Modelo de Expansión de Consultas basado en Análisis de Tópicos

Federico Zimmermann, Gerardo Enrique, Mariano Minoli, Carlos Pérez

Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional – Facultad Regional Resistencia
French 414. Resistencia (3500). Argentina
{fedezimm; geraenrique97; mariano_minoli; cperez}@ca.frre.utn.edu.ar

Resumen. Actualmente, la experiencia de usuario es un aspecto fundamental en el diseño de sistemas informáticos, que debe considerar tanto la eficacia como la eficiencia en el uso y los resultados que devuelven dichos sistemas. Más aún, cuando se trata de búsquedas en la Web, esos factores son críticos ya que, debido al gran volumen de datos, distinguir la información útil no suele ser un hecho trivial. Por ende, la precisión y utilidad de la información que los usuarios demandan como resultado de una búsqueda son factores críticos. En base a estas cuestiones, este artículo describe un modelo de expansión de consultas basado en el modelado de tópicos. Específicamente, el modelo infiere relaciones entre palabras que conforman la consulta con el objetivo de encontrar nuevos términos que están implícitos o no han sido considerados por el usuario, y que luego pueden ser utilizados en consultas más profundas y contextualizadas. La implementación del modelo se explica en detalle y se ejemplifica su utilización con tres casos de prueba.

Palabras Clave: modelado de tópicos, expansión de consultas, procesamiento de lenguaje natural, recuperación de información.

1 Introducción

Actualmente la enorme cantidad de información disponible en la Web direcciona el diseño de herramientas de administración de datos hacia la implementación de estrategias inteligentes que permitan distinguir aquello que es de relevancia para el usuario. De esta manera, la necesidad del desarrollo de aplicaciones que puedan ayudar en el análisis de la información se hace cada vez más necesaria. Un ejemplo claro de esta cuestión son las estrategias que se implementan en los buscadores Web, tanto para recuperación de información [1] como de sugerencias al usuario [2]. La recuperación de información tiene como objetivo la obtención de recursos relevantes en respuesta a las consultas de los usuarios que, por lo general, muchas veces es ineficiente debido a la mala definición de las palabras utilizadas en la consulta [2].

Debido a que la utilización adecuada de términos en una consulta es de suma importancia, las sugerencias, el refinamiento y la expansión se utilizan como tres alternativas

de modificación de consulta [2] y tienen como objetivo mejorar la calidad de los resultados obtenidos. En la expansión hay dos cuestiones importantes a analizar: la detección de los términos que contiene una consulta y la obtención de términos relacionados.

En primer lugar, detectar los términos que contiene una consulta requiere la utilización de técnicas de procesamiento de lenguaje natural (más conocido como NLP por sus siglas en inglés). Esos términos indican de qué se trata el tema de la consulta y pueden contener las características que son de relevancia para el usuario.

En segundo lugar, la obtención de términos en el dominio de lo que el usuario está buscando implica encontrar, de manera fiable, las relaciones entre conceptos. Un claro ejemplo de este proceso es el descubrimiento basado en la literatura [3] que busca términos y conceptos ocultos, principalmente en temas médicos. La idea detrás del descubrimiento basado en la literatura es utilizar una gran fuente de datos, como artículos científicos, y mediante un proceso iterativo e interactivo encontrar relaciones entre conceptos iniciales que define el usuario. Otros ejemplos de la obtención de términos basada en documentos de texto son los algoritmos de NLP Latent Semantic Indexing (LSI) y Latent Dirichlet Allocation (LDA) [4]. Estas técnicas detectan tópicos que contienen términos de relevancia y que sirven de base para la expansión de consultas.

En definitiva, la expansión eficiente implica un proceso de análisis que difiere de la típica consulta – respuesta y de la funcionalidad autocompletar que proveen los buscadores tradicionales dado que existe un proceso de realimentación que puede ser supervisado o no por el usuario. Este componente proactivo tiene como objetivo mejorar la calidad de los resultados obtenidos.

En este artículo se describe un modelo de expansión para evaluar términos asociados a una consulta y determinar la relevancia que tienen de acuerdo con el análisis de documentos Web utilizando técnicas de NLP. Para su validación se ha seleccionado un conjunto de claves de búsqueda que permite evidenciar la potencia del modelo propuesto. En la sección 2 se mencionan trabajos relacionados sobre modificación de claves de búsqueda, descubrimiento basado en literatura y los algoritmos LSI y LDA. Luego, en la Sección 3, se describe el modelo propuesto. Los resultados experimentales se muestran en la Sección 4 y, finalmente, algunas características del modelo propuesto se discuten en la Sección 5.

