Portland State University

PDXScholar

Engineering and Technology Management Faculty Publications and Presentations

Engineering and Technology Management

10-22-2018

Keyword-Based Patent Citation Prediction Via Information Theory

Farshad Madani
Portland State University

Martin Zwick
Portland State University, zwick@pdx.edu

Tugrul U. Daim
Portland State University, tugrul@etm.pdx.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/etm_fac

Part of the Computer and Systems Architecture Commons, Digital Communications and Networking Commons, and the Logic and Foundations Commons

Let us know how access to this document benefits you.

Keyword-based Patent Citation Prediction via Information Theory

Farshad Madani[a,], Martin Zwick[b], and Tugrul Daim[a]

a Department of Engineering and Technology Management, Maseeh College of Engineering and Computer Science, Portland State University, Portland, OR, USA;
b Systems Science PhD Program, Portland State University, Portland, OR, USA

Abstract

Patent citation shows how a technology impacts other inventions, so the number of patent citations (backward citations) is used in many technology prediction studies. Current prediction methods use patent citations, but since it may take a long time till a patent is cited by other inventors, identifying impactful patents based on their citations is not an effective way. The offered prediction method in this article predicts patent citations based on the content of patents. In this research, Reconstructability Analysis (RA), which is based on information theory and graph theory, is applied to predict patent citations based on keywords extracted from the abstracts of selected patents. After applying three classes of RA (variable-based analysis without and with loops and state-based analysis), nine specific IV states of a predicting model are extracted. These states involve the four keywords of 'chamber', 'hous', 'main', and 'return'. Lastly, the abstracts of the patents are examined to identify the technical subjects relevant to smart building technologies for which these keywords are proxies.

Keywords*: patent mining, patent citation analysis, patent citation prediction, information theory, Reconstructability Analysis, OCCAM

1. Introduction

Due to the rapid growth in the number of patents filed around the world, patent databases have become tremendous technological resources. To elicit technical knowledge from patents and use it in managerial and engineering decisions, three major types of patent analysis have been developed: bibliometrics analysis, citation analysis, and keyword-based analysis. These methods are based upon on the three major parts of patent documents: metadata, main body, and citations. In bibliometrics analysis, the relation between inventors, firms, research institutes, countries, etc. are analyzed. In citation analysis, citations are considered as proxies for technological impacts, which means that when a patent cites another patent, the citing patent is impacted by the cited patent. In keyword-based patent analysis, keywords are employed as proxy for patent content, extracted by text mining tools, and analyzed by machine learning methods, specifically by clustering analysis. This paper is an example of this third type of analysis which is keyword-based.

Patent citation analysis is applied for various purposes, because patent citations can represent technological changes. Assessing future technological impacts [1], monitoring technological trends [2], predicting emerging technologies [3], and exploring technology diffusion [4] are examples of such patent citation applications. But citation analysis inherently suffers from a drawback, namely that it is not able to rely on technical content of patents [5]. However, there is *correlation* between the number of citations and the technical richness of patents' contents; thus, we exploit the link between patent citations and patent content by using keyword-based analyses to make citation predictions. Predicting citations based on patent content may allow us to discover technologies that will be impactful in the future. Therefore, the research question of this paper is that if there are a set of keywords in the abstract of patents that the frequency number of the keywords can be applied to predict patent citation.

Reconstructability Analysis (RA) is chosen as the main methodology for citation prediction since RA is very well suited for predictive modeling that is exploratory as opposed to confirmatory. Other methodologies, such as network

analysis and cluster analysis, are unable to predict citations. In this work, RA is specifically applied to smart building technologies to predict which patents will be cited in the future.

2. Literature review

Patents are good representatives for technological events. When a patent cites another patent, this means the citing patent is impacted by the cited patent [1]. Patent citations are deployed to study many technological events such as linkage between science and technology [6][7], knowledge flow and diffusion [8], patent valuation[9][10], stock market valuation [11], technological convergence between industries [12], emerging research areas [13], emerging technologies [14][15], the future impact of current technologies [1] and technology diffusion [4]. Different methodologies like stochastic analysis [1], cluster analysis [3][4][15][16], and network analysis[3] are applied to patent citation analysis, but patent citation analysis is subject to some drawbacks [5]. First, patent citation analysis only discovers individual relations between two patents. There might be semantically relation between two patents without any citations between them, Therefore, patent citation analysis does not identify the overall relationships among all patents.. Basically, citations cannot cover the richness of potential information, so they limit the scope of analysis. Second, citation analysis is not able to consider semantic relationships between patents; it may even produce superficial or misleading indices.

To remedy the above-mentioned problem, patent researchers have developed keyword-based methods by applying text mining tools because keywords are good representatives of the content of patents [17]. The majority of keyword based methods are developed for the purpose of technology prediction [5][18] [19] [20] or other related topics such as technology monitoring [2], technology discovery [21], and roadmapping [22]. The dominant methods deployed in the keyword based methods are cluster analysis methods and network analysis. Clustering methods such as *k*-mean [19], principle component analysis [21] are utilized to group or to map patents based on their semantic similarity. Network analysis analyzes the relation between patents based on the intense of their similarity.

There is, however, a huge gap between patent citation analysis and keyword based patent analysis. As introduced above, many patent citation analysis methods and many keyword-based patent analysis methods have been developed, but no research has been done to bridge these two types of patent analysis. Bridging these two types of analysis would allow us to predict patent citation, which would improve significantly the accuracy of patent citation based methods [23].

None of the existing methodologies – network analysis or cluster analysis -- is able to predict patent citations based on the content of patents (keywords). Network analysis studies the structure of relationships between entities like citations, and cluster analysis groups a set of similar objects, like keywords. To improve patent citation prediction, we need to apply a method that predicts patent citations based on the frequency number of keywords. Such a capability is provided by Reconstructability Analysis (RA) which yields models that predict citations based on keyword frequencies. RA is introduced in the next section.

3. Methodology

3.1. Patents Extraction

Patents extraction provides the data set for analysis, so the more accurate the extraction, the more accurate analysis we will have. Since we are prediction patent citations based on patent contents, we extract patents based on keywords representing smart building technologies. It is very important to have the correct keywords that address smart building technologies. In this research, an initial dictionary of keywords is developed based on the literature and experts' judgment, and then it is expanded by applying Google Adword capabilities. To identify the keywords which address smart building patents, the concept of smart building is divided into three main categories: 1) energy, 2) efficiency, and 3) building (Table 1). The first column titled "energy" represents all forms of energy consumption in buildings. To cover all possible wordings in the patents, the asterisk symbol (*) is used as a wildcard. For instance, heat* covers all possible variations such as heat, heating, heater, etc. The second column addresses all possible words representing the concept of 'smart', such as saving, managing, reducing, etc. The last column signifies all possible words approximating 'building'. The key words are extracted through several rounds by applying the literature [24] and Google AdWords.

To extract the patents information from the LexisNexis database, the query shown in Figure 1 is designed based upon field tags used by the database. To gather smart building patents, the query parameters were set for US patents

issued between 1990 and 2013. All keywords mentioned in Table 1, were deployed to look in titles and abstracts of patents. After applying the query, 2,483 patents were recognized. In this case, 'abstract', 'application number', 'filing date', and 'cited patents' are represented by ABST, APPL-NO, FILED-DATE, and REF-CITED tags, respectively, to design the query shown in Figure 1. The output was in text file format.

