Expansión de Claves de Búsqueda: un Enfoque Basado en Análisis de Entidades

Patricio Costilla, Raúl Montiel, Jorge Roa

Centro de Investigación Aplicada en TIC (CInApTIC)
Universidad Tecnológica Nacional, Facultad Regional Resistencia
French 414, Resistencia 3500, Chaco, Argentina
patriciocostilla,raulmontiel,roajorge{@ca.frre.utn.edu.ar}

Abstract. La expansión de claves de búsqueda es una técnica que permite lograr mayor calidad en los resultados cuando se realizan búsquedas en Internet. Claramente, la elección de buenos términos es de suma importancia para obtener documentos acordes a las necesidades del usuario. Uno de los desafíos encontrados en el proceso de expansión de consultas, además de la elección de la fuente a consultar y cómo realizar esa consulta, es qué criterio utilizar para la selección de los términos que se consideren mejores candidatos para incluirlos en la expansión. En este artículo se describe un modelo que, a partir de una clave de búsqueda, detecta las entidades contenidas en ella, explora con qué otras entidades están relacionadas y construye grafos de conocimiento parciales basados en documentos de Wikipedia. Durante el proceso de creación de los grafos parciales se calcula la relevancia de cada nodo y se integran todos los nodos en un grafo final. Por último, los valores de relevancia obtenidos definen las mejores entidades que serán sugeridas como términos de expansión. Además de la descripción del modelo, se presentan dos ejemplos de utilización.

Keywords: expansión de consultas, grafos de conocimiento, análisis de información desestructurada, recuperación de información.

1 Introducción

Encontrar información útil sobre un tema puede resultar desafiante, sobre todo si es un tópico desconocido o nuevo para quien lo realiza [1]. La definición acertada de una clave de búsqueda cumple un rol fundamental para obtener resultados valiosos.

Las claves de búsqueda suelen ser simplificaciones de la información que se desea obtener, definidas en un lenguaje desestructurado de consulta y de fácil utilización [2]. La elección desacertada de vocabulario y consultas subespecíficas, que contienen solo una fracción de conceptos que representan el tema, equivalen a resultados incompletos e inexactos [3]. Por tanto, y de acuerdo con los resultados obtenidos las claves deben ser reformuladas. Una técnica que ayuda a lograr este objetivo es la expansión de consultas o de términos, que consiste en un proceso de reformulación de la consulta inicial agregando términos relacionados a la consulta original, obteniendo una mejora en la calidad de los resultados [1]. De esta forma, si un usuario ingresa la consulta en inglés

"Car", una herramienta de expansión de consultas podría sugerirle los siguientes términos de expansión: "vehicle", "automobiles", "auto", etc.

El desafío que se plantea es como elegir los nuevos términos para la expansión de una clave de búsqueda. Si bien hay varias alternativas para hacer la expansión, encontrar y administrar las relaciones entre los conceptos suele ser un problema importante. Para tratar esta dificultad, se pueden utilizar grafos de conocimiento (Knowledge Graphs - KG) que proveen la estructura adecuada para el manejo de relaciones, y pueden utilizarse para mejorar el proceso de expansión de términos [4]. Los KG se han convertido en una herramienta para representar conocimiento en forma de grafos etiquetados y para dar semántica a la información textual. Un KG es un grafo construido al representar cada elemento, entidad o usuario como nodos, vinculando los nodos que interactúan entre sí a través de arcos. Los KG tienen una estructura cercana al lenguaje natural, y con ello son una herramienta poderosa para esta área [5].

Además de la representación del conocimiento, otro factor a tener en cuenta en la expansión de claves de búsqueda es seleccionar la fuente de información de donde extraer los términos y las relaciones. Si bien existen muchas bases de conocimiento disponibles, Wikipedia es una de las más grandes y ricas en conocimiento semántico, altamente actualizada y de rápido acceso [1].

