ISSN 1870-4069

Análisis de representaciones vectoriales en bitácoras de mantenimiento en la Industria: Hacia un sistema de recuperación de información

Jesús Roberto Enrique León Carmona[1], Samuel González-López[1], Esaú Villatoro-Tello[2,3], Jesús Miguel García-Gorrostieta[4]

[1] Instituto Tecnológico de Nogales,
Sonora,
México

[2] Universidad Autónoma Metropolitana,
Unidad Cuajimalpa,
México

[3] Idiap Research Institute,
Martigny,
Switzerland

[4] Universidad de la Sierra,
Sonora,
México

17341003@itnogales.edu.mx, samuel.gl@nogales.tecnm.mx,
evillatoro@cua.uam.mx, jgarcia@unisierra.edu.mx

Resumen. La identificación de información útil en textos a través de aplicaciones con diferentes técnicas de minería de datos es poco utilizada en el contexto industrial [1]. En este artículo se analizan representación como Word2Vec, Doc2Vec y TF-IDF para determinar la más adecuada para la tarea de recuperación de información en bitácoras de mantenimiento. Además, se propone una metodología para la recuperación de información la cual brinde ayuda en el área de producción analizando el texto de los mantenimientos previos de esa área. La metodología propuesta ayudará a la toma de decisiones dándole resultados al técnico de mantenimientos posibles soluciones previas con el mismo problema. Se observaron resultados alentadores por parte del modelo Word2Vec Skip-Gram para representar los documentos.

Palabras clave: Modelos de representación textual, líneas de producción, procesamiento del lenguaje natural.

Analysis of Vector Representations in Maintenance Logs in the Industry: Towards an Information Retrieval System

Abstract. The identification of useful information in texts through applications with different data mining techniques is little used in the industrial context [1].

In this article, representations such as Word2Vec, Doc2Vec and TF-IDF are analyzed to determine the most suitable for the task of retrieving information in maintenance logs. In addition, a methodology for the recovery of information is proposed which provides help in the maintenance area by analyzing the text of the previous maintenance in the production area. The proposed methodology will help decision-making, giving the maintenance technician results possible previous solutions with the same problem. Encouraging results were seen from the Word2Vec Skip-Gram model to represent the documents.

Keywords: Textual representation models, production lines, natural language processing.

1. Introducción

La identificación de información útil en textos a través de aplicaciones con diferentes técnicas de minería de datos es poco utilizada en el contexto industrial. Esto abre una brecha para poder explorar la información y transformarla en conocimiento útil [1].

La recuperación de información (RI) se ha desarrollado desde finales de la década de 1950. Actualmente adquiere un rol más importante por el valor que tiene la información, disponer o no de la información en tiempo y forma puede resultar en el éxito o fracaso de una operación. RI es el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información [2].

La recuperación de información es un área amplia, donde se abarcan diferentes temas, algunos computacionales como el almacenamiento y la organización; y otros relacionados con el lenguaje y los usuarios como la representación, la recuperación y la interpretación [3]. En los últimos años el crecimiento de las tecnologías junto con el procesamiento del lenguaje natural abre las puertas a las empresas donde se pueda extraer una cantidad considerable de texto para poder utilizar técnicas que nos ayude a tomar decisiones. La incertidumbre de cuándo ocurrirán tiempos caídos, fallas en alguna máquina o un mantenimiento preventivo es frecuente en las líneas de producción.

Una línea de producción involucra estaciones de trabajo que pueden automatizar los procesos de ensamble de algún producto. Estas fallas pueden suceder en cualquier momento de un proceso de producción. En este artículo proponemos una primera etapa de la metodología para la recuperación de información la cual brinde ayuda en el área de mantenimiento analizando el texto de los mantenimientos previos en el área de producción. El método incluye técnicas de procesamiento del lenguaje natural para analizar textos de mantenimiento, la construcción de diferentes modelos con los datos textuales y un método para analizar la similitud entre documentos.

