Análisis de patentes basado en el modelo de espacio vectorial de características: un caso de IoT

VSM también ha visto amplias aplicaciones en el análisis de patentes y se utiliza como una herramienta fundamental para mapear los documentos de patentes a los datos estructurados. Sin embargo, el método VSM adolece de varias limitaciones cuando se aplica a las tareas de análisis de patentes, como pérdida de problemas semánticos de nivel de sentido y problemas de tridimensionalidad.

El FVSM es un espacio de vectores de características de patentes, donde el vector de características de una patente corresponde al obtenido de la capa de agrupación de una red neuronal convolucional (CNN).

**Trabajos Relacionados**

Modelo espacio vectorial: Una limitación de los VSM basados ​​en TF-IDF es que no se consideran las coincidencias de palabras clave. Los VSM anteriores sufren problemas de pérdida de semántica.

Red neuronal convolucional para PNL: CNN es una de las más influyentes en las innovaciones en el campo del procesamiento de imágenes. Una estructura CNN típica consta de tres tipos de capas desde la entrada hasta la convolución de salida, la agrupación y la conexión completa.

**MÉTODO PROPUESTO**

**A. CONSTRUCCIÓN FVSM**

**1) ENTRADA CNN**

Primero, los datos de patentes se recopilan en función de una regla especificada (por ejemplo, que tiene el mismo código de clase de patente), y todas las patentes seleccionadas forman un conjunto de datos. Luego, necesitamos preprocesar los documentos de patente y convertir los datos no estructurados, en datos estructurados mediante técnicas de minería de texto. Considere que hay H términos únicos en total, luego cada término recibe un índice que varía de 1 a H. Finalmente, un documento de patente en el conjunto de datos con N términos se convierte en un vector N-dimensional d = [t1, t2, · ··, tN], donde tn, n ∈ {1,2, ···, N} es el índice en el diccionario de términos del enésimo término en el vector d, es decir, tn ∈ {1,2, ·· ·, H}.

Los términos en el vector d están ordenados de acuerdo con su secuencia de aparición en el documento, de esta manera, CNN puede preservar la semántica a nivel de oración para etapas analíticas posteriores.

**2) ARQUITECTURA CNN**

La primera capa incorpora términos de texto de patente en vectores de baja dimensión. El próximo jugador extrae las características avanzadas del vector. Luego, la capa de conexión completa agrega la regularización de abandono y clasifica el resultado con una probabilidad usando una capa softmax. La entrada a CNN es el vector N-dimensional d que representa el texto de una patente particular como se discutió anteriormente. El resultado de CNN es el resultado de la clasificación de la patente.

**a: CAPA DE INCORPORACIÓN DE PALABRAS**

El propósito de la capa de incrustación de palabras es convertir el vector de entrada N-dimensional d para una patente con N términos en una matriz X (N × K) - dimensional, donde el vector de lanzamiento (n ∈ {1, · ··, N}) en X corresponde a la incrustación de palabras de dimensión K del enésimo término en d. En otras palabras, cada término en d se asigna a un vector K-dimensional por la capa de incrustación de palabras.

**b: CAPA CONVOLUCIONAL**: El propósito de la capa convolucional es minar la oración para graficar representaciones abstractas que contienen información semántica enriquecedora. Dado que un vector de incrustación de palabras es una entidad integral que no tiene sentido cuando se divide, se utiliza comúnmente un convolucional tridimensional para NLP en lugar del convolucional bidimensional para el procesamiento de imágenes.

**c: CAPA DE PISCINA:** El propósito de la capa de piscina es doble. En primer lugar, la resolución de la función se reduce para evitar el sobreajuste. En segundo lugar, resuelve el problema de que las dimensiones del vector de característica de capa convolucional son diferentes para diferentes documentos de patente.

**d: CAPA TOTALMENTE CONECTADA:** El vector de características de la capa de agrupación se utiliza como entrada a la capa totalmente conectada, cuya salida es el resultado de la clasificación de la patente correspondiente. Se utiliza una función de activación lineal para la capa completamente conectada.

**B. APLICACIÓN FVSM**

**1) SIMILARIDAD DE PATENTE**

Para este propósito, primero diseñamos una serie de patrones de patentes con cada tríada que consta de tres grupos de pacientes, a saber, P, P + y P−. P denota una patente base, mientras que los otros dos pacientes P + y P− se comparan con la patente base P. La diferencia entre P + y P− es que el primero es más similar a P que el segundo.

Finalmente, para todas las triadas de patentes, utilizamos la relación entre el número de calificación de consenso con el número total de calificaciones, es decir, la tasa de precisión de la discriminación de similitud de patentes, como un indicador de la precisión de FVSM

**2) AGRUPACIÓN DE PATENTES**

La agrupación se puede realizar mediante el método K-means, que es un aprendizaje no supervisado.

Algoritmo que tiene ventajas de operación sucinta y cálculo eficiente para obtener resultados de agrupamiento razonables Existen dos métodos comunes para encontrar el número apropiado de grupos dentro de un conjunto de datos, a saber, el método del codo y el método del coeficiente de contorno.

