UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Representación compacta de series de tiempo astronómicas

PROPUESTA DE TEMA DE MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

Benjamín Guerrero Caro

PROFESOR GUÍA:
Gonzalo Navarro
PROFESOR GUÍA 2:
Francisco Förster

SANTIAGO DE CHILE
2021

1. Introducción

A lo largo de la historia, la astronomía ha sido un tema de estudio interesante para la humanidad. Esta disciplina se ha visto afectada por los distintos avances tecnológicos y científicos, los cuales han ido aportando nuevas herramientas y enfoques desde los cuales investigar el cielo y el universo.

El exponencial avance tecnológico de los telescopios y el aumento de la diversidad de éstos ha revolucionado la forma en que se estudia la astronomía. Estos nuevos telescopios generan datos con mayor velocidad que antes, tomando en cuenta nuevas características y obteniendo como resultado un gran ecosistema de información que pretende aportar a comprender de forma más profunda nuestro universo. Todo esto trae consigo una mayor responsabilidad en torno a la agregación, procesamiento y clasificación de los datos, lo que es un rol que asumieron los agentes de alerta astronómica (astronomical alert brokers). Su principal función es entregar de manera rápida una clasificación de los distintos eventos observados y disponibilizar los datos para la comunidad científica.

En este contexto es que en Chile nace ALeRCE (Automatic Learning for the Rapid Clasification of Events), que tiene la intención de ser un agente de alertas (alert broker) que intermedie entre la astronomía y la ciencia de datos. ALeRCE está compuesto por un grupo interdisciplinario de científicos de distintas universidades e institutos de investigación que trabajan en colaboración con investigadores de universidades de otros países. [3]

ALeRCE trabaja con grandes cantidades de datos en forma de series de tiempo con distintos canales de frecuencia provenientes de las observaciones obtenidas por los telescopios. Se aplica una clasificación sobre los datos y luego se disponibilizan. Sin embargo, el gran tamaño de estas series de tiempo produce que las consultas que se deseen realizar sobre las mismas sean de una complejidad considerable y por consecuencia que el tiempo de demora de éstas sea significativo. [3]

El problema a abordar en esta memoria es encontrar una forma de optimizar estas consultas para que se puedan realizar de una manera eficiente. Para esto se propone representar las series de tiempo a través de una estructura de datos compacta, que entregue los resultados esperados de una manera eficiente.

2. Situación Actual

Sobre las series de tiempo

Cada vez existe más presencia de series de tiempo en distintas disciplinas, tales como economía, ciencias e incluso en videojuegos. Esto se ha visto acompañado por un creciente número de investigaciones relacionadas con este tipo de datos y específicamente, con representaciones compactas para series de tiempo.

Una serie de tiempo es una colección de datos que se observaron en un momento respectivo, siendo de esta forma ordenados cronológicamente. Las series de tiempo pueden estar separadas por intervalos de tiempo iguales o variables. Cada observación puede tener una o muchas características o atributos, siendo esto último una serie de tiempo multidimensional.

Como ejemplos de representaciones compactas sobre series de tiempo existen estudios que buscan resolver el problema de la multidimensionalidad para simplificar el trabajo sobre las series de tiempo. También se usan índices de mapas de bits (Bitmap indexes) para optimizar las consultas de rangos en series de tiempo. En ambos casos se puede ver como las representaciones compactas resuelven problemas particulares sobre las series de tiempo, lo que significa que según lo que se quiera optimizar o resolver se han usado ciertas representaciones y estructuras de datos. [4]

Wavelet Trees

Es una estructura de datos que se puede definir como un árbol binario, el cual se aplica sobre strings para dividirlos de una manera balanceada. Para esto codifica el string original en forma de bits según una partición binaria y jerárquica previamente definida del alfabeto (es decir, los elementos del primer subconjunto se representan como 0's y los del segundo subconjunto como 1's) y de esta forma en el primer nivel del árbol se divide el string en dos partes de tamaños similares. Esto se aplica se forma recursiva hasta que se obtienen las hojas del árbol, que corresponderán respectivamente a cada elemento del alfabeto. [2]

Sea un Wavelet Tree que almacena una secuencia $S[1, n]$ de elementos. Este árbol poseerá 3 operaciones básicas, que corresponden a: [1]

- *access*: Su notación corresponde a $S[i]$ y retorna el elemento que se encuentra en la posición i-esima del string S.
- *rank*: Se expresa de la forma $rank_c(S, i)$ y retorna el número de ocurrencias del símbolo c en el string $S[1, i]$
- *select*: Se expresa de la forma $select_c(S, i)$ y retorna la i-esima ocurrencia del símbolo c en S.

