Innovation Hierarchy Based Patent Representation Model

Weidong Liu[1,2,3,4,5], Haijie Zhang[1]
[1] College of Computer Science, Inner Mongolia University, Hohhot, China
[2] Institute of Scientific and Technical Information of China, Beijing, China
[3] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot, China
[4] Inner Mongolia Key Laboratory of Social Computing and Data Processing, Hohhot, China
[5] Inner Mongolia Engineering Laboratory for Big Data Analysis Technology, Hohhot, China
cslwd@imu.edu.cn, xiaojie@mail.imu.edu.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892513

***Abstract*—Patent representation is a critical task for patent data mining. However, the patent content has depth and layers, which is hard to understand. There are two challenges in patent representation, including: how to identify the patent innovations and demonstrate the innovation hierarchy. To solve above issues, we propose the Innovation Hierarchy Based Patent Representation Model (IHPRM). In this model, we use the attention mechanism to identify innovations and the nested hierarchical Dirichlet process to represent the innovation hierarchy. In this paper, 1,000 patents are used to test the effectiveness of IHPRM. The result shows that IHPRM has better performance than the comparative model.**

***Index Terms*—innovations, hierarchy, attention mechanism, nested hierarchical Dirichlet process**

I. Introduction

With the development of innovation-driven economy, intellectual property has received extensive attention from countries around the world. As an important part of intellectual property, the patent brings a multitude of benefits to enterprises, which causes an increasing number of patent applications. When patent examiners manually examine the patents, they expect a patent representation that can outstand the innovations and represent the innovation hierarchy, since the representation will be convenient for the patent examination.

The related research on patent representation is summarized into two categories:

- The research on topic discovery [1]–[5]. These studies mainly focus on discovering different topics and the word distribution of each topic. Most achieve dimensionality reduction of data by mapping sets of high-dimensional words to low-dimensional topic spaces, that is, topic discovery.
- The research on topic hierarchy [6]–[11]. These studies discover the knowledge structure through the potential hierarchical tree in the text, and mine the word weight in the hierarchy. Most models use word weight to hierarchize the topics.

Directly using these methods in patent representation does not outstand the patent innovation and does not reveal the innovation hierarchy.

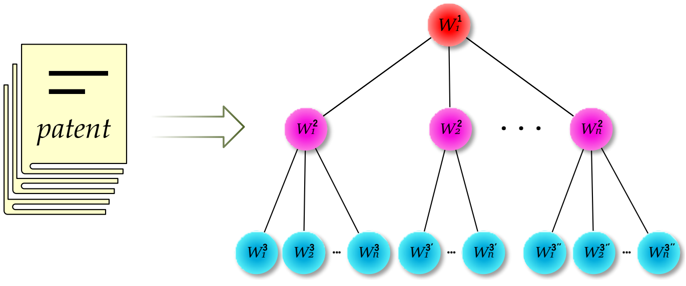

Fig. 1. The hierarchical representation of patent innovations.

After analyzing the patent structure, we found that the innovation degrees of words in different parts are varied. The innovation degrees of words in the patent title, abstract, claims and description decrease in turn. Herein, we propose IHPRM. In the model, we use the attention mechanism to calculate the innovative scores of words with the consideration of the connection between different parts of the patent [12]–[16]. Thereafter, the nested hierarchical Dirichlet process is used to discover the innovation hierarchy [6]. In $Fig.$1, each patent on the left is a flat text. The hierarchical representation of the patent innovations is shown on the right of $Fig.$1 [17]–[19], where W_1^1 denotes the first class of the first layer; W_1^2 denotes the first class of the second layer; W_1^3 denotes the first class of the third layer.

The remainder of the paper is organized as follows. In Sect.II, we introduce the preliminaries including some basic definitions and the problem definition. In Sect.III, we propose IHPRM. We detail the inference in Sect.IV. Experimental results are presented in Sect.V. We give the conclusion and future work in Sect.VI.

II. Preliminaries and problem statement

A. Primary knowledge

a) Attention mechanism: The attention mechanism is a method that calculates the weight of words. After weighting

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

and summing the words, the intermediate semantic transformation function is obtained [12], [13], as shown in $Eq.1$.

$$
\begin{aligned}
&\cdot\, address\ memory\ (score\ function)\\
&\qquad e_{ij} = F\left(s_{i-1}, h_j\right)\\
&\cdot\, normalize\ (aligment\ function)\\
&\qquad \gamma_{ij} = sofymax\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{K}\exp\left(e_{ik}\right)}\\
&\cdot\, read\ content\ (generate\ context\ vector\ function)\\
&\qquad c_i = \sum_{j=1}^{T_x}\gamma_{ij}h_j
\end{aligned}
\tag{1}
$$

where h_j denotes the hidden state of the decoder; s_{i-1} denotes the hidden state of the encoder; e_{ij} denotes the matching score between the hidden state of encoder i and the hidden state of decoder $j-1$. The higher the value of γ_{ij}, the more attention is allocated to the $i-th$ output from the $j-th$ input, which indicates that the $i-th$ output is more affected by the $j-th$ input, c_i is the weighted sum of attention.

b) Nested Chinese restaurant processes: Chinese restaurant processes (CRP) represent the uncertainty of the number of topics and nested Chinese restaurant processes (nCRP) represent the uncertainty of the topic hierarchy. When a customer comes to a restaurant on the first day, he/she chooses a table for dinner. On the next day, the customer will enter the next restaurant for dinner at one table as per the instructions on the previous table. If many customers repeat the above steps in many days, a tree-like structure with unlimited depth and unlimited branches is formed, as shown in $Fig.2$. For a text, each word traverses a path top-down, starting from the root node. Given the current node, the probability of a child node to be selected is proportional to the times that the child node was previously selected. If the child node has not been previously selected, its probability of being selected is proportional to α.

