International Journal of Advances in Computer Science & Its Applications– IJCSIA
Volume 5 : Issue 2 [ISSN : 2250-3765]
Publication Date: 30 October, 2015

Growable Cyber-I's Modeling with Increasing Personal Data

Song Zhang, Jianhua Ma, Runhe Huang, and Dongming Chen

***Abstract*— Cyber-Individual (Cyber-I), is a counterpart of Real-Individual (Real-I) in cyberspace, namely, a unique, digital and comprehensive description for a real individual. A Cyber-I is to gradually approximate to its Real-I by continuously collecting, processing and utilizing Real-I's personal data. The personal data in terms of data types and data amount is continuously increasing due to widely use of smartphones, sensors, and other devices as well as software tools. Such personal data makes Cyber-I's growable model possible. This paper is mainly focused on describing the initialization and growth of Cyber-I model. That is, how is a Cyber-I model initialized with the basic data generated at the Cyber-I birth stage and how can it grow with continuously collected incoming personal data to approximate to its Real-I's states, behaviors and characteristics. In this paper, a system prototype for supporting the model initialization and growth is illustrated, and a case study for showing the growable modeling is given.**

KeywordsCyber-I, Increasing personal data, growable modeling

I. Introduction

With rapid advances of computing and communication technologies, we are stepping into a completely new cyber-physical-social integrated world with digital explosions of data, connectivity, services and intelligence. Facing explosive information and services, we may not be aware of what are the most necessary or suitable things. However, the emergence of Cyber-Individual (Cyber-I) makes each of our human individual person to have a digital clone [1] [2] [3] which can help us to understand more and better about our needs as well as support us making better choices. The study on Cyber-I is an effort to re-examine and analyze human essence in the cyber-physical-social integrated world in order to assist individuals in dealing with the digital explosions and make them having an enjoyable life in this digital era.

In developing Cyber-I, one of the fundamental problems encountered is to figure out how a Cyber-I can approximate to characteristics and even mind of its real individual (Real-I) [3]. Obviously, a comprehensive and sophisticated Cyber-I model can't be built at once because there is no enough personal data at the initial modeling stage. Fortunately, more and more personal data can be continuously collected by means of various software tools and ubiquitous devices such as smartphones, sensors, wearable devices, and so on.

Such personal data comes increasingly and thus offers a possibility to make the models grow as if creatures in nature need nutriment to grow and the Cyber-I models can grow with the digital "nutrition", namely, increasing personal data. When more and more personal data is available, the next problem is how to generate Cyber-I's initial models and make them growable. The ultimate goal is to enable the models to successively approach towards individual's actual characteristics along with increasing personal data from various sources covering different human facets. This research is trying to put our effort on the initialization and growth of Cyber-I's models. The initial models are generated based on the personal data acquired at a Cyber-I's birth stage, while the models start to grow with the continuously incoming personal data collected after the birth. We proposed three mechanisms for Cyber-I modeling to enable the models growing more and more similar with and closely resemble to its Real-I.

II. Related Work

There are many user modeling researches, some representative user models are known as user profiles, personas or archetypes, which can be used by developers for other developers in their personalized services and applications [4]. With increasing necessity of developing personalized systems, like personalized e-learning systems, u-health care, and e-commerce service, collecting personal data has become an important process and challenging issue. In order to give appropriate advices or recommendations with individual identical personal features, user models have to be built in service systems [5][6]. However, almost all of those user models are application-specific or service-specific and are difficult to be shared by or used in other applications/services [7]. To overcome this barrier of user models among different applications, a generic user model system (GUMS) was proposed to support interoperability among different user modeling systems [8] . Life logging is utilized to automatically record user's life events in digital format. With continuously capturing contextual information from a user and the user's environment, personal data increases fast and becomes huge. The most of lifelog systems are putting more emphases on personal data collection, storage and management. Lifelong user modeling is trying to provide users such models accompanied with users' whole life [9]. This idea or vision is attractive, but no general mechanism has been made and no practical system has been built yet. Lifelong machine learning (LML), received great attention in recent years, is to enable an algorithm or a system to learn tasks from more domains over its lifetime [10].

All technologies in user modeling, life logging, lifelong user modeling and LML are closely related to Cyber-I

Song Zhang, Jianhua Ma, and Runhe Huang
Faculty of Computer and Information Sciences, Hosei University, Japan

Dongming Chen, Song Zhang,
Software College, Northeastern University, China

modeling. However, our emphases are modeling a human beyond a user and building growing models to approximate the human along with increasing personal data.

