Explorando el impacto de las consultas de texto por usuarios de carreras de ingeniería a través de la técnica TF-IDF
Exploring the impact of text queries by engineering career users through the TF-IDF technique

Juan José López Cisneros[1], Ana Lidia Franzoni[2]

[1] Universidad de Guadalajara – Centro Universitario de Ciencias Exactas e Ingenierías, Blvd. Marcelino García Barragán 1421, Olímpica, Guadalajara, Jalisco., 44430.
juan.lopez@academicos.udg.mx

[2] Instituto Tecnológico Autónomo de México – Río Hondo 1, Progreso Tizapán, Álvaro Obregón, 01080. Ciudad de México, CDMX.
analidia@itam.mx

Fecha de recepción: 29 de julio de 2023

Fecha de aceptación: 27 de septiembre de 2023

Resumen. El aumento en la cantidad de información disponible en la educación superior hace necesario un enfoque personalizado para la entrega de recursos educativos. El objetivo es explorar la relevancia de palabras en la búsqueda de recursos para los estudiantes en un sistema web, dándoles la oportunidad de elegir aquellos que mejor se adapten a sus necesidades. El método utilizado implica la evaluación de los términos de búsqueda de los estudiantes mediante la técnica de ponderación TF-IDF. Los resultados muestran que esta técnica puede ser eficaz para identificar la relevancia de los recursos educativos, lo que se traduce en un mejor acceso a la información para los estudiantes. La conclusión es que la personalización de los recursos educativos basados en consultas de texto y TF-IDF puede mejorar significativamente el acceso a la información relevante para los estudiantes universitarios, lo que les permite tomar decisiones más informadas en su proceso de aprendizaje.

Palabras claves: Procesamiento de lenguaje natural, TF-IDF, Búsquedas Personalizadas, Recursos educativos.

Summary. The increase in the amount of information available in higher education necessitates a personalized approach to the delivery of educational resources. The objective is to explore the relevance of words when searching for resources for students in a web system, giving them the opportunity to choose those that best suit their needs. The method used involves the evaluation of students' search terms using the TF-IDF weighting technique. The results show that this technique can be effective in identifying the relevance of educational resources, which results in better access to information for students. The conclusion is that personalization of educational resources based on text queries and TF-IDF can significantly improve access to relevant information for university students, allowing them to make more informed decisions in their learning process.

Keywords: Natural language processing, TF-IDF, Custom Searches, Educational resources.

1 Introducción

La identificación de las palabras clave que utilizan los estudiantes en su carrera es esencial para personalizar los recursos educativos y mejorar el proceso de enseñanza-aprendizaje. Al conocer las palabras específicas de cada disciplina, los docentes pueden diseñar materiales educativos que se adapten mejor a las necesidades de los estudiantes y les permitan comprender mejor los conceptos.

La personalización de los recursos educativos es una práctica cada vez más valorada, que implica adaptar la enseñanza a las necesidades individuales de cada estudiante, con el fin de mejorar su motivación, compromiso y rendimiento académico. Según Brusilovsky y Peylo (2003), "los sistemas educativos web adaptativos e inteligentes se basan en el uso de técnicas de minería de datos y aprendizaje automático para personalizar la experiencia de aprendizaje para cada estudiante" [1].

En un estudio realizado por Pinho et al. (2019) señala que temas como el aprendizaje automático, el aprendizaje profundo, la minería de datos, la inteligencia artificial, han hecho posibles recomendaciones más útiles, que es el predominio del estudiante como principal destinatario de las recomendaciones. Señalan que "no sorprende ya que la formación académica necesita de muchas lecturas y estudios y que, debido a la diversidad de información, recomendar contenidos fiables optimiza el tiempo de estudio y la eficacia" [2].

2 Estado del arte

A continuación, se presenta el estado del arte sobre la importancia de identificar las palabras más relevantes que utilizan los estudiantes según su carrera, utilizando técnicas de aprendizaje automático, inteligencia artificial o TF-IDF para personalizar recursos educativos:

En un estudio de 2020, Baidada y colaboradores propusieron un modelo para analizar las búsquedas de los estudiantes con el fin de determinar las palabras relevantes que reflejan sus intereses, con el objetivo de enriquecer sus perfiles, recopilaron las descripciones de los enlaces devueltos por el motor de búsqueda, que constituirán un corpus sobre el cual aplicaron el método TF-IDF (frecuencia de término-frecuencia inversa de documento) para determinar las palabras relevantes. Luego, utilizaron la técnica Word2vec para determinar palabras similares a estas palabras relevantes en la descripción de recursos educativos internos, de modo que pudieron recomendar aquellos que mejor se ajusten a las necesidades del alumno [3].