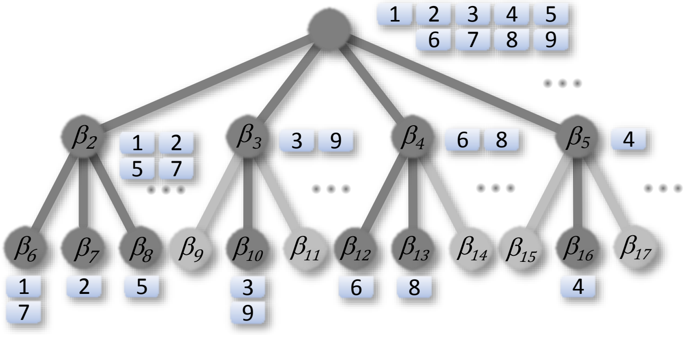

Fig. 2. Tree-like structure formed by the nCRP.

nCRP is constructed by the stick-breaking construction [20]. The probability of node i transitioning to node j is equal to the $j-th$ stick-breaking construction. The child nodes of node i_l follow the Dirichlet processes [20], as shown in $Eq.2$.

$$
\begin{aligned}
G_{i_l} &= \sum_{j=1}^{\infty}\beta_{i_l,j}\prod_{k=1}^{j-1}\left(1-\beta_{i_l,k}\right)\sigma_{\vartheta_{i_l,j}}\\
\beta_{i_l,j} &\sim Beta\left(1,\alpha\right)\\
\vartheta_{i_l,j} &\sim G_0
\end{aligned}
\tag{2}
$$

where $\beta_{i_l,j}\prod_{k=1}^{j-1}\left(1-\beta_{i_l,k}\right)$ is broken from a unit-length stick.

c) Hierarchical Dirichlet processes: The topics of each document share the topics with the Dirichlet processes G_0, which is constructed by the base distribution G_h with concentration parameter α_h. The Dirichlet processes G_j for each set of data is constructed by the base distribution G_0 with the concentration parameter α_0. The words are generated from the mixed topics and sampled from the distribution G_j, as shown in $Eq.3$ [7], [10].

$$
\begin{aligned}
W_{ji} &\sim F\left(\theta_{ji}\right)\\
\theta_{ji} \mid G_j &\sim G_j\\
G_j \mid G_0 &\sim DP\left(\alpha_0, G_0\right) forj = 1, 2, ..., J\\
G_0 &\sim DP\left(\alpha_h, G_h\right)
\end{aligned}
\tag{3}
$$

We use stick-breaking construction to sample [20], as shown in $Eq.4$.

$$
\begin{aligned}
\pi_k &= \beta_k\prod_{i=1}^{k-1}\left(1-\beta_i\right)\sigma_{\vartheta_k}\\
G &= \sum_{j=1}^{\infty}\pi_k\\
\beta_k &\sim Beta\left(1,\alpha\right)\\
\vartheta_k &\sim G_0
\end{aligned}
\tag{4}
$$

TABLE I
SYMBOLS AND THEIR MEANINGS.

Symbols	Meaning in this paper
s, α, U, X	concentration parameters / strength parameters
G_0	base distribution
θ, ϕ	parameter space
W	observable data
π, β	stick length
σ_{ϑ_k}	sampling points in G
n_i	number of customers on table i
e	unweighted attention score
γ	attention score is mapped to a value of $(0, 1)$
c	attention-weighted average
b	deviation of the network
h_t	hidden state of each word
s_t	semantic representation vector
y_{t-1}	predicted attention score at the time $t-1$
z	innovative class