III. Personal Data and Cyber-I Modeling

Cyber-I is a counterpart of Real-I in the cyber world. A Cyber-I model is not static, i.e., it is gradually approximating to its Real-I. Figure 1 shows the process of Cyber-I modeling in which the Cyber-I model grows to be bigger, higher, and closer to Real-I's actual features based on the initial model along with increasing personal data from applications and various data sources from Physical-Cyber-Social world.

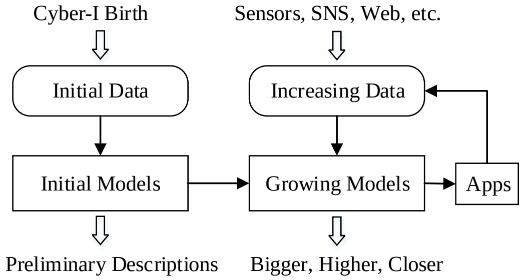

Fig. 1. Overview of Cyber-I modeling

Collected personal data, including the initial data generated at the Cyber-I birth stage, is to be used as a start-up growable model of the Cyber-I and it follows up to grow along with coming data, namely, increasing personal data captured and processed from various Web, ubiquitous sensing and other ubiquitous technologies. Based on available sets of the initial data, different kinds of initial models will be generated. The initial models, namely preliminary descriptions about Cyber-I, will act as seeds for the models' growth. The initial models are generated from the initial datasets including the user's basic information, profiles and preferences. By means of continuously collecting and integrating scattered and increasing personal data, models will grow to including more and more its Real-I features collected and mined from different life facets. The growing models keep refining themselves to proximate Real-I's actual features. The Cyber-I models can be applied to support personalized services and they are evaluated in the course of supporting personalized services. Their evaluation results in return are used for improving the Cyber-I models. Meanwhile, the personalized services and applications can generate more personal data for the Cyber-I model's further growth.

Personal data (PD) is the data concerned a person related data, the scattered pieces that the person left in real, cyber, social worlds via his/her activities. Since personal data collected and generated from devices, sensors or applications can become progressively bigger and bigger as time goes by, it is on the whole called increasing personal data. Personal data may increase in two ways: (1) increasing in size, i.e. the amount of data from the same source is getting more and more; and (2) increasing in type, i.e. the new types of data are continuously added from different sources. There are four categories of increasing data: continuous, periodic, aperiodic and burst as shown in Fig. 2.

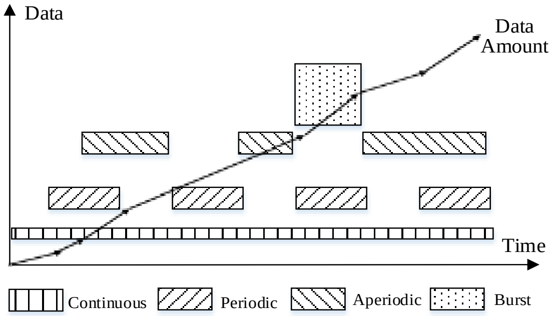

Fig. 2. Categories and features of increasing personal data

Along with time, the total amount of personal data may become more and more, namely, personal big data. In this research, the following types of personal data have been collected, as shown in Table I.

TABLE I. COLLECTED DATA AND FEATURES

Data Source	Data Type	Data Category	Data Used in
Cyber-I's Birth	Basic Info	--	Initial Models
Facebook/QQ	Profile	--	
Multiple Choices	Preference	--	
Twitter	Tweets	Aperiodic	Growing Models
Web	Web Pages	Aperiodic/Burst	
Browser	URL History	Aperiodic	
Jawbone UP	Movement	Continuous	
GPS	Location	Periodic/Aperiodic	
Manic Time	App/Act Log	Aperiodic	

IV. Initialization of Cyber-I Model

An Initial model is a Cyber-I start-up model which is generated from the user's basic information, preferences and profiles. The initial model stands for the primitive description or template, which provides a well-grounded foundation for the model to grow. It is importance as seeds to tree or plant. IM-C (Initial Model from the Core Data) is the seed of Cyber-I model. Like in nature, different "seeds" experience different stages of incubation and play different roles at different life stages. Apart from IM-C seed, there are other seeds and associated methods which support the Cyber-I growable models. As shown in Fig. 3, three kinds of primitive initial models IM-C, IM-S and IM-P are generated at different situations for covering a Real-I different facets and can be combined like IM-CS, IM-CPS and IM-CP.