Dado este contexto, se describe aquí un modelo de expansión de las claves de búsqueda mediante la incorporación de términos relacionados semánticamente. El modelo extrae entidades de la clave de búsqueda, obtiene conceptos relacionados en Wikipedia, que los representa mediante un grafo, y determina los términos relevantes para la expansión. Cabe destacar que, debido a que en Wikipedia existe una gran cantidad de artículos disponible, que cambia constantemente, el diseño del modelo propuesto permite trabajar con información actualizada sin necesidad de recolectarla constantemente. De esta manera, se adquiere los datos necesarios al momento de procesarlos.

El artículo se encuentra organizado de la siguiente manera: en la Sección 2 se describen algunos trabajos relacionados a la expansión de términos. La descripción del modelo propuesto se presenta en la Sección 3. En la Sección 4 se muestran ejemplos de utilización. Finalmente, en la Sección 5 se presentan las conclusiones y trabajos futuros.

2 Trabajos Relacionados

La expansión de claves de búsqueda es un tema que ha sido abordado desde diferentes perspectivas. Respecto a la utilización de grafos de conocimiento, en [5] se detalla cómo se estructuran las bases de conocimientos, y cómo es posible utilizar grafos para mapear las entidades y relaciones encontradas. Otra alternativa, presentada en [6], se basa en un ranking de grafos, combinando un algoritmo para la expansión de consultas y el resumen de varios documentos. De esta manera las relaciones entre oraciones y las relaciones de oración a palabra se utilizan para encontrar los términos de expansión. En [7] se describe un enfoque interactivo donde las claves de búsqueda se modelan mediante reglas de asociación y la transitividad de las relaciones con grafos. De esta manera se le presenta al usuario una serie de términos relacionados a su consulta que se utilizan en el proceso de expansión. En [3] se asignan pesos a los conceptos de acuerdo

con diferentes tipos de características, como ser las estadísticas de un grafo de conceptos. Luego se intenta encontrar los conceptos más eficaces para la expansión de términos, seleccionando, en un grafo, conceptos de manera secuencial en base a las distancias entre los conceptos y la consulta original. Existen, además, enfoques relacionados con el uso de distribuciones de probabilidad como el propuesto en [8]. En este enfoque se determina estadísticamente la relevancia de cada término respecto a la consulta.

En cuanto a la utilización de Wikipedia como base de conocimiento existen varios desarrollos que tienen enfoques diferentes. Por ejemplo, en [4] se realiza un análisis de las tendencias que aparecen en la estructura de Wikipedia. Se define que cada consulta está relacionada con un grupo de documentos que son resultados correctos. En base a esto se relaciona cada consulta del conjunto con un grafo de artículos y categorías de Wikipedia que permiten recuperar los documentos correctos para esa consulta en particular. En esta línea, Azad y Deepak proponen en [9] un esquema de ponderación doble para los términos de expansión. Primero, se puntúan individualmente las relaciones (para términos extraídos de Wikipedia). Luego, se puntúa los términos de expansión seleccionados con respecto a toda la consulta utilizando la puntuación de correlación. Usando un esquema basado en tf-idf (frecuencia de término - frecuencia inversa de documento), para términos extraídos de WordNet. Otro desarrollo interesante se describe en [1] donde se construye una red de Markov utilizando el cuerpo de un texto referido a un tema, y se suman los términos que son obtenidos de Wikipedia resultando una red de Markov más grande, rica y con más relaciones, esta red es utilizada como base para sugerir los nuevos términos.

En el método propuesto en [10] se ejecuta la consulta original contra la colección de Wikipedia y se recuperan los *N* principales artículos de Wikipedia. Luego, se devuelve una página interactiva para que el usuario valore los resultados, donde los títulos de estos artículos se organizan en base a las etiquetas de categoría de Wikipedia. Una vez seleccionada al menos una etiqueta de categoría, se utiliza esta para reformular la nueva consulta. En el trabajo de Bøhn y Nørvåg [11] se utilizan los contenidos de Wikipedia para generar automáticamente un diccionario de entidades y sinónimos que se refieren a la misma entidad. Este diccionario se puede utilizar luego para la expansión de consultas. Permite clasificar a las Entidades como personas, organizaciones y empresas.

