Parallel Spatio-textual Similarity Join with Spark

Saeed Shafiee, Jesus Alfonso Pereyra Duarte, Scott Wallace, Suprio Ray
Faculty of Computer Science, University of New Brunswick, Fredericton, New Brunswick, Canada

Introduction

Given a collection of objects that carry both spatial and textual information, a spatio-textual similarity join retrieves the pairs of objects that are spatially close and textually similar [1]. For instance, consider a social network with spatially and textually tagged persons.

- Friendship recommendation according to spatially being close and profiles overlap.

Figure 1. [4]

Problem Definition

According to [1], defines a spatio-textual object x as a triple $t(x.id, x.loc, x.text)$, modeling the identity, the location, and the textual description of x, respectively. The entry $x.loc$ takes values from the two-dimensional geographical space, while $x.text$ is a set of terms drawn from a finite global dictionary $T = \{t_1, t_2, \ldots t_n\}$. For every pair of spatio-textual objects x and y, [1] defines their spatial distance, $dist_l(x, y)$, with respect to $x.loc$ and $y.loc$, and their textual similarity, $sim_t(x, y)$, as the set similarity between sets $x.text$ and $y.text$, quantified with measures as (weighted) overlap, Jaccard or cosine similarity. It assumes that the spatial distance of objects x and y is the Euclidean distance of their locations, $dist_l(x, y) = dis_l(x.loc, y.loc)$ and that their textual similarity equals the Jaccard similarity $sim_t(x, y) = \frac{|x.text \cap y.text|}{|x.text \cup y.text|}$ [1].

Given a collection of spatio-textual objects R, the spatio-textual similarity join (ST-SJOIN) identifies pairs of objects in R that are both spatially close and textually similar. Formally, given a spatial distance threshold ε and textual similarity threshold θ, $ST\text{-}SJOIN(R, \varepsilon, \theta)$ retrieves all pairs (x, y) with $x, y \in R$, such that $dist_l(x, y) \leq \varepsilon$ and $sim_t(x, y) \geq \theta$ *[1]*.

Approach

- Dealing with the spatio-textual similarity join efficiently when the data size is large.
- Improving on an existing implementation of parallel spatio-textual similarity join as proposed by the Zhang et al. [2] using **Apache Spark™**.

Apache Spark™

MapReduce in Hadoop requires Low-level programming and manual optimization by the user to achieve higher performance [5]; however, **Apache Spark™** has been proposed as a general-purpose cluster computing engine with

Figure 2. Spark Runs Everywhere [5]

- APIs in Scala, Java and Python [3]
- Libraries for streaming, graph processing and machine learning [3]
- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk[5] [3].
- Accessing diverse data sources including HDFS, Cassandra, HBase, and S3 [3].
- Running in standalone cluster mode, on EC2, on Hadoop YARN, or on Apache Mesos [3].

Dataset introduction

Two collections of POIs and business listings for the state of California, USA and Australia, respectively, based on the SimpleGeo Places dataset[3] [1].

- POI-USCA (1,511,837 objects with a dictionary of 16,048 terms) [1].
- POI-AU (696,212 objects and 2,633 terms) [1].

References

1. Bouros, Panagiotis, Shen Ge, and Nikos Mamoulis. "Spatio-textual similarity joins." *Proceedings of the VLDB Endowment* 6.1 (2012): 1-12.
2. Zhang, Yu, Youzhong Ma, and Xiaofeng Meng. "Efficient Spatio-textual Similarity Join Using MapReduce." *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*. Vol. 1. IEEE, 2014.
3. http://spark.apache.org/

3 https://simplegeo.com/products/places/
4 http://www.investopedia.com/articles/markets/100215/twitter-vs-facebook-vs-instagram-who-target-audience.asp