Energy	Efficiency	Building
Heat*	Sav*	Home
Light*	Optimi*	House
Cool*	Manag*	Floor
Ventil*	Reduc*	
Refrig*	Control*	
Pump*	Smart	
HVAC	automat*	
	Sustainab*	
	Intelligent	

Table 1 'Smart Building' keywords used in the query shown in Figure 1

3.2. Patents data preprocessing

To do Reconstructability Analysis, we need to elicit two main groups of numbers from the patents data set. The groups are: 1) the number of citations of the patent (the DV), and 2) the frequencies of the keywords (the IVs). So the extracted patents data, which are in text format, need to be organized in two separate databases. The first database is provided to figure out the number of citations between the extracted patents. There might be patents citing the extracted patents, but they are not considered in the data set because they are not basically relevant to smart technologies, and consequently their keywords are not considered for the citation prediction. The second database is for text mining purposes to figure out the frequency numbers. The first database is a 2483x2483 matrix reflecting relations between the patents. If a cell contains '1', it means the patent located in the column is cited by the patent located in the row and if the cell shows '0', it means there is no citation. The data fields of the second database are 'patent number' which comprises the patent numbers of the extracted patents, and 235 fields that contain the frequency numbers of keywords in the abstracts.

Several tasks including syntax tagging, word stemming, and stop-word elimination are required to extract the keywords. In syntax tagging, words or terms are distinguished in the sentences, then their suffixes are removed via a stemming procedure. To omit words such as *the*, *is*, *at*, *which*, and *on* from the reserved words, a list of stop-words, given from RANKS.NL Website [25], are deployed.

To extract the most repeated keywords in the patents, we used Weka [26], written in Java and developed at the University of Wikato in New Zealand. We applied Weka's StringToWordVector filter which removed numbers and stop words. We also removed general keywords, such as 'Winter, 'mold', and 'dark,' that are not relevant to the technologies we are trying to predict. At the end of this filtering process, 235 keywords remained. With these keywords, we created a database including patent number, abstract, frequency number of the keywords, and also citation number of the patents. The citation numbers come from the matrix mentioned earlier.

With this database, we apply a two-step preprocess to the data set to make it ready for RA. First, we create a spreadsheet containing the IVs, which are the frequency numbers of the keywords extracted from text mining, and the DV, which is the number of citations of the patents. Then the frequencies of the IVs are binned, because, as mentioned before, only nominal variables or discretized quantitative variables converted to nominal variables are usable in RA. After binning, the binned data including IVs and DV are organized in a text file in OCCAM format [27] .

The DV is binned to two bins, because in the majority of cases (patents) citation number is zero. In this case, 92.9% of the patent citations are zero. Therefore, where DV = 1 the citation number is zero, meaning the patent is not cited, and DV = 2 means the citation number is more than zero, i.e., the patent is cited at least once.

Figure 1 The query used to search in USPTO database through 'LexisNexis Academic Universe'

3.3. Reconstructability Analysis

Reconstructability Analysis (RA), introduced first by Ashby [28], is a method based on information theory and graph theory. RA has similarities with log-linear methods applied in social sciences [29] and can be considered a machine learning technique such as those widely used in computer science [30] . RA, like log-linear and machine learning methods, is applied in many different fields, including time-series analysis, classification, decomposition, compression, pattern recognition, prediction, control, and decision analysis [31].

Basically, RA decomposes a probability or frequency distribution into component distributions [32] by applying statistical multivariate analysis similar to log-linear methods [33] and logistics regression . RA also overlaps mathematically with Bayesian networks. Despite these similarities, RA, log-linear, and Bayesian Networks methodologies each has its own unique capabilities [33]. For example, RA can analyze set-theoretic relations and arbitrary functions, is capable of state-based analysis [34][35], and has a Fourier version [36]. RA and log-linear, but not Bayesian Networks, can utilize models with loops in both directed and neutral systems. (In directed systems, IVs and DVs are distinguished, and IVs are used to predict DVs. In neutral systems, IVs and DVs are not distinguished. Neutral systems simply try to find the relations between all the variables.)

Three classes of RA models are applied in this research: 1) variable-based (VB) models without loops, 2) variable-based models with loops, and 3) state-based (SB) models (which typically have loops). These models give us coarse, refined, and ultra-refined models, schematically depicted in Figure 2 [33]. The bold lines in the figure show how complex a model might be accepted in each of the classes. In VB loopless models, the indicated bold line might be the most complex model that could be statistically acceptable. The dotted lines above the bold line indicate models that are statistically insignificant. Because VB models with loops make finer discriminations, a more complex such model might be statistically significant. Since SB models make even finer discriminations, they may allow a still more complex model to be accepted as statistically significant.

In loopless models, there is only a single 'predicting component'. For example, in the model ABCD:ABZ, A and B predict Z. (The first component, ABCD, allows for associations between the IVs.) Models with loops have more than one predicting component. For example, in ABCD:AZ:BZ model, two variables, A and B, separately predict Z, and these predictions are fused in the model. State-based models like $ABCD:A_1B_2Z:B_1C_3Z$ predict Z with more specific information about states of IVs. The states are indicated by subscripts.

A web-based program developed at Portland State University (Portland, Oregon) was used to perform the RA analysis. The program is named 'Organizational Complexity Computation and Modeling' (OCCAM), which is also an allusion to Occam's (sometimes spelled Ockham's) Razor, the principle of parsimony, important in modeling. The first precursor to OCCAM was written by Zwick and Hosseini [37] . A detailed overview of RA [31] , OCCAM manual [27], and a review of OCCAM architecture [38] are available at the RA web site, http://www.pdx.edu/sysc/research-discrete-multivariate-modeling.

Four basic functions of OCCAM were employed for this data patent analysis:

1- *Loopless analysis*, as the first step in RA, is used to discover what keywords (IVs) are individually the most predictive of patent citation (DV).

2- *With-loops analysis* find three best models for patent citation predicting. These best models involve multiple and/or interacting IVs are based on BIC (Bayesian Information Criterion), AIC (Akaike Information Criterion), and Incremental p-values criteria. The AIC and BIC best model criteria select models based on a linear tradeoff of the information captured in a model and its complexity, with the BIC criterion penalizing models more heavily for complexity. The Incremental-p model picks the highest information model whose difference from the reference (independence) model is statistically significant, where there is also a path from the reference to the model, where every incremental step of the path is statistically significant. In all of the analyses, significance level is chosen as 5%. Actually, models are chosen based on $\triangle$AIC and $\triangle$BIC, which are *differences* of AIC and BIC of the models from values for

the reference (the independence model); this results in good models having higher positive values of these differences. (Normally, when BIC or AIC values are given, they're given as values for particular models, not as differences from a reference model, and then good models are those that have lower or negative values.) These models with loops are the key outcomes of this research.