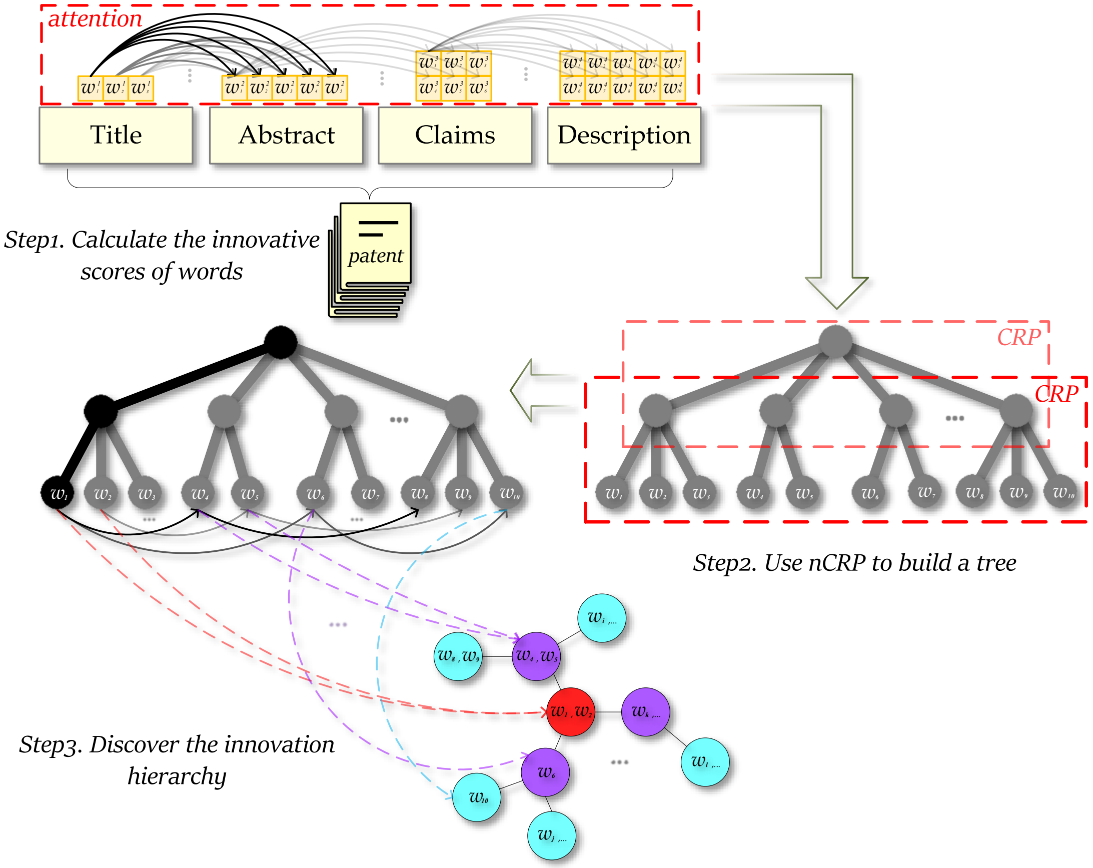
attention
Title
Abstract
Claims
Description
Step1. Calculate the innovative scores of words
patent
CRP
CRP
Step2. Use nCRP to build a tree
Step3. Discover the innovation hierarchy

Fig. 3. Discovering the innovation hierarchy by IHPRM.

B. Problem statement

For a patent, the task is to discover the innovation hierarchy. Through analyzing the patent structure, the innovative scores of title, abstract, claims and description are different. It is a hierarchical labeling task. To solve this problem, we proposed IHPRM, as shown in $Fig.3$.

$$T = W_1 \left(W_2 \left(W_4, W_5, \ldots \right), W_3 \left(W_6, \ldots \right), \ldots \right) \qquad (5)$$

The root node of the innovation hierarchy T is W_1, the child nodes of W_1 are $W_2, W_3, \ldots$, the child nodes of W_2 are $W_4, W_5, \ldots$, and the child nodes of W_3 are $W_6, \ldots$

We list the defined symbols in $Table$.I, which is used thoroughly in this paper.

III. Proposed Model

We use the attention mechanism to calculate innovative scores for the words, since the innovations are confirmed by multi-parts of the patent. Innovations in the title are likely confirmed by the abstract and claims. The innovations in the abstract are likely confirmed by the claims and description. The innovations in the claims are likely confirmed by the description. Therefore, we use the attention mechanism to calculate innovative scores of the words, as shown in $Fig.4$.

The weighting steps of attention mechanism are as follows:

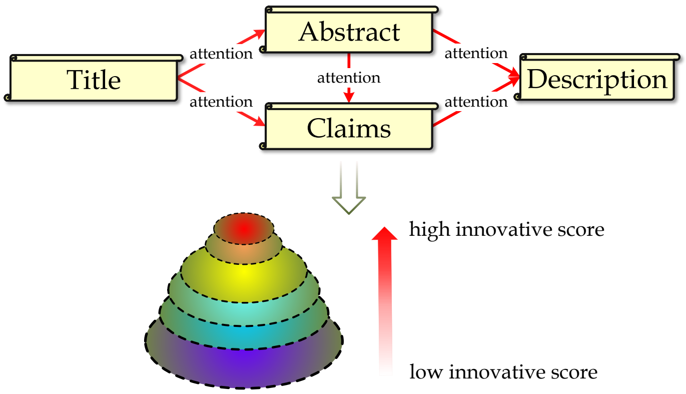

Fig. 4. Calculating the innovative scores of words by attention mechanism. According to the innovative scores, the words are divided into different levels.