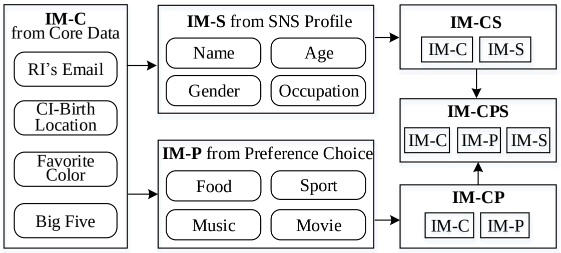

Fig. 3. Three different initial models and their combinations

A. *The Initial Model: IM-C*

The core initial data is generated in the process of the Cyber-I's birth. It includes Real-I (RI)'s email address, birth location (inferred from IP address), favorite color and Big Five personality traits (through personality test questionnaire). At the Cyber-I birth stage, an email input by a user could serve as a bridge between Real-I and Cyber-I for their info further information interaction. The current IP-address automatically recorded can be processed to obtain a rough location. Color as one of the preference attributes is selected for inclusion in IM-C since we believe that color can somehow reflect a user's inner property and it is also less sensitive to one's privacy so that it is relatively easy for a user to make color choice. Finally, the user is asked to fill up a Big Five questionnaire. These five factors provide a rich conceptual framework for integrating all the research findings and theory in personality psychology. After collecting and processing the core data and storing it into the personal database, the core seed of Cyber-I model is generated.

B. *The Initial Model: IM-S*

It is a fact that people spend more and more time and conduct activities on Web and social networking service (SNS), where is a good place to obtain useful data sources for extracting a user's profile content since people leave much of their footprints which contains personal information and as well as implicit personal features. The SNS profile makes important contributions to the initialization of Cyber-I modeling. IM-S is a kind of initialization method if a user is willing to provide his/her SNS account. After observation of some popular SNS websites (Facebook, LinkedIn and QQ), it is decided to use the four common elements: name, age, gender and occupation as compulsory elements for generating the IM-S.

C. *The Initial Model: IM-P*

We believe that some human preferences are innate and may reflect something deep inside the user like personality, characteristics or traits. Moreover, one's preferences could influence his/her decisions or action. In this research, we provide users some other preference choices as optional preferences. They are Foods, Sport, Movie and Music and for each of them 6 choices are listed as given in Tab. II. If a user is willing to choose one or some of them, the model will contain more aspects about the user. For each type of preference, the user interface provides predefined 6 choices for users to select. If there is no appropriate answer among the 6 choices, one can input his/her preferred choice. IM-P could also be a start to generate a higher level description about the user, which is one kind mechanism of growable modeling and will be explained in the next section.

TABLE II. PREFERENCE CHOICES

Preference	Choices
Food	salty, sweets, spicy, sour, meet, vegetable, etc.
Sport	swimming, athletics, team/person ball, outdoor,fight, etc.
Movie	animation, comedy, adventure, action, romance, war, etc.
Music	jazz, classical, folk, rock, pop, rap, etc.

There are three combinational models, IM-CS, IM-CP and IM-CPS. If a user is willing to input his/her SNS account and choose preferences of different aspects, we can know more properties about the user so that more exact information could be offered to him/her in personalized applications. Meanwhile, the applications can also generate some additional new data, which help the models to grow.

V. The Growth of Cyber-I Model

A growable model means the one that is able to successively approach to individual's actual features, along with increasing personal data from various sources. The general modeling process is illustrated in Fig. 4.

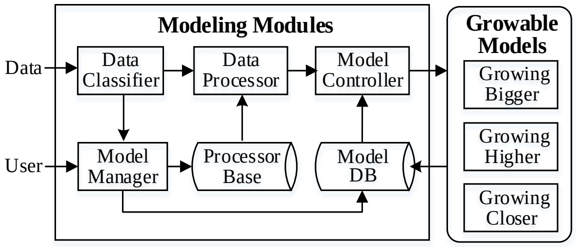

Fig. 4. Modeling modules for growing bigger, higher and closer

The data reserved in personal database is added to our modeling module, and data classifier is activated to make a classification and add a class *id*. A user can see data list and select a function from processor base to process the data. Data processor generates the processed results. Model controller is utilized to (1) examine the existing models, (2) make the model grow driven by time, user, or event, and (3) choose a corresponding function to grow bigger, higher or closer. In this way, the new generated grown model is reserved in model database for next cycle of growth.

