Dr. Eric Sadit Téllez Ávila

Doctor en Ciencias

Investigadora o Investigador Nacional, Nivel 2 (SNII-2)
(2022-2026)

Datos de contacto:

eric.tellez@infotec.mx

Algoritmos aproximados para búsqueda en espacios métricos con bajo consumo de memoria y aplicaciones a recuperación de información y reducción de dimensión no-lineal

Datos Generales

Resumen ejecutivo

Encontrar algoritmos eficientes para construir índices para búsqueda por similitud en memoria limitada; así mismo se diseñarán algoritmos para búsqueda que funcionan en condiciones similares. Se buscará que los índices preserven características de interés para aplicaciones particulares, tal es el caso de recuperación de información y reducción de dimensión no-lineal.

Línea General y Aplicación del Conocimiento (LGAC)

9. Combinatoria, modelado y análisis de algoritmos

Palabras clave

algoritmos aproximados; búsqueda en espacios métricos; construcción grafo de k-vecinos; reducción de

Objetivo General

Desarrollar algoritmos eficientes para búsqueda por similitud en memoria limitada con aplicaciones de análisis de datos, recuperación de información y reducción de dimensionalidad no lineal.

Objetivos especificos

- Analizar algoritmos para búsqueda de k vecinos cercanos, en particular, cuidando la relación calidad, tiempo y memoria, con especial atención en la memoria disponible.. - Diseñar y evaluar algoritmos aproximados para búsqueda de k-vecinos, comparar con el estado del arte en términos de eficiencia y calidad. - Explorar nuevas aplicaciones en recuperación de información y el aprendizaje automático.

Datos del proyecto

Descripción

La búsqueda de k-vecinos cercanos sobre grandes bases de datos métricas es un problema actual que tiene diferentes aplicaciones. En particular, es de utilidad en la recuperación de información, la reducción de dimensión no lineal y aprendizaje computacional. La búsqueda de vecinos cercanos es un problema que puede ser resuelto de manera eficiente en bajas dimensiones, pero a medida que la dimensión crece, el problema se vuelve impráctico y de esta manera se ataca con diferentes aproximaciones. Uno de sus grandes problemas son los requerimientos de memoria, que pueden ser imprácticos para muchas aplicaciones. En este proyecto se buscarán algoritmos que puedan funcionar en memoria limitada, y esto puede implicar grandes colecciones o dispositivos de cómputo pequeños.

Antecedentes del problema a resolver

Los algoritmos de búsqueda son centrales en múltiples áreas de computación; en particular en la Recuperación de Información. Este problema consiste en identificar rápidamente en una colección objetos que son cercanos a un ejemplo dado. La cercanía o similitud es representada mediante una función de distancia que da una perspectiva geométrica o espacial al problema. Una de las operaciones más importantes en la búsqueda por similitud es la recuperación de k vecinos cercanos que consiste en encontrar los k elementos más cercanos de una base de datos a una consulta dada. El problema consiste en preprocesar una colección de datos, bajo alguna función de semejanza, para que la identificación de los objetos cercanos a una consulta pueda hacerse en tiempo que no dependa linealmente del tamaño de la colección. En otras palabras, que no sea necesario revisar todos los elementos de la base de datos para responder la consulta. La búsqueda de los k-vecinos cercanos tiene aplicaciones en diferentes áreas, como parte operativa de algoritmos de agrupamiento [@SS2021] o como parte de la aceleración de dichos algoritmos [@YCC2020; @SKL2011]. En procesamiento de lenguaje natural, la técnica es usada para diferentes objetivos, tal es el caso de recuperación de argumentos [@WPAA2017], descubrimiento de estructuras semánticas [@DS2020], búsqueda semántica [@SPA2019; @FH2017], entre otras. En problemas de clasificación, el método de k vecinos cercanos ha sido usado como uno de los métodos más simples y populares, y a la vez, efectivo [@GCVR2018; @OTGMM2020]. El cálculo de grafos de todos los vecinos cercanos es uno de los componentes de las técnicas de reducción de dimensión no-lineales, como UMAP [@MHM2018], TriMap [@AW2019], o t-SNE [@VMH2018], además de las alternativas clásicas [@LV2017]. En particular, calcular todos los vecinos cercanos puede ser una tarea desafiante para bases de datos grandes o de alta dimensión. El proceso de reducción de dimensión es usado para incrementar el desempeño de algoritmos de aprendizaje computacional, así como en el proceso de análisis de los datos. Las dimensiones 2 y 3 nos permiten visualizar información de manera efectiva, y por tanto, una proyección fiel en baja dimensión siempre será de utilidad en el proceso de análisis de la información. Una de las partes más costosas del proceso de reducción de dimensión no-lineal es la búsqueda de k vecinos cercanos, los algoritmos eficientes de búsqueda de vecinos son fundamentales para obtener en tiempos prácticos información de valor. Así mismo, asegurar la calidad de los vecinos es importante para confiar en las proyecciones que se obtienen. Finalmente, es necesario mencionar que aunque el tiempo de construcción y el tiempo de búsqueda es determinante para la selección de un índice, la memoria disponible es uno de los limitantes más fuertes para el uso de índices de búsqueda por similitud. Este proyecto está dedicado a la creación de índices para búsqueda en condiciones de memoria limitada, cuidando que se preserve la velocidad y la calidad de los resultados para diferentes aplicaciones.