3- *State-based analysis* helps to more deeply scrutinize the IVs selected in prior steps. There might exist some interactions between the IVs selected in prior steps that are undetectable in variable-based analysis. State-based models can use these subtle interactions for more accurate predictions.
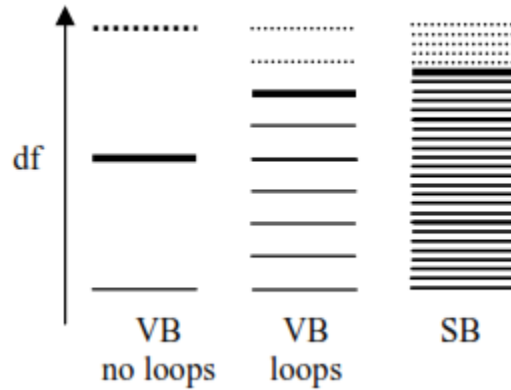
Figure 2 Degree of refinement of RA models [33]

4. RA Results

As mentioned before, RA is designed for nominal variables or for discretized quantitative variables converted to nominal variables. To make quantitative variables usable, it is necessary to bin them. Table 2 shows the binning of the four IVs that analysis (discussed below) revealed to be predictive. The four keywords shown in Table 2 are the best predictors in the variable-based analysis with loops (see section 4.2 below and Table 4). For instance, Db=2 means 'Chamber' is frequented at least one time, or Dg=3 means 'hous' is frequented between 2 and 15. Also, Z, which is the dependent variable (DV), is binned to 1 and 2. If Z=2, it means the patent is cited; otherwise it is not cited.

IV	Keyword	Values			
Db	chamber	Bin	1		2
		Frequency	0		1-12
		#Patents	2316		167
Dg	hous	Bin	1	2	3
		Frequency	0	1	2-15
		#Patents	1806	410	267
Iz	main	Bin	1		2
		Frequency	0		1-11
		#Patents	2346		137
Ja	return	Bin	1		2
		Frequency	0		1-11
		#Patents	2399		84

Table 2 relation between the bin numbers and the frequency of occurrence of the IVs (the keywords)

To bin the number of frequency of the keywords, a binning software program is applied. The binning program tries to create a set of bins that are equally sampled. Since for all key words, most cases have no citations of the key words, a uniform distribution is impossible to achieve, but the binning program attempts a reasonable compromise. The first bin is necessarily always for zero frequency of keywords (IVs) or citations (DV). The

number of additional bins then depends on how many cases are spread over 1 frequency/citation, 2 frequencies/citations, etc. For keywords *chamber*, *main*, and *return*, the number of cases where there are 1 or more is small (167, 137, and 84 compared to over 2000 for no frequency), so it makes sense only to add one extra bin for these keywords. For *hous*, however, there are 410 cases where this keyword is frequented once, and 267 where it is frequented more than once, so it makes sense to assign a separate bin for each of these situations. Other keywords are treated similarly.

4.1. Loopless Variable-based Analysis
The result of loopless analysis is tabulated in Appendix 2, where uncertainty reduction progressively increases as models go from the bottom (the independence model) to the top (the most complex model considered). Uncertainty reduction is the primary information theoretic measure of the goodness of a model; it is analogous to %variance explained, but because of the logarithm term in the Shannon expression of uncertainty even small uncertainty reductions (for example, even 8%) can represent large effect sizes. Since uncertainty reduction is specific to information theoretic analyses, a more general measure of goodness of models, namely %correct in prediction, is also offered in this paper. The independence model, IV:Z, by definition, has no uncertainty reduction in the DV since no IVs are associated with it, i.e., so %$\triangle$H(DV) is always zero in level 0. Also, %Correct(Data) starts from 94.7%, which means that a default prediction (the prediction without any knowledge of the presence or absence of any keywords) of no citation (Z=1) will be 94.7% correct. This is the default prediction, because that percentage of the patents is not cited by other patents in the sample. This means that we are struggling to predict very infrequent occurrence of Z=2 in this data set.

Among the IVs, three single IVs, namely Ac, Iz, and Ja, give the most uncertainty reduction. Of these three IVs, Ja, which is 'return', is a much stronger predictor than Iz ('main') and Ac ('light'), as shown in its much higher values of both uncertainty reduction and %Correct. Note that Iz and Ac do not improve %Correct over its default value, but their predictive effect is captured in their non-zero uncertainty reduction.

IV	%$\triangle$H(DV)	P	%correct
Ja	54	0.00	98.1
Iz	15	0.00	94.7
Ac	3.3	0.00	94.7

Table 3 Best single IVs prediction patent citations

As shown in Appendix 2, IV:IzJaZ is the best BIC model in the loopless analysis. Iz and Ja together predict Z with 83.33% uncertainty reduction and $\triangle$DF = 3. Also, IV:DbIzJaZ and IV:DgIzJaZ are the best AIC and incremental-p models, respectively.

4.2. Variable-based analysis with loops
Allowing loops delivers more powerful variable-based models, as shown in Table 4. Since BIC is more conservative than AIC and Incremental p-value, the BIC model is the most reliable result of this study. Not surprisingly, the first and second single IV predictors are in the BIC model. In addition, two more IVs, Db and Dg, are added in the BIC model as new predictors. The AIC and Incremental p-values models are identical, and they include most of the same predicting variables. In comparison to BIC, they keep Dg, Iz, and Ja and drop Db. Also, the AIC and the p-value models have more uncertainty reduction and prediction correctness, but the delta degrees of freedom, $\triangle$DF, has increased greatly from 5 to 15, which means these two models are 3X more complex. All variable-based models generated by OCCAM are available in Appendix 3.

Criterion	Model	%$\triangle$H(DV)	%correct	$\triangle$DF	p
BIC	IV:DbZ:DgZ:IZ:JaZ	86.6	98.8	5	0
AIC	IV:AmZ:AsZ:DgZ:EbZ:FdZ:IZ:JaZ	90.9	99.3	15	0
Incremental p	IV:AmZ:AsZ:DgZ:EbZ:FdZ:IZ:JaZ	90.9	99.3	15	0

Table 4 Best variable-based models selected based on three criteria

RA also gives specific information about any given model. The 'Do Fit' option in OCCAM examines the given model in detail, and states exactly what DV values it predicts for all possible values of the predicting IVs (see Appendix 4). Since we are interested in predicting whether a patent will be cited or not, we look at column

'Z=2' for the model probabilities. If the probability of Z=2 is considerably larger than its marginal probability of 5.27%, patent citation is more likely, assuming that the difference between the model probabilities and the independent model margins are statistically significant. This difference is examined by a Chi-squared test to see if its p-value is less than 0.05. We also consider only IV states where the calculated conditional probability distribution is different from uniform, and this difference is statistically significant with the same p-value cutoff. Eight combinations of the IVs that have passed both of the above criteria are shown in Table 5. For example, in the first row, having Db=1, Dg=1, Iz=1, and Ja=2, the model 100% predicts Z=2. Or for the second row, having Db=1, Dg=1, Iz=2, and Ja=1, the model predict Z=2 for 22.13%. The column named 'rule' is the DV (Z) prediction result. The prediction rule indicates which DV state one should predict for any particular IV state. The rule is determined from the conditional probability distribution for the model, namely q(DV|IV). Specifically, the rule indicates the *most probable* DV state, the DV state whose q(DV|IV) is maximum, for the particular IV. In the present case, if rule=2, it means that one should predict Z=2, i.e., that the patents containing the IVs (the keywords) will be cited. If rule=1, it means that one should predict Z=1, i.e., that the patents will not be cited. So, for example, in Table 5, for IV state (Db, Dg, Iz, Ja) = (1,1,2,1), the probability of Z=1 is .779, while the probability of Z=2 is .221, so one predicts Z=1. In addition to Table 5, OCCAM generates individual tables for each of the predicting IVs in the model. These tables, shown in Appendix 4, reveal that none of the other three IVs alone ever predict Z=2, but Ja predicts Z=2 100% when Ja=2. This shows that this analysis method can pick up predictive interaction effects involving several IVs even though the IVs are not individually predictive.