Buscamos relacionar las causas raíz de un problema de una maquina con la solución que el técnico le dio a dicho problema. La metodología ayudará a la toma de decisiones dándole resultados al técnico de mantenimientos posibles soluciones previas con el mismo problema y brindando la posibilidad de elegir la solución se ajuste a la situación, ayudando a agilizar el proceso de encontrar la falla y reducir tiempos caídos [1][4].

[1] Corresponde al tiempo que la estación de trabajo no realiza su actividad debido alguna falla

2. Trabajos relacionados

El procesamiento de lenguaje natural (PLN) es una opción viable para resolver problemas dentro de la industria. Nuestro trabajo ha recurrido a técnicas para calcular representaciones vectoriales continuas de palabras a partir de conjuntos de datos muy grandes. La calidad de estas representaciones se mide en una tarea de similitud de palabras y los resultados se comparan con las técnicas de mejor rendimiento basadas en diferentes tipos de redes neuronales

Los modelos vectoriales de palabras o Word Embedding (en inglés) son un recurso que puede ser utilizado como insumo para la resolución de varios de los problemas del área de PLN. Un modelo vectorial consiste en la codificación de las palabras y/o frases en un vector numérico de grandes dimensiones. Esta codificación permite tener un mapeo de "(palabra, vector)" que permite identificar cada palabra con su correspondiente vector. Contar con este mapeo es importante ya que la gran mayoría de los algoritmos utilizados están pensados para trabajar con números (sobre todo las redes neuronales) [5].

El proceso comienza con la recolección de los datos, así las empresas pueden usar sensores de bajo costo, conectividad inalámbrica y herramientas de procesamiento de bigdata para que sea más barato y fácil recopilar datos de rendimiento real y monitorear el estado del equipo. Esto mediante el uso de algoritmos basados en datos que analizan la información recopilada de una máquina determinada y su entorno ambiental, y luego la procesa de regreso a la máquina para un control adaptativo para una planificación de producción eficaz, eficiente y la programación de mantenimiento a tiempo [6]. In [7] se presenta un método para predecir fallas en el proceso de manufactura, utilizando atributos no categorizados.

El primer paso del método es el agrupamiento de datos de aquellos procesos similares, posteriormente aplicaron técnicas de aprendizaje, construyendo por cada grupo formado un clasificador. Para la predicción primero buscan clasificar el dato en un grupo y después es clasificado. Los autores reportan una AUC (área bajo la curva ROC) de 0.69, lo que revela la complejidad del problema. El uso de una red Bayesiana para predecir fallas ha sido estudiada en [8]. En este trabajo presentan una metodología que incluye 4 pasos.

El primero refiere a la recolección de los datos, el segundo se enfoca al aprendizaje y optimización de la red Bayesiana. En el paso 3 se extraen patrones de todos los tipos de fallas por separado y en el último paso se realiza la predicción de fallas recurriendo a un conjunto de reglas. Reportaron diferentes desempeños por cada una de las reglas establecidas y usaron el índice de predictibilidad (PI) para medir el rendimiento.

Por otro lado, el enfoque de las empresas que usan datos de éxito/falla (Ground Truth), puede utilizar un marco de referencia sobre un aprendizaje supervisado, se pueden crear modelos basados en los datos de exitosos o no para estudiar el comportamiento de las predicciones midiéndolos, utilizando métricas de clasificación como precisión, recuerdo, f-score, exactitud. El utilizar modelos que comparen diferentes técnicas da una perspectiva más amplia de cómo se comportan los datos dentro de una empresa y dar una visión diferente [9].

Algunos de los datos de mantenimiento, alejándonos de los sensores, están escritos por el personal experto al realizar servicio a las instalaciones y equipos de manufactura en Excel como una cadena de palabras. Una de las estrategias para revisar estos datos

Tabla 1. Ejemplo de fallas y soluciones del corpus.

