduction, and the initial classification accuracy.

Obviously, the average number of features is reduced from 25 to about 7.8, i.e., 69% features are removed. Based on the concept of case coverage and soft fuzzy rough set, 7.0% cases are reduced from the original dataset. But the classification performance for each of the four journals is enhanced dramatically for feature and case reduction. Using the reduced dataset and feature space, the average problem-solving accuracy reaches 0.921. The average enhancement is about 0.248 compared with the initial accuracy. Analyzing the classification results in (a)–(c) and the initial one, two aspects should be mentioned:

(i) Classification performance: All of the three mechanisms of feature reduction (a), case selection (b), and both the feature and case reduction (c) gain better performance than the original one. And the accuracies in (c) are the highest. It shows that the CBC model proposed in this paper does work. There are noises in the citation activities of papers. Both the features and cases are reduced to lessen the negative influences brought by noises.

(ii) Typical features: Compared the reduced feature subsets in (a) and (c), the feature subset achieved for each of the four journals is the same. And the typical features for different journals are little difference from each other. Features of $\{x_4, x_{17}, x_{19}, x_{20}, x_{22}, x_{24}, x_{25}\}$ are extracted out for all of the journals. Features of $\{x_{10}, x_{14}, x_{18}\}$ are also existed as the informative features, but each of them is just emerged once. Table 6 shows the typical ten forecasting features, along with their occurrence frequencies in the four journals. Since the goal of this research is to find the fair and reliable predictors for highly cited papers, which should be general enough to be applicable to a wide range of fields. The features occurs frequently in the four journals should be detected out to approach our goal. Thus, among the original twenty-five features, only seven features of $\{x_4, x_{17}, x_{19}, x_{20}, x_{22}, x_{24}, x_{25}\}$ could be better to predict a paper's citing trend in future, including one external feature and six quality features.

Table 5
Accuracies after feature and case reduction.

Journals	Reduced feature subset	Reduced feature	Reduced case	Accuracy	ΔAccuracy
ITAC	$\{x_4, x_{24}, x_{25}, x_{17}, x_{20}, x_{19}, x_{22}, x_{18}\}$	0.68	0.06	0.898	0.221
JAP	$\{x_4, x_{17}, x_{25}, x_{24}, x_{20}, x_{19}, x_{22}, x_{14}\}$	0.68	0.08	0.95	0.27
JEM	$\{x_{25}, x_4, x_{24}, x_{17}, x_{20}, x_{19}, x_{22}\}$	0.72	0.04	0.938	0.236
JMP	$\{x_4, x_{24}, x_{17}, x_{20}, x_{25}, x_{19}, x_{22}, x_{10}\}$	0.68	0.1	0.898	0.265
Aver.	–	0.69	0.07	0.921	0.248

Table 6
Typical features for HCPs.

Feature	Definition	Frequencies	Source
x_4	The h index of the first author before this paper	4	External features
x_{17}	The first-cited age of this paper	4	Quality features
x_{19}	The total citations to this paper in the first five years after publication	4	Quality features
x_{20}	The number of countries citing this paper in the first five years after publication	4	Quality features
x_{22}	The number of institutions citing this paper in the first five years after publication	4	Quality features
x_{24}	The number of journals citing this paper in the first five years after publication	4	Quality features
x_{25}	The number of subjects citing this paper in the first five years after publication	4	Quality features
x_{10}	The maximum h index of the authors before this paper	1	External features
x_{14}	The kind of this paper	1	External features
x_{18}	The first-citations of this paper	1	Quality features

Here, the question is whether the predictors chosen can well correspond with our preliminary results, or whether they are the factors researches like to probe. In order to answer this question, a feature-specific analysis was performed.

(i) The h index of the first author $\{x_4\}$: The h-index is an index proposed by Hirsch (2005) to evaluate the quality of scientific research from a micro viewpoint. A larger h-index indicates that an author has gained considerable research capabilities or reputations in science. Van Dalen and Henkens (2001, 2005) suggested that authors with high reputations could receive disproportionately more citations than authors with low reputations. Our preliminary work also examined the significance of the first author's h index (Wang et al., 2011).