IVs				Data			Model		
					obs. p(DV|IV)		calc. q(DV|IV)		
Db	Dg	Iz	Ja	freq	Z=1	Z=2	Z=1	Z=2	rule
1	1	1	2	56	0	100	0	100	2
1	1	2	1	81	77.8	22.2	77.9	22.1	1
1	1	2	2	4	0	100	0	100	2
1	2	1	2	4	0	100	0	100	2
1	2	2	1	14	85.7	14.3	85.56	13.4	1
1	3	1	2	8	0	100	0	100	2
1	3	2	1	13	23.1	76.9	22.1	77.9	2
2	1	1	2	5	0	100	0	100	2
2	3	2	1	7	0	100	3.8	96.2	2
--	--	--	--	2483	94.7	5.3	94.7	5.3	1

Table 5 Variable based analysis for the BIC model (IV:DbZ:DgZ:IZ:JaZ) (Probabilities p & q are shown as %)

4.3. State-based analysis

In state-based analysis, only the IVs (Db, Dg, Iz, and Ja) are used; the remaining IVs are ignored. The result of state-based analysis is given in Appendix 5. The best model for all three criteria (BIC, AIC, and incremental-p) is IV:Db1Z:Dg3Iz2Ja1Z:Iz1Ja1Z:Ja2Z:Z. This state-based model has four specific states added in the order listed to the independence model DbDgIzJa:Z. As shown in Table 7, this model is slightly better in terms of $\triangle$DF, %$\triangle$H(DV), and $\triangle$BIC than the (BIC) best variable-based model with loops, namely IV:DbZ:DgZ:IZ:JaZ, previously given in Table 4.

	Model	$\triangle$DF	P-value	%$\triangle$H(DV)	$\triangle$AIC	$\triangle$BIC	Inc.p-value	%C(Data)
State-based	IV:Db$_1$Z:Dg$_3$Iz$_2$Ja$_1$Z:Iz$_1$Ja$_1$Z:Ja$_2$Z:Z	4	0	87.3	887.0	863.8	0.00	98.8
Variable-based	IV:DbZ:DgZ:IZ:JaZ	5	0	86.6	878.8	849.7	0	98.8

Table 6 Comparing the models resulted from variable-based analysis and state-based analysis

The conditional probability distribution of the state-based model is shown in Table 7, which also provides other detail omitted in Table 5 above. p-rule is the p-value for testing the deviation of the calculated conditional probability distribution q(DV|IV) from equal likelihood of Z=1 or 2, and p-margin is the p-value for testing its deviation from the marginal p(DV) distribution, which is (94.7%, 5.3%). Six IV states (not shaded) have a lower probability of being cited (Z=2) and nine IV states (shaded) have a higher probability of being cited. Components of the state-based model, IV:Db1Z : Dg3 Iz2 Ja1 Z : Iz1 Ja1 Z : Ja2 Z : Z, point to important IV states: (Dg = 3, Iz =2, Ja = 1) and (Ja = 2) states always predict citation, while (Iz = 1, Ja = 1) states never predict citation.

IV				Data			Model						
					obs. p(DV|IV)		calc. q(DV|IV)				%correct		p-margin
Db	Dg	Iz	Ja	freq	Z=1	Z=2	Z=1	Z=2	rule	#correct		p-rule	
1	1	1	1	1583	100	0	100	0	1	1583	100	0.00	0.00
1	1	1	2	56	0	100	0	100	2	56	100	0.00	0.00
1	1	2	1	81	77.8	22.2	79.4		1	63	77.7	0.00	0.00
1	1	2	2	4	0	100	0	100	2	4	100	0.05	0.00
1	2	1	1	343	100	0	100	0	1	343	100	0.00	0.00
1	2	1	2	4	0	100	0	100	2	4	100	0.05	0.00
1	2	2	1	14	85.7	14.3	79.4		1	12	85.7	0.03	0.01
1	3	1	1	208	100	0	100	0	1	208	100	0.00	0.00
1	3	1	2	8	0	100	0	100	2	8	100	0.00	0.00
1	3	2	1	13	23.1	76.9	21.2	78.8	2	10	76.9	0.04	0.00
2	1	1	1	65	100	0	100	0	1	65	100	0.00	0.06
2	1	1	2	5	0	100	0	100	2	5	100	0.03	0.00
2	2	1	1	42	100	0	100	0	1	42	100	0.00	0.13
2	3	1	1	27	100	0	100	0	1	27	100	0.00	0.22
2	3	2	1	7	0	100	3.4	96.6	2	7	100	0.01	0.00
--	--	--	--	2483	94.7	5.3	94.7	5.3	1	2454	98.8	--	--

All of the IV states are statistically significant (p < 0.05).

Table 7 State-based BIC model IV:Db_1Z:$Dg_3Iz_2Ja_1$Z:Iz_1Ja_1Z:Ja_2Z:Z (Probabilities p & q are shown as %)

4.4. Prediction Rules

As pointed out earlier, each of the three classes of RA analysis has a different degree of refinement. Variable-based (VB) analysis without loops, VB with loops, and state-based (SB) analyses allow coarse, refined, and ultra-refined modeling respectively [39]. In this case, four variables with the largest effects in DV prediction, for VB without loops analysis, are Db, Dg, Iz, and Ja. Furthermore, more detailed models made of the four variables are recognized in VB with loops and SB analyses. Db, Dg, Iz, and Ja represent 'chamber', 'hous', 'main', and 'return' keywords correspondingly. While both Table 5 and Table 7 contain detailed information about Db, Dg, Iz, and Ja as predictors of Z, Table 7 is based on the more refined state-based model, so we use Table 7 as the basis of the prediction. Table 8, based on Table 7, shows the citation prediction of the IV states whose p-values are less than 0.05.

To display which combination of keyword frequencies predicts the patent citations, we need to convert the bin numbers of the four IVs in Table 7 into their corresponding frequency numbers or ranges; this is done in Table 8.The relation between the bin numbers and the frequency numbers are given above in Table 2. So, for example bin number 1 for Db, the upper left number in Table 7, corresponds to frequency 0, the upper left number in Table 8. Table 8, shows how we can predict the patent citations via the keywords frequencies. Specifically, it shows nine predictions from keywords frequencies of whether patents with these frequencies would be cited. Note that two of the IV states ({Db=1, Dg=1, Iz=2, Ja=1} and {Db=1, Dg=2, Iz=2, Ja=1}) in Table 7 which have a prediction rule of

1 (which predict not being cited), their $q(DV=2|IV)$ are 20.6%, almost 4 times bigger than 5.276%, which is the default probability of being cited in the data.

#	Keyword Frequency				Freq.	Probability (%)	
	Chamber (Db)	Hous (Dg)	Main (Iz)	Return (Ja)		Not cited	Cited
1	0	0	0	>0	56	0	100
2	0	0	>0	0	81	79.3	20.6
3	0	0	>0	>0	4	0	100
4	0	1	0	>0	4	0	100
5	0	1	>0	0	14	79.3	20.6
6	0	>1	0	>0	8	0	100
7	0	>1	>0	0	13	21.2	78.7
8	>0	0	0	>0	5	0	100
9	>0	>1	>0	0	7	3.4	96.5
--	--	--	--	--	2483	94.7	5.3

All of the IV states are statistically significant ($p < 0.05$).