**3) MAPA DE PATENTES**

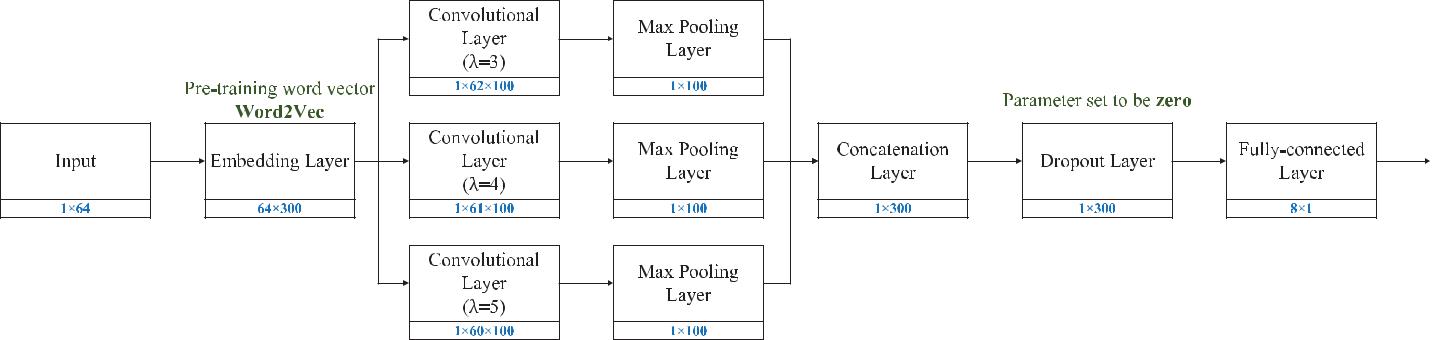
Utilizamos métodos de reducción de dimensionalidad para generar un mapa de patentes y visualizar el resultado de la agrupación por K-means. En términos generales, hay dos categorías amplias de métodos de reducción de dimensionalidad, uno es el método lineal como el Análisis de componentes principales (PCA), mientras que los otros métodos lineales de neón como la incrustación de vecinos estocásticos distribuidos en T (T-SNE).

**ESTUDIO DE CASO: INTERNET DE LAS COSAS**

En esta sección, las patentes relacionadas con la tecnología de Internet de las cosas (IoT) se eligen para ser analizadas como un caso de estudio con el fin de demostrar la efectividad y confiabilidad de los análisis de patentes basados ​​en FVSM. La investigación sobre patentes de IoT se lleva a cabo principalmente a través de tres pasos, es decir, recopilación de datos y preprocesamiento, construcción de FVSM y aplicación de FVSM. En el último paso, se establece una evaluación cuantitativa para cada aplicación a través de resultados de puntaje específicos. Los detalles de cada paso se dan de la siguiente manera.

**A. RECOPILACIÓN Y PREPROCESO DE DATOS**

Solo las patentes bajo la Clase IP de H04 están incluidas para el estudio. Todas las patentes en H04 se filtran para generar el conjunto de datos de patentes para experimentos. Se utilizó LDA para resolver el problema de etiquetado.

LDA es un modelo probabilístico generativo para colecciones de datos discretos, como corpus de texto, en el que cada elemento de una colección se modela como una mezcla finita sobre un conjunto subyacente de temas.

**FIGURA 2.** CNN utiliza para la construcción FVSM en nuestros experimentos.

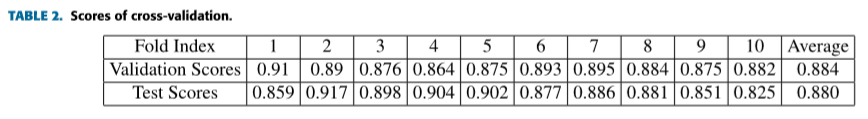
Un método más simple y directo es utilizar el número de clase IPC para la clasificación. IPC proporciona un sistema jerárquico de símbolos independientes del lenguaje para la clasificación de patentes y modelos de utilidad de acuerdo con las diferentes áreas de tecnología a las que pertenecen.

**B. CONSTRUCCIÓN FVSM**

Los pesos de la capa de incrustación de palabras se inicializan con los vectores de palabras previamente entrenados dados por Word2Vec.

Después de dividir cada valor que proviene de la salida de la capa de agrupación, obtenemos los vectores de características de 300 dimensiones de la entrada en el clasificador.

Los vectores de características pasan a través de una capa completamente conectada y la probabilidad de un número dado de categorías se genera como salida. La validación cruzada es un método estadístico utilizado para estimar la habilidad de los modelos de aprendizaje automático.



**C. APLICACIONES FVSM**

**SIMILARIDAD DE PATENTE:** Al comparar los dos valores de medida de similitud, podemos determinar qué patente, es decir, P + o P−, es más similar a la patente de base P para cada triada de patentes.   
Los resultados de la discriminación se comparan con los resultados del etiquetado manual de todas las triadas de patentes para obtener la tasa de precisión.

**AGRUPACIÓN DE PATENTES**

Primero encontramos el número apropiado de grupos κ usando el método del codo. Con el aumento de κ, la tasa descendente de SSE (Suma de cuadrados para el error) se vuelve menor y menor en el rango de 5 a 50 grupos.

LDA se utiliza para extraer las palabras clave representativas contenidas en todas las patentes para definir la definición de las tecnologías en ese grupo

**3) MAPA DE PATENTES**

Los resultados de la agrupación de K-means de las patentes relacionadas con IoT se pueden visualizar generando mapas de patentes.

Por lo tanto, para obtener un mapa de patente con un mejor efecto de visualización, se utiliza un algoritmo de reducción de dimensionalidad no lineal, es decir, T-SNE.

**Conclusión**

Este paper propone un modelo de espacio vectorial de características basado en el aprendizaje profundo para extraer características de los textos de patentes. Se propone, además un método de análisis de patentes, pero también se ha aplicado para analizar las patentes en el campo de la tecnología IoT. Con base en un número de patentes de IPCH04 de la base de datos de patentes de USPTO, se estableció un gran conjunto de datos de patentes a través de la eliminación de ruido de texto. En términos de comparación de similitudes de patentes, se muestra que se puede lograr una precisión de hasta aproximadamente el 90%. Al mismo tiempo, utilizando el método de reducción de la dimensionalidad de T-SNE.