Modelo para Expansión de Consultas basado en Información Desestructurada

Raúl Montiel
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
raulmontiel@ca.frre.utn.edu.ar

Marcelo Karanik
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
marcelo@frre.utn.edu.ar

Patricio Costilla
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
patriciocostilla@ca.frre.utn.edu.ar

Federico Zimmermann
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
fedezimm@ca.frre.utn.edu.ar

Gerardo Enrique
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
geraenrique97@ca.frre.utn.edu.ar

Mariano Minoli
Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional
Fac. Regional Resistencia
Resistencia, (3500) Chaco. Argentina.
mariano_minoli@ca.frre.utn.edu.ar

***Abstract*.** Nowadays, getting meaningful results in the Internet search process is not a trivial task. This is partly because the huge amount of information available on the Web can hide valuable results if the appropriate terms are not written in the query. In this context, query expansion is a mechanism that improves the quality of the results. The use of reliable data sources and the strategy to select the new terms are key aspects of the expansion process. This article describes an unsupervised query expansion model based on Wikipedia. To get terms, topic modeling and knowledge graphs are used.

***Keywords*.** Query Expansion; Natural Language Processing; Knowledge Graphs; Topic Modeling.

***Resumen*.** Hoy día la obtención de resultados útiles en el proceso de búsqueda en Internet no es una tarea trivial. Esto se debe en parte porque la inmensa cantidad de información disponible puede ocultar resultados valiosos si no se usan los términos adecuados en la consulta. En este contexto, la expansión de consultas es un mecanismo que permite mejorar la calidad de los resultados. Tanto la utilización de una fuente fiable de datos como de la estrategia para seleccionar los nuevos termino son aspectos clave en el proceso de expansión. En este artículo se describe un modelo de expansión de consultas no supervisado basado en Wikipedia. Para la obtención de términos se utiliza modelado de tópicos y grafos de conocimiento.

Palabras Clave. Expansión de consultas; Procesamiento de lenguaje natural; Grafos de Conocimiento; Modelado de Tópicos.

I. Introducción

En la actualidad, debido al gran volumen de datos disponibles, encontrar información en Internet que sea de utilidad es uno de los más grandes desafíos al que se enfrentan los diseñadores de motores de búsqueda. Así, las técnicas que dan soporte al usuario en la redefinición de las consultas para los buscadores han obtenido una gran relevancia como área de investigación [1]–[3]. Una de esas técnicas de modificación es la expansión de consultas (Query Expansion - QE), que consiste en la reformulación de la consulta original agregando términos relacionados a la misma [4].

Un punto clave en el proceso de QE es la selección de los términos que se agregan a la consulta original y, por lo general, esta selección está condicionada por la poca información del contexto. Una alternativa para enfrentar este problema es utilizar bases de conocimiento como Wikipedia o DBPedia para obtener documentos y conceptos relacionados a los términos originales de la consulta [5]. Claramente, el análisis del contenido de los documentos requiere la utilización de algoritmos de Procesamiento del Lenguaje Natural (Natural Language Processing - NLP) [6]. NLP provee una gran cantidad de técnicas que pueden usarse en el reconocimiento de texto para la obtención de términos útiles para QE [4].

Si bien la utilización de algoritmos de NLP sirve para la identificación de la estructura de las oraciones, para la detección de entidades o para el modelado de tópicos en documentos Web, hay que considerar qué relación tienen los términos obtenidos con los términos de la consulta original. En otras palabras, se debe establecer la relevancia de las relaciones entre los términos de expansión respecto a la consulta original. En este sentido, una alternativa válida de representación y evaluación de dichas relaciones es la utilización de Grafos de Conocimiento (Knowledge Graphs - KG) [7], donde los términos se representan como nodos y las relaciones como arcos que los conectan.