A. *Becoming Bigger (GM-B)*

The growable modeling for becoming bigger, short for GM-B, has two meanings. One is that the model could grow to cover more aspects of its Real-I just as a tree will grow with more branches. The other is that the model grows finer in one specific aspect by showing more exact and more detailed descriptions. The process of growing bigger is shown in Fig. 5. Suppose data D1, D2 and D3 are three kinds of data incoming from different sources. It is found that D1 and D2 cover the aspects A1 and A2, which are not included in the present model and the model controller adds A1 and A2 to the current model. D3 contains some detailed information that is related to A2, therefore A2 is extended with A21 and A22 extracted from D3. In this way, the model has grown with more aspects and finer in some specific aspect. Whether generating more aspects or extending the exist aspects to a deep level is decided by the model controller and the current model.

For example, one of the keywords "Kobe Bryant" is contained in user's tweet. After analyzing by some text mining tool, it is found that "Kobo Bryant" is a basketball player so that this aspect could be classified under basketball. If the aspect of basketball doesn't exist in the current model, then the aspect of basketball is added to the model. If it exists, the user's favorite basketball player may be added under the aspect of basketball.

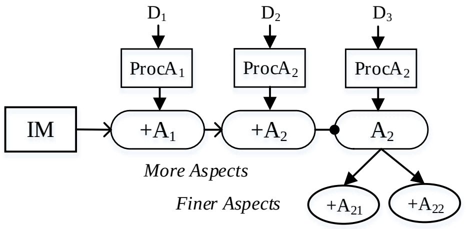

Fig. 5.Growable modeling for becoming bigger

B. *Becoming Higher (GM-H)*

The growable Modeling for becoming to a higher level means an abstract refinement of description based on Real-I's interest, behavior and trait. There are four levels: Level 0 is the data level; Level 1 stands for state level; Level 2 is the behavior level; Level 3 is characteristic level. As time goes by, more and more data can be added into level 0, which lays a foundation for the model to grow higher. Three mechanisms are defined for the model to grow. The first one is growing step by step so that the data of a specific aspect gradually grows to a higher level (one level at a time). The second one relies on the combination of multiple aspects to generate a new feature at a higher level. The model could leap to grow in the third way where it may grow from the state level directly to the characteristic level.

C. *Becoming Closer (GM-C)*

The growable Modeling for becoming closer is a process, which make the model less errors and adaptable to the sudden changes in the user's attributes. In Fig. 6, the process is shown and two kinds of mechanisms (error reduction and change tracking) for GM-C are illustrated. In order to reduce the errors occurred in the model generation, more data concerning the specific aspects is needed for gradually approximating to the real value. A user's trait may have a or some sudden change(s) at some situations and the model should be able to detect such changes and tracking them. For example, a user created his/her model as a student but later he graduated and went to work as an employee. The model should be able to notice such changes, capture his company related information mined from SNS conversations, or other active media and data sources, and update the model with a new occupation property.

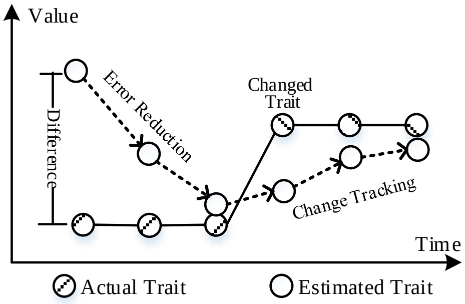

Fig. 6. Growable modeling for becoming closer

VI. Case Study

The system mainly consists of three parts: data processing, growable model, and a user interface. A user can choose corresponding functions to process the data, and the proper time to grow and view the grown model stored in the model DB. Our model is designed in a hierarchical structure which can be easily viewed and modified. The database is designed using Adjacency list model that contains three columns, i.e. id, parent and name. The user interface is designed using HTML, CSS3 and JavaScript. The system is implemented in Java language, and MySQL database is utilized to store the models. In this case study, we collect user's data from a SNS profile, Twitter, preference choices, Web content, app usage records, browser history, movement logs, and GPS-based Location.

As for the data processing part, we analyzed a user's interest, location or even sentiment concerning a specific word processing tweets; obtained the user's possible occupation and other information by analyzing the user's Web contents; and extracted the user's behaviors by mining App usage data and browsing history. We also use the user's movement log and location data for understanding the user's life pattern.

