An Approach for Predicting Hype Cycle Based on Machine Learning

Zhijun Ren
Institute of Scientific and Technical Information of China,
Beijing 100038, China
The China Patent Information Center, the State
Intellectual Property Office, Beijing 100088, China
renzhijun@cnpat.com.cn

Shuo Xu
Institute of Scientific and Technical Information of China,
Beijing 100038, China
xush@istic.ac.cn

Xiaodong Qiao
Institute of Scientific and Technical Information of China,
Beijing 100038, China
qiaox@istic.ac.cn

Hongqi Han
Institute of Scientific and Technical Information of China,
Beijing 100038, China
bithhq@163.com

Kai Zhang
The China Patent Information Center, the State
Intellectual Property Office, Beijing 100088, China
zhangkai@cnpat.com.cn

ABSTRACT

Analyzing mass information and supporting insight based on analysis results are very important work but it needs much effort and time. Therefore, in this paper, we propose an approach for predicting hype cycle based on machine learning for effective, systematic, and objective information analysis and future forecasting of science and IT field. Additionally, we execute a comparative evaluation between the suggested model and Hype Cycle for Big Data, 2013 for validating the suggested model and generally used for information analysis and forecasting.

Keywords

Hype Cycle; Data Mining; Technology Predict; KNN

1. OVERVIEW

Hype Cycle is a conceptual model widely used by Gartner, Inc., which can reveal the basic laws of the evolution of technological innovation chain and is a powerful tool to get the overall grasp of technological innovation development trend, to get the objective assessment of the maturity of technological innovation, to make the reasonable choice on innovative intervention time and to seek the technological innovation late-mover advantage.

Published at Ceur-ws.org

Proceedings of the Second International Workshop on Patent Mining and its Applications (IPAMIN). May 27–28, 2015, Beijing, China.

Some scholars believe that hype means advertised in publicity and exaggerated propaganda[1], and the best measure is the expectation, which means people's expectation for technology innovation. Enterprises should use hype cycle to target the emerging technologies, and use the concept of digital business transformation to predict future business trends.

Gartner believes that hype cycle is a qualitative decision-making tool, like other management methods, it relies mainly on the judgment of experts. And to complete the hype cycle assessment and prediction of a set of technologies in a certain field, we need to use a variety of evaluation methods. Other scholars have explored how to measure expectations for innovation, and quantitative indicators are mainly the number of participants[2], the number or the ratio of the technological innovation documents[3], patent statistical data[4] and the search flow of Google and other search engines[5]. When using the network measurement method, other tools may need to be supplemented: USPTO patents, news reports of Google news archive and the official website market share of a certain product, etc., these methods mainly create the hype cycle by adopting artificial methods.

The InSciTe[6,7] developed by Korea Institute of Science and Technology Information adopts the decision tree and statistical feature analysis method, based on Gartner research, to provide the technical life cycle diagram and adopts the emerging technologies discovering model to provide the key technologies on the life cycle diagram. In the particular judgment process, there might be problems of force compliance in certain stage. For example, a technology is during the plateau of productivity between the year from 2000 to 2005, but its data between 2006 to 2007 is in line with the stage of slope of enlightenment, so the force compliance is needed to judge the technology life cycle of the last two years[1]. Related reference didn't include how specific technology term coordinates are obtained.

Therefore, the paper uses papers and patent information to realize emerging technology discovering method by machine learning; then acquires some features by feature selection and uses machine learning algorithm to classify and locate the coordinates, and produces the hype cycle according to the prediction.

The paper is structured as follows. The prediction frame of technology maturity is described in Chapter 2. The learning method and model of technology maturity is introduced in Chapter 3. The prediction method and model of technology maturity is illustrated in Chapter 4. Experiment and analysis are conducted in Chapter 5 and conclusion and forecasting are made in the last chapter.

2. TECHNOLOGY MATURITY PREDICTION FRAME

The model of approach for predicting hype cycle based on machine learning includes the learning part and the prediction part. The learning model mainly relates to data training. The data acquiring and learning includes acquiring training data, data annotation and feature calculation. The prediction model means identification and discovering method of emerging technology in certain field and process to discover technology system innovation by producing hype cycle and predict maturity and discover an emerging technology, as well as to make sure the input ratio of partial innovation and overall innovation. It includes the term selection, feature calculation, stage classification, technology position and visible information module. The hype cycle prediction module is show in Fig. 1.