Claramente la variedad de alternativas que se presentan en los distintos trabajos indica que en un modelo de expansión debe tenerse en cuenta dos cuestiones principales: la relación y la relevancia de los términos candidatos con la clave original. En línea con estos aspectos, se presenta en la Sección siguiente la descripción del modelo propuesto.

3 Modelo Propuesto

El proceso completo de expansión se muestra en la Fig. 1. Se captura la consulta del usuario y, al mismo momento que se realiza la búsqueda, se generan términos que estén asociados a ella. Para esto, el motor de expansión analiza la consulta, extrayendo las entidades contenidas en ella y genera un grafo de conocimiento con los términos relacionados a las entidades como nodos que están relacionados (o no) entre sí dependiendo del análisis del contenido Web que los describa.

4

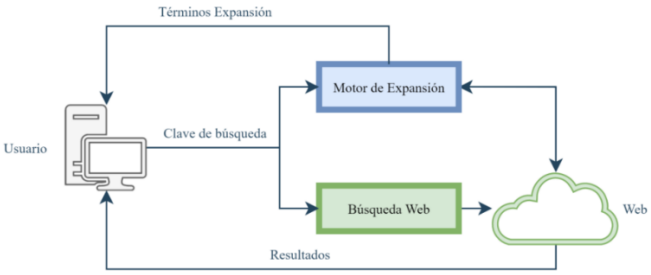

Fig. 1. Modelo de expansión de claves de búsqueda.

Una vez que se ha generado el grafo se computa el peso de cada nodo y finalmente se muestra al usuario los mejores términos encontrados.

El proceso general de expansión se muestra en el Algoritmo 1. La función `expandirConsulta()` recibe la consulta original del usuario que será analizada con el objeto de seleccionar los términos que serán agregados. En la línea 02 se define la cantidad máxima de niveles que tendrá el grafo que se genere durante el análisis. Las entidades contenidas en la consulta se extraen llamando a la función `obtenerEnt()` que explora la consulta y busca aquellas que serán la base de la expansión. Para este proceso se utiliza un extractor de entidades (línea 03). Luego se generan las tuplas (o pares) de entidades (línea 04) que serán utilizadas para construir los grafos parciales (líneas 05-07). Estos grafos parciales se combinan en un grafo resultante (línea 08) y se calculan los valores de relevancia de cada una de las entidades que lo componen (línea 09). Finalmente se devuelven los términos más relevantes encontrados (línea 10).

Algoritmo 1. Expansión de consulta.

```
01.Función expandirConsulta(consulta)
02.  limite = N
03.  entidades = obtenerEnt(consulta)
04.  tuplasEnt = generarTuplas(entidades)
05.  grafosParciales = Lista_de_grafos_vacía
06.  Para cada tupla en tuplasEnt
07.    grafosParciales = grafosParciales + generarGrafo(tupla)
08.  grafoFinal = combinarGrafos(grafosParciales)
09.  resultadosFinales = calcularResultados(grafoFinal)
10.Retornar resultadosFinales
```

Para encontrar los términos relacionados a la consulta del usuario, se construyen grafos que se expanden a medida que se analizan las entidades extraídas de la consulta. Estos grafos se construyen en base a tuplas de entidades que se combinan en pares sin repetición. La generación de cada grafo parcial comienza con un par de entidades en una tupla, las cuales se ordenan lexicográficamente y se computan para generar el primer nivel de expansión de cada grafo.

El computo de un nivel del grafo parte de una lista de tuplas para cada nivel de expansión, explorando las entidades asociadas a las que están contenidas en cada tupla. Este procesamiento se hace por cada tupla de entidades actuales, generando una lista de nuevas entidades relacionadas a las contenidas en la tupla. En base a estas nuevas listas de entidades se generan nuevas tuplas que, de manera recursiva, incorporarán

nuevos nodos al grafo. La condición para finalizar el proceso recursivo es alcanzar la cantidad máxima de niveles del grafo. Una vez que se alcanzó el último nivel de generación de cada grafo, los grafos parciales se combinan (Algoritmo 1, línea 08).

