Big Data Research ●●● (●●●●) ●●●–●●●

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

ELAN: An Efficient Location-Aware Analytics System ☆

Yaxiao Liu [a,*], Henan Wang [a], Guoliang Li [a], Junyang Gao [a], Huiqi Hu [a], Wen-Syan Li [b]

[a] *Department of Computer Science, Tsinghua University, Beijing, China*
[b] *SAP Labs, Shanghai, China*

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Location-aware analytics
Spatial index
Incremental algorithm
Big data

ABSTRACT

We demonstrate an Efficient Location-Aware aNalytics system (ELAN), aiming to provide users with location-aware data analytics services. For each user-selected spatial region, ELAN can instantly identify the most important functionality features of the region (e.g., business zones and residential areas) by efficiently analyzing the user-generated content (UGC) within the region. For each feature, ELAN can efficiently calculate the spatial boundary of the functional zone (denoted by a convex hull) in order to help users better understand the feature and furthermore we can identify the influential range of a certain feature. ELAN has many real-world applications, e.g., choosing business locations and popular regions discovery. There are two main challenges in designing a location-aware data analytics system. The first is to achieve high performance, as the region may contain a large amount of location-based UGC data. The second is to support continuous queries as users may continuously change the region by zooming in, zooming out, and panning the map. To address these challenges, we propose effective spatio-textual indexes and efficient incremental algorithms to support instant location-aware data analytics. We have implemented and deployed a system, which has been commonly used and widely accepted.

1. Introduction

Nowadays, the rapid growth of Mobile Internet has witnessed the widespread use of location-aware services, e.g., Google map search and Yelp local search. Meanwhile, with the popularity of location-based social networks (LBSN), e.g., Foursquare and Qzone, users are generating more and more user-generated content (UGC) with location information, e.g., check-ins and reviews at venues. Obviously the location-based UGC data plays an important role in location-based services, and we can take advantage of them to provide users with various location-aware analytics services. Although traditional applications can improve the functionality of the online map, such as region-based keyword queries [1] and finding places with specific category within user selected regions [2], they cannot fully support online location-aware data analytics. Obviously we can instantly identify the most important functionality features of users' interested region, e.g., business zones and residential areas, to help users better understand the selected region, by efficiently analyzing the UGC data. To this end, in this paper we demonstrate an Efficient Location-Aware aNalytics system (called ELAN), which can instantly analyze location-based UGC data.[1]

Fig. 1 shows a screenshot of our ELAN system. Given a user-selected spatial region, ELAN efficiently analyzes the location-based UGC data in this region and shows the important functionality features (e.g., *attorney* and *restaurants*) to the user using a word-cloud-style interface. For each feature, e.g., *attorney*, the system calculates the spatial boundary of the functional zone to the feature, as illustrated in Fig. 1. ELAN provides a new strategy to discover the map – it is not only a simple tool to find location-based information, but also a way to explore the aggregated features of different regions on the map. ELAN has many real-world applications. (1) Map Navigation: Suppose a user is visiting a strange place and wants to know the most important functionalities in the region. Obviously it is difficult for her to get such information using traditional map applications. Alternatively, ELAN can show the functionalities on the map, and the user can browse and compare different regions, and choose the most interesting region. (2) Business Location Selection: If a businessman wants to open a new supermarket in a region, she can use our ELAN system to browse the map and check whether there are competitive supermarkets in the region and whether there are potential customers close to the

☆ This article belongs to Analytics and Applications.

* Corresponding author.
E-mail addresses: liuyaxiao12@mails.tsinghua.edu.cn (Y. Liu), whn13@mails.tsinghua.edu.cn (H. Wang), liguoliang@tsinghua.edu.cn (G. Li), gaojy11@mails.tsinghua.edu.cn (J. Gao), hhq11@mails.tsinghua.edu.cn (H. Hu), wen-syan.li@sap.com (W.-S. Li).

[1] Our system is available at http://tsingnus.cs.tsinghua.edu.cn/elan/ (Beijing) and http://tsingnus.cs.tsinghua.edu.cn/elan/usa.html (USA).

http://dx.doi.org/10.1016/j.bdr.2016.08.001
2214-5796/© 2016 Published by Elsevier Inc.

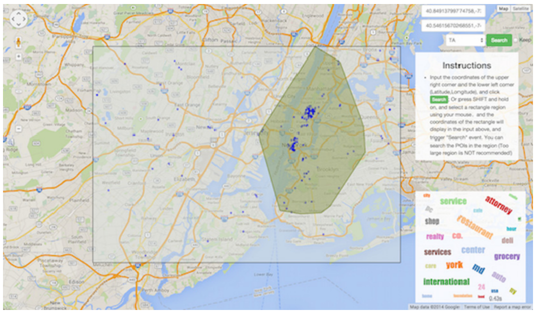

Fig. 1. A screenshot of ELAN.

region. (3) Public Facility Planning: ELAN can also provide suggestions for the government to do public facility planning. Suppose the government wants to build a sport center. The government can use ELAN to find the regions that are short of sport centers and have a large population.