The first step is the encoding process of Recurrent Neural Networks (RNN). For a sentence $W = (w_1, w_2, \ldots, w_t)$, input the words one by one into the RNN to obtain the hidden state h_t of each word in the loop layer, as shown in $Eq.6$.

$$h_t = f_1 \left(U h_{t-1} + X w_t + b \right) \qquad (6)$$

The second step is the decoding process of RNN. The hidden state s_t in the decoder is calculated according to the semantic representation vector s_t, the predicted y_{t-1} and the hidden state s_{t-1}, as shown in $Eq.7$.

$$s_t = f_2\left(y_{t-1}, s_{t-1}, c_t\right) \tag{7}$$

The third step obtains e_{ij} given s_t and h_t, and calculates the weighted innovative score c_i, as per $Eq.8$.

$$\begin{aligned} e_{ij} &= F\left(s_{i-1}, h_j\right) \\ \gamma_{ij} &= \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{K}\exp\left(e_{ik}\right)} \\ c_i &= \sum_{j=1}^{T}\gamma_{ij}h_j \end{aligned} \tag{8}$$

where e_{ij} denotes the attention score which has not been normalized; γ_{ij} denotes the alignment model; γ_{ij} compares the hidden state h_j of each word w_j in the encoder with the hidden state s_{i-1} of the previous word y_i in the decoder, calculates each input word w_j and generates the matching score between the words y_i; c_i denotes the normalized attention score.

With the innovative score of each word, we get a tree with n subtrees which is an innovation hierarchy, a word with a higher innovative score is more likely to be assigned to the front subtree, and the probability of each subtree being selected decreases gradually. For patent d, we have tree T_d, and for each tree T_d, we get $G_{i_l} \in T_d$ like $Eq.2$, as shown in $Eq.9$.

$$G_{i_l}^{(d)} = DP\left(\alpha, G_{i_l}\right) \tag{9}$$

In the patent, each word traverses with CRP, starting from the root node of T_d, and updates its probability of the selected node. When traversing to the leaf node i_l, we use a Beta distribution to determine the path to stop or jump to another subtree [6]., as shown in $Eq.10$.

$$U_{d,i_l} \sim Beta\left(\alpha_1, \alpha_2\right) \tag{10}$$

The above implies a stick-breaking construction. If the $i_l - th$ word is not selected, $G_{i_l}^{(d)}$ continues looking for the $i_{l+1} - th$ word. The probability of word n in patent d selecting the word $\phi_{d,n}$ is shown in $Eq.11$ [6].

$$\begin{aligned} &Pr\left(\phi_{d,n} = \theta_{i_l} \mid T_d, U_d\right) = \\ &[\prod_{k=0}^{l-1} G_{i_k}^{(d)}\left(\{\theta_{i_{k+1}}\}\right)][U_{d,i_l}\prod_{m=1}^{l-1}\left(1 - U_{d,i_k}\right)] \end{aligned} \tag{11}$$

where $Pr\left(\phi_{d,n} = \theta_{i_l} \mid T_d, U_d\right)$ denotes the probability of selecting the path i_l, which is equal to the probability of selecting the $l - th$ path instead of the previous $l - 1$ paths, and all paths share the same distribution. We used the nested hierarchical Dirichlet process based on attention mechanism to discover the innovation hierarchy, as shown in $Alg.1$.

Algorithm 1 Generating the innovation hierarchy by IHPRM.

```
Use the attention mechanism to calculate innovative score of each word in the document.
Generate a global tree-like structure $T$ by nCRP.
for each document $d$ do
    for each $DP$ in $T$ do
        Draw $G_{i_l}^{(d)}$ by $Eq.9$.
    end for
    for each node in $T_d$ do
        Draw $U_{d,i_l}$ by $Eq.10$.
    end for
    for each word $n$ in document $d$ do
        Sample atom $\phi_{d,n}$ by $Eq.11$.
        Sample word $W_{d,n}$ from class $\phi_{d,n}$.
    end for
end for
```

IV. Inference

We need to estimate all potential variables: β', π', c, z, ϕ. We follow the standard mean-field variational steps and use a fully factored variational distribution [21], [22]:

$$\begin{aligned} &q\left(\beta', \pi', c, z, \phi\right) = q\left(\beta'\right)q\left(\pi'\right)q\left(c\right)q\left(z\right)q\left(\phi\right) \\ &q\left(\beta'\right) = \prod_{k=1}^{K} q\left(\beta' \mid u_k, v_k\right) \\ &q\left(\pi'\right) = \prod_{j=1}^{M}\prod_{t=1}^{T} q\left(\pi_{jt} \mid \alpha_{1jt}, \alpha_{2jt}\right) \\ &q\left(c\right) = \prod_{j=1}^{M}\prod_{t=1}^{T} q\left(c_{jt} \mid \rho_{jt}\right),\ c_{jt} \sim Mult\left(\beta\right) \\ &q\left(z\right) = \prod_{j=1}^{M}\prod_{n=1}^{N} q\left(z_{jn} \mid \zeta_{jn}\right) \\ &q\left(\phi\right) = \prod_{k=1}^{K} q\left(\phi_k \mid \lambda_k\right) \end{aligned} \tag{12}$$

where u_k, v_k, α_{1jt} and α_{2jt} denote the variational parameters of the Beta distribution, they control the Beta distribution of the stick-breaking on the corpus level; ρ_{jt} and ζ_{jn} denote the variational parameters of the Multinomial distribution; ρ_{jt} controls the probability that the $t - th$ innovative class of the $j - th$ document selects a corpus level innovative class; ζ_{jn} controls the probability that the document level assigns an innovative class.

