CAP: Community Activity Prediction Based on Big Data Analysis

Yin Zhang, Min Chen, Shiwen Mao, Long Hu, and Victor C. M. Leung

Abstract

Crowd sensing harnesses the power of the crowd by mobilizing a large number of users carrying various mobile and networked devices to collect data with the intrinsic multi-modal and large-volume features. With traditional methods, it is highly challenging to analyze the vast data volume generated by crowd sensing. In the era of big data, although several individual-oriented approaches are proposed to analyze human behavior based on big data, the common features of individual activity have not been fully investigated. In this article, we design a novel community-centric framework for community activity prediction based on big data analysis. Specifically, we propose an approach to extract community activity patterns by analyzing the big data collected from both the physical world and virtual social space. The proposed approach consists of community detection based on singular value decomposition and clustering, and community activity modeling based on tensors. The proposed approach is evaluated with a case study where a real dataset collected over a 15-month period is analyzed.

As a novel means of generating data, crowd sensing has already become prevalent and attracted extensive attention from both academia and industry. Traditionally, the research on crowd sensing has been focused on the programming framework [1], incentive mechanism [2], and its applications in various networking systems (e.g., vehicular social networks [3]). However, the problem of effective analysis of crowd sensing data has not been well addressed. This is largely because the scale of crowd sensing data collected in limited ways is not large, and such data can still be processed via traditional methods.

The proliferation of mobile devices and their enhanced onboard sensing capabilities are some of the major forces that drive the explosion of mobile sensing data. Furthermore, advances in social networking and cyber-physical systems are making mobile sensing data "big" and increasingly challenging for analysis with existing processing methods. There is a compelling need to develop effective big data analysis techniques for crowd sensing data, which is the focus of this article.

With its fast growing scale, crowd sensing data will soon exhibit the 4V characteristics of big data, making such datasets drastically different from the traditional datasets [4]:

- *Volume*: The volume of such a dataset will be so big that it cannot be processed by traditional information technology (IT) and software/hardware tools within a tolerable time [5].
- *Variety*: Crowd sensing data have various modalities with respect to time, position, and track information.
- *Velocity*: Thanks to the development of mobile device and mobile networks, mobile sensing data can be generated rapidly in real time.
- *Veracity*: Raw mobile sensing data may include numerous noise signals, as well as redundant and erroneous information.

In order to extract maximum values through effective analysis of crowd sensing data, typically, various techniques such as machine learning, information transmission, social networking, and graph clustering methods can be utilized. By analyzing crowd sensing data, individual behavior patterns can be extracted, which could be useful for guiding and improving people's daily life. Most of the existing crowd sensing research is focused on data collection [1, 6] or individual human behavior analysis [7–9]. In [9], Sohn *et al.* propose a Bayesian network model to predict the individual activity rates of nascent entrepreneurship and new business ownership. In [10], Peng *et al.* propose a scheme based on conditional random fields (CRF) to model the forwarding behavior of microblog users. In [11], Fatima *et al.* develop a unifying framework for individual activity recognition-based behavior analysis and action prediction.

Yin Zhang, Min Chen (corresponding author), and Long Hu are with Huazhong University of Science and Technology.

Shiwen Mao is with Auburn University.

Victor C. M. Leung is with The University of British Columbia.

This work was supported in part by China National Natural Science Foundation under Grants 61300224, the International Science and Technology Collaboration Program (2014DFT10070) funded by China Ministry of Science and Technology (MOST), Hubei Provincial Key Project under grant 2013CFA051, and the US National Science Foundation (NSF) under Grants CNS-0953513, CNS-1247955, and IIP-1266036.