Justificación y pertinencia

En un índice para búsqueda por similitud el tiempo de construcción y el tiempo de búsqueda es determinante, sin embargo, para grandes colecciones y alta dimensionalidad, la memoria disponible se convierte en una de las mayores limitantes. Por ejemplo, los embeddings vectoriales que producen los grandes modelos de lenguaje y visión (basados en deeplearning) utilizan de cientos a miles de componentes, haciendo que la memoria sea un recurso preciado. Este proyecto está dedicado a la creación de índices para búsqueda en condiciones de memoria limitada.

Metas

- Desarrollar algoritmos para construcción eficiente de índices en memoria limitada. - Desarrollar algoritmos de búsqueda que funcionen en memoria limitada. - Explorar y crear demostraciones de aplicaciones

Metodologías

La naturaleza del proyecto es la exploración de algoritmos para la construcción de índices para espacios métricos y de algoritmos de búsqueda de k vecinos, así como diversas aplicaciones. Se comparará la calidad, velocidad y la memoria necesaria por los algoritmos que se generarán y los algoritmos tradicionales. Se tendrá especial atención en la memoria. En general, se seguirá la siguiente metodología para los objetivos propuestos: -Revisión del estado del arte. -Propuesta de mejora a los algoritmos establecidos. -Prueba y comparación contra las alternativas anteriores sobre conjuntos de datos estándar en la literatura. -Validación, descarte o ajuste a los algoritmos o técnicas desarrolladas según los resultados. -Escritura de artículos.

Resultados esperados

-Creación de algoritmos competitivos para construcción de índices que funcionen en memoria limitada. -Creación de algoritmos competitivos para búsqueda en memoria limitada. -Implementación de código abierto en biblioteca de los algoritmos. -Publicaciones científicas y de divulgación. -Desarrollo de demostraciones susceptibles de transformarse en desarrollos tecnológicos.

Cronograma de trabajo

#	Entregable(s) comprometido(s)	Fecha inicio	Fecha fin
1	**Artículo de investigación en revista especializada**	01/08/2024	31/12/2024
2	**Impartición de docencia.**	15/01/2024	07/06/2024
3	**Impartición de docencia.**	02/08/2024	20/12/2024
4	**Impartición de docencia.**	02/08/2024	20/12/2024
5	**Participación en comité tutoral, dirección o codirección de trabajo de titulación.**	15/01/2024	07/06/2024
6	**Participación en comité tutoral, dirección o codirección de trabajo de titulación.**	15/01/2024	07/06/2024
7	**Participación en comité tutoral, dirección o codirección de trabajo de titulación.**	02/08/2024	20/12/2024

8	Participación en comité tutoral, dirección o codirección de trabajo de titulación.	02/08/2024	20/12/2024
9	Participación en la creación o diseño de Planes de estudio.	15/02/2024	12/03/2024
10	Participación en la creación o diseño de Planes de estudio.	16/05/2024	01/08/2024
11	Divulgación	01/11/2024	01/11/2024