2 Trabajos relacionados

2.1 Modificación de claves de búsqueda: refinamiento, expansión y sugerencia

Las técnicas de modificación tienen el objetivo de mejorar la precisión, la tasa de recuperación de las búsquedas y también eliminar la ambigüedad en la consulta original [2]. Las principales técnicas dentro de la modificación de claves de búsqueda son la expansión, el refinamiento y las sugerencias de claves de búsqueda. Específicamente la expansión de consultas es un proceso de reformulación de una consulta inicial para mejorar el rendimiento en las operaciones de recuperación de información [5][6]. La hipótesis de expansión de consultas consiste en entender que uno de los principales motivos de inexactitud de los sistemas de recuperación de información se debe a que los usuarios tienden a emplear claves de búsqueda breves, y en consecuencia incompletas

o mejorables [7]. Por ello, de acuerdo con esta hipótesis, su expansión potencialmente permitirá mejorar los resultados, reflejando la necesidad de información del usuario con mayor precisión [8]. Por otro lado, el refinamiento es un proceso de transformación de una consulta en una nueva consulta que refleja la necesidad de información del usuario con mayor precisión [9]. Si bien es similar a la expansión, se diferencia en el hecho de que el énfasis se pone en mejorar la precisión y no necesariamente en la expansión. Las sugerencias consisten en proponer modificaciones a las claves de búsqueda al usuario mientras éste ingresa la clave [10]. Estas técnicas se han convertido en una característica fundamental de los motores modernos de búsqueda Web de uso general. Hoy día, es muy común que un usuario reformule su consulta cuando no recibió el resultado ideal de su consulta original. Se puede optimizar el esfuerzo de búsqueda del usuario proporcionando sugerencias antes de que se ejecute la búsqueda, previendo la intención del usuario, de acuerdo con el comportamiento pasado del mismo o el contexto o el comportamiento de otros usuarios [11]. En los últimos años, las propuestas en este ámbito se han enfocado en la utilización de fuentes de conocimiento externo como Wikipedia o DBpedia [12][13][14][15][16]. Este trabajo se enmarca en esta categoría, aunque a diferencia de los trabajos encontrados, combina la fuente externa con el uso de algoritmos de topic modeling y técnicas de descubrimiento de literatura.

2.2 Descubrimiento basado en la literatura

El descubrimiento basado en la literatura busca descubrir nuevos conocimientos a partir de la literatura existente de forma automatizada o semi-automatizada [3] y comúnmente es utilizado en el ámbito de la literatura científica. El gran volumen de publicaciones ha dado lugar a repositorios de literatura especializados que no interactúan, creando silos de conocimiento en los que los descubrimientos en un área no se conocen fuera de ella [17]. A medida que crece la literatura científica, este modelo se está convirtiendo en una herramienta cada vez más necesaria para facilitar la investigación [18][3]. La mayor parte de los sistemas establecidos de descubrimiento basados en la literatura utilizan el modelo de coocurrencias ABC de Swanson [17]. En este modelo, el conocimiento explícito está en el texto al disponer de las relaciones "A implica B" y "B implica C". A partir del conocimiento explícito, el conocimiento implícito se descubre con una conclusión del tipo "por lo tanto, A implica C". Esta es la base empírica sobre la cual construimos nuestro modelo.

2.3 Topic Modeling y los algoritmos LSI y LDA

El modelado de tópicos es una estrategia que permite identificar los principales temas subyacentes en una colección no estructurada de documentos. Específicamente, un tópico se presenta como un patrón recurrente de palabras concurrentes e incluye un grupo de palabras clave que suelen aparecer juntas. El modelado de tópicos puede vincular palabras con el mismo contexto y diferenciar los usos de palabras con diferentes significados [20]. Estos algoritmos normalmente se aplican a colecciones masivas de documentos [4]. Los algoritmos LSI y LDA pertenecen a esta categoría.

El objetivo de LSI [21] (también conocido como Latent Semantic Analysis o LSA) es aprovechar la estructura implícita de orden superior en la asociación de términos con documentos, su estructura semántica, para mejorar la detección de documentos relevantes en base a términos encontrados en las consultas. La técnica utilizada para ello es la descomposición de valores singulares, que se aplica a una matriz de términos del documento descomponiéndola en un conjunto de factores ortogonales a partir de los cuales la matriz original puede aproximarse mediante una combinación lineal.

Por su parte, Latent Dirichlet allocation (LDA) es uno de los algoritmos más utilizados en esta área, y a la vez uno de los más simples [22]. La intuición detrás de LDA es que los documentos presentan múltiples temas que pueden ser obtenidos usando un modelo no supervisado basado en la estadística relacionada a su contenido. El modelo LDA es una herramienta poderosa para descubrir y explotar la estructura temática oculta y es uno de los componentes principales del modelo propuesto en este artículo.

3 Modelo Propuesto

El modelo de expansión propuesto tiene dos niveles de análisis de los tópicos derivados de la consulta original. De esta manera se intenta obtener información de relevancia asociada a los términos de la consulta del usuario en más de un nivel de análisis de tópicos (Fig. 1). El procesamiento comienza con la recepción de una consulta de usuario construida de manera lógica con términos relacionados al tema de interés del usuario. Esta consulta es procesada con el fin de extraer los términos que la componen. Luego se recuperan los documentos de Wikipedia correspondientes a los términos extraídos de la consulta y se procesan para la obtención de los tópicos contenidos en ellos (Fig. 1–a).