Diagnóstico del Operador	Diagnóstico del Técnico	Solución
Problemas con cortos	Programa movido	Se ajustó programa
Problemas con cortos	Boquilla no tira flux	Se ajustó flux y limpio boquilla
Por cortos	Turno anterior	Turno anterior
Cortos e insuficiencias en ws2	No sale fluxer suficiente	Se purgo fluxer
Se está apagando la maquina cada rato	Precalentador	Se revisó precalentador y reseteo equipo
Problemas con la flaxeadora	Bomba pierde presión	Válvula dañada
Baja temperatura en la olla	Maquina bloqueada	Se reseteo maquina

es proporcionar un análisis de las palabras claves para facilitar la identificación de tendencias, así el aprendizaje automático sirve como puente entre los datos textuales extraídos de los equipos y el personal ya que se utilizarán en el mapeo para clasificar los textos [10].

3. Métodos y materiales

3.1. Corpus

El corpus proviene de 6853 registros escritos por el técnico del área, el cual se almacena en una hoja de Excel, en él se encuentra el diagnóstico del operador "la observación del operador del problema", la descripción del problema por el técnico y la solución que este mismo le dio, la mayoría de estos casos contienen la descripción de dichos mantenimientos otros contienen casillas en blanco y otros con palabras o símbolos sin sentido. La Tabla 1 muestra algunos ejemplos de los registros escritos en mayúsculas.

Después de observar los datos provenientes de Excel en formato. xlsx se observan inconsistencias, posteriormente se lleva a un formato csv para ajustarlo para leer las palabras separadas por comas y no por casillas esto facilita a word2vec a vectorizar los datos. Donde cada coma "," representa una celda y cada salto de línea representa una nueva fila.

Salida del archivo csv:

- Fallas con el conveyor, conveyor se atora, se ajustó tornillos.
- Pallets atorados en sello, cambio de turno, cambio de turno.
- Problemas con la flaxciadora, problemas con la flaxer, se ajustó sensor.
- Los registros de mantenimiento a menudo se registran de manera desordenada y no estandarizada: Estos registros se pueden escribir a mano y luego transcribir o ingresar directamente desde el entorno operativo a través de dispositivos de entrada

limitados, computadoras. Los datos de entrada subsiguientes, aunque nominalmente son lenguaje natural, exhiben muchas características que hacen difícil el procesamiento de la entrada.

- El vocabulario utilizado en las descripciones es inconsistente: por lo general, no existe un conjunto estándar de términos utilizados para los nombres de las piezas mecánicas o las actividades que se realizan en ellas. Por ejemplo, un registro puede contener "cambiar el aceite" mientras que otro puede contener "reemplazar el fluido", ambos se refieren a la misma acción.
- El vocabulario puede no corresponder directamente a sistemas o componentes de interés: el personal de reparación puede referirse a un componente o parte de un sistema sin mencionar explícitamente el sistema que la empresa está interesada en monitorear. Por ejemplo, “problemas con cortos" en realidad, puede ser "pre-calentador no calienta" que debe clasificarse como una falla del sistema de coordenadas y no como una falla de cortos en sí.
- La entrada no está bien formada: Debido a las limitaciones de longitud de los caracteres y al probable tratamiento de las entradas del registro como tarea secundaria, las descripciones del texto se limitan a frases cortas o fragmentos de oraciones. Además, normalmente no se presta atención a la ortografía y otras reglas del lenguaje y, por lo tanto, los datos exhiben errores gramaticales y ortográficos. Por ejemplo, “tiempo caído no funciona”, contiene un error tipográfico, mientras que “se corrió pzas. Durante el tiquetse ajusto rpm altura y flux. Tiene 30 min run”, No menciona el tópico de la falla ni el porqué de ella.
- Son comunes la jerga y las abreviaturas: Debido al espacio de entrada limitado y las presiones de tiempo de los técnicos de mantenimiento, las entradas de registro suelen estar acompañadas de abreviaturas y contienen una gran cantidad de jerga. Por ejemplo, " limpieza de pot t/c 20 " podría ser de “tiempo caído”. El uso de la jerga puede variar desde términos generalmente conocidos hasta términos muy locales para máquinas o herramientas particulares.
- No se dispone de una gran cantidad de datos: Se trabaja con datos proporcionados de una empresa maquiladora, estos datos se han recopilado por unos meses de los registros de mantenimiento de alrededor de 6853 registros para una sola clase de máquina.