Los Wavelet Trees son una estructura de datos muy versátil y que puede funcionar en múltiples situaciones más allá de solamente representar de una manera compacta un string. Cada año aparecen nuevas soluciones basadas en Wavelet Trees, desde secuencias numéricas hasta grafos, esta estructura de datos ha demostrado ser útil en cada uno de esos casos. [2]

3. Objetivos

Objetivo General

Desarrollar una representación compacta de series de tiempo y aplicarla sobre una base de datos para que responda eficientemente las consultas de interés. Esta solución debe ser utilizada mediante una API que se pueda añadir de forma modular al ecosistema de software actual de ALeRCE.

Objetivos Específicos

1. Definir las consultas que se van a resolver a través de la representación compacta. Esto implica que las consultas deben ser ampliamente usadas por ALeRCE.
2. Modelar la nueva estructura de datos a usar para representar las series de tiempo astronómicas. Se considerará el Wavelet Tree como una alternativa promisoria para implementar la estructura.
3. Diseñar e implementar un algoritmo que reciba series de tiempo y obtenga su representación a través de la nueva estructura de datos.
4. Diseñar e implementar los algoritmos para resolver las consultas elegidas sobre la representación elegida. Para esto se usarán las operaciones de la estructura de datos a implementar.
5. Diseñar e implementar una API para poder aplicar la representación compacta sobre la base de datos de ALeRCE. Esto permitirá incluir la solución implementada al software actual de ALeRCE.

Evaluación

El trabajo será evaluado comparando los resultados entregados por la solución implementada con los resultados esperados. Para esto primero se realizarán experimentos sobre una réplica de la base de datos de ALeRCE para evaluar la correctitud de los resultados y el espacio ocupado por la implementación. Luego se harán pruebas sobre la API ya vinculada con el software de ALeRCE, de forma similar a lo anteriormente mencionado.

4. Solución Propuesta

Se propone construir un Wavelet Tree para representar las series de tiempo. Esta estructura de datos ofrece una gran variedad de operaciones que, en un primer análisis, la hace promisoria para implementar varias operaciones de interés que se han considerado.

Se escogerán consultas que tengan un alto uso y que sean de interés por parte de ALeRCE para homologarlas usando las operaciones de los Wavelet Trees. Algunos ejemplos de consultas que pueden ser interesantes son:

- Buscar todas las series de tiempo en un rango de valores dados.
- Buscar todas las series de tiempo en un rango de cuantiles dados.

Sea un Wavelet Tree que almacena una secuencia S y el alfabeto de símbolos $[1, \sigma]$, se definen las siguientes operaciones [1] que se consideran útiles pensando en consultas con rangos dados como las anteriormente presentadas:

- $range\_quantile(S, i, j, k)$: Retorna el k-ésimo valor más pequeño en $S[i, j]$.
- $range\_next\_value(S, i, j, x)$: Retorna el menor $S[r] \geq x$ tal que $i \leq r \leq j$.

donde ambas operaciones se resuelven en un tiempo $O(log\ \sigma)$, con σ . Existen más operaciones tales como $range_i ntersect(S, i_1, j1, ..., i_k, j_k)$ que retorna los distintos valores comunes en $S[i_1, j_1]$, ..., $S[i_k], j_k$.

Se usará el lenguaje C++ debido a su eficiencia de ejecución y a que la librería SDSL - Succinct Data Structure Library en donde se encuentra la implementación del Wavelet Tree con todas sus operaciones, está implementada en este lenguaje.

Se diseñará un algoritmo que use como input series de tiempo de una base de datos que se extraerá desde la información de ALeRCE y obtenga de esta forma el Wavelet Tree diseñado. Junto con lo anterior se implementarán las consultas escogidas usando las operaciones asociadas al Wavelet Tree.

Se implementará una API que se conecte con la infraestructura de software existente en ALeRCE y que tenga acceso a la base de datos en tiempo real, para que se puedan realizar consultas con la información actualizada. Para esto se sabe que en ALeRCE se cuenta con acceso a recursos de AWS (Amazon Web Services) por lo que se planea usar esto para facilitar el trabajo.

Referencias

[1] Gagie, Travis, Gonzalo Navarro y Simon J. Puglisi: *New algorithms on wavelet trees and applications to information retrieval*. Theor. Comput. Sci., 426:25–41, 2012. https://doi.org/10.1016/j.tcs.2011.12.002.

[2] Navarro, Gonzalo: *Wavelet trees for all*. J. Discrete Algorithms, 25:2–20, 2014. https://doi.org/10.1016/j.jda.2013.07.004.

[3] Sánchez-Sáez, P., I. Reyes, C. Valenzuela, F. Förster, S. Eyheramendy, F. Elorrieta, F. E. Bauer, G. Cabrera-Vives, P. A. Estévez, M. Catelan y et al.: *Alert Classification for the ALeRCE Broker System: The Light Curve Classifier*. The Astronomical Journal, 161(3):141, 2021.

[4] Zacharatou, Eleni Tzirita: *Compact Representations and Indexing for Time Series Data*. 2014. https://scholar.google.com/citations?view_op=view_citation&hl=sv&user=NOwiCrQAAAAJ&citation_for_view=NOwiCrQAAAAJ:Tyk-4Ss8FVUC.