0890-8044/14/$25.00 © 2014 IEEE

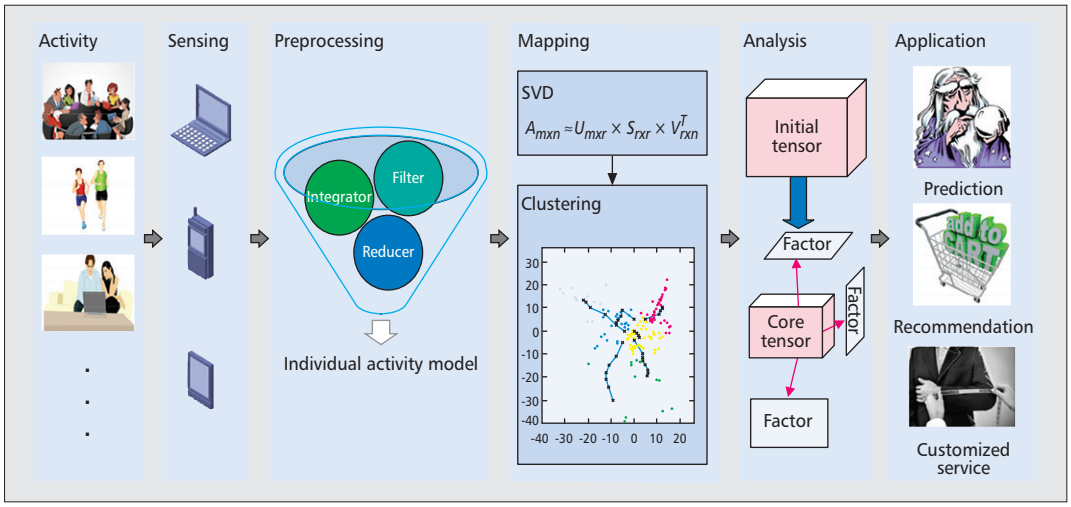

Figure 1. The proposed CAP framework.

Since 2012, in addition to individual activity, collective human behaviors have attracted considerable attention in the research community. There are several interesting works on this aspect, although they just analyze data in a particular domain. In [12], Yuan *et al.* present a recommendation system called T-finder based on the probability sequence to collect both passengers' mobility patterns and taxi drivers' pick-up/drop-off behaviors, and provide drivers and people intending to take a taxi with useful recommendations. In [13], Calabrese *et al.* propose a new model to predict the location of a person over time based on individual and collective behaviors.

Although these earlier works demonstrate the value of collective behavior analysis, little research has been done that is focused on community activity prediction (CAP) based on big data analysis. Indeed, CAP is extremely valuable in the paradigm of big data for the following reasons:

- *Low complexity*: With CAP, it is not necessary to create activity models for each person; only one model is needed for a community consisting of several individuals with similar behavior patterns. CAP reduces the complexity of big data modeling and analysis.
- *High efficiency*: Although it is hard to guarantee absolute correctness of customized service, CAP can quickly discover the majority's demands according to the analysis of a community activity model. In other words, CAP is more concerned with efficiency than accuracy.

The crowd sensing data from numerous participating individuals are usually complex and with multiple dimensions. With CAP, we can predict community activity and customize services for the community that may replace individualized service in some scenarios. However, the following challenging issues have to be addressed in CAP.

- *Community-centric*: Within the large amount of data collected by mobile sensing, a human is an abstract logic unit comprising digital content in the cyber space, location in the physical space, and activities in society. It is challenging to analyze the complex relationship among the high-dimensional multimodal variables. CAP should exploit community-centric classifications of individual activities to simplify the analysis of crowd sensing data.
- *Unified presentation*: There are many differences in encoding, format, structure, and other aspects among various datasets. It is still an open problem to find an efficient model for a unified presentation of all community-centric data sensed by different platforms and technologies from multiple fields.
- *Efficient analysis*: Due to the high generation rate of crowd sensing data, the analysis will consume a large amount of computational resources. Hence, as in other big data applications, it is a great challenge to efficiently explore the hidden masses of heterogeneous data to identify the scientific value in CAP.

Through the community activity features in our society, this article proposes an approach to analyze and predict community activity. With the proposed approach, activity data from individuals are first collected. Then we create models for analyzing the relationships among individual, community, and activity with singular value decomposition (SVD), clustering, and tensors. Finally, we predict future activities of the community and provide customized services. More specifically, this article makes the following contributions to crowd sensing data analysis:

- We propose a method based on SVD to discover communities in high-dimensional multimodal data. It is a comprehensive community-centric classification according to the features of individual activities in the physical world and social networking.
- We establish a fusion and analytic model for crowd sensing big data, which breaks through the barriers between the physical world and virtual cyber space. It provides a unified presentation for individual activity data, as well as a tensor-based model for integration from three domains: space-time, community, and activity.
- We develop a novel theory to extract feature information and identify activity patterns from community activity data by utilizing tensor decomposition. Furthermore, we evaluate the performance of CAP with a case study to demonstrate its efficacy in predicting community activity.