Using Jensen's inequality in standard variational inference, we get:

$$\begin{aligned}
&log\, p\left(w \mid \gamma, \alpha_0, \eta\right) \\
=&log \int_{\beta', \pi', \phi} \sum_{c,z} p\left(\beta', \pi', c, z, \phi, w \mid \gamma, \alpha_0, \eta\right) \\
=&log \int_{\beta', \pi', \phi} \sum_{c,z} p\left(\beta', \pi', c, z, \phi, w \mid \gamma, \alpha_0, \eta\right) \frac{q\left(\beta', \pi', c, z, \phi\right)}{q\left(\beta', \pi', c, z, \phi\right)} \\
\geq&\mathbb{E}_q[log\, p\left(\beta', \pi', c, z, \phi, w \mid \gamma, \alpha_0, \eta\right)] - \\
&\mathbb{E}_q[log\, q\left(\beta', \pi', c, z, \phi\right)]
\end{aligned} \tag{13}$$

$$\begin{aligned}
\mathcal{L} =&\mathbb{E}_q[log\, p(\beta', \pi', c, z, \phi, w \mid \gamma, \alpha_0, \eta)] - \\
&\mathbb{E}[log\, q(\beta', \pi', c, z, \phi)] \\
=&\sum_{j=1}^{M} \big(\mathbb{E}_q[log\, p(w_j \mid c_j, z_j, \phi)] + \mathbb{E}_q[log\, p(c_j \mid \beta')] + \\
&\mathbb{E}_q[log\, p(z_j \mid \pi_j)] + \mathbb{E}_q[log\, p(\pi'_j \mid \alpha_0)] + \mathbb{E}_q[log\, p(\phi \mid \eta)] + \\
&\mathbb{E}_q[log\, p(\beta' \mid \gamma)] - \sum_{j=1}^{M} (\mathbb{E}_q[log\, q(\pi_j)] + \mathbb{E}_q[log\, q(c_j)] + \\
&\mathbb{E}_q[log\, q(z_j)]\big) - \mathbb{E}_q[log\, q(\beta')] - \mathbb{E}_q[log\, q(\phi)]
\end{aligned} \tag{14}$$

$Eq.14$ is variational function that reaches a constant equivalent to true posterior Kullback-Leibler. By taking the derivative of this lower limit for each variable parameter, we can derive the following coordinate ascending update.

Document-level Updates: At the document level, we update the parameters for each document stick, word class indices and document class indices [22], [23]:

$$\begin{aligned}
\alpha_{1jt} =&1 + \sum_{n=1}^{N} \zeta_{jnt} \\
\alpha_{2jt} =&\alpha_0 + \sum_{n=1}^{N} \sum_{s=t+1}^{T} \zeta_{jns}
\end{aligned} \tag{15}$$

$$\begin{aligned}
\mathcal{L}_\rho =&\sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{k=1}^{K} \big(\sum_{n=1}^{N} \zeta_{jnt} \rho_{jtk} \mathbb{E}_q[log\, p(w_{jn} \mid \phi_k)] + \\
&\rho_{jtk} \mathbb{E}_q[log\, \beta_k] - \rho_{jtk} log\, \rho_{jtk} \big) + \sum_{j=1}^{M} (\sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}) - \\
&log(\sum_{n=1}^{N} \sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni})
\end{aligned} \tag{16}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \rho_{jtk}} =&\sum_{n=1}^{N} \zeta_{jnt} \mathbb{E}_q[log\, p(w_{jn} \mid \phi_k)] + \\
&\mathbb{E}_q[log\, \beta_k] - 1 - log\, \rho_{jtk} + \sum_{n=1}^{N} \zeta_{jnt} - \\
&\frac{\prod_{m=1}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jml})}{\sum_{n=1}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni})} \times \\
&\frac{\sum_{n=1}^{N} \frac{\zeta_{jni}}{\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni}}}{\sum_{n=1}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni})}
\end{aligned} \tag{17}$$

$$\begin{aligned}
\zeta_{jnt} \propto \exp\big(&\sum_{k=1}^{K} \rho_{jtk} \mathbb{E}_q[log\, p(w_{jn} \mid \phi_k)] + \mathbb{E}_q[log\, \pi_{jt}] + \rho_{jt} \\
&- \frac{\prod_{n=1, n \neq now}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni}) \cdot \sum_{e=1}^{K} \rho_{jie}}{\prod_{n=1}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni})} \big)
\end{aligned} \tag{18}$$

Corpus-level Updates: At the corpus level, we update the parameters for the top-level sticks and class [22], [23]:

$$\begin{aligned}
u_k =&1 + \sum_{j=1}^{M} \sum_{t=1}^{T} \rho_{jtk} \\
v_k =&\gamma + \sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \rho_{jtl}
\end{aligned} \tag{19}$$

$$\lambda_{ki} = \sum_{j=1}^{M} \sum_{t=1}^{T} \rho_{jtk} (\sum_{n=1}^{N} \zeta_{jnt}[w_{jn} = i]) + \eta \tag{20}$$

We Summarize the variational IHPRM algorithm in Alg.2.

