PERGAMON

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Information Systems 28 (2003) 929–948

www.elsevier.com/locate/infosys

Efficient data mining for calling path patterns in GSM networks☆

Anthony J.T. Lee*, Yao-Te Wang

Department of Information Management, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 106, Taiwan, ROC

Received 7 December 2001; received in revised form 30 September 2002; accepted 5 December 2002

Abstract

In this paper, we explore a new data mining capability that involves mining calling path patterns in global system for mobile communication (GSM) networks. Our proposed method consists of two phases. First, we devise a data structure to convert the original calling paths in the log file into a frequent calling path graph. Second, we design an algorithm to mine the calling path patterns from the frequent calling path graph obtained. By using the frequent calling path graph to mine the calling path patterns, our proposed algorithm does not generate unnecessary candidate patterns and requires less database scans. If the corresponding calling path graph of the GSM network can be fitted in the main memory, our proposed algorithm scans the database only once. Otherwise, the cellular structure of the GSM network is divided into several partitions so that the corresponding calling path sub-graph of each partition can be fitted in the main memory. The number of database scans for this case is equal to the number of partitioned sub-graphs. Therefore, our proposed algorithm is more efficient than the PrefixSpan and a priori-like approaches. The experimental results show that our proposed algorithm outperforms the a priori-like and PrefixSpan approaches by several orders of magnitude.

Keywords: Data mining, Sequential pattern, Calling path pattern, GSM network

1. Introduction

With the increasing use of computing for various applications, the importance of mining knowledge from large databases is growing at a rapid pace recently. There is a large amount of valuable information embedded in databases or data warehouses which is useful for analyzing customer's buying behavior and thus improving the business decisions.

Data mining is an application-specific issue and various mining techniques have been developed to solve different application problems, such as mining association rules [1–12], classification [13–16], clustering [17–22], sequential patterns [23–26], partial periodic patterns [27], and path traversal patterns in World Wide Web [28].

To the best of our knowledge, there are no data mining techniques specially designed to analyze the sequential patterns of users' calling paths in a

☆Recommended by Dr. Nick Koudas, Area Editor.

*Corresponding author. Tel.: +886-2-2363-0231x2978; fax: +886-2-2362-1327.

E-mail address: jtlee@im.ntu.edu.tw (A.J.T. Lee).

0306-4379/03/$ - see front matter © 2002 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4379(02)00112-6

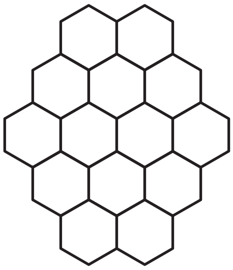

Fig. 1. Cellular structure of a GSM network.

TID	Calling path
T100	*a b c d e f*
T200	*a b c d e f g h*
T300	*b c d e*
T400	*d e f g h i j k l*
T500	*e f g h i j*
T600	*g h i j k*

Fig. 2. The database of calling path patterns.

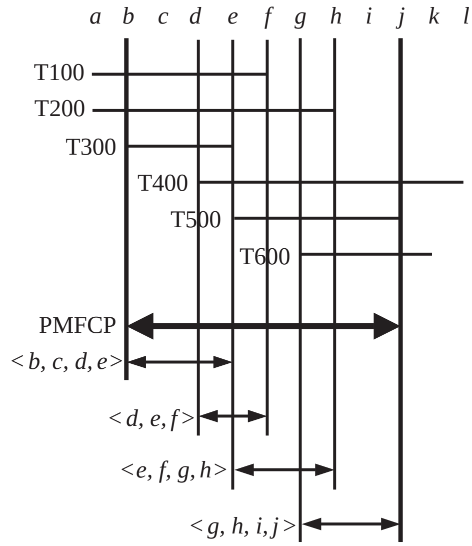

Fig. 3. The calling path patterns represented in line segments.

global system for mobile communication (GSM) network, and we believe that it is an interesting issue especially on providing mobile broadband services. A GSM network is based on the cellular radio technology [29]. A particularly characteristic feature of cellular radio is that each hexagonal (six-sided polygon) cell (the radius of a base station) is surrounded by at most six neighboring cells. Fig. 1 illustrates the cellular structure of a GSM network.

Velez and Correia [30] have shown that the scenarios of mobility in GSM networks can be characterized by a triangular distribution with average velocities 0, 1, 10, 15 and 22.5 m/s, for the static, pedestrian, urban, main road, and highway, respectively. By 2010 AD, the cell radii will be at the limit (75 m for CBD, 600 m for urban cells, and 2 km for sub-urban cells). In fact, the cell radii have already been less than the limit in dense populated areas such as a big city. Consider that a mobile phone user may frequently make a phone call on the way home or to work. It is very likely that such phone calls take place on a moving vehicle. If the user makes a 3-min phone call on a main road, he/she drives $15\,\text{m/s} \times 60\,\text{s/min} \times 3\,\text{min} = 2700\,\text{m}$ during the phone call. Since the cell radius in the urban area is 600 m, he/she may pass through 5–6 cells. As the utilization of mobile broadband services increases, a phone call may last longer and thus the calling path may be longer.

A mobile phone user may make a phone call at one cell and then move to the other cells during the phone call. The sequence of visited cells during the phone call is termed a *calling path*. In a calling path database, each transaction is a calling path. We say that a transaction supports a calling path P if P is contained in the transaction. The *support* of P is the ratio of transactions in the database that support P. A calling path with support no less than the user-specified *minimum support threshold* is termed a *frequent calling path*.

Let us consider an example. A mobile phone user may make phone calls as the patterns shown in Fig. 2. Suppose the minimum support is 50%. As shown in Fig. 3, the frequent calling paths are $\langle b,c,d,e \rangle$, $\langle d,e,f \rangle$, $\langle e,f,g,h \rangle$, $\langle g,h,i,j \rangle$, $\langle b,c,d \rangle$, $\langle c,d,e \rangle$, $\langle e,f,g \rangle$, $\langle f,g,h \rangle$, $\langle g,h,i \rangle$, $\langle h,i,j \rangle$, $\langle b,c \rangle$, $\langle c,d \rangle$, $\langle d,e \rangle$, $\langle e,f \rangle$, $\langle f,g \rangle$, $\langle g,h \rangle$, $\langle h,i \rangle$, and $\langle i,j \rangle$. However, it is more meaningful to extract the calling path pattern $\langle b,c,d,e,f,g,h,i,j \rangle$ since the user frequently makes phone calls along this path. The calling