Framework for Community Activity Prediction

Framework

CAP includes data acquisition, data preprocessing and transmission, data storage, and data analysis and application as other big data applications. More specifically, CAP shall:

- Acquire and analyze a large amount of individual activity data with mobile sensing.
- Cluster the acquired crowd sensing data into different communities according to activity features.
- Analyze and mine the activity data of each community.

Figure 1 illustrates the five layers of the CAP framework (i.e., data sensing, data preprocessing, relation mapping, community activity analysis, and application):

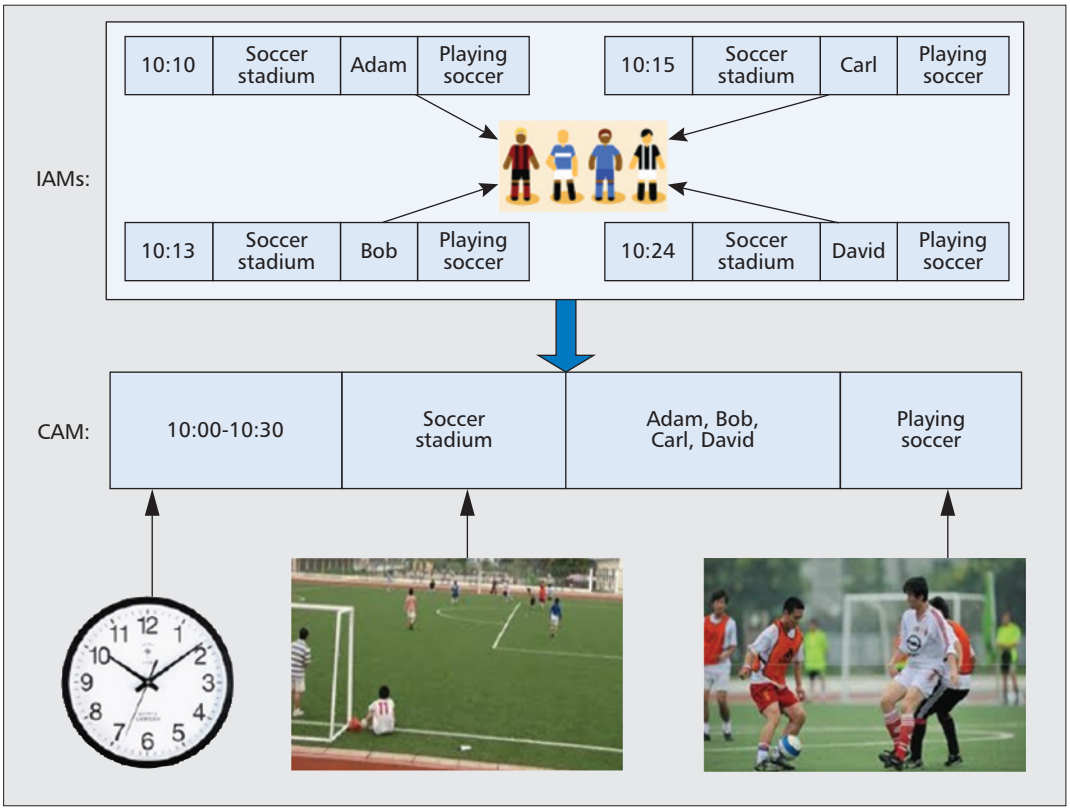

Figure 2. Four individual activity models are integrated into one community activity model.

- *Sensing*: To collect individual activity data by crowd sensing.
- *Preprocessing*: This layer consists of three modules: a filter, a reducer, and an integrator. After raw data collection, this layer will filter out invalid data, reduce redundant data, and preliminarily integrate individual activities into communities.
- *Mapping*: This layer consists of two modules, SVD and clustering. It first utilizes SVD to reduce the dimension and compresses the volume of data in order to reduce the complexity of data processing. It then generates not only an integrated community activity dataset, but also an individual-community relation mapping after clustering.
- *Analysis*: This layer utilizes tensors and related mathematical methods to analyze the community activity data and extract the features.
- *Application*: According to the analytical results, we can gain useful understanding of the activity features of each community, and provide further services such as individual activity prediction and customized recommendation.