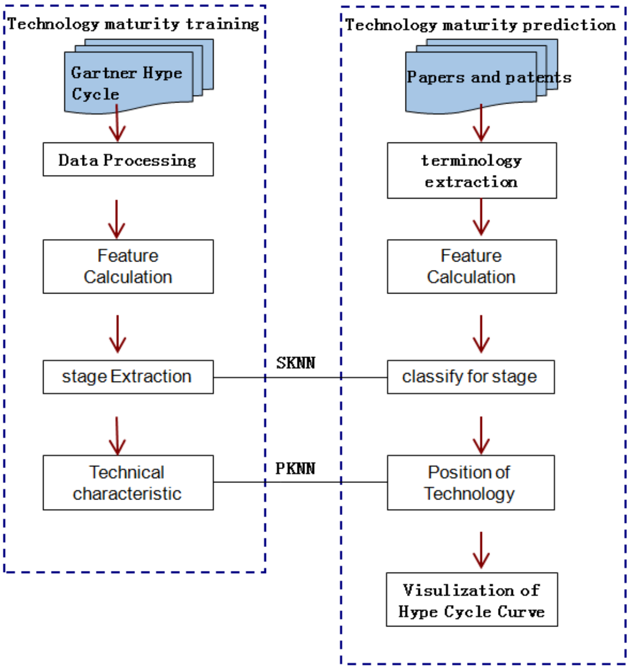

Fig.1 Technology maturity prediction frame

3. TECHNOLOGY MATURITY LEARNING

3.1 Data Acquirement

Since 1995, Gartner began to pay attention to the hype and disillusionment along with every appearance of new technologies and innovations, to track the trends of the technology life cycle, to study the common pattern between them, in order to provide guidance of when and where all types of organizations make technology deployed.

Data from Internet between 2001 and 2014 about Gartner Hype Cycle is manually collected on terminology, stage and coordinate to be used as training data.

3.2 Feature Calculation

Technical features are the foundation of technology life cycle discovering model. The technical features calculation uses the papers and patents as data, and uses paper index $S(Pp)=\{Pp_1,Pp_2,\ldots,Pp_n\}$, patent index $S(Pt)=\{Pt_1,Pt_2,\ldots,Pt_n\}$ and combined index of paper and patent $S(Ppt)=\{Ppt_1,Ppt_2,\ldots,Ppt_n\}$ as calculation objects. It can study the interaction and exclusion between papers and patents, and explore the rule of development between science and technology.

Paper index includes paper growth rate

$$Pp_{GrowthRate}=\frac{AN_{Pp}^{k}-AN_{Pp}^{k-1}}{AN_{Pp}^{k-1}}$$

, paper relative growth rate

$$Pp_{RelativeGrowthRate}=\frac{N_{Pp}^{k}-N_{Pp}^{k-1}}{N_{Pp}^{k-1}}$$

, paper author rate

$$Pp_{AuthorRate}=\frac{A_{Pp}^{k}}{AA_{Pp}^{k-1}}$$

, paper author growth rate

$$Pp_{AuthorGrowthRate}=\frac{A_{Pp}^{k}-A_{Pp}^{k-1}}{A_{Pp}^{k-1}}$$

, paper institution rate

$$Pp_{InstitutionRate}=\frac{I_{Pp}^{k}}{AI_{Pp}^{k-1}}$$

, and paper institution growth rate

$$Pp_{InstitutionGrowthRate}=\frac{I_{Pp}^{k}-I_{Pp}^{k-1}}{I_{Pp}^{k-1}}$$

.

Patent index includes patent growth rate

$$Pt_{GrowthRate}=\frac{AN_{Pt}^{k}-AN_{Pt}^{k-1}}{AN_{Pt}^{k-1}}$$

, patent relative growth rate

$$Pt_{RelativeGrowthRate}=\frac{N_{Pt}^{k}-N_{Pt}^{k-1}}{N_{Pt}^{k-1}}$$

, inventor rate

$$Pt_{InventorRate}=\frac{I_{Pt}^{k}}{AI_{Pt}^{k-1}}$$

, inventor growth rate

$$Pt_{InventorGrowthRate}=\frac{AI_{Pt}^{k}-AI_{Pt}^{k-1}}{AI_{Pt}^{k-1}}$$

, application rate

$$Pt_{ApplicantRate}=\frac{A_{Pt}^{k}}{AA_{Pt}^{k-1}}$$

, and application growth rate

$$Pt_{ApplicantGrowthRate}=\frac{AA_{Pt}^{k}-AA_{Pt}^{k-1}}{AA_{Pt}^{k-1}}$$

Combined index of paper and patent includes paper and patent relative growth rate

$$Ppt_{\mathrm{R}elativeGrowthRate} = \frac{(N_{Pp}^{k} - N_{Pp}^{k-1}) + (N_{Pt}^{k} - N_{Pt}^{k-1})}{N_{Pt}^{k-1} + N_{Pp}^{k-1}}$$