Para el primer análisis de tópicos (Fig. 1–b) se asume que hay un tópico por cada término de la consulta y se realiza el proceso de extracción utilizando LDA [22]. Como resultado de este proceso se obtienen las palabras clave asociadas a los tópicos con un valor de pertenencia cada uno de ellos. Esas palabras clave, siempre que no estén contenidas parcial o totalmente en la consulta original, se utilizan para realizar una exploración Web (en este caso Wikipedia) para obtener sus documentos asociados. A partir de esos documentos se realiza un segundo procesamiento para la obtención de tópicos (Fig. 1– c).

Luego se realiza el segundo análisis de tópicos (Fig. 1–d). Este análisis difiere del explicado anteriormente en una característica sustancial, aquí no se determina la cantidad de tópicos que deben estar contenidos en los documentos, sino que se hace un análisis de performance. Dicho análisis consiste en determinar la cantidad óptima de tópicos a partir del cálculo de coherencia de los tópicos resultantes, con las métricas UCI [23] y UMass [24], las cuales se explican al final de esta sección.

Finalmente, las palabras claves cuyo valor de pertenecía a cada tópico es el más alto y que no estén contenidas en la consulta original, se le devuelven al usuario como sugerencias de términos de expansión. Si bien el modelo general mostrado en la Fig. 1 contiene la secuencia de procesos que se realizan en los dos niveles de extracción de tópicos, es conveniente detallar cada uno de ellos.

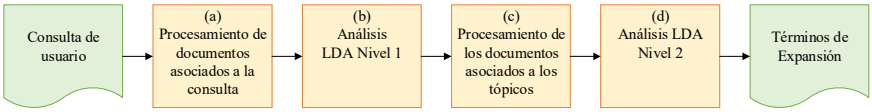

Fig. 1. Modelo general de obtención de términos de expansión.

Las etapas de procesamiento de documentos asociados a la consulta y a los tópicos (Fig. 1– a y c) son similares, la única diferencia es que el primero recibe los términos de la consulta y el segundo, agrega las palabras clave de los tópicos que se extrajeron en el análisis LDA de Nivel 1. Teniendo esto en cuenta, por cuestiones de simplicidad, en la Fig. 2 se muestra el detalle de los procesos a y c de la Fig. 1 como si fuera uno solo.

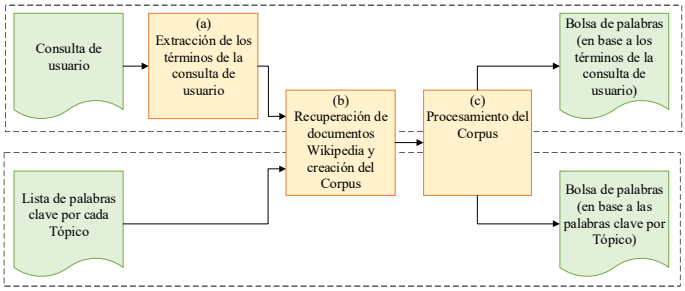

Fig. 2. Procesamiento de documentos.

El primer paso (Fig. 2-a) de este proceso consiste en identificar y extraer los términos de la consulta del usuario. Para esto se realiza un proceso de etiquetado sobre cada una de las palabras de la consulta. Este etiquetado permite identificar la clase de la palabra tal como sustantivo, verbo u otro. Teniendo en cuenta estas etiquetas se arma un árbol sobre la consulta y siguiendo un conjunto de patrones provistos por una librería externa, se identifican los términos claves.

Los términos de la consulta, o los términos obtenidos del análisis LDA de Nivel 1 (dependiendo de la etapa de análisis que se trate), se utilizan como entradas al proceso de recuperación de documentos Web y creación del Corpus (Fig. 2-b). Aquí se realiza la búsqueda de los términos en Wikipedia y luego se extrae el contenido de la página correspondiente para cada término, esta información se unifica en un mismo Corpus y luego se separa el contenido total por párrafos. Este Corpus de documentos debe procesarse (Fig. 2-c). El tratamiento del Corpus se inicia con un proceso de tokenización de cada documento. Esta tokenización consiste en separar cada documento en palabras elementales, llevándolas a todas a minúscula y sacando aquellas palabras sin significado que no aportan información. Luego, se identifican los bigramas del corpus, los cuales son términos de dos palabras que tienen un significado distinto al de las palabras por sí solas, y se los agrega al corpus como términos individuales.

El siguiente paso es la creación del diccionario del corpus que tiene el objetivo de asignar un código numérico a cada token único que exista en el corpus. Con el uso de

este diccionario, se lleva cada documento del Corpus a una representación de bolsa de palabras que consiste en representar a cada documento como un conjunto de pares, donde el primer componente de cada tupla representa el código del token presente y el segundo componente representa la cantidad de veces que ese token se repite en el documento. De esta manera, se obtiene como primera salida el corpus de los términos identificados en el Nivel 1, y como segunda salida el corpus de estos términos más los nuevos surgidos como palabras claves de los tópicos hallados en el proceso LDA de Nivel 1, ambos en una representación de tipo bolsa de palabras.