Mapping and analysis are the two key components of CAP. We provide more detailed discussions of these two components later.

The dataset used in this article is the individual activity data collected by the EPIC Lab of Huazhong University of Science and Technology through the five-layer framework discussed above.[1] The dataset includes various activities from both the physical world and social network. We use the data from a period of 12 months, from July 2012 to July 2013, as training data for the analyzer; and the data from a period of three months, from August 2013 to October 2013, as verification data. We then check whether the analytical results and the predicted activities through the model are consistent with the actual measured data.

We design two models for storage and processing of this dataset, which are discussed in the following.

Individual Activity Model

In social activities, each activity can be identified by three elements (i.e., time, arena, and individual). Each item of the individual activity data recorded by mobile sensing should include these three pieces of information. We then define the individual activity model (IAM) as

IAM: <Time, Arena, Individual, Activity>,

where *Time* records the time of the activity, *Arena* records the location of the activity, *Individual* records the activity participant, and *Activity* records the activity type. In CAP, the metadata acquired is based on the IAM model.

Community Activity Model

Because the raw data includes numerous invalid or redundant data, they need to be preprocessed to increase the veracity and reduce the volume. Through preprocessing, the IAMs are integrated into the community activity model (CAM), which is defined as

CAM: <Time, Arena, Community, Activity>.

In the CAM model, the meanings of *Arena* and *Activity* are the same as that in IAM. *Time* does not mean a time instance, but a time window, while *Community* records all the individuals who conduct the same activity in the same time period at the same location. Figure 2 illustrates that four IAMs, each of which stores an individual's activity, can be integrated into one CAM record recording the activity of a community consisting of these four individuals.

[1] *This datatest is available at http://epic.hust.edu.cn/minchen/files/dataset.zip*

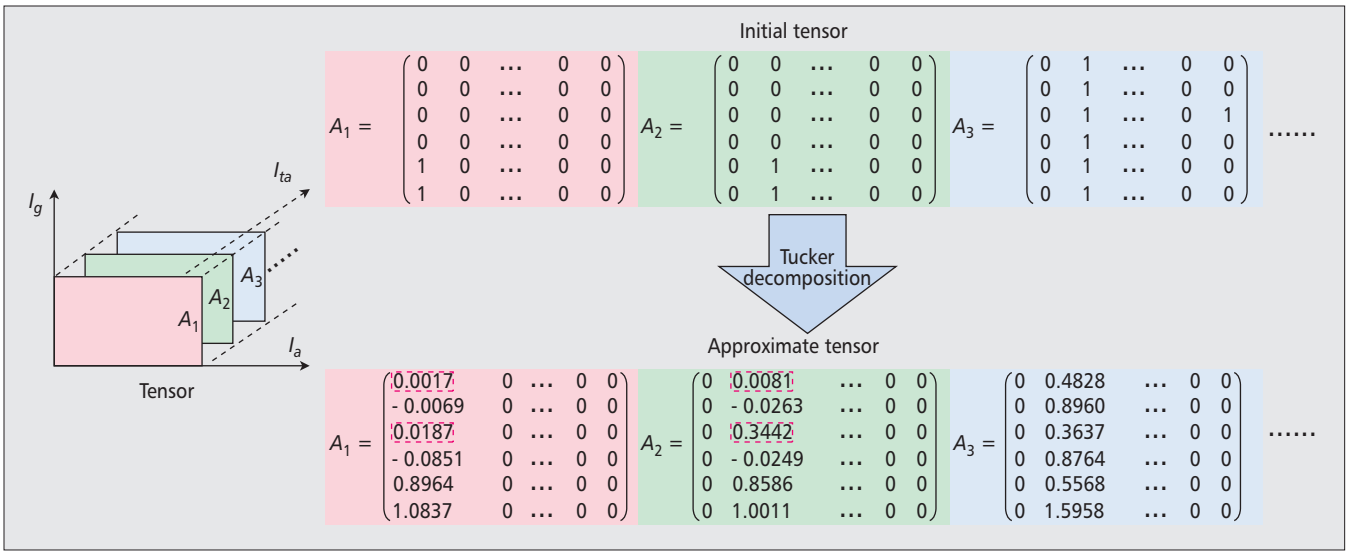

Figure 3. Big data analysis based on tensors.