En otro estudio Gómez y colaboradores propusieron un perfilador para recursos de aprendizaje disponibles en una plataforma Web basado en el algoritmo TF-IDF utilizado para identificar las palabras clave más relevantes e implementarlo en un cálculo de similitud de coseno. Los autores concluyen "que es posible predecir la relevancia de recursos de aprendizaje utilizados por el estudiante, además de que con esa información se puede generar un perfil del recurso y ubicarlo en cierto campo de conocimiento, área o materia según su relevancia" [4].

En un estudio de 2018, Fan y Qin propusieron un algoritmo TF-IDF mejorado (TF-IDCRF) que toma en cuenta las relaciones entre clases para completar la clasificación de textos [5].

En 2018, Shahzad y Ramsha realizan una investigación que sugiere que TF-IDF es una técnica útil para examinar la relevancia de las palabras clave en un corpus de documentos y puede ser aplicada en varios campos, incluyendo la minería de textos y la recuperación de información [6].

3 Metodología

La metodología utilizada para aplicar TF-IDF a las consultas de texto realizadas por los usuarios para recuperar los recursos educativos consiste en los siguientes pasos:

- Recopilación
- Preprocesamiento
    - Tokenización
    - Eliminación de palabras vacías ‘stopwords’
- Cálculo de TF
- Cálculo de IDF
- Multiplicación de TF y IDF
    - Normalización
    - Ordenamiento
- Consulta

Aplicación de los pasos de la metodología ejemplificando el proceso para los usuarios de carreras de ingeniería.

Al inicio se realiza la recopilación de las consultas de texto a los recursos educativos disponibles en forma de documentos y textos (esto es llamado corpus) como se puede observar en la Figura 1.

```
Los elementos por carreras:
0 -- Licenciatura en Ingeniería en Computación (CUCEI) tiene 579 elementos
1 -- Ingeniería en Computación (CUCEI) tiene 298 elementos
2 -- Licenciatura en Ingeniería en Comunicaciones y Electrónica (CUCEI) tiene 88 elementos
3 -- Licenciatura en Informática (CUCEI) tiene 104 elementos
4 -- Ingeniería Informática (CUCEI) tiene 282 elementos
5 -- Licenciatura en Ingeniería Mecánica Eléctrica (CUCEI) tiene 409 elementos
6 -- Ingeniería Mecánica Eléctrica (CUCEI) tiene 233 elementos
7 -- Ingeniería Industrial (CUCEI) tiene 172 elementos
8 -- Licenciatura en Ingeniería Industrial (CUCEI) tiene 286 elementos
9 -- Ingenieria Fotónica (CUCEI) tiene 52 elementos
10 -- Licenciatura en Diseño para la Comunicación Gráfica (CUAAD) tiene 32 elementos
11 -- Ingeniería Biomédica (CUCEI) tiene 47 elementos
12 -- Ingeniería en Logística y Transporte (CUCEI) tiene 16 elementos
13 -- Licenciatura en Ingeniería Química (CUCEI) tiene 86 elementos
14 -- Licenciatura en Física (CUCEI) tiene 82 elementos
15 -- Ingeniería Química (CUCEI) tiene 143 elementos
16 -- Licenciatura en Químico Farmacobiólogo (CUCEI) tiene 14 elementos
17 -- Licenciatura en Química (CUCEI) tiene 27 elementos
18 -- Licenciatura en Ingeniería en Alimentos y Biotecnología (CUCEI) tiene 12 elementos
19 -- Licenciatura en Ingeniería Biomédica (CUCEI) tiene 13 elementos
20 -- Ingenieria en Computación (ITAM) tiene 10 elementos
```

Figura 1. Recopilación de las consultas en texto realizadas por carrera (corpus).

Después se hace el preprocesamiento que al igual que en el caso de los documentos, se debe realizar un preprocesamiento de la consulta para eliminar signos de puntuación, caracteres especiales, y convertir todo el texto a minúsculas.