Table 8 Patent citation predictions: the probabilities of being cited for the 9 keyword combinations all exceed the marginal probability (on the last line) of 5.3%

Of the nine prediction rules (rows) in Table 8, seven are combinations of the keywords that predict that patents will be cited. These seven can be grouped to summary rules I & II which are:

I: **Return > 0** **predicts $P_{citation}$ = 100%, #patents = 77**

This encompasses individual rules 1, 3, 4, 6 and 8 in the above table. The second summary rule has primary and secondary components which encompass individual rules 7 and 9, respectively, in the table. This summary rule is

For Return = 0,

II: **Hous > 1 & Main > 0** **predicts $P_{citation}$ = 78.7%, #patents = 13**
IIa: **Hous > 1 & Main > 0 & Chamber > 0** **predicts $P_{citation}$ = 96.5%, #patents = 7**

5. Discussion

To see how these four keywords enhance patent citation prediction, one must dig into the patents to understand to what the keywords are referring in the abstracts. To do this, rules I and IIa are considered as the basis of the discussion (see section 4.4). Rule I expresses the fact that if a smart building patent contains 'return' or its derivatives in its abstract, the patent will definitely be cited (100% probability). Rule IIa indicates that a smart building patent will be cited with 96.5% probability if it contains 'main' and 'chamber' and at least two derivatives of 'house' in its abstract.

'Return' has two main roles in the abstracts of smart building patents. First, 'return' occurs in the abstract of smart building patents when something such as air, or water is circulating in a HVAC system and also occurs when someone such as a user or an occupant is going back to a building. For example, 'return air' refers to how 'return air' may be used as a cooling source in HVAC systems in US20080265046; 'return water' acts as a part of pool heater in US5560216, an automatic washer in US5241843, a water heater system in US20080265046, and a steam heating system in US20080223947. In addition, 'occupants return' and 'user return' are mentioned in the abstract of patents related to communicating control systems (such as US20080217419, US8386082 and US20110125329)

9

where the system manages the energy consumption based the presence or absence of a user or an occupant. Second, 'return' is used in the abstract of smart building patents when a main part of the invention functions to return something. 'A return plenum pressure controller' and 'a return fan control system' in an HVAC system (US8326464, US20130096722, and US20100057258), 'return line connecting an output of the chiller to an input of the cooling tower' in a climate control system (US20100201125), 'return ducts' in a dehumidifier system (US20060086112), and 'non-return valve' in a water supply system (US7066197) are examples of the main parts that function to return something.

'Main', 'chamber', and derivatives of 'house' in the smart building patents provide inventions that generate a form of energy, more specifically heat, or convert heat to another form of energy. A group of patents including US7978468, US7969735, US20110235270, and US20100188814 explain "a power converter that converts input power to a specific type of power and outputs the power resulting from the conversion". In this technology, a semiconductor module is housed inside a cooling chamber with the peripheral wall thereof constituted of a thermally conductive material to reduce the effect of heat from the semiconductor module on the other components. In addition, patent US20060236906 describes another invention that provides an automatic waste burning heating device. This technology is applied in poultry industry to burn poultry litter to generate and consume heat in buildings for different purposes.

Prediction Rules	Key Phrases	Inventions
Rule I: (return >0)	**return** users, **return** occupants	Communication Control Systems
	return fan control system	HVAC Systems
	return line connecting an output of the chiller to an input of the cooling tower	Climate Control Systems
	return duct	Dehumidifier Systems
	non-**return** valve	Water Supply Systems
Rule IIa: (main>0, hous>1, chamber>0)	- A casing **houses**: semiconductor modules, constituting a **main** circuit for power conversion;… - Within the casing, a cooling **chamber** including a coolant passage is formed, and a **chamber** wall of the cooling chamber is formed with a thermally conductive material. - At least the semiconductor modules, are **housed** inside the cooling **chamber**, and at least the capacitor and the control circuit are disposed outside the cooling **chamber**.	Power Converters
	- An automatic biomass fuel burning heating device and method comprising a burn **chamber** having … - A burn **chamber** having a feed auger opening in a burn **chamber** wall. - The control maintain supply air into the brooder **house** at a constant set temperature by varying the volume rate of air flow based on air inlet temperature and the temperature in the burn **chamber**.	Waste litter heater

Table 9 Predicted inventions that will be cited

In summary, it is expected that smart building inventions related to communication control systems, HVAC systems, climate control systems, dehumidifier systems, and water supply systems will be cited, according to rule I. In addition, smart building inventions related to power converters and waste litter heater will be cited, according to rule IIa. The summary of predictions and their related rules are illustrated in Table 9.

6. Conclusion

Patent citation is used as a proxy for technological impact studies, but is subject to the limitations of each patent's contents [5]. The importance of patent citation prediction is revealed when we see 94% of smart building patents of this study are not cited. In this research, a keyword-based method is developed to predict patent citations. The keywords are extracted by applying Weka, a text mining software program developed by the University of Waikato (New Zealand) [26]. Keywords are analyzed by Reconstructability Analysis (RA) [28] to discover keyword patterns in promising patent citations. This method enables us to both predict patent citations as proxies of technological impacts and to find out which aspects of technologies cause the impacts by interpreting the associated keywords.

Three different classes of RA searches are applied: 1) variable-based models without loops, 2) variable-based models with loops, and 3) state-based models. These models give us coarse, refined, and ultra-refined models, as shown in Figure 2 [33]. As a result of RA analysis, four keywords, including 'chamber', 'house' and its derivatives, 'main', and 'return', emerged as the keywords whose frequency in a patent related to smart building technologies which were the most predictive indicators of likelihood to be cited by other patents. The specific combinations of frequency of the keywords are summarized in Table 8. According to the keywords investigated in the abstracts, some smart building inventions, shown in Table 9, are predicted to be cited.

Practitioners can take the advantage of our method to not only explore in patent databases to find those patents leveraging technological changes in their industry like smart buildings, but also dig into the patents more efficiently by considering the keywords recognized by RA to identify the technologies. In keyword-based studies, like ours, having a comprehensive thesaurus or dictionary of the research area is ideal for the researchers, but professional thesaurus or dictionaries are rarely available for specific emerging technologies, such as smart building technologies. Semantic analysis can be employed as a remedy in the future studies to compensate for the lack of professional thesaurus or dictionaries. WordNet, a lexical database made by Princeton University [40] enables researchers to analyze their corpus semantically, e.g., patents, to extract all possible keywords related to the main concept of their research.

The use of Reconstructability Analysis in this research makes several theoretical contributions. First, the analytical procedures described in this paper show how data can be analyzed in an exploratory mode, eliminating the need to hypothesize and then test specific predictive relations whose form is explicitly specified. Second, these procedures allow one to detect interaction effects involving multiple predictors, even when the individual IVs do not have significant predictive effects. Third, the predictive models are conceptually transparent, being simple conditional probability distributions. This differentiates these methods from other data mining techniques, such as neural networks, which are 'black boxes,' where predictive models are not as easy to interpret and are not strongly associated with statistical measures. Fourth, this work outlines a hierarchical procedure where exploratory analysis is done at different degrees of refinement; this allows flexibility in the analysis under varying conditions of sample size and computational capabilities. (RA requires sample sizes that are larger than those required for standard linear regression analyses.)