Mapping and Analysis in CAP

Mapping

Due to the large dataset, analyzing the record of each individual independently is a highly complicated task. We need to merge the data items with the same activity features using CAM and form an individual-community mapping relationship, thus reducing the complexity of data analysis.

Take the following 119 × 47 matrix as an example, which is generated in the case study discussed later. In this matrix, $a_{ij} = 1$ means that individual j is classified as belonging to community i, and $a_{ij} = 0$ indicates that individual j does not belong to community i.

$$\begin{array}{c} \\ \textit{Community} \end{array} \begin{array}{c} \text{Individual} \\ \begin{pmatrix} 0 & 1 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & 0 \\ 1 & 0 & \cdots & 0 & 1 \end{pmatrix} \end{array}$$

After SVD, we compress the 119 × 47 matrix into three matrices with dimensions 119 × 3, 3 × 3, and 3 × 47. From the singular value matrix, we can see that the first column of the left singular matrix and the first row of the right singular matrix feature the highest importance. Through further analysis, the absolute value of the first column in the left singular matrix approximately reflects the quantity of individuals included in each community, and the absolute value of the first row in the right singular matrix approximately reflects the quantity of times each individual appears in different communities.

From the singular value matrix, we find that the second and third singular value better approximate the initial matrix. Therefore, we select the last two dimensions of the left singular vectors U and the right singular vectors V, and project them into a second-order coordinate system. It is important to note that the reason for making individuals participate in the clustering analysis is because each independent individual can also be regarded as a community with only one element. After clustering, all the individuals can be classified into six communities in this example. This kind of community relation mapping can further compress the data storage space and reduce the complexity of data analysis.

Analysis

Figure 3 illustrates the three steps of CAP analysis. We discuss these steps in the following:

- *Tensor initialization*: According to the definition of CAM, the data obtained from the earlier steps are in the form of fourth-order tensors. Since in the social activity time and arena can be merged, we can further simplify the fourth-order CAMs to third-order tensors.
- *Tucker decomposition*: Based on the initial tensor, we can obtain a core tensor and three projection matrices $U^{(g)}$, $U^{(a)}$, $U^{(ta)}$ after Tucker decomposition.
- *Approximate tensor*: Finally, we obtain the approximate tensor. The physical meaning of the approximate tensor lies in that it can represent the community activity rules approximately.

Compared to the initial tensor, the values of many elements in the approximate tensor have been changed from zero to non-zero values, representing whether a community conducts an activity in a certain time period and arena. Our prediction about the trend of changes also reflects an assessment value of a community engaging in a certain activity in the next future period. The higher the assessment value, the larger the probability of this community engaging in the activity, and vice versa. The elements marked in Fig. 3 are all greater than 0.1, and each one's value is different from that in the initial tensor.

Utilizing these assessment values, we can create a model for the community activity, and apply the model for prediction and individualized recommendation. For example, if the assessment value of a certain community going to a shopping mall is extremely high, we can send not only the preference and discount information of the relevant sellers to the individuals in the community, but also the traffic status to the shopping mall to each user, and can even order a taxi and food for such an individual in advance.

Case Study

In this section, we evaluate the performance of the proposed CAP approach from two prospectives, i.e., the ratio of reducing the dimension and complexity of raw data, and the accuracy of prediction.

