ISBN: 978-987-4998-33-0
DOI:https://doi.org/10.33414/AJEA.4.407.2019

Estrategias de Abordaje y Visualización de Grandes Datos Mediante Grafos Semánticos

Approach and Visualization Strategies of Big Data Using Semantic Graphs

Presentación: 22/11/2019

Doctorando:

Valerio Frittelli

Universidad Tecnológica Nacional – Facultad Regional Córdoba
vfrittelli@gmail.com

Director/es:

Mario José Diván

Resumen

La propuesta que aquí se realiza es diseñar y desarrollar una herramienta de medición de similitud documental, contemplando el análisis semántico y permitiendo la visualización de su esquema de relaciones, para facilitarle al usuario final la selección desde una base documental de carácter científico. Para esto, serán analizadas y comparadas diversas posibilidades existentes en cuanto al diseño de las estructuras de datos de soporte interno del motor de búsqueda (modelo vectorial, modelo probabilístico, variantes del modelo de rankeo por popularidad, o incluso variantes basadas en aplicación de meta-buscadores que procesen los resultados entregados por dos o más motores pre-existentes), a la vez que se propondrán nuevas variantes que puedan ajustarse mejor al contexto específico de esta investigación. En el área de la visualización de relaciones entre documentos, se trabajará en el diseño de un modelo de interfaz gráfica que represente a los documentos recuperados como objetos visualmente manejables, destacando también en forma visual las relaciones de similitud entre ellos.

Palabras claves: Similitud Documental, Análisis Semántico, Motores de Búsqueda, Visualización de Relaciones, Objetos Visualmente Manejables.

Abstract

The proposal made here is to design and develop a document similarity measurement tool, contemplating the semantic analysis and allowing the visualization of its relationship scheme, to facilitate the final user the selection from a scientific documentary base. For this, various existing possibilities will be analyzed and compared in terms of the design of the internal support data structures of the search engine (vector model, probabilistic model, variants of the ranking model by popularity, or even variants based on application of meta- search engines that process the results delivered by two or more pre-existing engines), while new variants will be proposed that can better fit the specific context of this research. In the area of visualization of relations between documents, work will be done on the design of a graphical interface model that represents the recovered documents as visually manageable objects, also highlighting visually the similarity relations between them.

Keywords: Documentary Similarity, Semantic Analysis, Search Engines, Relationship Visualization, Visually Manageable Objects.

Introducción

La búsqueda, recuperación y ordenamiento por relevancia de documentos cuyo contenido sea coincidente con una consulta realizada por un usuario es un problema del campo de la Recuperación de Información y tiene plena

vigencia. Esencialmente, representa el problema que deben resolver los motores de búsqueda sea que se apliquen para explorar la Web o bien para llevar adelante una búsqueda dentro de un contexto local. Existen diversos modelos matemáticos clásicos en los que se podría basar el diseño de los motores de búsqueda [1], entre los que se puede citar a los modelos clásicos como el modelo booleano, el modelo vectorial y el modelo probabilístico , y hasta llegar a uno de los más recientes y exitosos, el modelo de rankeo por popularidad (conocido como PageRank) que se utiliza en el motor Google [2].

La mayor parte de los modelos clásicos se basan en medir la similitud de dos documentos y luego la relevancia de un documento frente a una consulta, usando aproximaciones vectoriales, estadísticas y/o probabilísticas, pero sin considerar necesariamente la similitud semántica entre los contenidos de los documentos. Es decir, se aplican diversas técnicas algorítmicas para encontrar sinónimos, lematización de palabras (reducción de dos palabras con la misma raíz a una palabra común única) y cálculos de distancia de edición entre palabras (entre otras técnicas) [1] que podrían ayudar a un motor de búsqueda a recuperar documentos similares, pero el análisis del significado o semántica del texto no forma parte de los alcances del modelo.

Alternativamente, se han planteado modelos basados en el análisis semántico [3], e incluso modelos basados en grafos de relaciones semánticas para directamente plantear el diseño de motores de búsqueda indexando los documentos como partes de un gran grafo semántico [4]. El análisis semántico pretende considerar el significado del texto contenido en los documentos y eso posibilitaría que los motores de búsqueda sean capaces de obtener resultados más precisos, esto es, que sean capaces de recuperar documentos más relevantes con respecto a una consulta que lo que se obtendría utilizando otros modelos (aunque posiblemente al costo de una merma en la velocidad en la respuesta) [4]. El problema de la incorporación de técnicas de medición de similitud semántica sigue abierto en el contexto de la recuperación de información. Típicamente, una consulta se realiza en base a palabras clave que luego son chequeadas en un índice, siguiendo elementos de algún modelo vectorial, probabilístico o de popularidad, pero las relaciones de significado entre una consulta y un documento (relaciones semánticas) siguen siendo difíciles de captar y modelar. Esto es un problema porque puede afectar a la precisión de los resultados: una búsqueda que no contemple alguna forma de medir relevancia semántica produce como resultado conjuntos de documentos (normalmente muy grandes) con documentos que de alguna forma estadística se relacionan con la consulta (y entre ellos), pero posiblemente sin capturar otros documentos que no contengan las mismas palabras pero toquen los mismos temas.