La entrada para cada proceso de análisis LDA de Nivel 1 y Nivel 2 (Fig. 1– b y d) son las bolsas de palabras generadas en el procesamiento de los documentos Wikipedia. Si bien esas entradas son del mismo tipo y el núcleo del procesamiento LDA es el mismo, ambos procesos son diferentes.

El detalle del proceso de análisis LDA de Nivel 1 (Fig. 1– b) se muestra en la Fig 3. Inicialmente se toma la bolsa de palabras generada en base a los documentos Wikipedia recuperados que están asociados a los términos de la consulta original y, a partir de allí, se utiliza el algoritmo LDA de extracción de tópicos. Además del corpus, otra entrada que se debe definir es la cantidad de tópicos a obtener. Para este caso, se toman tantos tópicos como términos se hayan identificado en la consulta del usuario.

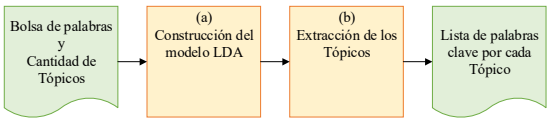

Fig. 3. Análisis LDA Nivel 1.

Como salida del algoritmo LDA se tienen los tópicos del corpus. Cada tópico se representa como un conjunto ordenado de pares que contienen una palabra y un valor asociado que indica la probabilidad de pertenencia de esta palabra al tópico. Se usa este valor para ordenar al conjunto de pares de mayor a menor, de manera que la primera palabra del tópico es la que tiene mayor probabilidad de pertenencia.

A diferencia del proceso de Nivel 1, el análisis LDA de Nivel 2 (Fig. 1– d) tiene como objetivo encontrar la cantidad óptima de tópicos en un rango de 1 a 10 tópicos sobre el agregado de las bolsas de palabras obtenidas a partir de los términos de la consulta de usuario original y de las palabras clave de cada tópico en el proceso LDA de Nivel 1. En la Fig. 4 puede observarse el detalle del análisis LDA de Nivel 2. El proceso consiste en utilizar nuevamente el algoritmo LDA como en la etapa del Nivel 1. Aquí los parámetros que se utilizan son la bolsa de palabras construidas a partir de los términos de la consulta y de las palabras de mayor probabilidad de pertenencia a los tópicos de primer nivel, el diccionario correspondiente a este último corpus y el parámetro correspondiente al número de tópicos con el cual se decide generar los modelos (en este caso de 1 a 10, Fig. 4– a). Así, se obtienen diez modelos, de 1 a 10 tópicos, y cada tópico con sus correspondientes términos asociados (Fig. 4– b).

Para cada uno de estos modelos se calcula el nivel de coherencia de tópicos infiriendo el grado de similitud semántica de las palabras. Dicho cálculo se basa en las métricas UCI y UMass para evaluar el número óptimo de tópicos.

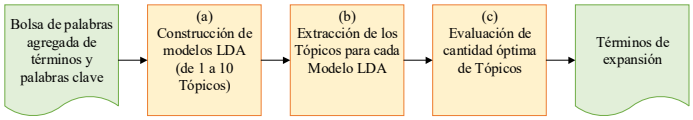

Fig. 4. Análisis LDA Nivel 2.

Estas métricas han demostrado emparejar bien con el juicio humano en calidad de tópicos [25]. Además, ambas se calculan considerando un conjunto N de mejores palabras de un tópico, donde mejores se refiere a las palabras que tienen una mayor probabilidad de correspondencia al tópico [25]. Para el modelo propuesto, se define $N = 10$ en un conjunto de 50 ítems. Así, la medida de coherencia UCI [22] se define como:

$$C_{\text{UCI}} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{PMI}(w_i, w_j) \tag{1}$$

donde PMI es el puntaje de información mutua, que, basado en distribuciones conjuntas e individuales de las palabras w_i y w_j, cuenta la coocurrencia de palabras en una ventana deslizante sobre un corpus externo, y se define como:

$$\text{PMI}(w_i, w_j) = log \frac{P(w_i, w_j) + \varepsilon}{P(w_i) \cdot P(w_j)} \tag{2}$$

La métrica UMass [23], en cambio, contabiliza ocurrencias sobre el corpus original utilizado para entrenar el modelo y se calcula como:

$$C_{\text{UMass}} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \tag{3}$$

donde el factor de suavizado ε se utiliza para evitar el logaritmo de cero. Finalmente, el promedio de estos resultados devuelve el valor de coherencia del modelo. Una vez obtenida la coherencia de los N modelos, se elige el mejor. Es decir, aquél que identifica el óptimo número de tópicos. Por último, los términos incluidos en estos tópicos de la iteración ganadora presentan la máxima coherencia y se utilizan para la expansión.

Con la estrategia de exploración de dos niveles, el modelo propuesto tiene como objetivo la recuperación de términos que puedan contextualizar la consulta de usuario más allá del análisis de los documentos asociados a los términos de la consulta original del usuario. Este aspecto es importante para sugerir términos al usuario ya que muchas veces durante el proceso de generación de las consultas hay aspectos que el usuario considera están incluidos o bien no los había considerado al momento de escribirla.