Está limitada la cantidad de datos provenientes de informes de mantenimientos en el área de producción, están en formato XLSX: se trata de un libro de Excel que no cuenta con ningún tipo de macros, y para mantener una integridad de los datos se pasan a formato CSV estos (del inglés comma-separated values) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma en donde la coma es el separador decimal como en “se, corrió, pzas.

Durante, el, tiquetse, ajusto, rpm, altura, y, flux., Tiene, 30, min, run”,) y las filas por saltos de línea. El formato CSV es muy sencillo y no indica un juego de caracteres concreto, ni cómo van situados los bytes, ni el formato para el salto de línea. Estos puntos deben indicarse muchas veces al abrir el archivo, por ejemplo, con una hoja de cálculo.

La idea básica de separar los campos con una coma es muy clara, pero se vuelve complicada cuando los valores del campo también contienen comillas dobles o saltos de línea. Las implementaciones de CSV pueden no manejar esos datos, o usar comillas

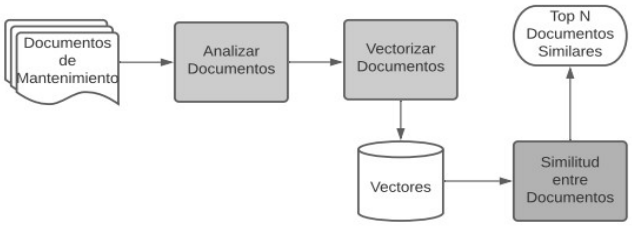

Fig. 1. Metodología propuesta.

de otra clase para envolver el campo. Algunos campos también necesitan embeber estas comillas, así que las implementaciones de CSV pueden incluir caracteres o secuencias de escape. Además, el término "CSV" también denota otros formatos de valores separados por delimitadores que usan delimitadores diferentes a la coma (como los valores separados por tabuladores).

Un delimitador que no está presente en los valores de los campos (como un tabulador) mantiene el formato simple. Es importante indicar que por el momento el corpus contiene información sensible de la empresa por la que no puede ser compartido en su estado actual.

3.2. Metodología

Se realizará un análisis utilizando técnicas de procesamiento de lenguaje natural, el primer paso es procesar los datos que provienen de los mantenimientos correctivos del área de producción que los técnicos realizan, para resolver dicho problema se planteara utilizar técnicas como Word2vec, entre otras y analizar dichos modelos para obtener el modelo con mejores resultados.

Se propone una comparación con cuatro técnicas de procesamiento de lenguaje para atender el problema desde diferentes puntos de vista. En la Fig. 1 se observa la metodología propuesta en la cual primeramente se analizan los documentos del corpus eliminamos palabras vacías y convertimos a minúsculas. Posteriormente vectorizamos los documentos para generar un vector para cada documento, para esto se probaron técnicas tradicionales como TF-IDF y técnicas basadas en Word embeddings como Word2vec y Doc2vec.

Una vez construidos los vectores de los documentos se utilizó la distancia coseno para determinar la similitud coseno entre la consulta y los documentos buscados. En esta primera etapa del estudio se busco identificar cual modelo de representación se adecuaba mejor a los datos. En una etapa futura del estudio se busca presentar y evaluar los Top N documentos similares encontrados.

4. Diseño experimental

Los experimentos realizados consisten en analizar representaciones como Word2vec, Doc2vec, TF-IDF para determinar cuál modelo de representación se adecua mejor a los datos. Para ello dividimos nuestro corpus en 90% datos de entrenamiento

Tabla 2. Resultados Similitud Coseno.

Modelo	Resultado
TF-IDF	0.54499435
Word2Vec-CBOW	0.89598442
Word2Vec-SKIP-GRAM	**0.94781154**
Doc2Vec-DBOW	0.80774255
Doc2Vec-PV-DM	0.78620946

(5,900) y 10% para datos para prueba (590), con el fin de comparar los vectores promedio resultantes del conjunto de entrenamiento contra el conjunto de prueba.