En este artículo se describe un modelo de QE, actualmente en fase de desarrollo, que utiliza Wikipedia (como base de conocimiento) y que combina varias técnicas de NLP para extracción de términos y KG para representar y evaluar las relaciones entre dichos términos y la consulta original. En la Sección II se describen conceptos preliminares de QE y KG: En la Sección III se describe el modelo propuesto. En la Sección IV se plantean algunos puntos de discusión del modelo y, finalmente, en la Sección V, se mencionan las líneas de trabajo futuro.

II. Conceptos Preliminares

El proceso de QE consta de 4 pasos: (a) procesamiento previo de fuentes de datos y extracción de términos, (b) ponderaciones y clasificación de términos, (c) selección de términos y (d) reformulación de consultas [4].

En el paso (a) se utilizan diferentes técnicas como división del flujo de texto en palabras, eliminación de palabras de uso frecuente, por ejemplo, artículos, adjetivos, preposiciones, etc. También la derivación de palabras es un proceso de reducción de palabras derivadas a su palabra base. Una vez identificados estos términos se deben obtener las relaciones entre los objetos, que se utilizan en el paso (b) para luego en base a las ponderaciones obtenidas seleccionar los términos en (c), para finalmente agregarlos en (d) [4].

En los diferentes pasos es necesario consultar fuentes de datos, es decir, recursos y colecciones de textos externos (Internet, Corpus Externos, etc), recursos de conocimiento hechos a mano (Diccionarios, Ontologías, Wikipedia, etc.), documentos utilizados en el proceso de recuperación [4].

Una manera natural de organizar el conocimiento es utilizar grafos, donde los conceptos se representan como nodos y las relaciones como arcos que los conectan. Específicamente KG es una representación estructurada del conocimiento, que consta de entidades, relaciones, atributos y descripciones semánticas[8].

Una entidad es un objeto, una relación describe la interacción y la influencia entre dos entidades; un atributo describe las características de una entidad y una descripción semántica incluye la cadena del nombre de la entidad, el valor numérico, la información literal y el valor del atributo de la cadena, etc. [7].

Debido a que un KG brinda esta estructura es posible explorar relaciones no solamente entre los términos de una consulta sino a partir de nuevos que se puedan generar a partir de ella, analizando nuevas relaciones que puedan aflorar en el KG. Además, otra ventaja de su utilización es que pueden utilizarse métricas referidas a grafos, aportando elementos cuantitativos para la evaluación de nuevos términos.

Una las métricas para grafos es la que devuelven los algoritmos de centralidad. Estos algoritmos se utilizan para comprender cuál es papel individual que cumplen los nodos en un grafo y su impacto en la estructura total. Son útiles porque identifican los nodos más importantes y su dinámica de interacción con el resto de los nodos. Muchos de estos algoritmos fueron propuestos inicialmente para el análisis de redes sociales, desde entonces han encontrado usos en una variedad de industrias y campos incluyendo diversas áreas del procesamiento del lenguaje natural [9].

III. Modelo propuesto.

La idea principal en la que se basa la propuesta de QE que se describe en este artículo, es la utilización del conocimiento disponible en Internet, específicamente en Wikipedia. De esta manera y a partir de una consulta de usuario, se intenta encontrar términos relacionados mediante la exploración del contenido Web que describe los conceptos implicados. El proceso completo de expansión se muestra en la Fig. 1 que, por fines descriptivos, se ha divido en seis subprocesos.