4 Resultados experimentales

A fin de ejemplificar el funcionamiento del modelo propuesto, en esta sección se muestra el proceso de expansión de tres consultas en distintos contextos. El objetivo de las

pruebas es analizar los tópicos derivados de las consultas y los términos sugeridos para la expansión. Para cada uno de los tres casos se utiliza el procedimiento descripto en la Sección 3, se muestran los resultados parciales obtenidos después de cada análisis LDA (Nivel 1 y Nivel 2), se mencionan algunos aspectos relacionados con los tópicos obtenidos y se presentan los términos de expansión. Para el desarrollo del modelo se utilizó Python 3 y librerías Gensim 3.6.0 y NLTK 3.2.5.

4.1 Prueba 1. Consulta: "covid19 and recession"

La primera consulta evaluada es "covid19 and recession" que está orientada a dos áreas distintas como son la salud y la economía. Luego de identificar los términos `covid19` y `recession`, se recuperan los documentos asociados a ellos desde Wikipedia. Con los contenidos recuperados se construye el corpus y se realiza el análisis LDA de Nivel 1. Aquí, como se detectaron dos términos, el número de tópicos es dos (Tabla 1).

Tabla 1. Resultados análisis LDA Nivel 1 – "covid19 and recession".

Términos identificados en la consulta	Términos agrupados por tópico
`covid19`	[(`virus`, 0.0218), (`sars`, 0.0215), (`sars cov`, 0.0183), (`disease`, 0.0175), (`infection`, 0.0161), (`risk`, 0.0146), (`symptoms`, 0.0135), (`severe`, 0.0125), (`respiratory`, 0.0119), (`infected`, 0.0103)]
`recession`	[(`economic`, 0.0315), (`recessions`, 0.0212), (`rate`, 0.0180), (`gdp`, 0.0173), (`spending`, 0.0157), (`rates`, 0.0149), (`decline`, 0.0147), (`people`, 0.0138), (`government`, 0.0136), (`economy`, 0.0134)]

Luego se toman los diez primeros términos y se vuelve a recuperar los documentos Web asociados a ellos, y se inicia la segunda etapa del proceso. Para ello se añade al corpus existente, los documentos de Wikipedia de los diez primeros términos de cada tópico. Es decir, se realiza la búsqueda en Wikipedia los documentos de `virus`, `sars cov`, etc. del tópico 1 y de `economic`, `recessions`, etc. correspondientes al tópico 2. Cabe aclarar que existe la posibilidad de que algunos términos no posean contenido en Wikipedia y no se sumen al corpus. Cuando el corpus está preparado, se lo utiliza para construir diez modelos utilizando el algoritmo LDA, variando la cantidad de tópicos a obtener de 1 a 10. Estos diez modelos se evalúan utilizando las métricas UCI y UMASS para establecer cuál es el modelo de mayor coherencia (Tabla 2).

Tabla 2. Mejores modelos asociados a la consulta "covid19 and recession"

Modelo (cantidad de tópicos)	UCI	U_Mass
1	-5.7627	-6.9797
2	-3.8204	-4.0667
3	-2.5516	-3.3218
4	-2.3712	-2.7182
5	**-2.3192**	**-2.5767**
6	-3.2200	-3.2400
7	-5.4321	-6.4316
8	-4.5817	-6.1375
9	-6.5742	-6.0660
10	-6.8483	-5.7402

El número de tópicos que posee el mayor valor de coherencia con ambas métricas es cinco. Esto implica que en el corpus final se detectaron cinco temáticas. En la Tabla 3 se observa cada uno de los tópicos obtenido en el análisis LDA de Nivel 2.

Tabla 3. Resultados análisis LDA Nivel 2 – "covid19 and recession".

Tópico	Términos agrupados por tópico
1	[(air, 0.04), (pressure, 0.02), (oxygen, 0.02), (lungs, 0.02), (blood, 0.019), (kpa, 0.013), (respiratory, 0.012), (fig, 0.012), (water, 0.012), (alveoli, 0.012)]
2	[(gdp, 0.024), (economic, 0.021), (recession, 0.016), (economy, 0.016), (consumption, 0.014), (income, 0.009), (goods, 0.009), (government, 0.008), (growth, 0.007), (services, 0.007)]
3	[(risk, 0.053), (people, 0.015), (health, 0.011), (risks, 0.009), (anxiety, 0.007), (social, 0.007), (different, 0.007), (human, 0.007), (associated, 0.006), (management, 0.006)]
4	[(disease, 0.037), (infectious, 0.024), (diseases, 0.022), (infection, 0.017), (medical, 0.013), (person, 0.010), (agent, 0.008), (infected, 0.008), (transmission, 0.007), (spread, 0.007)]
5	[(viruses, 0.034), (virus, 0.027), (symptoms, 0.017), (host, 0.015), (disease, 0.014), (viral, 0.014), (infection, 0.012), (signs, 0.011), (cell, 0.011), (cells, 0.011)]