,

$$Ppt_{RatioRate} = \frac{N_{Pp}^{k}}{N_{Pt}^{k}}$$

paper and patent ratio rate , paper and patent people growth rate

$$Ppt_{PeopleGrowthRate} = \frac{(A_{Pp}^{k} - A_{Pp}^{k-1}) + (I_{Pt}^{k} - I_{Pt}^{k-1})}{A_{Pp}^{k-1} + I_{Pt}^{k-1}}$$

,

paper and patent people ration rate

$$Ppt_{PeopleRatioRate} = \frac{A_{Pp}^{k}}{I_{Pt}^{k}}$$

, paper and patent institution growth rate

$$Ppt_{InstitutionGrowthRate} = \frac{(I_{Pp}^{k} - I_{Pp}^{k-1}) + (A_{Pt}^{k} - A_{Pt}^{k-1})}{I_{Pp}^{k-1} + A_{Pt}^{k-1}}$$

,

paper and patent institution ration rate

$$Ppt_{InstitutionRatioRate} = \frac{I_{Pp}^{k}}{A_{Pt}^{k}}$$

.

4. TECHNOLOGY MATURITY PREDICTION

4.1 Terminology Extraction

Technology terminology refers to terms used in a certain field, which means concepts, features or relationships in the field. In this paper, terminology extraction is based on keywords of papers and templates. The keywords are word or phrase extracted from papers to meet the needs of literature indexing or retrieval work. It is used to express the literature subject and therefore can be used as emerging technology terminology.

Template technology is a common method for terminology recognition. By analyzing the characteristics of papers and patents, some fixed sentence structure are found, for example, ‘The development application of XX technology in XY ’, so ‘3D printing’ can be recognized in ‘The development application of 3D printing technology in medical science’. After certain strings are extracted from the template, frequency and the subjection degree can be used to perform terminology recognition. If a collocation is found in corpus, it must appear more than once, thus the frequency is an important index in terminology extraction. Only when the term frequency in the corpus exceeds a certain threshold value, it is believed that it has reached a technological terminology standard and awareness in the field is relatively high. Subjection degree refers to a relevant degree of the terminology and its filed. It represents the degree of a term belonging to a field. While meeting the frequency and subjection degree at the same time, the string is a technology terminology.

4.2 Classify for Stage

By calculating technical terminology and technical features extracted from patents and papers, TLCD model can determine the stage of the emerging technology adopting the five stages of classification of TSKNN algorithm, and the specific stages refer to the Gartner's Hype cycle, which includes Technology Trigger, Peak of Inflated, Expectations Trough of Disillusionment, Slope of Enlightenment, Plateau of Productivity.

SKNN algorithm is an improvement of KNN algorithm. KNN algorithm, by computing the distance between the training point from training set and test point from test set, considers the closest distance having the most similar feature and can be classified into the same group, obtaining test markers characteristic points and the same tag feature training points. SKNN mainly considers the time sequence issue of terminologies to be classified, so the data of next year need to be larger than the data of last stage, in order to avoid the force classification problem. The specific algorithm is as follows.

(1) Redescribe training technology terminology and feature vector, according to feature set.

(2) When the technology terminology feature vector reaches, 18 features should be calculated respectively to establish feature vector according to age.

(3) Select K technical terminologies which are most similar to new technical year that is to be calculated from the training technical terminology set following the formula below,

$$Sim(t_i, t_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^{2})(\sum_{k=1}^{M} W_{jk}^{2})}} \qquad (1)$$

(4) Among K neighbors of new terminology, weight of each classification is calculated respectively as follows,

$$p(\bar{x}, C_j) = \sum_{d_i \in STKNN} Sim(\bar{x}, \bar{d_i}) y(\bar{d_i}, C_j) \qquad (2)$$

In the formula, x refers to feature vector of the new terminology, Sim(x,di) refers to similarity calculation formula which is the same as mentioned in the last calculation formula, y(di,Cj) is categorical attributes function, if di belongs to Cj, then the function is equals to 1, otherwise is 0.

(5) Compare the weight of various classifications; distribute the computable terminologies to the stage with the greatest weight. Record the stage.