Finalmente analizamos la similitud coseno entre el vector promedio de entrenamiento y el vector promedio de prueba de los modelos creados. Para el modelo con Word2vec se obtiene una representación vectorial densa de palabras que capturan algo sobre su significado utilizando redes neuronales. La construcción del modelo se logra con la librería Gensim obteniendo vectores de tamaño 150 por cada palabra. Para obtener la representación vectorial de cada consulta se calcula el promedio de los Word embeddings. Finalmente se calcula la distancia coseno de vector promedio de entrenamiento y el vector promedio de prueba utilizando la herramienta sklearn. Para el modelo de Doc2Vec el cual genera una representación vectorizada de un grupo de palabras tomadas como una sola unidad.

El modelo modifica el algoritmo de word2vec para el aprendizaje no supervisado de representaciones continuas para bloques de texto más grandes, como oraciones, párrafos o documentos completos. Para la construcción de los vectores para cada documento utilizamos la librería de gensim con vectores de tamaño 150. Finalmente se calcula la distancia coseno de vector promedio de entrenamiento y el vector promedio de prueba utilizando la herramienta sklearn. El modelo de frecuencia de términos-frecuencia inversa de documentos TF-IDF es un modelo de bolsa de palabras con pesado.

El tamaño de los vectores generados está en función del tamaño del vocabulario del corpus. La generación de los vectores TF-IDF para cada documento se construyó utilizando la herramienta gensim. Finalmente se calcula la distancia coseno de vector promedio de entrenamiento y el vector promedio de prueba utilizando la herramienta sklearn.

5. Resultados

Los resultados de los diferentes modelos son comparados para determinar cuál modelo de representación se adecua mejor a los datos de los registros de mantenimientos correctivos.

Para ello utilizamos calculamos la distancia coseno entre el vector promedio del conjunto de entrenamiento contra el vector promedio del conjunto de prueba. En la

Tabla 2 observamos los resultados de la similitud coseno para cada modelo de representación. Al analizar los resultados notamos que algunos de los modelos son más adecuados para capturar la información textual y semántica de los registros de mantenimiento.

Suponemos que la similitud coseno entre los vectores promedio del conjunto de entrenamiento y prueba nos brinda una noción de que una representación que captura mejor la información textual de los registros logra una mayor cercanía. En base a esto al revisar los resultados de los diferentes modelos, notamos que no están tan alejados unos de otros, esto podría interpretarse de que se está reteniendo algo de información y que el mejor modelo para representar los documentos es Word2Vec Skip-Gram al tener una mayor puntuación.

6. Conclusiones y trabajo a futuro

Con las representaciones utilizadas pudimos observar que tan similares son dos documentos como si fuera un buscador tradicional al hacer búsqueda de concatenación de palabras, ya que nuestro corpus no nos permite trabajar con otras métricas por la falta de datos y su escasez en el léxico con los que los técnicos describen los mantenimientos realizados.

Los técnicos expertos realizan estos tickets de mantenimientos con su propia experiencia y los describen con sus palabras dejando por un lado la estandarización con la que se podrían escribir los tickets y así obtener un mejor análisis y utilizar otras métricas que ayuden a comprender el comportamiento de los datos textuales descritos por el personal de producción y los técnicos.

Esta investigación se utiliza la métrica de coseno inverso para mantener la integridad de los datos redactados por el personal intentando encontrar una relación entre lo descrito por los técnicos y las búsquedas que el usuario realiza al querer buscar una solución a un mantenimiento en común, al buscar una concatenación de palabras se espera obtener respuestas cercanas a lo descrito ya que la consulta realizada estará cercana "vectorialmente cercana" a la respuesta deseada intentando mantener las jergas, mal escrituras, mal formuladas, signos y diferentes formas de describir problemas y soluciones , como material para nutrir las búsquedas.