En el tópico 1, se observan términos como `air`, `oxygen`, `lungs`, `respiratory` y `alveoli` que están claramente relacionados al aparato respiratorio (es el sistema más comprometido por el covid-19). Para el caso del tópico 2, se obtienen términos que están estrechamente relacionados a una recesión económica como `gdp`, `consumption`, `income`, `goods` y `services`. En el tópico 3 aparecen términos como `people`, `health`, `risk`, `anxiety`, `social` y `human`, los cuales permiten inferir que está asociado al aspecto social y a la salud mental de las personas. Lo interesante es que esto puede estar relacionado tanto a `covid19` como a `recession`, ya que el aspecto social se puede ver afectado como consecuencia de ambas problemáticas. Éste es un tema que no estaba explícitamente planteado al inicio. Respecto al tópico 4 se observa que se trata de una característica principal de la enfermedad relacionada con la propagación. Esto se puede inferir por términos como `infectious`, `person`, `agent`, `transmission` y `spread` y, como en el caso anterior, este aspecto tampoco estaba planteado en inicialmente. En relación con el tópico 5 se obtiene términos tales como `virus`, `symptoms`, `host`, `viral`, `signs` y `cell` que hacen una mayor referencia al virus en sí, donde reside y lo que causa. Este tópico sí está relacionado con el término `covid19`.

Como resultado, la cantidad de términos que se sugieren varía en función al número de tópicos. En esta prueba, los términos sugeridos son aquellos que tienen mayor valor de probabilidad de pertenencia en cada tópico (`air`, `gdp`, `risk`, `disease`, `viruses`).

4.2 Prueba 2. Consulta: "meditation and sleep"

La segunda prueba se realiza con la consulta "meditation and sleep" como entrada. Una vez identificados los términos y luego del análisis LDA de Nivel 1, el modelo generado devuelve los resultados que se muestran en la Tabla 4. Con los nuevos términos de cada tópico se construye el nuevo corpus y se ejecutan los diez modelos LDA. Las medidas de coherencia obtienen el valor óptimo de tres tópicos (Tabla 5).

Tabla 4. Resultados análisis LDA Nivel 1 – "meditation and sleep".

Términos identificados en la consulta	Términos agrupados por tópico
meditation	[(practice, 0.0250), (practices, 0.0237), (meditative, 0.0196), (spiritual, 0.0152), (god, 0.0146), (techniques, 0.0143), (christian, 0.0139), (yoga, 0.0137), (mind, 0.0134), (buddhist, 0.0130)]
sleep	[(rem, 0.0201), (night, 0.0183), (brain, 0.0177), (circadian, 0.0171), (time, 0.0166), (hours, 0.0164), (people, 0.0162), (memory, 0.0158), (body, 0.0143), (quality, 0.0115)]

Tabla 5. Mejores modelos asociados a la consulta "meditation and sleep"

Modelo (cantidad de tópicos)	UCI	U_Mass
1	-5.6803	-6.1041
2	-2.9941	-3.0318
3	**-2.6135**	**-2.8696**
4	-3.5599	-2.9391
5	-3.9063	-3.5661
6	-4.2438	-3.6587
7	-4.8696	-4.0800
8	-6.3297	-6.0144
9	-6.7357	-6.2123
10	-7.5702	-6.9310

Los términos contenidos en esos tres tópicos se pueden observar en la Tabla 6. Al analizar estos resultados se pueden inferir tres temáticas distintas. En el tópico 1, términos como memory, brain, time, hours, day y circadian identifican al sueño, su duración y quizás, las partes de nuestro organismo que intervienen en esta actividad. En este caso este concepto se encontraba explícito en la consulta original. En ambos tópicos restantes se evidencia el tema de la meditación, pero con una diferencia que los distingue. El tópico 2 contiene términos como buddhism, yoga, million, asia y traditions que hacen referencia a la meditación como una tradición, o como parte de una religión. En cambio, en el tópico 3 se tienen los términos mind, mental, nature, existence, self y consciousness que revelan a la meditación como una actividad mental y quizás, los beneficios que conllevan su práctica. Si bien el término meditation estaba explícito en la consulta original, esta distinción en dos subtemáticas distintas no lo estaba. Para este caso, el modelo propuesto sugiere los términos memory, buddhist y mind para expandir la consulta de búsqueda, ya que son estos los que tienen mayor probabilidad de pertenencia.

Tabla 6. Resultados análisis LDA Nivel 2 – "meditation and sleep".

Tópico	Términos agrupados por tópico
1	[(memory, 0.028), (brain, 0.019), (sleep, 0.014), (term, 0.010), (time, 0.009), (information, 0.008), (hours, 0.008), (day, 0.008), (long, 0.007), (circadian, 0.007)]
2	[(buddhist, 0.039), (buddhism, 0.028), (buddhists, 0.015), (yoga, 0.014), (million, 0.012), (asia, 0.011), (traditions, 0.010), (meditation, 0.008), (like, 0.008), (modern, 0.008)]
3	[(mind, 0.028), (time, 0.021), (mental, 0.015), (meditation, 0.008), (states, 0.008), (different, 0.007), (nature, 0.007), (existence, 0.006), (self, 0.006), (consciousness, 0.006)]

4.3 Prueba 3. Consulta: "obesity and fast foods”

Para la tercera prueba se utiliza con la consulta “obesity and fast foods”. Los resultados tras el análisis LDA de Nivel 1 se muestran en la Tabla 7.