Para esto se hacen los siguientes pasos:

- Tokenización: la consulta debe ser tokenizada, es decir, se debe dividir en palabras individuales o términos.
- Eliminación de palabras vacías “stopwords”: se deben eliminar las palabras comunes que no aportan significado adicional a la consulta, como "un", "el", "de", etc., que no aportan información relevante.

Después se realiza el cálculo de la frecuencia (TF) de cada término en la consulta, es decir, cuántas veces aparece cada término en la consulta y el cálculo del valor IDF (Del inglés, ‘Inverse Document Frequency’) de cada término en la consulta.

El valor IDF se calcula utilizando la fórmula:

$$IDF = \log(N / n\_t)$$

Donde N es el número total de documentos en el corpus y n_t es el número de documentos que contienen el término t.

Para terminar, se hace la multiplicación de TF y IDF: Multiplicar el valor TF de cada término por su valor IDF correspondiente. Se normalizan los valores TF-IDF resultantes dividiendo cada valor por la norma Euclidiana del vector de términos en la consulta.

Esto asegura que la longitud del vector no afecte los resultados de la similitud. Se ordenan los términos según su valor TF-IDF y se seleccionan los términos más relevantes para su análisis. Y al final se utilizan los términos relevantes seleccionados para buscar documentos en el corpus que contengan esos términos.

4 Resultados experimentales

En esta sección se muestra el desempeño de la aplicación de la técnica TF-IDF que se realiza a las búsquedas de texto que ejecutan los usuarios de las carreras de ingeniería a un repositorio Web para recuperar recursos educativos de interés. A continuación, se explica el proceso siguiendo la metodología mencionada:

Del repositorio Web se recuperó el periodo de tiempo del 14 de marzo de 2021 al 26 de abril de 2023, las consultas de texto que han realizado 377 usuarios con carrera definida en un archivo en formato ‘*json*’. En Figura 2 se muestra las consultas de texto de un usuario específico.

```
{'usuario_id': abc,
 'carrera': 'Ingeniería Mecánica Eléctrica (CUCEI)',
 'busquedas': [
  {'fecha': '2023-01-22 20:41:00', 'frase': 'lenguajes programacion'},
  {'fecha': '2023-01-22 20:41:00', 'frase': 'electricidad'},
  {'fecha': '2023-01-22 20:41:00', 'frase': 'programacion aplicada'},
  {'fecha': '2023-01-22 20:43:00', 'frase': 'instalaciones electricas'},
  {'fecha': '2023-01-30 21:36:00', 'frase': 'peliculas ciencia tecnologia'},
  {'fecha': '2023-01-30 21:37:00', 'frase': 'enigma'},
  {'fecha': '2023-01-30 21:39:00', 'frase': 'código enigma'},
  {'fecha': '2023-02-25 13:10:00', 'frase': 'peliculas'}]}
```

Figura 2. Recuperación de la información de un usuario

Lo siguiente fue construir un diccionario con el identificador de la carrera como llave y los valores de cada llave, los cuales son obtenidos de las consultas de texto realizadas por los usuarios identificados con carrera. Durante este proceso se fue realizando el preprocesamiento, tokenización y eliminación de *'stopwords'*.

Se obtuvo la indexación de 21 carreras en el diccionario y a partir de ello, se generó una lista con las cadenas (*tokens*) de cada carrera (ver Figura 3).

```
['token1, token2, token, …', 'token4, token2, token67, …', …]
```

Figura 3. Organización de las cadenas por carrera en una lista

Posteriormente la lista se procesó bajo un modelo de '*bag-of-words*' (ver Figura 4), es decir, que no se codifica la información relativa a la posición de los '*tokens*' ni su contexto, solo información sobre si aparecen y su frecuencia. A lo que se obtiene un vocabulario de un total de 1360 'tokens'.

	token1	token2	token3	token4	token5	token6	token7	…
1	0	1	0	1	1	0	1	…
2	1	0	1	0	0	0	2	…
3	0	1	0	0	1	1	1	…

Figura 4. Organización de una matriz con la frecuencia de términos por carrera

Obtenida la matriz de frecuencia de los términos por cada carrera (cálculo de TF), se realiza el cálculo de IDF y la multiplicación de TF e IDF, y para finalizar se normalizan los vectores generados mostrando la relevancia de términos por carrera (ver Figura 5).