Compression Ratio

In three layers of the five-layer CAP framework — preprocessing, relation mapping, and analysis — the crowd sensing data is transformed as follows:

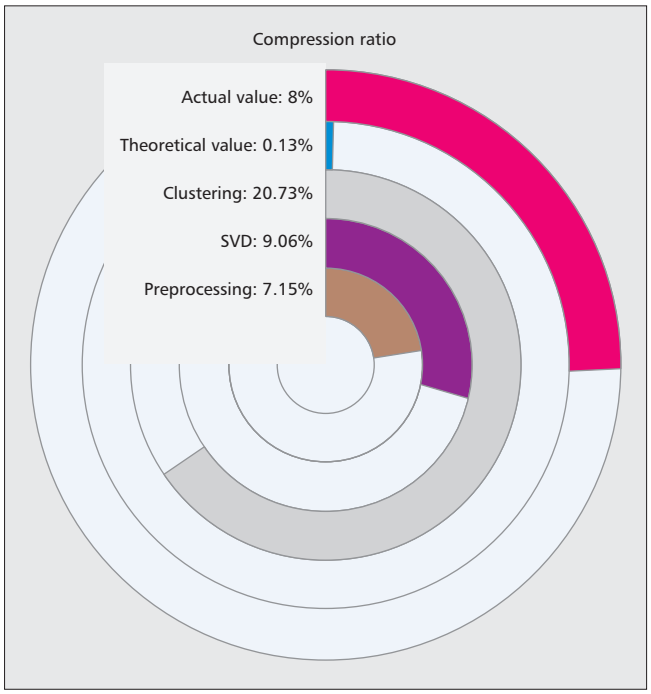

Figure 4. Compression ratio in each step.

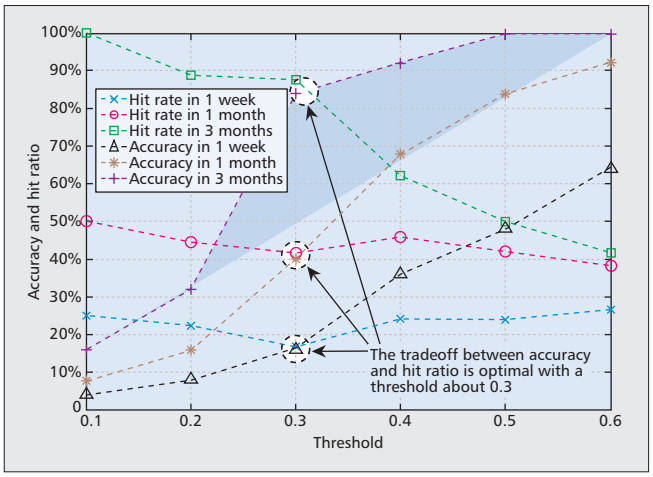

Figure 5. Accuracy and hit rates in different periods.

- Preprocessing: As introduced earlier, the raw dataset consists of 5342 IAMs, which are integrated into 382 CAMs after preprocessing, wherein there are 357 training data and 25 test data. The compression ratio is 7.15 percent.
- Relation mapping: In the training dataset, 47 individuals appear in 119 different communities. There are two modules in this layer. In the SVD module, the relationships between individuals and communities can be represented by a 119 × 47 matrix for the training data. After SVD, the matrix is compressed into three matrices with dimensions 119 × 3, 3 × 3, and 3 × 47, respectively. Thus, the compression ratio of SVD is 9.06 percent. In the cluster module, the 357 training data items based on CAMs are integrated into 74 CAMs from six communities, and the compression ratio of the cluster module is 20.73 percent.
- Analysis: In the data analysis layer, the size of the approximate tensor is equal to the initial tensor after Tucker decomposition.

Theoretically, the volume of raw data can be reduced to as low as 0.13 percent. However, data structures are transformed through the five-layer framework of CAP. According to the memory consumption during the data analysis, the actual compression ratio is about 8 percent. The compression ratio results discussed above are presented in Fig. 4.

Accuracy

In order to describe the CAP of approximate tensors more precisely, we set a threshold value. When the assessment value is greater than the threshold value, this element predicts that the community will conduct the corresponding activity, and vice versa. We take different threshold values, and compare the obtained results with the verification dataset.

There are 25 samples in the verification dataset. The accuracy rate and hit rate of the CAP results at different threshold values can be computed as follows:

$$\text{accuracy rate} = \frac{\text{hit number}}{\text{prediction number}}$$

$$\text{hit rate} = \frac{\text{hit number}}{\text{sample number}}.$$

From the CAP results presented in Table 1 and Fig. 5, it can be seen that with a smaller threshold value, there are more positive predictions. If the threshold value is increased, the accuracy rate will be lower but the hit rate will be higher. Moreover, over a longer period of time, both accuracy rate and hit rate are considerably improved. In particular, when the threshold value is 0.3, both the accuracy rate and hit rate are acceptable.