(ii) The citation properties in the first five years after a paper's publication $\{x_{19}, x_{20}, x_{22}, x_{24}, x_{25}\}$: These five features represent papers' knowledge distribution properties in the scientific community, which were also extracted out to be the typical features associated with papers' citation counts in our preliminary work (Wang et al., 2011). The wider citation distributions in various journals, subjects, countries and institutions increase a paper's visibility to a larger extent, and so to their citation counts.

(iii) The first-cited age $\{x_{17}\}$ of one paper: The first-cited age shows the rate of a paper to be accepted after publication. It indicates that the easier a paper is cited, the larger the probability of it being cited frequently. Van Dalen and Henkens (2005) showed that the status of uncitedness of a paper becomes a stigma and the longer a paper is uncited, the lower its quality and the less inclined researchers will be to cite it. Glänzel, Shilemmer, and Thijs (2003) stated that the probability of a paper's uncitedness increases dramatically with belated first citations, and the probability of being frequently cited later on decreases to the same extent. In fact, the negative influences brought about by a paper's uncitedness reflect the importance of the position of a paper's first citation on its later citation life. Our preliminary experiment also showed the important influences brought by papers' first-cited properties (Wang et al., 2011). But compared with the feature of papers' first-citations $\{x_{18}\}$, papers' first-cited age plays a more important role in citation count prediction.

Therefore, our results show that it is feasible to predict future highly cited papers with high predictivity. Compared with the prior endeavors to seek the possible factors influencing on papers' citation activities, we not only consider the external information about the journal and paper and authors, but also the information about papers' quality. Moreover, the predictive features extracted out are almost accordance with those obtained in our preliminary work, though different data set and learning methods were performed. It indicates that these features could provide reliable prediction information even in distinct areas of science. The only difference between our two works lies in the feature of $\{x_{11}\}$, the impact factor of the journal. Because only one journal in one field is considered to be the data source, this feature was not identified as the core indicator. But it did not conceal the significance role of the journal on papers' citation impact. In our preliminary work (Wang et al., 2011), we found that the reputation of journals plays an overriding role in gaining attention in science. The similar conclusions have also been suggested by Van Dalen and Henkens (2001, 2005).

5. Conclusions

In this paper, we show a quantitative support of prediction future highly cited papers by accessible bibliometric indicators.

Compared with our first attempt on prediction highly cited papers, the experiments in the present manuscript performed two important and additional analyses. First, an independent data set was performed. Models in this paper were trained on papers from four different journals in different fields, which could provide a more powerful prediction. Second, a different learning method was used. A novel framework of case-based classifier (CBC) was proposed based on the case-based reasoning (CBR) and the soft fuzzy rough set (SFRS). The application of SFRS in CBC model can better eliminate the negative influences brought by noises in the actual citation process. The CBC model operates by first reducing the number of features using the concept of soft fuzzy dependency in SFRS. Then, it selects cases with a case selection approach basing on the hybrid use of SFRS and the concept of case coverage in CBR. Finally, it classifies the unseen cases with two classification techniques of case coverage based prediction and case score based prediction. The features extracted out by the new data set and the

CBC model are highly consistent with those in our preliminary work. It shows that the predictors depends more on the choice of features rather than the choice of data sets and classifiers. And those following factors were proved to be the most relevant predictors: (a) The research capabilities of the first author, represented by his/her *h* index. (b) The papers' quality, represented by papers' earlier knowledge diffusion properties and the first-cited properties in the scientific community. (c) The reputation of journal, represented by its impact factor.