Algorithm 2 Variational Inference for IHPRM.

```
Initialize all the variational parameters.
while not converged or within MAX iteration do
  E Step:
    Upata per document stick by $Eq.15$.
    Upata per document class indices by $Eq.16$
     and $Eq.17$.
    Upata per word class indices $\zeta$ by $Eq.18$.
  M Step:
    Upata corpus level stick by $Eq.19$.
    Upata class mixture by $Eq.20$.
end while
```

V. Experiments

To validate the efficiency of IHPRM, we design the following experiments.

A. Experimental datasets

The 1,000 patents used in our experiments as shown in $Table.$II.

TABLE II
Description of the experimental data.

Type of data	Description				
Source	SIPO				
IPC code	A61	C04	D03	G07	H03
Total number	200	200	200	200	200

Collect the dataset through the following steps:

- Download 1,000 patents with the International Patent Classification (IPC) secondary classification numbers A61,C04,D03,G07 and H03 from the State Intellectual

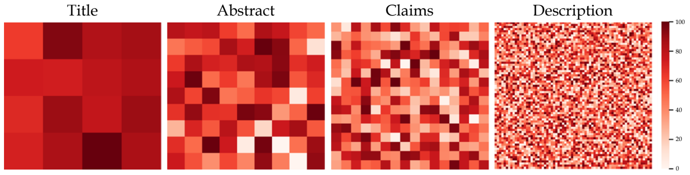

Fig. 5. Using the attention mechanism to calculate innovative scores for different parts, it can be observed that the average innovative scores in the patent title, abstract, claims and description decrease in turn.

Index	Description
(1, 1)	nature, disease, hepatiti, blood, part, medicine, efficacy, material
(1, 1, 1)	liver-protecting, relate
(1, 1, 1, 1)	preparation, sea
(1, 1, 1, 2)	extraction, dependent, dosage
(1, 1, 1, 3)	chinaberry, rhizome, pubescent
(1, 1, 1, 4)	cypress
(1, 1, 2)	patient, marasmu, cold, toxic
(1, 1, 3)	recipe, step, lithospermum, sophora, heat
(1, 1, 4)	parasite, armyworm
(1, 1, 5)	effect, promot

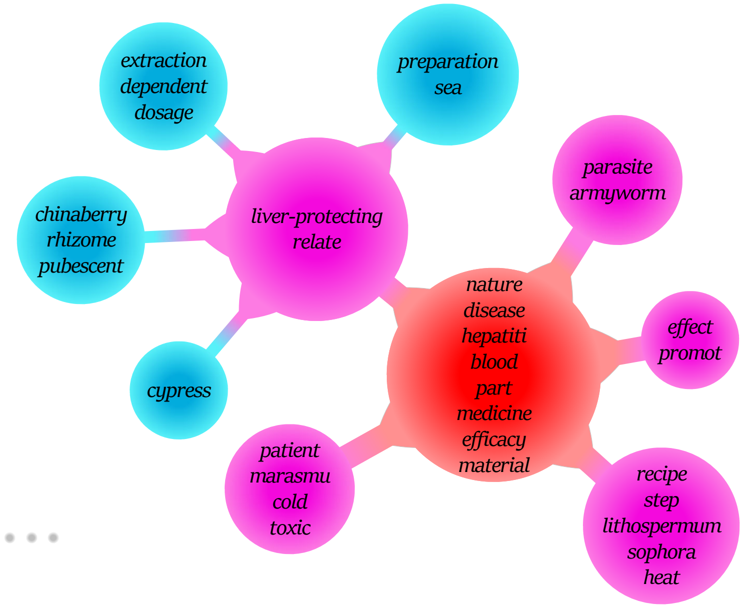

Fig. 6. The innovation hierarchy has three levels: red is the first level, purple is the second level, and blue is the third level. The lines denote the abstract relationship of the hierarchy. The higher the level, the higher the abstraction. For example, the colors of hepatiti, liver-protecting and cypress are red, purple and blue, which means that the patent innovation in the field of hepatiti, and the liver-protection drug have been developed, of which cypress is a raw material for the drug.

Property Office of China (SIPO) as a dataset. The secondary patent classification of IPC is shown in $Table$.III.

- Each patent is divided into four parts: title, abstract, claims and description, and each part is manually annotated.

TABLE III
INTERNATIONAL PATENT CLASSIFICATION AND THE CORRESPONDING PATENTS.

No. of Secondary Classification	Description	No. of patent
A61	medical or veterinary science	CN101773551A
		CN108721452A
	hygiene	CN104297441A
	...	...
C04	cements	CN113968682A
	concrete	CN113979707A
	...	...
D04	braiding	CN113943996A
	lace-making	CN212472638U
	...	...
G09	education	CN215450852U
	cryptography	CN213483196U
	...	...
H03	basic electric elements	CN114008872A
		CN114026672A
		...