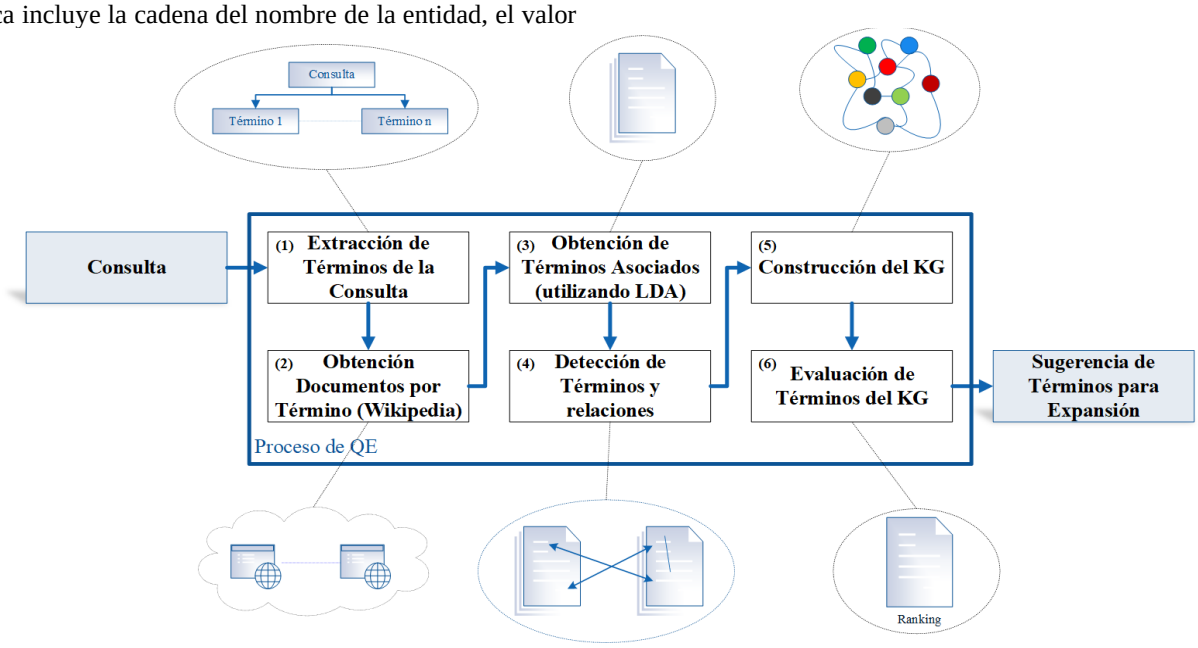

Fig. 1. Modelo de Expansión de Consultas de Búsqueda.

Al inicio del proceso, la consulta original del usuario se analiza buscando los términos que la componen (Fig. 1 (1)). Para ello se etiqueta cada una de las palabras para determinar si es un sustantivo, adjetivo, artículo, verbo, etc. Luego de esto, se eliminan las palabras sin significado (stop words) y se devuelven los términos en una lista. Por cada elemento de la lista se realiza la búsqueda en Wikipedia y se extrae el contenido correspondiente para cada término (Fig. 1 (2)).

El contenido de la página relacionada a cada término se procesa para crear el Corpus. El tratamiento del Corpus se inicia con la tokenización de cada documento que consiste en separar cada documento en palabras elementales, llevándolas a todas a minúscula y quitando los stop words.

Luego, se identifican los n-gramas que son términos que llevan más de una palabra y su significado es distinto al de las palabras por sí solas. Después se crea el diccionario del Corpus donde se asigna un código numérico a cada token único. En base al diccionario se transforma cada documento en una bolsa de palabras compuesta como un conjunto de pares, donde el primer componente representa el código del token y el segundo representa las repeticiones del token en el documento.

Con las bolsas de palabras se realiza un proceso de modelado de tópicos utilizando el algoritmo Latent Dirichlet allocation (LDA) [10] (Fig. 1 (3)). 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 útil para descubrir y explotar la estructura temática oculta en documentos de texto. Como resultado del algoritmo LDA se obtienen las palabras clave relacionadas al tópico de cada término de la consulta. Cada tópico contiene un conjunto ordenado de pares palabra clave con su valor de pertenencia al tópico.