Tabla 7. Resultados análisis LDA Nivel 1 – “obesity and fast foods”.

Términos identificados en la consulta	Términos agrupados por tópico
obesity	[(weight, 0.0305), (obese, 0.0202), (people, 0.0169), (bmi, 0.0166), (energy, 0.0166), (rates, 0.0145), (increased, 0.0142), (risk, 0.0141), (children, 0.0135), (fat, 0.0133)]
fast foods	[(restaurants, 0.0248), (united, 0.0142), (world, 0.0136), (states, 0.0123), (united states, 0.0122), (restaurant, 0.0121), (served, 0.0113), (billion, 0.0098), (health, 0.0093), (cooked, 0.0088)]

Luego del análisis LDA de Nivel 2 se halla el modelo óptimo con dos tópicos para ambas métricas (Tabla 8). En este caso los dos tópicos que se obtienen (Tabla 9) están relacionados en forma general a los términos originales de la consulta, aunque su estrecha relación hace que la diferenciación sea dificultosa. Términos del tópico 1 como trans, fat, fatty, acids, trans fats y consumption se pueden identificar a la temática de restaurantes de comida rápida, pero también son aplicables a la obesidad. Además, los términos energy, health, world, weight, people y child del tópico 2 se identifican indudablemente a la obesidad, en donde sucede y a qué población. Es interesante analizar la presencia de la palabra risk en ambos tópicos, lo que podría significar un posible nexo entre ambas temáticas. Finalmente, los términos trans y energy son los que el modelo propone añadir a la consulta original.

Tabla 8. Mejores modelos asociados a la consulta “obesity and fast foods”.

Modelo (cantidad de tópicos)	UCI	U_Mass
1	-4.2923	-2.6195
2	**-2.9119**	**-2.1745**
3	-5.0362	-4.1781
4	-3.0992	-2.7154
5	-3.0813	-2.6091
6	-5.1137	-4.6045
7	-5.2788	-4.5489
8	-5.3402	-4.6099
9	-5.7430	-5.1144
10	-7.1513	-5.9605

Tabla 9. Resultados análisis LDA Nivel 2 – “obesity and fast foods”.

Tópico	Términos agrupados por tópico
1	[(trans, 0.040), (fat, 0.029), (fats, 0.026), (fatty, 0.018), (acids, 0.015), (trans_fat, 0.015), (trans_fats, 0.013), (risk, 0.013), (consumption, 0.010), (acid, 0.009)]
2	[(energy, 0.021), (risk, 0.020), (obesity, 0.013), (health, 0.013), (world, 0.011), (children, 0.011), (weight, 0.010), (people, 0.006), (states, 0.006), (child, 0.005)]

5 Discusión y trabajos futuros

En este artículo se describe un modelo de sugerencias de términos de expansión de consultas para el proceso de búsqueda de información en la Web. El modelo propuesto trabaja con algoritmos de modelado de tópicos y términos que se extraen de Wikipedia y genera sugerencias que se le presentan al usuario para expandir su consulta.

Las pruebas realizadas reflejan que, además de términos relacionados con las palabras contenidas en la consulta original, el modelo propuesto encuentra asociaciones de términos en tópicos que están implícitos en la consulta. Esto es de suma importancia ya que uno de los problemas en la recuperación de información mediante búsquedas es que el usuario puede asumir cuestiones que no están explícitas o incluso omitir algunas otras que considera relevantes.

El modelo propuesto tiene la ventaja de no necesitar ejemplos, ni requiere una etapa de entrenamiento. Por la naturaleza no supervisada de la obtención de tópicos con LDA, su utilización puede extenderse a cualquier dominio de búsqueda.

La utilización de Wikipedia como fuente de información tiene la desventaja de la posible inconsistencia de los documentos. No obstante, la constante actualización de esos contenidos permite que los resultados obtenidos por el modelo reflejen los cambios sin necesidad de implementar técnicas específicas de manejo de conocimiento.

El procesamiento implementado en el modelo es incremental y adaptativo. Esto significa que se le sugieren nuevos términos al usuario y, según los que elija, los requerimientos de búsqueda pueden adaptarse a las variaciones que se produzcan ya sea por omisión o descubrimiento de aspectos no contemplados en la consulta original.

Es factible implementar un proceso de búsqueda continua con el modelo propuesto. De esta manera, mientras el usuario decida mantenerlo activo, recibirá las sugerencias que se obtengan mediante la adaptación de los resultados.

5.1 Trabajos futuros

Actualmente se está trabajando en cuatro aspectos específicos del modelo, el primero de ellos es la generación de grafos de conocimiento (Knowledge Graphs - KG) a partir de los términos descubiertos. La idea es explorar relaciones no solamente entre la consulta de usuario y los términos que se puedan generar a partir de ella, sino también analizar relaciones entre dichos términos. Además, la ventaja de trabajar con KG es que se pueden utilizar diversas métricas como herramienta complementaria.