Sea cual sea el modelo matemático empleado, en términos generales el proceso inicia con una consulta o requerimiento del usuario, el motor explora la base documental para extraer los documentos que mejor responden a la consulta, y se presenta al usuario el conjunto de documentos recuperados, pero ordenados en función de su relevancia. En tal sentido, es importante destacar que es el usuario quien determina y selecciona los documentos útiles a partir de aquellos sugeridos por el motor, descartando los demás. Así, la decisión final del usuario suele basarse en una exploración lineal de los primeros documentos de la lista sugerida por el motor, sin mayores herramientas de apoyo que el propio orden de presentación y la visualización por parte del motor de algún pequeño fragmento de texto o snippet (extraído del documento) que contiene las palabras de la consulta [5]. Dado este contexto, el usuario explora o analiza los documentos recuperados y selecciona alguno de ellos pero probablemente sin ir más allá de la primera página de resultados [6].

Lo anterior sugiere que en el proceso final de decisión entre los documentos obtenidos por un motor se podrían plantear mecanismos de apoyo que ayuden al usuario a hacer una selección más certera con más información. Algunos sistemas del mundo del Big Data (tales como Qlik Sense y Tableau) permiten manejar grandes volúmenes de información de distintas fuentes simplificando la interpretación del subconjunto en el cual se enfoque un usuario, presentándolo en forma de tablas, gráficos estadísticos, grafos, y modelos visuales diversos. Otros sistemas (como Gephi) directamente toman los documentos de la base documental y los presentan visualmente como nodos de un grafo en el cual los arcos representan algún tipo de relación de contenidos entre dos documentos, y a partir de ese grafo se pueden extraer subconjuntos de documentos con algún tipo de relación de proximidad.

El problema del diseño de la interfaz visual para incorporar elementos que faciliten el manejo del conjunto de documentos recuperados, sigue abierto. Se busca que el proceso de selección por parte del usuario final sea más intuitivo, con apoyo de elementos de resalte visual de relaciones, capacidad de manejar en forma visual grupos de documentos similares o la posibilidad de destacar una o más propiedades de un documento con sólo seleccionarlo, La mayor parte de los motores de búsqueda se basan en interfaces de usuario poco interactivas, de forma que la carga de la selección final de los documentos recae en la decisión personal del usuario. Pero la falta de elementos visuales que destaquen la relevancia de un documento frente a otros y/o las relaciones entre distintos documentos, hacen que el proceso de selección tienda a ser secuencial, chequeando uno por uno los documentos recuperados y en el orden en que fueron devueltos. Esto es un problema desde el punto de vista del tiempo insumido en ese proceso, y por ende en los costos. Y en esa dirección, la propuesta que aquí se realiza es diseñar y desarrollar una herramienta

de medición de similitud documental, contemplando el análisis semántico y permitiendo la visualización de su esquema de relaciones, para facilitarle al usuario final la selección desde una base documental de carácter científico.

Hipótesis y Objetivos

Considerando lo expuesto en la sección anterior, se pueden plantear las siguientes preguntas de investigación:

- ¿En qué forma podrían aplicarse técnicas de procesamiento de lenguaje natural en el planteo de modelos matemáticos de recuperación de información, que lleven al diseño de motores de búsqueda más eficientes en la medición de la relevancia de un documento frente a una consulta?
- ¿Cuáles son las variantes que pueden plantearse en el diseño de la estructura de un motor de búsqueda, de forma que se combinen técnicas clásicas existentes, con técnicas de análisis semántico, y se incorporen además ideas de otros campos (como el de la búsqueda en espacios métricos)?
- ¿En qué forma podría diseñarse la interfaz visual general de un motor de búsqueda, de forma que se rescaten a primera vista las relaciones semánticas entre los documentos recuperados, y se pueda gestionar el acceso a esos documentos (o a sus propiedades) en forma simple e intuitiva?

En este contexto, la hipótesis de trabajo es:

i. Si un motor de búsqueda aplicado en una base documental científica en idioma inglés digitalizada se plantea en base al análisis de relaciones semánticas entre esos documentos,

ii. y por otra parte se diseñan herramientas de manipulación y visualización que permitan tomar los documentos recuperados por el motor y presentarlos en la interfaz de usuario de forma de resaltar en forma gráfica las relaciones entre los documentos sugeridos,

entonces el trabajo del usuario que formule una consulta, en cuanto a la selección final de los documentos más relevantes para esa consulta, podría realizarse en forma más rápida, precisa y confiable frente al mismo tipo de consulta realizada con motores de búsqueda y herramientas de apoyo tradicionales.