Ideally, these methods should be applied in conjunction with cross-validation techniques, i.e., subjecting models obtained from training data to separate test data; this has not been done in this study. Future research should include cross-validation assessment of predictive models. It should expand % correct calculations to include the explicit analysis of sensitivity and specificity, i.e., the subdivision of incorrect predictions into false negatives and false positives. Also, more specifically, as an expansion of this particular study, analysis should consider the predictive keywords found in models favored by the less conservative model selection criteria (AIC and IncrP) that are not found in models favored by the preferred model selection criterion (BIC). As shown in Table 4, these keywords include variables Am, As, Eb, and Fd, which correspond to keywords control, dev, head, and flow. In addition, although state-based calculations take considerable computer time, state-based runs can be done with additional predictors, i.e., 5 or 6, as opposed to the 4 reported in this paper.

Abbreviations

Abbreviation	Keyword	Description
RA	Reconstructability Analysis	The methodology used to predict patent citations based on keywords extracted from abstract of patents.
IV	Independent Variable	IVs are the inputs (keywords) of the prediction model.
DV	Dependent Variable	DV is the output of the model, predicted by the IVs.
VB	Variable-based	Three different classes of RA models are applied in this research: 1) variable-based (VB) model without loops, 2) variable-based (VB) model with loops, and 3) state-based (SB) models (which typically have loops).
SB	State-based	
BIC	Bayesian Information Criterion	BIC and AIC are generally used to define the best models. These criteria trade off uncertainty reduction and model simplicity in different ways [39]. BIC, the more conservative criterion, is favored in this study. A third criterion, Incremental-p, is also used; it picks the highest information model whose difference from independence is statistically significant and for which a path from independence exists where each increment of complexity is also statistically significant,
AIC	Akaike Information Criterion	
DF	Degree of Freedom	The number of parameters in a model, the measure of its complexity.
H	Uncertainty	Information theoretic measure of variable unpredictability. The reduction of uncertainty of a DV achieved by knowing the IVs is roughly analogous to the %variance explained, except small values of uncertainty reduction can indicate big effect sizes.

Appendix 1- The extracted keywords from the patents

IV	Keyword	IV	Keyword	IV	Keyword	IV	Keyword	IV	Keyword	IV	Keyword
aa	power	bp	panel	de	threshold	eu	Condim	gj	sampl	ia	follower
ab	energ	bq	valv	dg	hous	ev	Scor	gk	condenser	ib	gener
ac	light	br	reflect	dh	coil	ew	Beam	gl	wir	ic	assemb
ad	ga	bs	switch	di	room	ex	Test	gm	model	id	wast
ae	member	bt	van	dj	dur	ey	Cool	gn	dat	ie	enclosur
af	bal	bu	cit	dk	select	ez	Condit	go	inform	if	hom
ag	golf	bv	electron	dl	timer	fa	Funct	gp	st	ig	load
ah	hydr	bw	fir	dm	door	fb	Cover	gq	ba	ih	steam
ai	finger	bx	thermostat	dn	lock	fc	Cast	gr	saving	ii	3-d
aj	lift	by	transfer	do	fuel	fd	Flow	gs	stat	ij	mot
ak	food	bz	port	dp	liquid	fe	Ha	gt	rot	ik	picture
al	cel	ca	carbur	dq	treatm	ff	Hav	gu	imag	il	cabinet
am	control	cb	commun	dr	semiconduc	fg	Invent	gv	hot	im	fig
an	il	cc	transformer	ds	structur	fh	Melt	gw	crop	in	exchang
ao	circuit	cd	inst	dt	firebox	fi	Metal	gy	exerc	io	loc
ap	wiper	ce	park	du	detect	fj	Molt	gz	provid	ip	point
aq	system	cf	wind	dv	upper	fk	NoZl	ha	interfac	iq	motor
ar	aud	cg	air	dw	air-m	fl	Par	hb	sh	ir	rf
as	dev	ch	shel	dx	posit	fm	Pat	hc	lower	is	stick
at	pump	ci	cabl	dy	mach	fn	Pool	hd	diffuser	it	light-em
au	channel	cj	sol	dz	sid	fo	Prov	he	composit	iu	guest
av	endotrach	ck	weather	ea	el	fp	Strip	hf	layer	iv	chain
aw	tub	cl	em	eb	head	fq	Substr	hg	outer	iw	user
ax	compon	cm	ar	ec	print	fr	Surface	hh	protect	ix	appl
ay	opt	cn	alarm	ed	fireplac	fs	U	hi	period	iy	dish
az	hvac	co	brush	ee	chair	ft	uniform	hj	led	iz	main
ba	pressur	cp	water	ef	sen	fu	weir	hk	acoust	ja	return
bb	electr	cq	launch	eg	temper	fv	step	hl	reson	Z*	Citation
bc	heat	cr	vacuum	eh	oper	fw	filter	hn	refriger		
bd	network	cs	ozon	ei	concentr	fx	vibr	ho	se		
be	hydraul	ct	sc	ej	displ	fy	clamp	hp	barrel		
bf	sign	cu	therm	ek	tank	fz	box	hq	ccfl		
bg	unit	cv	oil	el	greenh	ga	build	hr	bod		
bh	modl	cw	miner	em	bask	gb	lin	hs	oil-st		
bi	level	cx	subst	en	remov	gc	sewer	ht	tissu		
bj	setpoint	cy	brak	eo	row	gd	mater	hu	mirror		
bk	compr	cz	dist	ep	cut	ge	veloc	hv	clean		
bl	sect	da	fluid	eq	cl	gf	infrar	hw	fixtur		
bm	extern	db	chamber	er	roof	gg	zon	hx	receiv		
bn	wal	dc	dril	es	combust	gh	forml	hy	sub-chamber		
bo	laser	dd	spac	et	Floor	gi	siz	hz	bottom		

* Z is the dependent variable (DV).

Appendix 2- OCCAM results – loopless variable-based analysis

ID	MODEL	Level	△DF	p-value	%△H(DV)	△AIC	△BIC	%C(Data)
13	IV:AqBgIzJaZ	4	95	0.00	88.92	722.10	169.46	98.5
12	IV:BaDgIzJaZ	4	35	0.00	88.38	836.60	633.00	98.9
11	IV:AmDgIzJaZ	4	71	0.00	88.35	764.27	351.25	98.9
10*	IV:DgIzJaZ	3	11	0.00	85.75	857.58	793.59	98.6
9	IV:BgIzJaZ	3	15	0.00	85.34	845.41	758.15	98.1
8*	IV:DbIzJaZ	3	7	0.00	84.98	857.75	817.03	98.5
7*	IV:IzJaZ	2	3	0.00	83.33	848.76	831.31	98.1
6*	IV:DbJaZ	2	3	0.00	58.46	593.67	576.22	98.1
5*	IV:DgJaZ	2	5	0.00	57.11	575.77	546.69	98.1
4*	IV:JaZ	1	1	0.00	54.89	561.03	555.21	98.1
3*	IV:IZ	1	1	0.00	15.46	156.56	150.75	94.7
2*	IV:AcZ	1	4	0.00	3.32	26.08	2.81	94.7
1*	IV:Z	0	0	1.00	0.00	0.00	0.00	94.7

'*' indicates those models whose difference from their lower level progenitor is statistically significant

Best Model(s): by △BIC is 7*, by △AIC is 8*, and by Information, with all Inc. p-value < 0.05 is 10*.