Conclusion

In this article, we have proposed a CAP method based on big data analysis. The proposed approach consists of community detection based on SVD and clustering, and community activity modeling based on tensors. The proposed scheme has been validated with a case study using a dataset captured over a 15-month period. Not only can the CAP approach achieve good prediction performance, but it can also effectively reduce the complexity of data. Furthermore, we have shown that a moderate threshold of assessment value in the approximate tensor can improve both the reliability and accuracy of CAP. With the advantages of low complexity and acceptable reliability and accuracy, CAP can be applied in multiple fields, such as customized recommendation services, smart grid, intelligent city, and other big data applications.

CAP in this article refers to the offline modeling of historical data within a period. For future work, we will investigate how to process the individual activity data in an online manner and update the existing model dynamically.

References

[1] M.-R. Ra *et al.*, "Medusa: A Programming Framework for Crowd-Sensing Applications," *Proc. ACM MobiSys '12*, Low Wood Bay, Lake District, U.K., June 2012, pp. 337–50.
[2] D. Yang *et al.*, "Crowdsourcing to Smartphones: Incentive Mechanism Design for Mobile Phone Sensing," *Proc. MobiCom '12*, Istanbul, Turkey, Aug. 2012, pp. 173–84.
[3] X. Hu *et al.*, "Social Drive: A Crowdsourcing-Based Vehicular Social Networking System for Green Transportation," *Proc. ACM MSWiM '13*, Barcelona, Spain, Nov. 2013, pp. 85–92.
[4] M. Chen *et al.*, "Big Data: Related Technologies, Challenges and Future Prospects," *Springer Briefs in Computer Science*, 2014.
[5] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Springer Mobile Networks and Applications Journal*, vol. 19, no. 2, Apr. 2014, pp. 171–209.
[6] X. Hu *et al.*, "Vita: A Crowdsensing-Oriented Mobile Cyber Physical System," *IEEE Trans. Emerging Topics Computing*, vol. 1, no. 1, July 2013, pp. 148–65.
[7] D. Huang, T. Xing, and H. Wu, "Mobile Cloud Computing Service Models: A User-Centric Approach," *IEEE Network*, vol. 27, no. 5, Sept./Oct. 2013, pp. 6–11.

Threshold	Prediction	Hit number			Accuracy rate			Hit rate		
		1 week	1 month	3 months	1 week	1 month	3 months	1 week	1 month	3 months
0.6	4	1	2	4	25.00%	50.00%	100.00%	4.00%	8.00%	16.00%
0.5	9	2	4	8	22.22%	44.44%	88.89%	8.00%	16.00%	32.00%
0.4	24	4	10	21	16.67%	41.67%	87.50%	16.00%	40.00%	84.00%
0.3	37	9	17	23	24.32%	45.95%	62.16%	36.00%	68.00%	92.00%
0.2	50	12	21	25	24.00%	42.00%	50.00%	48.00%	84.00%	100.00%
0.1	60	16	23	25	26.67%	38.33%	41.67%	64.00%	92.00%	100.00%

Table 1. Accuracy and hit rate over different time periods.