Since several distinct areas were considered, the results have provided reliable and widely applicative prediction information. Our future research will investigate whether contextual information of papers could be used to find additional factors that allow differentiating between successful and lower-valued papers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71003020 and 70973031), the special funds of Central College Basic Scientific Research Bursary (Grant No. DL11CB09), and the Postdoctoral Science Foundation of Heilongjiang Province.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159–170.
Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.
Bornmann, L., & Leydesdorff, L. (2012). Which are the best performing regions in information science in terms of highly cited papers? Some improvements of our previous mapping approaches. *Journal of Informetrics*, *6*(2), 336–345.
Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, *52*, 3–12.
Burrell, Q. L. (2002a). On the nth-citation distribution and obsolescence. *Scientometrics*, *53*, 309–323.
Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*, 232–235.
Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, *54*(5), 372–378.
Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, *51*(7), 635–645.
Cloete, I., & Zyl, J. V. (2006). Fuzzy rule induction in a set covering framework. *IEEE Transactions on Fuzzy Systems*, *14*, 93–110.
Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology*, *62*(1), 50–60.
Dubois, D., & Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *General Systems*, *17*, 191–209.
Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*, 257–270.
Gates, G. W. (1972). The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, *18*, 431–433.
Glänzel, W. (1997). On the reliability of predictions based on stochastic citation processes. *Scientometrics*, *40*(3), 481–492.
Glänzel, W., Rinia, E. J., & Brocken, M. G. M. (1995). A bibliometric study of highly cited European physics papers in the 80s. *Research Evaluation*, *5*(2), 113–122.
Glänzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation processes. *Information Processing & Management*, *31*(1), 69–80.
Glänzel, W., Shilemmer, B., & Thijs, B. (2003). Better later than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, *58*(3), 571–586.
Hart, P. E. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, *14*, 515–516.
Hewings, A., Lillis, T., & Vladimirou, D. (2010). Who's citing whose writings? A corpus based study of citations as interpersonal resource in English medium national and English medium international journals. *Journal of English for Academic Purposes*, *9*(2), 102–115.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.
Hu, Q. H., An, S., & Yu, D. R. (2010). Soft fuzzy rough sets for robust feature evaluation and selection. *Information Sciences*, *180*, 4384–4400.
Hu, Q. H., Xie, Z. X., & Yu, D. R. (2007). Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. *Pattern Recognition*, *40*, 3509–3521.
Hu, Q. H., Yu, D. R., & Xie, Z. X. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition*, *27*, 414–423.
Jensen, R., & Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *IEEE Transactions on Fuzzy Systems*, *17*, 824–838.
Jin, X. B., Liu, C. L., & Hou, X. W. (2010). Regularized margin-based conditional log-likelihood loss for prototype learning. *Pattern Recognition*, *43*, 2428–2438.
Jorgenson, M. (2004). A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, *70*, 37–60.
Kim, K. (2004). The motivation for citing specific references by social scientists in Korea: The phenomenon of co-existing references. *Scientometrics*, *59*(1), 79–93.
Kupinski, M. A., & Giger, M. L. (1999). Feature selection with limited datasets. *Medical Physics*, *26*, 2176–2182.
Kwak, N., & Choi, C. H. (1994). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 1667–1671.
Laband, D. N., & Piette, M. J. (1994). Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, *102*, 194–203.
Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology*, *59*(12), 1973–1984.
Li, Y., Shiu, S. C. K., & Pal, S. K. (2006a). A rough set-based case-based reasoner for text categorization. *International Journal of Approximate Reasoning*, *41*, 229–255.
Li, Y., Shiu, S. C. K., & Pal, S. K. (2006b). Combining feature reduction and case selection in building CBR classifiers. *IEEE Transactions on Knowledge and Data Engineering*, *18*, 415–429.
Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Density based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 734–747.
Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277.
Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, *11*, 341–356.
Penas, C. S., & Willett, P. (2006). Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, *32*(5), 480–485.
Rong, T., & Martin, A. S. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation*, *64*(2), 246–272.
Symth, B., & Keane, M. (1998). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Application*, *28*, 127–135.

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science*, *34*(5), 680–687.
Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*, 448–452.
Van Dalen, & Henkens, H. P. K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics*, *50*, 455–482.
Van Dalen, & Henkens, H. P. K. (2005). Signals in science – On the importance of signaling in gaining attention in science. *Scientometrics*, *64*(2), 209–233.
Wang, M. Y., Yu, G., & Yu, D. R. (2011). Mining typical features for highly cited papers. *Scientometrics*, *87*(3), 695–706.
Xia, J. F., Myers, R. L., & Wihoite, S. K. (2011). Multiple open access availability and citation impact. *Journal of Information Science*, *37*(1), 19–28.