El Algoritmo 2 muestra el detalle del procesamiento de cada tupla. La idea básica es que, para las dos entidades en la tupla (`tupla[0]` y `tupla[1]`), se explora el contenido de la página de Wikipedia relacionada y se extraen las entidades desde ese recurso Web de forma separada (líneas 02 y 03). En este caso la función `obtenerEntRelac()` retorna una lista de entidades ordenada por el valor de relevancia de cada una de ellas en el documento Web analizado. En las dos listas obtenidas se buscan las entidades que se repiten (línea 04), y se genera una nueva lista que contiene las intersecciones entre los términos de la tupla. Esas intersecciones son las que se utilizan para generar los nuevos nodos del grafo que se almacena en la línea 05. En caso de que la lista no contenga las entidades originales, se agregan a fin de poder computarlas en el próximo nivel de construcción del grafo (líneas 06 - 09). Finalmente, se vuelve a construir una lista de tuplas combinando, de a pares, todas las entidades obtenidas en el análisis.

Algoritmo 2. Procesar una tupla de entidades

```
01.Función procTuplaEnt(grafo,tupla)
02.  entRelacTupla1 = obtenerEntRelac(tupla[0])
03.  entRelacTupla2 = obtenerEntRelac(tupla[1])
04.  nueListaEnt = intersección(entRelacTupla1,entRelacTupla2)
05.  almResultParc(grafo,tupla,nueListaEnt)
06.  Si tupla[0] no está en nueListaEnt entonces
07.    nueListaEnt = nueListaEnt + tupla[0]
08.  Si tupla[1] no está en nueListaEnt entonces
09.    nueListaEnt = nueListaEnt + tupla[1]
10.  nueListaDeTuplas = generarTuplas(nueListaDeEntidades)
11.Retornar nueListaDeTuplas
```

Como se mencionó anteriormente, una vez que se ha generado el grafo, se calcula la valoración de cada nodo. La idea es que cada término relacionado a un nodo del grafo sea considerado de acuerdo con su importancia dentro de ambos documentos web analizados. Para ello, cuando se generan las listas ordenadas de entidades relacionadas por cada tupla (Líneas 02 y 03 del Algoritmo 2) se captura el valor de relevancia y la posición que ocupa esa entidad en cada lista. Por cada entidad coincidente detectada a partir de las entidades de la tupla en el Algoritmo 2, se generan dos relaciones (Fig. 2) una por cada entidad de la tupla original, que se asocian con la entidad coincidente.

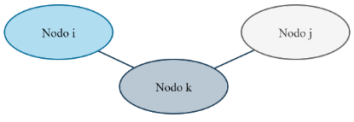

Fig. 2. Relaciones de coincidencia del nodo *k*.

La valoración de estas dos relaciones es igual, y se obtiene utilizando la Ecuación 1.

$$vRel_{ik} = vRel_{jk} = \left(relev_i \cdot \frac{1}{pos_i} \right) + \left(relev_j \cdot \frac{1}{pos_j} \right) \quad (1)$$

donde $vRel_{ik}$ y $vRel_{jk}$ son los valores de la relevancia de la relación entre los nodos *i*-*k* y *j*-*k* respectivamente, es decir, se corresponden a los valores de relevancia de dos

relaciones diferentes, pero cuyo valor es el mismo. $relev_i$ y $relev_j$ son los valores de relevancia de la entidad *k* dentro del documento Web de las entidades *i* y *j*, y pos_i y pos_j son las posiciones de las entidades *i* y *j* en las listas ordenadas devuelta por la función `obtenerEntRelac()` del Algoritmo 2, `entRelacTupla1` y `entRelacTupla2`, de las cuales se dispone al momento de hacer el cálculo. Finalmente, para computar la valoración de cada nodo, se revisan las *n* relaciones que tiene asociadas y se calcula:

$$vNodo_k = \sum_{p=1}^{n} vRel_{pk} \quad (2)$$

donde $vNodo_k$ es la valoración final del nodo *k* y *n* es la cantidad de relaciones asociadas al nodo *k*. Una vez que se calculan todas las valoraciones de los nodos en el grafo se genera la lista de entidades ordenadas por dicho valor. Aquellas que tengan mejor puntuación son las que se señalan como términos posibles para la expansión.