	TF_IDF 0	TF_IDF 1	TF_IDF 2	TF_IDF 3	TF_IDF 4	TF_IDF 5	TF_IDF 6	…
python	0.302717	0.430268	0.072387	0.320658	0.240541	0.428821	0.307034	…
películas	0.188701	0.291618	0.047174	0.208970	0.156759	0.052399	0.114338	…
redes	0.058023	0.265556	0.000000	0.000000	0.000000	0.029538	0.000000	…
java	0.124003	0.216203	0.000000	0.188819	0.125904	0.105213	0.000000	…
				…				
ética	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	…
óptica	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	…

Figura 5. Cálculo de TF-IDF de términos por carrera.

Para visualizar los términos que son más relevantes por carrera cada columna fue ordenada de mayor a menor para su presentación. En la Fig. 6 se visualiza una representatividad de los términos ordenados obtenidos e identificados para las carreras de Ingeniería en Computación, Ingeniería Industrial y la de Diseño y Comunicación Gráfica.

Ingeniería en computación		Ingeniería industrial		Diseño Gráfico	
python	0.430268	ergonomía	0.243311	diseño	0.718458
películas	0.291618	python	0.221099	marca	0.364380
redes	0.265556	industrial	0.201613	illustrator	0.242920
java	0.216203	funciones	0.146490	ilustrador	0.242920
programación	0.179457	ingeniería	0.146490	photoshop	0.201289
datos	0.174063	comunicación	0.121656	grafico	0.121460
algoritmo	0.174063	aplicada	0.119936	ilustracion	0.121460
estructura	0.136928	biblioteca	0.109868	after	0.121460
ciencia	0.118945	programación	0.104792	effects	0.121460
javascript	0.113810	cabezón	0.100806	naming	0.121460

Figura 6. Relevancia de términos por carreras identificadas

La identificación de la relevancia de los términos promovió la realización de una vista de recomendación de recursos que contienen esas palabras según se haya identificado el usuario.

5 Conclusiones y dirección para futuras investigaciones

En este artículo, propusimos utilizar la técnica TF-IDF para obtener la relevancia de los términos utilizados por los usuarios de distintas carreras en un repositorio de recursos educativos. De esta manera, los responsables de la gestión del repositorio pueden identificar los recursos más importantes y mejorar su organización, el sistema puede recomendar recursos relevantes que aún no han sido explorados por un usuario y los profesores pueden personalizar su enseñanza y adaptarla a las necesidades y habilidades específicas de sus estudiantes. Como trabajo futuro será importante experimentar haciendo más preprocesamiento de la información obtenida, identificar y reemplazar *'n_gramas'* para abstraer/simplificar términos compuestos del vocabulario, además combinar el análisis de datos y el aprendizaje automático para ofrecer una experiencia educativa altamente personalizada.

Referencias

[1] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems", International Journal of Artificial Intelligence in Education, vol. 13, no. 2-4, pp. 159-172, 2003.

[2] P. C. R. Pinho, R. Barwaldt, D. Espíndola, M. Torres, M. Pias, L. Topin y M. Oliveira. Developments in Educational Recommendation Systems: a systematic review, IEEE Frontiers in Education Conference (FIE), IEEE, pp. 1-7, 2019.

[3] M. Baidada, K. Mansouri and F. Poirier, "Development of an Automatic Process for Recommending Well Adapted Educational Resources in an E-learning Environment," 2020 6th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, 2020, pp. 231-235, doi: 10.1109/CiSt49399.2021.9357199.

[4] M. Gómez, L. Mendoza, y M. Valverde, Detección de estilos de aprendizaje y recomendación personalizada de contenido. Congreso nacional de investigación educativa (CNIE), 2021. (https://www.comie.org.mx/congreso/memoriaelectronica/v16/doc/2490.pdf)/)

[5] H. Fan, Y. Qin. *Research on Text Classification Based on Improved TF-IDF Algorithm, Advances in Intelligent Systems Research*, vol. 147, International Conference on Network, Communication, Computer Engineering (NCCE 2018), (https://www.atlantis-press.com/proceedings/ncce-18/25896557)

[6] Q. Shahzad & A. Ramsha. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications, pp. 181, 2018. DOI: 10.5120/ijca2018917395 (https://www.ijcaonline.org/archives/volume181/number1/qaiser-2018-ijca-917395.pdf)