[8] X. Ge *et al.*, "Spectrum and Energy Efficiency Evaluation of Two-Tier Femtocell Networks with Partially Open Channels," *IEEE Trans. Vehic. Tech.*, vol. 63, no. 3, Mar. 2014, pp. 1306–19.
[9] S. Y. Sohn and A. S. Lee, "Bayesian Network Analysis for the Dynamic Prediction of Early Stage Entrepreneurial Activity Index," *Int'l. J. Expert Systems with Applications*, vol. 40, no. 10, Aug. 2013, pp. 4003–09.
[10] H.-K. Peng *et al.*, "Retweet Modeling Using Conditional Random Fields," *Proc. ICDM Wksps.*, Vancouver, BC, Canada, Dec. 2011, pp. 336–43.
[11] I. Fatima *et al.*, "A Unified Framework for Activity Recognition-Based Behavior Analysis and Action Prediction in Smart Homes," *Sensors*, vol. 13, no. 2, Feb. 2013, pp. 2682–99.
[12] N. J. Yuan *et al.*, "T-Finder: A Recommender System for Finding Passengers and Vacant Taxis," *IEEE Trans. Trans. Knowl. Data Eng.*, vol. 25, no. 10, Oct. 2013, pp. 2390–2403.
[13] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human Mobility Prediction Based on Individual and Collective Geographical Preferences," *Proc. IEEE ITSC '10*, Washington, DC, Sept. 2010, pp. 312–17.
[14] E. Henry and J. Hofrichter, "Singular Value Decomposition: Application to Analysis of Experimental Data," *Essential Numerical Computer Methods*, M. Johnson, Ed., Ch. 6, pp. 81–138, Elsevier, 2010.
[15] S. Rendle and L. Schmidt-Thieme, "Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation," *Proc. ACM WSDM '10*, New York, NY, Feb. 2010, pp. 81–90.

Biographies

YIN ZHANG is a postdoctoral fellow in the School of Computer Science and Technology at Huazhong University of Science and Technology (HUST). He is a lead Guest Editor for *New Review of Hypermedia and Multimedia*. He serves as a reviewer for *IEEE Network* on information sciences. He is a Technical Program Committee (TPC) Co-Chair for CloudComp '14. He is Local Chair for TRIDENTCOM '14), as he was for CloudComp '13.

MIN CHEN [M'08, SM'09] is a professor in the School of Computer Science and Technology at HUST. He was an assistant professor in School of Computer Science and Engineering at Seoul National University (SNU) from September 2009 to February 2012. He worked as a postdoctoral fellow in the Department of Electrical and Computer Engineering at the University of British Columbia (UBC) for three years. Before joining UBC, he was a postdoctoral fellow at SNU for one and half years. He has more than 180 paper publications. He received the Best Paper Award at IEEE ICC '12 and Best Paper Runner-up Award from QShine '08. He has been a Guest Editor for *IEEE Network*, *IEEE Wireless Communications*, and other publications. He was Symposium Co-Chair for IEEE ICC '12 and IEEE ICC '13. He was General Co-Chair for IEEE CIT '12 and TRIDENTCOM '14. He is a TPC member for IEEE INFOCOM 2014 and TRIDENTCOM '14. He was the Keynote Speaker for CyberC '12 and Mobiquitous '12.

SHIWEN MAO [SM]is the McWane Associate Professor in the Department of Electrical and Computer Engineering, Auburn University, Alabama. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society Class of 2014. He received the 2013 IEEE ComSoc MMTC Outstanding Leadership Award and the NSF CAREER Award in 2010. He was a co-recipient of the IEEE ICC 2013 Best Paper Award and the 2004 IEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is on the Editorial Boards of *IEEE Transactions on Wireless Communications*, *IEEE Internet of Things Journal*, *IEEE Communications Surveys & Tutorials*, *Elsevier Ad Hoc Networks Journal*, and *Wiley International Journal of Communication Systems*.

LONG HU is a Ph.D. candidate in the School of Computer Science and Technology at HUST. He received his B.Sc. and M.Eng. degrees from HUST. His research interests include 5G mobile communication systems, marine-ship communication, the Internet of Things, and multimedia transmission over wireless networks. He was Publication Chair for Cloudcomp '13. He received the Best Paper and Outstanding Service Awards at Cloudcomp '13.

VICTOR C. M. LEUNG [S'75, M'89, SM'97, F'03] is a professor of electrical and computer engineering and holder of the TELUS Mobility Research Chair at the University of British Columbia (UBC). He has contributed some 700 technical papers, 29 book chapters, and eight books in the areas of wireless networks and mobile systems. He was a Distinguished Lecturer of the IEEE Communications Society. He has been serving on the Editorial Boards of *IEEE Wireless Communications Letters* and several other journals, and has contributed to the organizing committees and TPCs of numerous conferences. He was a winner of the 2012 UBC Killam Research Prize and the IEEE Vancouver Section Centennial Award. He is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.