Finalmente, se plantean a continuación los objetivos del trabajo:

Objetivo general: El objetivo general del presente trabajo es el abordaje de una estrategia basada en la estructura semántica de las relaciones entre documentos en inglés científicos digitalizados con un formato dado respecto a sus contenidos, para mejorar la precisión y profundidad de la búsqueda en relación a los requerimientos del usuario, facilitando la presentación, manipulación y selección de documentos en la interfaz visual de usuario.

Objetivos específicos:

1. Proponer modelos escalables de medición de similitud y detección de relaciones semánticas entre documentos en el contexto citado, con el objetivo de lograr una nueva perspectiva sobre las técnicas de rankeo y medición de distancias, en busca de mejoras de eficiencia y de presentación final de esos documentos.
2. Diseñar modelos escalables de presentación visual de representaciones abstractas de documentos y de las relaciones semánticas entre ellos, para así dar fundamento al núcleo algorítmico de un sistema gráfico de presentación de relaciones.
3. Diseñar y desarrollar la o las herramientas que integren los modelos de similitud y detección de relaciones semánticas junto con su presentación visual, a los efectos de interactuar con el usuario final y facilitar la selección de material relevante.

Metodología

El plan de trabajo supone una primera etapa de investigación general acerca del estado del arte en el campo de la recuperación de información, abarcando el estudio de modelos clásicos y una inmersión en propuestas más recientes.

Luego corresponde una exploración acerca de técnicas (clásicas y recientes) para comparación de similitud entre documentos y eventual medición de distancia y relevancia. Independientemente de la intención de producir y proponer nuevas técnicas, existen numerosos avances en este sentido que pueden adaptarse a los requerimientos de este trabajo.

Adicionalmente, al planteo de esas nuevas propuestas se le incorporará el diseño, desarrollo y prueba de las herramientas propuestas para facilitar la visualización de relaciones, lo que posiblemente insumirá gran parte del tiempo previsto en el plan.

Las distintas etapas previstas en el plan de trabajo son las siguientes:.

- Realización de cursos de posgrado.
- Revisión de bibliografía asociada
- Análisis de modelos para recuperación de información (RI).
- Análisis de algoritmos de similitud.
- Diseño de nuevas propuestas
- Análisis de resultados.
- Publicación en congresos.
- Diseño y documentación del prototipo.
- Desarrollo e implementación del prototipo.
- Escritura de la Tesis Doctoral.

Resultados

Como resultado de la investigación y el trabajo a realizar, se espera diseñar e implementar un motor de búsqueda funcional, basado en algoritmos de medición de similitud documental y las técnicas de soporte de información que sean necesarias para la inclusión de esos algoritmos, incluyendo además una interfaz visual que facilite el trabajo de selección final de los documentos recuperados.

Se pretende a partir de esto, lograr mejoras de rendimiento en diversos aspectos referidos a la eficiencia del motor de búsqueda (frente al rendimiento esperado de otros motores existentes). Los elementos en los que podría lograrse un mejor rendimiento son:

a.) La posible mejora en el tiempo de ejecución y consumo de memoria (interna y externa) a partir del diseño de técnicas de soporte de información que incluyan combinaciones de técnicas ya conocidas y/o variantes específicamente propuestas.

b.) La posible mejora en la calidad de respuesta del motor (o sea, un posible aumento en la precisión de los resultados que obtenga en cuanto a que los documentos que se recuperen para una consulta dada, sean efectivamente los más útiles), a partir de incorporar algoritmos de medición de similitud documental basados en procesamiento de lenguaje natural combinados con técnicas estadísticas típicas, de forma que el diseño del motor de búsqueda aproveche las ventajas de ambas fuentes.

c.) La posible mejora en cuanto a la usabilidad (y con ello, en la eficiencia de uso por parte del usuario final) del motor a partir del diseño de la interfaz de usuario incorporando elementos visuales para facilitar la gestión de los documentos recuperados y la selección de los más relevantes, tomando aportes e ideas de herramientas de software ya existentes en el campo de la minería de datos o de la gestión de grandes datos (*Clic Cense*, *Tablea*, *Ge phi*), pero para ser adaptadas y aplicadas en el contexto de un motor de búsqueda.

Referencias

[1] G. Navarro, “Spaces, trees, and colors: The algorithmic landscape of document retrieval on sequences,” ACM Comput. Surv., vol. 46, no. 4, Apr. 2014.

[2] V. Frittelli and M. J. Diván, “Clasificación de Modelos para Recuperación de Información,” in 6to. Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI 2018), 2018.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka, “A web search engine-based approach to measure semantic similarity between words,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 7, pp. 977–990, Jul. 2011.

[4] S. V. Figueroa and V. N. Lozano, “Organización de documentos mediante grafos de relaciones semánticas,” Intel. Artif., vol. 19, no. 57, pp. 1–21, 2016.

[5] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in Proceedings of the 15th International Conference on World Wide Web, 2006, pp. 377–386.

[6] R. Baeza-Yates and B. Ribeiro-Neto, “Modern information retrieval: the concepts and technology behind search,” Choice Rev. Online, vol. 48, no. 12, pp. 48-6950-48–6950, 2011.