B. Baseline model

In our comparative experiments, the nested hierarchical Dirichlet process (nHDP) [6] and the Negative Binomial-Neural Topic Model (NB-NTM) [5] are used as the baseline. Based on the patent structure, we use attention mechanism to calculate the innovative scores of words. After joining the attention mechanism, the accuracy of the discovered innovation hierarchy has been improved significantly.

C. Evaluation measurements

Some widely used evaluation measurements were used in our experiments. The Precision, Recall, F-measure, ROC curve and AUC score are used to measure the model.
Precision is calculated by,

$$P = \frac{TP}{TP + FP} \tag{21}$$

	nHDP	NB-NTM	IHPRM
Precision	0.3542857142857	0.4133333333333	0.4545454545455
Recall	0.3351351351351	0.4189189189189	0.4807692307692
F-measure	0.3444444444444	0.4161073825503	0.4672897196262

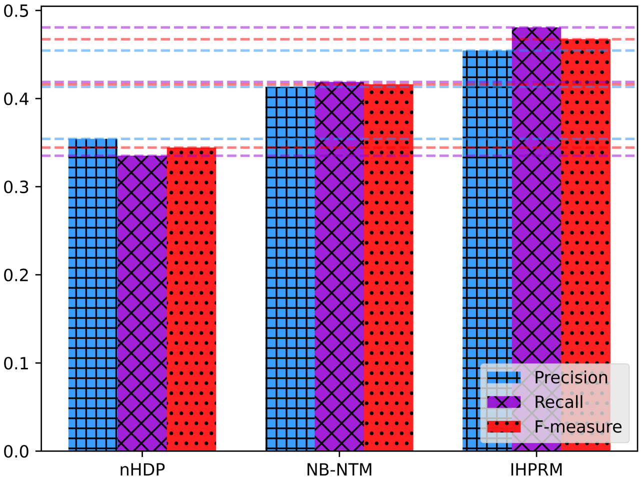

Fig. 7. Comparison of the Precision, Recall, and F-measure between IHPRM with baseline.

Recall is calculated by,

$$R = \frac{TP}{TP + FN} \tag{22}$$

F-measure is calculated by,

$$F = \frac{2 \times P \times R}{P + R} \tag{23}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative. The horizontal axis of the ROC curve is $\frac{FP}{FP+TN}$ and the vertical axis is $\frac{TP}{TP+FN}$. The ROC curve is made up of numerous points. These points denote the model hierarchizing effect under different thresholds. From left to right, the ROC curve can be seen as the change of the threshold from 0 to 1.0. The AUC score measures the total area under the ROC curve, which is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example. The AUC score range from 0.5 to 1.0, and the higher the AUC score, the higher the application value.

D. Experimental setups

For each patent in the dataset, we conduct the following experiment.

- Calculate the innovative scores of words by using the attention mechanism.
- Build a tree-like structure by using the nCRP.
- Discover the innovation hierarchy. Each word follows the CRP and traverses the tree-like structure. When traversing to the leaf node, the Beta distribution determines that the path stops or jumps to another subtree.
- Compare the results between IHPRM with the baseline.

False positive rate	**nHDP**	**NB-NTM**	**IHPRM**
0.00000000000	0.00000000000	0.00000000000	0.00000000000
0.00815276605	0.01043187581	0.01237951237	0.01299781299
0.01672647765	0.02063321294	0.02339552339	0.02546642546
...	...	...	...
0.96422829129	0.97243141473	0.97610497610	0.97888597888
0.97585394401	0.98126873126	0.98420498420	0.98588438588
0.98793680281	0.99195399195	0.99310959310	0.99998078844
1.00000000000	1.00000000000	1.00000000000	1.00000000000

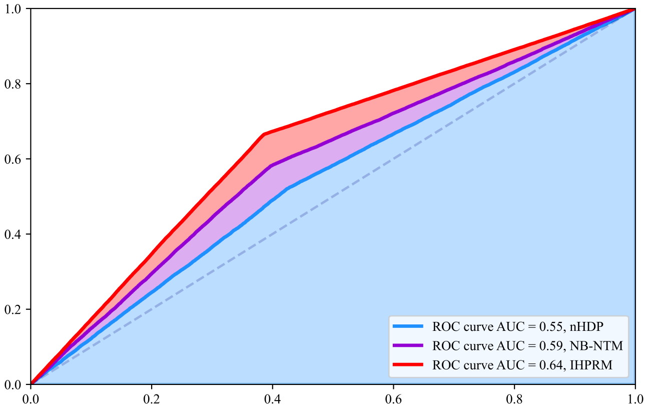

Fig. 8. Comparison of ROC curve and AUC score between IHPRM with baseline.

- Evaluate the results by using the evaluation measurements.

E. Experimental results

- The innovative scores of words in the patent are shown in Fig.5, using the data from $Table$.II in IHPRM. The result demonstrates that the innovative scores of words in the patent title, abstract, claims and description decrease in turn, which means that the abstraction of innovations in the four parts also decreases in turn.
- Fig.6 shows the innovation hierarchy discovered from innovative scores data using IHPRM. The result demonstrates that IHPRM can effectively represent the patent.
- Comparing IHPRM with the baseline, the Precision, Recall and F-measure obtained through the experimental results are shown in Fig.7. After joining the attention mechanism, the Precision, Recall and F-measure of IHPRM have been improved significantly, which demonstrates that IHPRM has significant effectiveness in discovering the innovation hierarchy.
- Comparing IHPRM with the baseline, ROC curve and AUC score of IHPRM and baseline are shown in Fig.8. The result demonstrates that IHPRM has a high application value.