En comparación con otros trabajos que utilizan diferentes características de Wikipedia para la expansión de términos como en [1, 9–11], en este modelo se realiza un proceso recursivo. De esta manera a partir de términos extraídos de la consulta original se exploran los términos relacionados de Wikipedia generando los grafos de primer nivel. A diferencia de los trabajos [6, 8, 10], el modelo propuesto analiza la relación entre los documentos que devuelven los términos mediante el procesamiento de las entidades de los mismos. Estas entidades luego son ponderadas en función a la cantidad de relaciones asociadas a cada una de ellas, de manera similar a los modelos presentados en [5, 6, 9].

Resumiendo, el modelo propuesto realiza la exploración de entidades de la consulta, explora las entidades asociadas, analiza las relaciones entre ellas, las pondera y devuelve aquellas con mayor valor. Para ver el descubrimiento de conceptos relacionados, en la Sección siguiente se presentan algunos ejemplos de utilización.

4 Ejemplos de utilización

En esta sección se muestran dos ejemplos de utilización del modelo propuesto. Primero se obtienen las entidades contenidas en la consulta, a partir de las cuales se generan y agregan los grafos parciales, y se calculan las valoraciones de cada nodo. Finalmente, se muestran los términos a sugerir y sus valoraciones.

Ejemplo 1

A partir de la consulta "***Optimization of deep neural networks for natural language processing***" se obtienen las siguientes entidades iniciales utilizando la herramienta de extracción: "Natural language processing", "Neural network", "Deep learning", "language". La Fig. 3 muestra el grafo de términos obtenido.

Luego, el grafo de la Fig. 3 se utiliza para computar la valoración de cada nodo. Los resultados finales de las entidades y su valoración son: (1) "Artificial neural network" 401,56, (2) "Machine learning" 179,13, (3) "Deep learning" 155,86, (4) "Backpropagation" 110,39, (5) "Neural network" 85,63, (6) "Feedforward neural network" 79,54, (7) "Speech recognition" 75,37, (8) "Convolutional neural network" 75,08, (9) "Recurrent neural network" 74,57, (10) "Perceptron" 70,28.

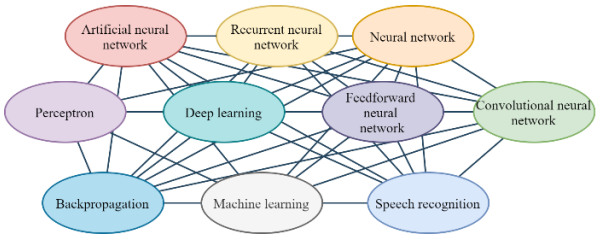

Fig. 3. Grafo final de la consulta "Optimization of deep neural networks for natural language processing".

Ejemplo 2

Se utiliza la clave de búsqueda "***Pesticide residue treatment in agriculture***", obteniendo los siguientes términos: (1) "Pesticide" 110,23, (2) "Insecticide" 70,63, (3) "Agriculture" 46,25, (4) "Silent Spring" 45,54, (5) "DDT" 39,04, (6) "Pesticide residue" 37,76, (7) "Rachel Carson" 34,33, (8) "Pest control" 33,49, (9) "Pesticide poisoning" 25,94, (10) "Toxicity" 17,64.

Para ambos ejemplos, todos los términos sugeridos tienen relación con el área de búsqueda. Se puede resaltar que el modelo encuentra tanto términos más genéricos como más específicos que los que están incluidos en la consulta. También, se obtienen términos que tienen conexión implícita con el tema de búsqueda. Tal es el caso del ejemplo 2 donde se puede observar la aparición de la entidad "Rachel Carson", quien escribió el libro "Silent Spring", donde advirtió los efectos perjudiciales sobre el medio ambiente ocasionados por el uso de pesticidas, y que también aparece como término sugerido. Este es un ejemplo de que, para una consulta sin información del contexto, el modelo retorna entidades que ayudan al usuario a buscar en alguna dirección específica.