A este punto se tiene una lista de palabras (con su respectivo valor de pertenencia) que están relacionadas a un término de la consulta. Este proceso de obtención de listas se repite con las 50 primeras palabras por tópico. Esto genera una cantidad de listas de términos ordenados que deben ser analizadas para encontrar relaciones entre ellas (Fig. 1 (4)). Ese análisis consiste en buscar en las listas, tomadas de a pares y de manera exhaustiva, las palabras que estén contenidas en ambas. Estas coincidencias denotan relaciones entre los términos que generaron las listas obtenidas por el algoritmo LDA. Genéricamente, dadas dos listas asociadas a dos términos distintos, t_i y t_j, si un término t_k está contenido en ambas listas con valor de pertenencia v_{ki} en la lista de ti y valor de pertenencia v_{kj} en la lista de tj, se almacena la tupla (t_i, t_j, t_k, v_{ki}, v_{kj}).

Una vez obtenidas todas las coincidencias, la lista de tuplas (t_i, t_j, t_k, v_{ki}, v_{kj}) se utilizan para construir el KG (Fig. 1 (5)). La construcción genera un nodo por cada término y se los relaciona como se muestra en la Fig. 2.

Fig. 2. Relaciones entre términos del KG.

Una vez que se han conectado todos los nodos se procede al cálculo de la importancia de cada relación utilizando el valor v_{kj} (para la relación entre los nodos t_i y t_k) y el valor v_{kj} (para la relación entre los nodos t_j y t_k). Para el cálculo de la relevancia de cada relación, además de los valores de pertenencia v_{ki} y v_{kj}, se utilizan las posiciones del término coincidente en cada una de las listas para generar una ponderación asociada a cada relación:

$$w_{ki} = \frac{1}{pos_{ki} \cdot \left(1 + e^{-1/pos_{ki}}\right)} \quad (1)$$

$$w_{kj} = \frac{1}{pos_{kj} \cdot \left(1 + e^{-1/pos_{kj}}\right)} \quad (2)$$

donde w_{ki} y w_{kj} son los pesos asociados a cada conexión, pos_{ki} y pos_{kj} son las posiciones del término t_k en las listas de t_i y t_j respectivamente. Finalmente, los valores de las relevancias de las relaciones se obtienen como:

$$rel_{ki} = w_{ki} \cdot v_{ki} \quad (3)$$

$$rel_{kj} = w_{kj} \cdot v_{kj} \quad (4)$$

Se puede observar en las ecuaciones anteriores que la relevancia de una conexión no solamente depende de los valores de pertenencia de un término a las listas que devuelve el algoritmo LDA, sino que también se tiene en cuenta la posición que ocupa en dichas listas. Esto es importante ya que los valores de pertenencia son relativos a cada tópico analizado.

Como se mencionó antes, los nodos del KG son los términos coincidentes a partir del análisis de las listas de palabras relacionadas a los tópicos que retorna el algoritmo LDA. Es decir, son los términos derivados de la consulta original del usuario que tienen relación con el tema de búsqueda. Dependiendo de ese tema, la cantidad de nodos varía y, en base a los resultados preliminares que se han obtenido hasta el momento, pueden ser varios cientos de nodos.

Para poder analizar la importancia de cada nodo hay que establecer de qué manera está relacionado con el resto de los nodos en el KG. Para ello se utilizan métricas de centralidad para obtener el ranking de los mejores nodos respecto a su influencia (Fig. 1 (6)). Para este modelo se utilizan dos algoritmos de centralidad (actualmente en fase de implementación), Closeness Centrality [9] que detecta los nodos que pueden difundir información eficientemente a través de un subgrafo y Betweenness Centrality [11] para analizar la cantidad de influencia que un nodo tiene sobre el flujo de información o recursos.

Los mejores nodos que se obtienen al aplicar las medidas de centralidad son los que finalmente se sugieren al usuario, en forma de ranking de términos (Fig. 1 (6)), para completar el proceso de QE.

IV. Discusión.

En este artículo se describe un modelo de sugerencias de términos de expansión de consultas (QE) para el proceso de búsqueda de información en la Web. El modelo descripto, actualmente en fase de implementación, genera un grafo de conocimiento (KG) con términos obtenidos del modelado de tópicos sobre documentos de Wikipedia asociados a la consulta de usuario. Finalmente se obtiene un ranking de los términos contenidos en el grafo que se le presenta usuario para expandir su consulta.