Bibliografía relevante

[@AFN2004] Alber, J., Fellows, M. R., & Niedermeier, R. (2004). Polynomial-time data reduction for dominating set. Journal of the ACM (JACM), 51(3), 363-384. [@AW2019] Amid, E., & Warmuth, M. K. (2019). TriMap: Large-scale dimensionality reduction using triplets. arXiv preprint arXiv:1910.00204. [@CNBYM2001] Chávez, E., Navarro, G., Baeza-Yates, R., & Marroquín, J. L. (2001). Searching in metric spaces. ACM computing surveys (CSUR), 33(3), 273-321. [@CT2010] Chávez, E., & Tellez, E. S. (2010, September). Navigating k-nearest neighbor graphs to solve nearest neighbor searches. In Mexican Conference on Pattern Recognition (pp. 270-280). Springer, Berlin, Heidelberg. [@DS2020] DS, D. (2020). A simple solution for the taxonomy enrichment task: Discovering hypernyms using nearest neighbor search. [@FH2017] Faessler, E., & Hahn, U. (2017, July). Semedico: a comprehensive semantic search engine for the life sciences. In Proceedings of ACL 2017, System Demonstrations (pp. 91-96). [@GCVR2018] Gallego, A. J., Calvo-Zaragoza, J., Valero-Mas, J. J., & Rico-Juan, J. R. (2018). Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation. Pattern Recognition, 74, 531-543. [@GLS2008] Goyal, N., Lifshits, Y., & Schütze, H. (2008, February). Disorder inequality: a combinatorial approach to nearest neighbor search. In Proceedings of the 2008 international conference on web search and data mining (pp. 25-32). [@HN2014] Houle, M. E., & Nett, M. (2014). Rank-based similarity search: Reducing the dimensional dependence. IEEE transactions on pattern analysis and machine intelligence, 37(1), 136-150. [@Hetland2020] Hetland, M. L. (2020, September). Optimal Metric Search Is Equivalent to the Minimum Dominating Set Problem. In International Conference on Similarity Search and Applications (pp. 111-125). Springer, Cham. [@LV2017] Lee, J. A., & Verleysen, M. (2007). Nonlinear dimensionality reduction (Vol. 1). New York: Springer. [@MHM2018] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv: 1802.03426. [@MPLK2014] Malkov, Y., Ponomarenko, A., Logvinov, A., & Krylov, V. (2014). Approximate nearest neighbor algorithm based on navigable small world graphs. Information Systems, 45, 61-68. [@MYD2018] Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence, 42(4), 824-836. [@OTGMM2020] Ortiz-Bejar, J., Téllez, E. S., Graff, M., Moctezuma, D., & Miranda-Jiménez, S. (2020). Improving k Nearest Neighbors and Naïve Bayes Classifiers Through Space Transformations and Model Selection. IEEE Access, 8, 221669-221688. [@RCGT2015] Ruiz, G., Chávez, E., Graff, M., & Téllez, E. S. (2015, October). Finding near neighbors through local search. In International Conference on Similarity Search and Applications (pp. 103-109). Springer, Cham. [@RUB2018] Rubinstein, A. (2018, June). Hardness of approximate nearest neighbor search. In Proceedings of the 50th annual ACM SIGACT symposium on theory of computing (pp. 1260-1268). [@SKL2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830. [@SPA2019] Soto, A. J., Przybyła, P., & Ananiadou, S. (2019). Thalia: semantic search engine for biomedical abstracts. Bioinformatics, 35(10), 1799-1801. [@SS2021] Sharma, K. K., & Seal, A. (2021). Spectral embedded generalized mean based k-nearest neighbors clustering with S-distance. Expert Systems with Applications, 169, 114326. [@TR2022] Tellez, E. S., & Ruiz, G. (2022). Similarity search on neighbor's graphs with automatic Pareto optimal performance and minimum expected quality setups based on hyperparameter optimization. arXiv preprint arXiv:2201.07917. [@TRCG2021] Tellez, E. S., Ruiz, G., Chavez, E., & Graff, M. (2021). A scalable solution to the nearest neighbor search problem through local-search methods on neighbor graphs. Pattern Analysis and Applications, 24(2), 763-777. [@VMH2018] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11). [@WPAA2017] Wachsmuth, H., Potthast, M., Al Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., ... & Stein, B. (2017, September). Building an argument search engine for the web. In Proceedings of the 4th Workshop on Argument Mining (pp. 49-59). [@YCC2020] Yu, Q., Chen, K. H., & Chen, J. J. (2020, September). Using a set of triangle inequalities to accelerate k-means clustering. In International Conference on Similarity Search and Applications (pp. 297-311). Springer, Cham.