5 Conclusiones y Trabajos Futuros

En este artículo se presenta un modelo de sugerencia de términos para expansión de claves de búsqueda. El modelo utiliza técnicas de extracción de entidades de documentos Web y encuentra relaciones entre las mismas que luego se representan en grafos. Una vez construido el grafo se calcula la valoración de cada nodo teniendo en cuenta las relaciones que tiene asociadas.

Los ejemplos ilustrativos presentados muestran que ninguno de los términos obtenidos se encuentra por fuera del ámbito de las búsquedas realizadas.

Aunque la cantidad de relaciones indica la importancia de los conceptos, el modelo considera también la calidad de esas relaciones. Es decir, al utilizar la posición que ocupa una entidad en una lista ordenada por relevancia, el cálculo de la valoración final considera los valores de importancia de las demás entidades obtenidas.

Si bien el modelo utiliza Wikipedia como repositorio, puede trabajar con cualquier otra fuente de documentos Web. Es decir, se puede utilizar en áreas de conocimiento específicas (publicaciones científicas, documentación técnica, noticias, etc.).

Actualmente, el equipo de investigación se encuentra trabajando en el diseño de mecanismos para incluir la especificidad de los términos en la ponderación de estos. Además, se está trabajando en modelos de evaluación cuantitativa que midan la calidad de los resultados devueltos tanto con la consulta original como la consulta expandida.

8

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).

Referencias

1. Gan, L., Tu, W.: Improving query expansion using wikipedia. In: Proceedings - 2014 International Conference on Management of e-Commerce and e-Government, ICMeCG 2014 (2014). https://doi.org/10.1109/ICMeCG.2014.37.
2. Bobed, C., Mena, E.: QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems. Inf. Sci. (Ny). (2016). https://doi.org/10.1016/j.ins.2015.09.013.
3. Balaneshin-Kordan, S., Kotov, A.: Sequential query expansion using concept graph. Int. Conf. Inf. Knowl. Manag. Proc. 24-28-Octo, 155–164 (2016). https://doi.org/10.1145/2983323.2983857.
4. Guisado-Gámez, J., Prat-Pérez, A.: Understanding graph structure of wikipedia for query expansion. In: 3rd International Workshop on Graph Data Management Experiences and Systems, GRADES 2015 - co-located with SIGMOD/PODS 2015 (2015). https://doi.org/10.1145/2764947.2764953.
5. Duan, Y., Shao, L., Hu, G., Zhou, Z., Zou, Q., Lin, Z.: 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). https://doi.org/10.1109/SERA.2017.7965747.
6. Zhao, L., Wu, L., Huang, X.: Using query expansion in graph-based approach for query-focused multi-document summarization. Inf. Process. Manag. 45, 35–41 (2009). https://doi.org/10.1016/j.ipm.2008.07.001.
7. Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N.: Concept-based interactive query expansion. In: International Conference on Information and Knowledge Management, Proceedings (2005). https://doi.org/10.1145/1099554.1099726.
8. Dahir, S., El Qadi, A., Bennis, H.: Query expansion based on term distribution and DBpedia features. Expert Syst. Appl. 176, 114909 (2021). https://doi.org/10.1016/j.eswa.2021.114909.
9. Azad, H.K., Deepak, A.: A new approach for query expansion using Wikipedia and WordNet. Inf. Sci. (Ny). 492, 147–163 (2019). https://doi.org/10.1016/j.ins.2019.04.019.
10. 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.
11. Bøhn, C., Nørvåg, K.: Extracting named entities and synonyms from Wikipedia. In: Proceedings - International Conference on Advanced Information Networking and Applications, AINA (2010). https://doi.org/10.1109/AINA.2010.50.