Un punto destacado de la propuesta es que, dado el enfoque no supervisado del modelo, no se requiere ejemplos ni etapa de entrenamiento para para la construcción del KG.

Esto es de suma importancia porque la utilización del modelo puede extenderse a cualquier ámbito.

Otra característica del modelo es que se generan relaciones entre los términos que no necesariamente estuvieron contenidas en la consulta original. Así, se pueden sugerir términos que el usuario no haya contemplado inicialmente o, incluso, haya asumido como implícitos en su consulta.

Al utilizar medidas de centralidad, el modelo considera tanto la calidad como la cantidad de relaciones asociadas a cada nodo. Esto es posible dado que los valores de relevancia de las relaciones inciden directamente sobre la valoración de los términos respecto a los demás.

A pesar de las posibles inconsistencias, la utilización de Wikipedia como fuente de datos provee un mecanismo de actualización constante de los conceptos que se reflejará en los resultados obtenidos sin la necesidad de contar con técnicas adicionales de manejo de conocimiento.

Finalmente, es destacable que el modelo puede utilizar cualquier fuente de documentos Web. Es decir, se puede utilizar el modelo para áreas de conocimiento específicos (publicaciones científicas, documentación técnica, noticias, etc.) solamente cambiando la fuente de obtención de los documentos Web.

V. Trabajos Futuros.

Como se mencionó, este modelo se encuentra actualmente en fase de implementación. Específicamente, se está trabajando en tres líneas de desarrollo. La primera línea es la evaluación de los KG a partir de los términos descubiertos utilizando medidas de centralidad. Con esta evaluación se validarán los resultados en base a la precisión de los resultados que se obtengan.

La segunda línea de trabajo es el desarrollo de una estrategia de contextualización de las búsquedas. Esto permitirá poder seleccionar fuentes de datos especializadas relacionadas al tema de búsqueda y, de esta manera, obtener resultados más precisos.

Finalmente, la tercera línea está relacionada a la interacción del modelo con el usuario. La idea aquí es proveer mecanismos de feedback que permitan mejorar la especificidad de los resultados.

Agradecimientos.

Este artículo fue desarrollado en el marco del proyecto “Diseño de algoritmos inteligentes para el análisis de información desestructurada” (SIUTIRE5276TC). Universidad Tecnológica Nacional – Facultad Regional Resistencia (Argentina).

Referencias

[1] J. Ooi, X. Ma, H. Qin, and S. C. Liew, “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.

[2] C. Bobed and E. Mena, “QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems,” *Inf. Sci. (Ny).*, 2016.

[3] S. Balaneshin-Kordan and A. Kotov, “Sequential query expansion using concept graph,” in *International Conference on Information and Knowledge Management, Proceedings*, 2016.

[4] H. K. Azad and A. Deepak, “Query expansion techniques for information retrieval: A survey,” *Inf. Process. Manag.*, vol. 56, no. 5, pp. 1698–1735, Sep. 2019.

[5] J. Guisado-Gámez and A. Prat-Pérez, “Understanding graph structure of wikipedia for query expansion,” *3rd Int. Work. Graph Data Manag. Exp. Syst. GRADES 2015 - co-located with SIGMOD/PODS 2015*, 2015.

[6] E. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research,” *IEEE Computational Intelligence Magazine*. 2014.

[7] K. Zeng, C. Li, L. Hou, J. Li, and L. Feng, “A comprehensive survey of entity alignment for knowledge graphs,” *AI Open*, vol. 2, pp. 1–13, 2021.

[8] Y. Duan, L. Shao, G. Hu, Z. Zhou, Q. Zou, and Z. Lin, “Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph,” in *Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017*, 2017.

[9] F. Boudin, “A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction,” in *International Joint Conference on Natural Language Processing (IJCNLP)*, 2013.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, 2003.

[11] L. C. Freeman, “A Set of Measures of Centrality Based on Betweenness,” *Sociometry*, 2006.