(6) Back to step 3 and calculate for next year.

4.3 Technical Position

Based on the classification results, PKNN algorithm calculates the position of terminologies on the hype cycle curve and S-curve. The algorithm is to find K most similar technologies in all classifications and the position of the technical terminology is the average position of these terminologies.

Input data includes trained technical terminology, feature vector and position under the classification, technical terminology and feature vector to be calculated.

Output data includes technical terminology position.

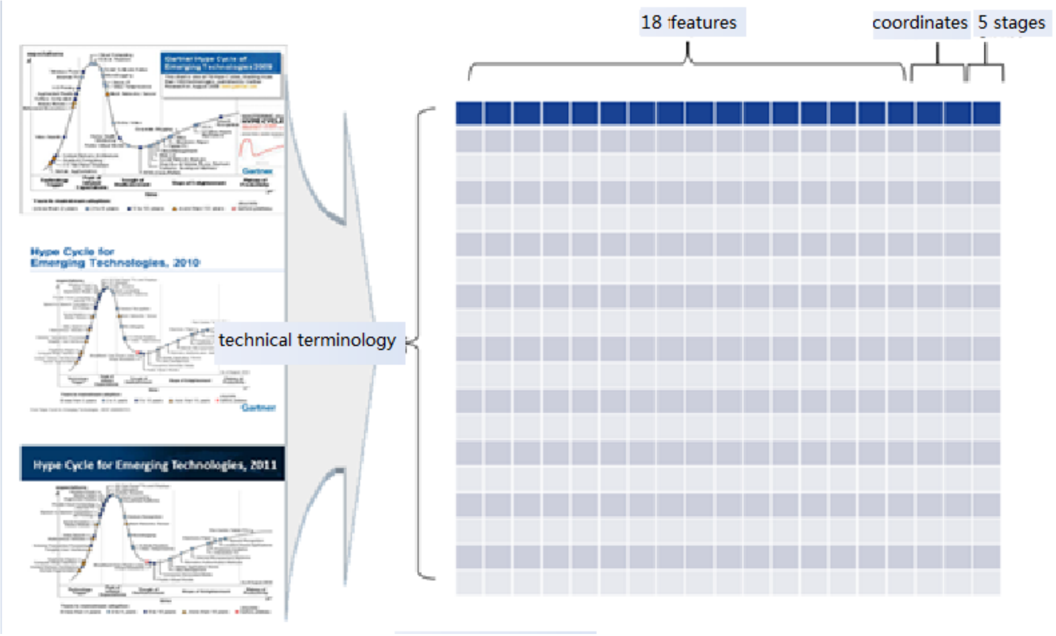
18 features
coordinates
5 stages
technical terminology
Hype Cycle for Emerging Technologies, 2010
Hype Cycle for Emerging Technologies, 2011

Fig.2 Feature Extraction

Algorithm steps are listed as follows.

(1) Select K technical terminologies which are most similar to the emerging technical terminology in the year to be calculated from the training technical terminology set.

$$\mathrm{pred}(y)=\frac{\sum_{k=1}^{M}[sim(x_k,y)*r_k]}{\sum_{k=1}^{M}sim(x_k,y)} \tag{3}$$

wherein K is 10.

(2) Among K neighbors of new terminologies, consider the deviation between the most similar K terminologies and the emerging technical terminology, and calculate the prediction position of new terminology.

4.4 Visualization of Hype Cycle Curve

The paper uses fitting algorithm to produce Hype cycle. The process is as follows.

(1) Draw coordinates of horizontal and vertical axis, and mark the expectation and time information.

(2) Produce Hype Cycle with curve fitting formula.

(3) Draw 5 stages line and label the text.

(4) Draw the technical terminology position according to the position calculated by maximum similarity algorithm.

5. EXPERIMENT AND ANALYSIS

Gartner Emerging Technologies report describes some technologies that become famous because of hype, or technologies that Gartner believes will have significant impact. In order to validate the technical life cycle discovery model, 960 training data released by ‘Hype Cycle Report’ from 2001 to 2013 is used for training technical maturity. Hype Cycle for Big Data, 2013 is used to evaluate validation set of prediction results.

Using SKNN method to perform Hype Cycle model stage test, Table 1 shows the experiment results of Hype Cycle for Big Data, 2013 data and technical life cycle model. Compared with Gartner, the technical life cycle model achieves the precision of 67.24% and recall rate of 68.46%. The reason why the accuracy and recall rate in the fifth stage and the fourth stage is lower than that of other stages is that the sample size in the fifth stage and the fourth stage is too small and the problem data has greater impact.