Con respecto a los resultados de Word2Vec skip-grams, este nos parece el más adecuado comparado con los demás modelos creados con Gensim por el hecho de realizar consultas de concatenación de palabras. Este modelo intentara predecir las palabras "vectores" vecinas de esta búsqueda, ya que estas palabras se estarán asociando con sus palabras vecinas y dar una respuesta más cercana a la buscada. En este modelo no se interesa del todo las entradas y salidas de la red, sino que el objetivo es simplemente aprender los pesos de la capa oculta que son en realidad los vectores de las palabras que se está intentando aprender de ellas.

La tarea para el modelo skip-gram seria, dada una palabra intentar predecir las palabras vecinas, y esto está definido por la ventana que en nuestro caso fue de 10. Podemos concluir que son viables las técnicas utilizadas en este trabajo para aplicarse en diferentes áreas dentro de las industrias tales como en los datos textuales de mantenimiento. En trabajos futuros se planea completar el proceso de recuperación de información al identificar los tops N mejores resultados para ser presentados al usuario.

Además, se pretende experimentar con técnicas de aprendizaje profundo como seq2sec y BERT. Finalmente se pretende también crear una aplicación amigable con el usuario que muestre las respuestas de las consultas, con el propósito de suministrar al personal de mantenimiento una herramienta más visual y entendible para el personal de esa área.

Agradecimientos. Durante la realización de este trabajo, Esaú Villatoro-Tello fue apoyado parcialmente por Idiap Research Institue, la UAM-Cuajimalpa, y el SNI-CONACyT México. Samuel González-López y Jesús Miguel Garcia-Gorrostieta fueron apoyados parcialmente por el SNI-CONACyT México.

Referencias

1. Ur-Rahman, N., Harding, J. A.: Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications, vol. 39, NO. 5, pp. 4729–4739 (2012) doi: 10.1016/j.eswa.2011.09.124
2. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., Trigg, L.: Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook Springer, Boston, pp. 1269–1277 (2009) doi: 10.1007/978-0-387-09823-4_66
3. Tolosa, G. H., Bordignon, F. R.: Introducción a la recuperación de información (2008)
4. Chakraborty, G., Krishna, M.: Analysis of unstructured data: Applications of text analytics and sentiment mining. In SAS global forum, pp. 1288–2014 (2014)
5. Claassen, M., Grill, P.: Aprendizaje semisupervisado de rasgos de temporalidad en el léxico del español (2017)
6. Al-Abassi, A., Karimipour, H., HaddadPajouh, H., Dehghantanha, A., Parizi, R. M.: Industrial big data analytics: Challenges and opportunities. In Handbook of Big Data Privacy, Springer, Cham. pp. 37–61 (2020) doi: 10.1007/978-3-030-38557-6_3
7. Zhang, D., Xu, B., Wood, J.: Predict failures in production lines: A two-stage approach with clustering and supervised learning. In: Proceedings of IEEE International Conference on Big Data (Big Data), Washington, DC, pp. 2070–2074 (2016) doi: 10.1109/Big Data.2016.7840832
8. Abu-Samah, A., Shahzad, M. K., Zamai, E., Said A.: Failure prediction methodology for improved proactive maintenance using Bayesian approach. IFAC-PapersOnLine, vol. 48, no. 21, pp. 844–851 (2015) doi: 10.1016/j.ifacol.2015.09.632
9. Brito, J. H., Pereira, J. M., da Silva, A. F., Angélico, M. J., Abreu, A., Teixeira, S.: Machine learning for prediction of business company failure in hospitality sector. In Advances in Tourism, Technology and Smart Systems Springer, Singapore, pp. 307–317 (2020) doi: 10.1007/978-981-15-2024-2_28
10. Cadavid, J. P. U., Lamouri, S., Grabot, B., Pellerin, R., Fortin, A.: Machine learning applied in production planning and control: A state-of-the-art in the era of industry 4.0. Journal of Intelligent Manufacturing, vol. 38, pp. 1531–1558 (2020) doi: 10.1007/s10845-019-01531- 7