VI. Conclusions and Future Work

When researchers are researching the innovation hierarchy in patents, it is difficult for them to discover the hierarchy, since the patent innovations have not been identified and the innovation hierarchy has not been demonstrated. To solve above issues, we proposed IHPRM, in which we used the attention mechanism to identify the patent innovations and nested hierarchical Dirichlet process to demonstrate the innovation hierarchy. The results demonstrate that IHPRM can effectively identify the patent innovations and demonstrate the innovation hierarchy. In the future, we plan to join a timeline to the model in order to explore the innovation hierarchy based on the timeline.

ACKNOWLEDGMENT

The research work in this paper was supported by the National Science Foundation of China (grant no. 62161036 and grant no. 61801251), Postdoctoral Science Foundation of China (grant no. 2020M680643) and University-Industry Collaborative Education Program (grant no. 202002325034).

REFERENCES

[1] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 856–864.

[2] J. Gan and Y. Qi, "Selection of the optimal number of topics for LDA topic model - taking patent policy analysis as an example," *Entropy*, vol. 23, no. 10, p. 1301, 2021.

[3] C. Yang, L. Xie, and X. Zhou, "Unsupervised broadcast news story segmentation using distance dependent chinese restaurant processes," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 2014, pp. 4062–4066.

[4] C. Li, S. Rana, D. Q. Phung, and S. Venkatesh, "Data clustering using side information dependent chinese restaurant processes," *Knowl. Inf. Syst.*, vol. 47, no. 2, pp. 463–488, 2016.

[5] J. Wu, Y. Rao, Z. Zhang, H. Xie, Q. Li, F. L. Wang, and Z. Chen, "Neural mixed counting models for dispersed topic discovery," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 6159–6169.

[6] J. W. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 256–270, 2015.

[7] T. Omoto, K. Eguchi, and S. Tora, "Hybrid parallel inference for hierarchical dirichlet processes," *IEICE Trans. Inf. Syst.*, vol. 97-D, no. 4, pp. 815–820, 2014.

[8] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 7:1–7:30, 2010.

[9] A. Ahmed and E. P. Xing, "Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream," in *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, P. Grünwald and P. Spirtes, Eds. AUAI Press, 2010, pp. 20–29.

[10] Y. Kim, M. Chae, K. Jeong, B. Kang, and H. Chung, "An online gibbs sampler algorithm for hierarchical dirichlet processes prior," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, ser. Lecture Notes in Computer Science, P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, Eds., vol. 9851. Springer, 2016, pp. 509–523.

[11] L. S. Tekumalla, P. Agrawal, and I. Bhattacharya, "Nested hierarchical dirichlet processes for multi-level non-parametric admixture modeling," *CoRR*, vol. abs/1508.06446, 2015.

[12] Q. Liu, X. Cheng, S. Su, and S. Zhu, "Hierarchical complementary attention network for predicting stock price movements with news," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, and H. Wang, Eds. ACM, 2018, pp. 1603–1606.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[14] Y. Burda, R. B. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[15] H. Xu, J. Winnink, Z. Yue, Z. Liu, and G. Yuan, "Topic-linked innovation paths in science and technology," *J. Informetrics*, vol. 14, no. 2, p. 101014, 2020.

[16] D. Nolasco and J. Oliveira, "Detecting knowledge innovation through automatic topic labeling on scholar data," in *49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, HI, USA, January 5-8, 2016*, T. X. Bui and R. H. S. Jr., Eds. IEEE Computer Society, 2016, pp. 358–367.

[17] H. Yu, J. Lu, and G. Zhang, "Topology learning-based fuzzy random neural networks for streaming data regression," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 2, pp. 412–425, 2022.

[18] Y. Hang, L. Jie, and Z. Guangquan, "Morstreaming: A multioutput regression system for streaming data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021.

[19] Y. Hang, L. Jie, L. Anjin, W. Bin, L. Ruimin, and Z. Guangquan, "Real-time prediction system of train carriage load based on multi-stream fuzzy learning," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[20] J. W. Paisley, A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin, "A stick-breaking construction of the beta process," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, J. Fürnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 847–854.

[21] C. Zhang, C. H. Ek, X. Gratal, F. T. Pokorny, and H. Kjellström, "Supervised hierarchical dirichlet processes with variational inference," in *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 254–261.

[22] N. Manouchehri, N. Bouguila, and W. Fan, "Batch and online variational learning of hierarchical dirichlet process mixtures of multivariate beta distributions in medical applications," *Pattern Anal. Appl.*, vol. 24, no. 4, pp. 1731–1744, 2021.

[23] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, ser. JMLR Proceedings, G. J. Gordon, D. B. Dunson, and M. Dudík, Eds., vol. 15. JMLR.org, 2011, pp. 752–760.