Appendix 3- OCCAM results- variable-based analysis with loops

ID	MODEL	Level	△DF	p-value	%△H(DV)	△AIC	△BIC	Inc. p-value	%C(Data)
22*	IV:AmZ:AsZ:DgZ:EbZ:FdZ:IZ:JaZ	7	15	0.00	90.9	902.5	815.2	0.00	99.3
21*	IV:AmZ:AsZ:CbZ:DgZ:EbZ:IZ:JaZ	7	15	0.00	90.9	902.2	814.9	0.00	99.2
20*	IV:AmZ:AsZ:BgZ:CbZ:DgZ:IZ:JaZ	7	17	0.00	90.8	897.5	798.6	0.00	99.3
19*	IV:AmZ:AsZ:DgZ:EbZ:IZ:JaZ	6	13	0.00	89.9	895.7	820.1	0.00	99.2
18*	IV:AmZ:AsZ:DcZ:DgZ:IZ:JaZ	6	13	0.00	89.6	893.0	817.4	0.00	99.1
17*	IV:AmZ:AsZ:CbZ:DgZ:IZ:JaZ	6	14	0.00	89.6	890.8	809.4	0.00	99.1
16*	IV:AmZ:AsZ:DgZ:IZ:JaZ	5	12	0.00	88.5	883.4	813.6	0.00	99.0
15*	IV:AmZ:CbZ:DgZ:IZ:JaZ	5	11	0.00	88.2	882.4	818.4	0.00	98.9
14*	IV:AmZ:DgZ:EbZ:IZ:JaZ	5	10	0.00	88.1	883.9	825.7	0.00	99.1
13*	IV:AmZ:DgZ:IZ:JaZ	4	9	0.00	86.9	873.6	821.2	0.00	98.9
12*	IV:BgZ:DgZ:IZ:JaZ	4	7	0.00	86.7	875.3	834.6	0.00	98.7
11*	IV:DbZ:DgZ:IZ:JaZ	4	5	0.00	86.6	878.8	849.7	0.00	98.8
10*	IV:DgZ:IZ:JaZ	3	4	0.00	85.3	867.4	844.1	0.00	98.7
9*	IV:BgZ:IZ:JaZ	3	5	0.00	84.9	861.1	832.0	0.00	98.2
8*	IV:DbZ:IZ:JaZ	3	3	0.00	84.6	861.8	844.3	0.00	98.6
7*	IV:IZ:JaZ	2	2	0.00	83.0	847.1	835.4	0.00	98.1
6*	IV:DbZ:JaZ	2	2	0.00	58.4	595.3	583.7	0.00	98.1
5*	IV:DgZ:JaZ	2	3	0.00	57.1	579.6	562.1	0.00	98.1
4*	IV:JaZ	1	1	0.00	54.9	561.0	555.2	0.00	98.1
3*	IV:IZ	1	1	0.00	15.5	156.6	150.7	0.00	94.7
2*	IV:AcZ	1	4	0.00	3.3	26.1	2.8	0.00	94.7
1*	IV:Z	0	0	1.00	0.0	0.0	0.0	0.00	94.7

'*' indicates those models whose difference from their lower level progenitor is statistically significant

Best Model(s): by △BIC is 11*, by both △AIC and Incremental-p is 22*.

Appendix 4- OCCAM results- variable-based analysis with loops: the distributions corresponding to the four separate components of IV:DbZ:DgZ:IZ:JaZ Model

IV	Data			rule	#correct	%correct
	freq	obs. p(DV|IV)				
Db		Z=1	Z=2			
1	2316	95.50	4.49	1	2212	95.50
2	167	83.83	16.16	1	140	83.83
	2483	94.72	5.27	1	2352	94.72

IV	Data			rule	#correct	%correct
	freq	obs. p(DV|IV)				
Dg		Z=1	Z=2			
1	1806	94.96	5.03	1	1715	94.96
2	410	97.31	2.68	1	399	97.31
3	267	89.13	10.86	1	238	89.13
	2483	94.72	5.27	1	2352	94.72

IV	Data			rule	#correct	%correct
	freq	obs. p(DV|IV)				
Iz		Z=1	Z=2			
1	2346	96.675	3.325	1	2268	96.675
2	137	61.314	38.686	1	84	61.314
	2483	94.724	5.276	1	2352	94.724

IV	Data			rule	#correct	%correct
	freq	obs. p(DV|IV)				
Ja		Z=1	Z=2			
1	2399	98.04	1.95	1	2352	98.04
2	84	0	100	2	84	100
	2483	94.72	5.27	1	2436	98.10

Appendix 5- OCCAM results- state-based-based analysis for Db, Dg, Iz, and Ja variables

ID	MODEL	Level	△DF	P-value	%△H(DV)	△AIC	△BIC	Inc. p-value	Prog.	%C(Data)
16	IV:Db1Dg1Iz1Ja2Z:Db1Z:Dg3Iz2Ja1Z:Iz1Ja1Z:Ja2Z:Z	5	5	0	87.25	885.01	855.93	0.91	13	98.8
15	IV:Db1Dg1Iz1Ja2Z:Db1Z:Dg3Iz2Z:Iz1Ja1Z:Ja2Z:Z	5	5	0	87.25	885.01	855.92	0.89	12	98.8
14	IV:Db1Dg1Iz1Ja2Z:Db1Z:Dg3Ja1Z:Iz1Ja1Z:Ja2Z:Z	5	5	0	87.23	884.83	855.74	0.93	11	98.8
13*	IV:Db1Z:Dg3Iz2Ja1Z:Iz1Ja1Z:Ja2Z:Z	4	4	0	87.25	887.03	863.76	0.00	9	98.8
12*	IV:Db1Z:Dg3Iz2Z:Iz1Ja1Z:Ja2Z:Z	4	4	0	87.25	887.02	863.76	0.00	10	98.8
11*	IV:Db1Z:Dg3Ja1Z:Iz1Ja1Z:Ja2Z:Z	4	4	0	87.23	886.83	863.56	0.00	8	98.8
10*	IV:Dg3Iz2Z:Iz1Ja1Z:Ja2Z:Z	3	3	0	85.82	874.32	856.87	0	7	98.6
9*	IV:Dg3Iz2Ja1Z:Iz1Ja1Z:Ja2Z:Z	3	3	0	85.82	874.32	856.87	0	7	98.6
8*	IV:Dg3Ja1Z:Iz1Ja1Z:Ja2Z:Z	3	3	0	85.76	873.70	856.25	0	7	98.6
7*	IV:Iz1Ja1Z:Ja2Z:Z	2	2	0	83.42	851.71	840.08	0	2	98.1
6*	IV:Db1Iz2Ja1Z:Iz1Ja1Z:Z	2	2	0	82.80	845.30	833.67	0	4	98.5
5*	IV:Iz1Z:Ja2Z:Z	2	2	0	82.79	845.27	833.64	0	2	98.1
4*	IV:Iz1Ja1Z:Z	1	1	0	71.83	734.83	729.01	0	1	96.6
3*	IV:Db1Iz1Ja1Z:Z	1	1	0	55.03	562.46	556.64	0	1	94.7
2*	IV:Ja2Z:Z	1	1	0	54.88	560.92	555.10	0	1	98.1
1*	IV:Z	0	0	1	0	0	855.93	0	0	94.7

'*' indicates those models whose difference from their lower level progenitor (Prog.) is statistically significant

Best Model(s): by △BIC is 13*, by △AIC is 13*, and by Information, with all Inc. p-value < 0.05 are 13*, 12*, and 11*.