El segundo aspecto es la prueba de algoritmos alternativos al LDA (como LSA y Word Embeddings). Esto permitirá realizar evaluaciones cuantitativas de los modelos.

El tercer aspecto es el desarrollo de una estrategia que permita determinar el área de conocimiento de la búsqueda (tema) para utilizar el modelo propuesto con otras fuentes de datos más especializadas.

Finalmente, como cuarto aspecto, se está trabajando en la incorporación de elementos que permitan la interacción entre el usuario y el modelo propuesto. De esta manera la selección de términos que realice el usuario realimentará el proceso de expansión contribuyendo a potenciar la característica de adaptación mencionada anteriormente.

Agradecimientos

Este artículo fue desarrollado dentro del marco del proyecto “Diseño de algoritmos inteligentes para el análisis de información desestructurada” - SIUTIRE5276TC de la Universidad Tecnológica Nacional – Facultad Regional Resistencia (Argentina). Queremos hacer un especial reconocimiento en memoria de nuestro amigo Carlos Pérez.

Referencias

1. Van Rijsbergen, C.J.: Information Retrieval, 2nd edition. Butterworths. (1979).
2. Ooi, J., Ma, X., Qin, H., Liew, S.C.: A survey of query expansion, query suggestion and query refinement techniques. In: 2015 4th International Conference on Software Engineering and Computer Systems, ICSECS 2015: Virtuous Software Solutions for Big Data (2015). https://doi.org/10.1109/ICSECS.2015.7333094.
3. Henry, S., McInnes, B.T.: Literature Based Discovery: Models, methods, and trends, (2017). https://doi.org/10.1016/j.jbi.2017.08.011.
4. Blei, D.: Introduction to Probabilistic Topic Modeling. Commun. ACM. (2012). https://doi.org/10.1145/2133806.2133826.
5. Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: A survey. Inf. Process. Manag. 56, 1698–1735 (2019). https://doi.org/10.1016/j.ipm.2019.05.009.
6. Singh, J., Sharan, A., Siddiqi, S.: A Literature Survey on Automatic Query Expansion for Effective Retrieval Task. Int. J. Adv. Comput. Res. (2013).
7. Zeng, Y., Lin, W., Lei, K., Huang, L.: Improving retrieval performance with Wikipedia’s category knowledge. In: Proceedings - 4th International Conference on Computational and Information Sciences, ICCIS 2012 (2012). https://doi.org/10.1109/ICCIS.2012.174.
8. Sharma, D.K., Pamula, R., Chauhan, D.S.: Semantic approaches for query expansion. Evol. Intell. (2021). https://doi.org/10.1007/s12065-020-00554-x.
9. Vélez, B., Weiss, R., Sheldon, M.A., Gifford, D.K.: Fast and effective query refinement. SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval). (1997). https://doi.org/10.1145/278459.258528.
10. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008). https://doi.org/10.1145/1401890.1401995.
11. Sethi, S., Dixit, A.: An efficient personalized query suggestion technique for providing relevant results. In: Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016 (2016).
12. Anand, R., Kotov, A.: An empirical comparison of statistical term association graphs with dbpedia and conceptnet for query expansion. In: ACM International Conference Proceeding Series (2015). https://doi.org/10.1145/2838706.2838715.
13. Aggarwal, N., Buitelaar, P.: Query expansion using wikipedia and DBpedia.

In: CEUR Workshop Proceedings (2012).
14. Al Masri, M., Berrut, C., Chevallet, J.P.: Wikipedia-based semantic query enrichment. In: International Conference on Information and Knowledge Management, Proceedings (2013). https://doi.org/10.1145/2513204.2513209.
15. Arguello, J., Elsas, J.L., Callan, J., Carbonell, J.G.: Document representation and query expansion models for blog recommendation. In: ICWSM 2008 - Proceedings of the 2nd International Conference on Weblogs and Social Media (2008).
16. Li, Y., Luk, W.P.R., Ho, K.S.E., Chung, F.L.K.: Improving weak ad-hoc queries using wikipedia asexternal corpus. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07 (2007). https://doi.org/10.1145/1277741.1277914.
17. Swanson, D.R.: Medical literature as a potential source of new knowledge. Bull. Med. Libr. Assoc. (1990).
18. Kostoff, R.N., Briggs, M.B., Lyons, T.J.: Literature-related discovery (LRD): Potential treatments for Multiple Sclerosis. Technol. Forecast. Soc. Change. (2008). https://doi.org/10.1016/j.techfore.2007.11.002.
19. Weeber, M., Klein, H., De Jong-Van Den Berg, L.T.W., Vos, R.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. J. Am. Soc. Inf. Sci. Technol. (2001). https://doi.org/10.1002/asi.1104.
20. Barde, B.V., Bainwad, A.M.: An overview of topic modeling methods and tools. In: Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017 (2017). https://doi.org/10.1109/ICCONS.2017.8250563.
21. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. (1990). https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
22. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. (2003). https://doi.org/10.1016/b978-0-12-411519-4.00006-9.
23. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2011).
24. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the ACM International Conference on Digital Libraries (2010). https://doi.org/10.1145/1816123.1816156.
25. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference (2012).