Table 1. Technical life cycle discovering model experiment result

Stage	Result				
	Gartner	Suggested approach	Number of results same in the both	Precision	Recall
Technology Trigger	11	11	10	91%	91%
Peak of Inflated Expectations	14	15	11	73.33%	78.6%
Trough of Disillusionment	11	9	8	88.89%	72.7%

Slope of Enlightenment	2	3	1	33%	50%
Plateau of Productivity	2	2	1	50%	50%
Total	40	40	33	67.24%	68.46%

Using SKNN method to perform Hype Cycle model position prediction, according to the Hype cycle visualization methods, Hype Cycle for Big Data, 2013 prediction result is produced and given as follows (This paper does not cover how each technology reaches the Plateau, and Gartner's result is used in the following visualized graph).

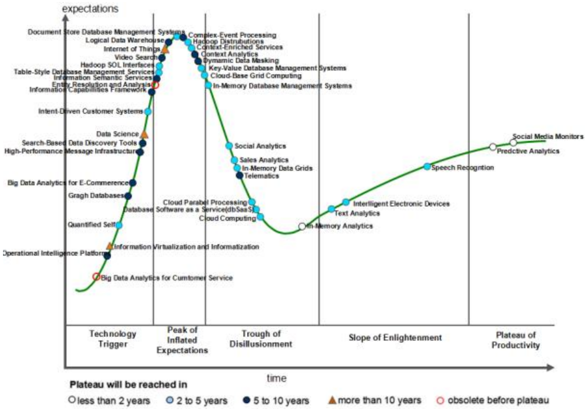

Fig.3 Gartner's Hype Cycle for Big Data,2013

6. CONCLUSION AND FORCASTING

The approach for predicting Hype Cycle based on machine learning discussed in this paper can effectively analyze paper and patent information, forecast and produce Hype Cycle. The approach consists of two processes of learning and training and provides users with a more diverse information analysis and forecasting tool. Compared with traditional methods and services, the approach is more systematic and objective.

Experimental results show that, compared to Hype Cycle for Big Data, 2013, the approach for predicting Hype Cycle based on machine learning achieves the accuracy of 68.46% and recall rate of 67.24% in forecasting stages, respectivly. Therefore, this model can provide more accurate information analysis and forecasting information to interested users，can locate the position of technology accurately, and produce the Hype Cycle automatically.

As for future work, considering the features of paper data and patent data, more features will be extracted and experiments in different databases to predict the experiment will be conducted; In addition, more machine learning data means more accurate results, so simulation experiments and different data sets will be used to further improve the accuracy of the prediction model.

This paper is funded and supported by the China Postdoctoral Science Foundation（2013M540125）.

7. REFERENCES

[1] Fenn J, Raskino M. Mastering the hype cycle: How to choose the right innovation at the right time[M]. Boston: Harvard Business School Press, 11-19, 2008.

[2] Guo Daoquan. The Study of Technology Maturity Model and Assessment based on TRL[D]. Changsha: National University of Defense Technology, 2010.

[3] J ärvenp ää H M, M äkinen S J. An empirical study of the existence of the Hype Cycle: A case of DVD technology[C]. IEEE International Engineering Management Conference. Estoril, Europe, 257-261, 2008.

[4] Budde B, Alkemadeb F, Hekkert M. On the relation between communication and innovation activities: A comparison of hybrid electric and fuel cell vehicles[J]. Environmental Innovation and Societal Transitions, 101:1-15, 2013.

[5] Steinert M, LeiferL. Scrutinizing Gartner's hype cycle[C]. Portland International Center for Management of Engineering and Technology, Portland, 2010.

[6] Kim, J., Lee, S., Lee, J., Lee, M., & Jung, H. Design of TOD Model for Information Analysis and Future Prediction. Communications in Computer and Information Science, 264(1): 301-305, 2011.

[7] Jinhyung Kim, Myunggwon Hwang, Do Jeong , Hanmin Jung.Technology trends analysis and forecasting applicati on based on decision tree and statistical feature analysis [J]. Expert Systems with Applications, 39, 2012.

[8] Wang Xin, Qiao Xiaodong, Xu Shuo, Han Hongqi, The Overview of Technology Life Cycle Analysis Method Based on Factual Database[J]. Digital Library Forum.

[9] Bin Sun. A Summarization of Information Extraction (2) (In Chinese). Terminology Standardization & Information Technology, 2003.