References

[1]   C. Lee, Y. Cho, H. Seol, and Y. Park, "A stochastic patent citation analysis approach to assessing future technological impacts," *Technol. Forecast. Soc. Change*, vol. 79, no. 1, pp. 16–29, Jan. 2012.

[2]   C. Lee, J. Jeon, and Y. Park, "Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach," *Technol. Forecast. Soc. Change*, vol. 78, no. 4, pp. 690–702, May 2011.

[3]   P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, "Prediction of emerging technologies based on analysis of the US patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, Jun. 2012.

[4]   S.-B. Chang, K.-K. Lai, and S.-M. Chang, "Exploring technology diffusion and classification of business methods: Using the patent citation network," *Technol. Forecast. Soc. Change*, vol. 76, no. 1, pp. 107–117, 2009.

[5]   B. Yoon and Y. Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *J. High Technol. Manag. Res.*, vol. 15, pp. 37–50, 2004.

[6]   R. Li, T. Chambers, Y. Ding, G. Zhang, and L. Meng, "Patent citation analysis: Calculating science linkage based on citing motivation," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 5, pp. 1007–1017, May 2014.

[7]   K. Kousha and M. Thelwall, "Patent citation analysis with Google," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 1, pp. 48–61, Jan. 2017.

[8]   B. Yoon and S. Lee, "Patent analysis for technology forecasting: Sector-specific applications," in *2008 IEEE International Engineering Management Conference*, 2008, pp. 1–5.

[9]   X. Hu, R. Rousseau, and J. Chen, "A new approach for measuring the value of patents based on structural indicators for ego patent citation networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 9, pp. 1834–1842, Sep. 2012.

[10]   D. Harhoff, F. Narin, F. M. Scherer, and K. Vopel, "Citation Frequency and the Value of Patented Inventions," *Rev. Econ. Stat.*, vol. 81, no. 3, pp. 511–515, Aug. 1999.

[11]   B. H. Hall, A. Jaffe, and M. Trajtenberg, "Market Value and Patent Citations," *RAND J. Econ.*, vol. 36, no. 1, pp. 16–38, 2005.

[12]   M. Karvonen and T. Kässi, "Patent citations as a tool for analysing the early stages of convergence," *Technol. Forecast. Soc. Change*, vol. 80, no. 6, pp. 1094–1107, Jul. 2013.

[13]   Y. Kajikawa and Y. Takeda, "Citation network analysis of organic LEDs," *Technol. Forecast. Soc. Change*, vol. 76, no. 8, pp. 1115–1123, 2009.

[14]   G. Kim and J. Bae, "A novel approach to forecast promising technology through patent analysis," *Technol. Forecast. Soc. Change*, 2016.

[15]   A. Breitzman and P. Thomas, "The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems," *Res. Policy*, vol. 44, no. 1, pp. 195–205, 2015.

[16]   D.-R. Liu and M.-J. Shih, "Hybrid-patent classification based on patent-network analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 246–256, 2010.

[17]   F. Madani and C. Weber, "The evolution of patent mining: applying bibliometrics analysis and keyword network analysis," *World Pat. Inf.*, vol. 46, pp. 32–48, 2016.

[18]   S. Jun, "A Technology Forecasting Method using Text Mining and Visual Apriori Algorithm," *Appl. Math. Inf. Sci.*, vol. 8, no. 1L, pp. 35–40, 2014.

[19]   G. J. Kim, S. S. Park, and D. S. Jang, "Technology Forecasting using Topic-Based Patent Analysis," *J. Sci. Ind. Res. (India).*, vol. 74, no. 5, 2015.

[20]   B. Yoon and Y. Park, "Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information," *IEEE Trans. Eng. Manag.*, vol. 54, no. 3, pp. 588–599, Aug. 2007.

[21]   S. Lee, B. Yoon, and Y. Park, "An approach to discovering new technology opportunities: Keyword-based patent map approach," *Technovation*, vol. 29, pp. 481–497, 2009.

[22]   S. Lee, S. Lee, H. Seol, and Y. Park, "Using patent information for designing new product and technology: keyword based technology roadmapping," *R&D Manag.*, vol. 38, no. 2, pp. 169–188, Mar. 2008.

[23]   R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, 2011, p. 1247.

[24]   H. Kim, A. Stumpf, and W. Kim, "Analysis of an energy efficient building design through data mining approach," *Autom. Constr.*, vol. 20, no. 1, pp. 37–43, Jan. 2011.

[25]   "Stopwords," *"Ranks" Company*, 2014. [Online]. Available: http://www.ranks.nl/stopwords.

[26]   M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor.*, vol. 11, no. 1, 2009.

[27]   J. Fusion, K. Willet, and M. Zwick, "OCCAM: A Reconstrctability Analysis Program," *Portland State University*, 2012. [Online]. Available: http://www.pdx.edu/sysc/research-discrete-multivariate-modeling.

[28]   W. R. Ashby, *Constraint Analysis of Many-Dimentional Relations*. General Systems Yearbook, 1964.

[29]   Y. Bishop, S. Feinberg, and P. Holland, *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press, 1978.

[30]   M. Perkowski, M. Marek-Sadowska, L. Jozwiak, T. Luba, S. Grygiel, M. Nowicka, R. Malvi, Z. Wang, and J. Zhang, "Decomposition of many-valued relations," in *ISMVL*, 1997, pp. 13–18.

[31]   M. Zwick, "An overview of reconstructability analysis," *Kybernetes*, vol. 33, no. 5, pp. 877–905, 2004.

[32]   G. J. Klir, *The Architecture of Systems Problem Solving*. NewYirk: Plenum Press, 1985.

[33]   M. Zwick, "Reconstructability Analysis of Epistasis," *Ann. Hum. Genet.*, vol. 75, no. 1, pp. 157–171, 2011.

[34]   M. Zwick, "State-based Reconstructability Analysis," *Kybernetes*, vol. 33, pp. 1041–1052, 2004.

[35]   J. Johnson, "State-based Systems Modeling: Theory, Implementation, and Application," Portland State University, 2005.

[36]   M. Zwick, "Reconstructability Analysis with Fourier Transforms," *Kybernetes*, vol. 33, pp. 877–905, 2004.

[37]   J. Hosseini and M. Zwick, "Segment Congruence Analysis Via Information Theory," in *International Society for General Systems Research*, 1986, pp. G62–G77.

[38]   K. Willet and M. Zwick, "A Software Architecture for Reconstructability Analysis," *Kybernetes*, vol. 33, pp. 997–1008, 2004.

[39]   P. Kramer, S. K. Westaway, M. Zwick, and S. Shervais, "Reconstructability analysis of genetic loci associated with Alzheimer disease," *6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst.*, pp. 2104–2110, Nov. 2012.

[40]   "Princeton University "About WordNet.," *WordNet. Princeton University*, 2010. [Online]. Available: http://wordnet.princeton.edu.