It is rather challenging to design an efficient location-aware analytics system because (1) the user-selected region can be rather large and there may be a large amount of location-aware UGC data in the region; and (2) the user may continuously change the region (e.g., zooming in, zooming out, or panning the map) to make a comparison between different regions to find which region is more interesting. To address these two challenges, we propose effective index structures and efficient incremental algorithms to support instant location-aware analytics. We extend R-Tree [3] by incorporating textual information into the R-Tree nodes and propose KR-Tree. Compared with existing spatio-textual indexes, the objective of KR-Tree is to support location-aware analytics instead of location-based search, and KR-Tree only maintains some statistics spatio-textual information and has much smaller index sizes than existing spatio-textual indexes. When analyzing the features of each region, we calculate the importance of the features using the KR-Tree index instead of extracting all UGC data in the region, which can filter a large amount of irrelevant data. To support continuous queries, we propose efficient incremental algorithms that avoid many unnecessary computations. To demonstrate the effectiveness of our algorithms, we compare with the baseline algorithms, and the experimental results show that our method significantly outperforms the baseline approaches.

We have implemented and deployed a system based on some real-world datasets, which has been commonly used and widely accepted.

2. System overview

In this section, we first define the location-aware analytics query and a related problem – functional region analyzing, and then introduce the system architecture.

2.1. Location-aware analytics query

The underlying data in our system is location-based UGC data, which includes both textual descriptions and geographical locations, e.g., check-ins, points of interest (POIs), and user reviews at venues. Given a query region, the location-aware analytics query aims to identify the top-k important keywords sorted by keyword score (e.g., keyword frequency or inverse document frequency) from the UGC data in the region. Moreover, for each user selected keyword, we analyze all locations related to the keyword, and calculate the spatial boundary of the functional zone (denoted by the convex hull). We call this functionality *functional zone visualization*, which can help users better understand a region and a feature, and furthermore can be used to identify the influential range of a certain keyword or POI. The location-aware analytics query and functional zone visualization are two parts of the location-aware analytics system, and the system can extend its functions based on KR-Tree.

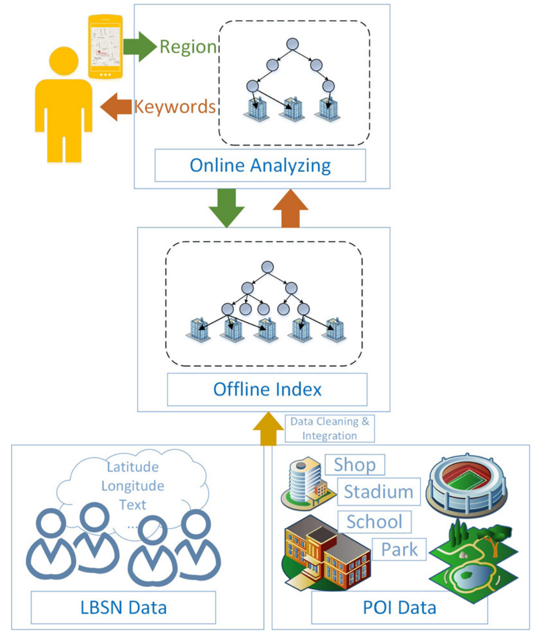

Fig. 2. Architecture.

2.2. System architecture

Our system is composed of four main components – data cleaning and integration, offline index, online analytics and user interface, as shown in Fig. 2.

Data cleaning and integration. ELAN collects millions of real world POI data, and millions of user check-ins from the LBSNs (Tencent and Foursquare), cleans and integrates the unstructured data to structured data, which mainly includes locations and textual descriptions. For POI data, the text information is the description of the POI, e.g. name and category, and for user check-ins, it is the keywords of the check-ins.

Offline index. We propose KR-Tree as the offline index structure, which organizes locations using the R-Tree. In the internal and leaf nodes, we create textual indexes for location-aware UGC data in the region, which maintain the keyword statistics information. For each keyword, we keep its importance (or score) in the region. When analyzing the features of a region, we calculate the importance of the features using the index instead of the keywords of all locations. For each keyword, we also keep a list of data objects in the region that contain the keyword. When searching for locations related to a keyword, we both check the spatial information and the keyword, thus can filter irrelevant nodes in the KR-Tree, which can significantly improve the efficiency. KR-Tree shares the similar ideas with some other extended R-Tree structures, such as BR-Tree and IR-Tree in [4,5], but KR-Tree is designed to not only search objects with textual descriptions like other structures, but also efficiently and incrementally analyze the keywords of a region without retrieving all objects (see Sections 3.2 and 3.3).

Online analytics. The main task of answering a location-aware analytics query is to analyze the importance of the keywords appearing in the text information of the user-selected region. The top-k keywords are displayed to the user. In addition to the basic algorithm (see Section 3.1), we also propose an improved algorithm (see Section 3.2) based on the KR-Tree; and for the continuous queries, we propose an incremental algorithm based on

Please cite this article in press as: Y. Liu et al., ELAN: An Efficient Location-Aware Analytics System, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2016.08.001