At the Cyber-I birth stage, the user first registered the system through an email address, then chose his favorite color blue, and filled in the Big Five questionnaire. The rough location, Japan, was also generated from the current IP address. By processing these data, an IM-C hierarchical initial structure was generated as shown on the Cyber-I modeling interface in Fig. 7.

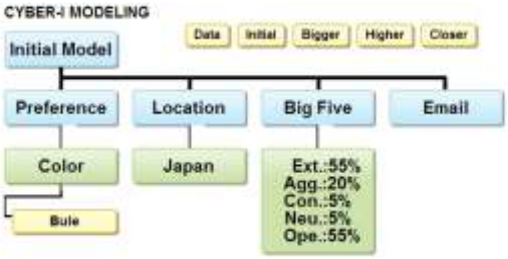

Fig. 7. an example of IM-C

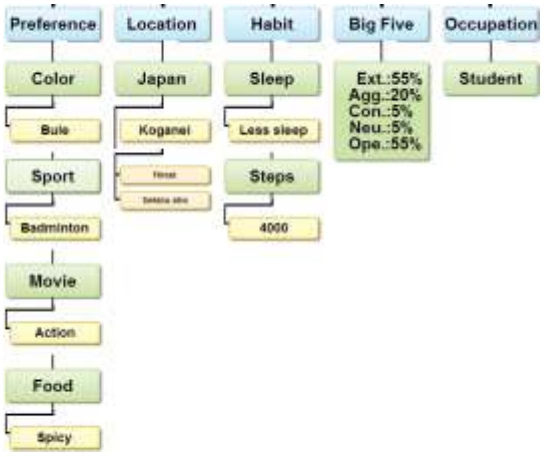

Fig. 8. Examples of GM-B and GM-C

As shown in Fig. 8, the model grows in the five aspects, including preference, location, habit, Big Five and occupation. For example, in the preference part, three more aspects (Sport, Movie and Food) are grown from multi-preference choices in IM-P. From one week's tweets collected, it was inferred that this user might love swimming and red color. By analyzing one month's data, it showed that badminton didn't appear but swimming could be found from every week's tweets. Then the badminton aspect is deleted by model controller. Such change means the model grows closer to reduce error.

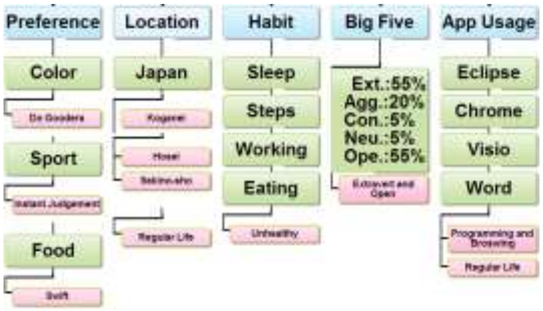

Fig. 9. Examples of GM-H

Three mechanisms for growing higher are shown in Fig. 9. The first one is growing up step by step. The app usage data indicated the four apps (Eclipse, Chrome, Visio and Word) were frequently used. The second one is that the model can directly grow based on Big Five, from which we could infer that the user may appear open and extravert. The third way relies on the combination of multiple aspects to generate new feature at a higher level. The combination of sport and habit showed that he liked playing badminton in weekend, while seemed an unhealthy life in the weekday.

In this case study, an actual user's personal data was collected for one month to establish his own Cyber-I model using our series of modeling approaches. The system prototype and the case study were to implement and verify our growing model mechanisms, which are quite different from traditional user modeling. We try to represent a user's model in a vivid manner and put emphasis on the process of how the model grow. Although our prototype modeling can dynamically show different growth processes from initial model, the modeling is not yet to approximate to the user always as expected due to relatively simple abilities in processing data and trait inference.

VII. Conclusions

A This research has been focused mainly on growable Cyber-I modeling based on increasing personal data. Three basic initial models of IM-C, IM-S and IM-P are generated from the Cyber-I core data, a user's Facebook profile, and preference choices, respectively. We proposed three modeling mechanisms for Cyber-I models to become bigger, higher and closer. A system prototype was built to manage personal data, and models' initializations and growths. A case study was conducted to concretely show how models were initialized and grown. Though our study has shown the basic ability for Cyber-I models to grow up, much research work still remains. Our future work is mainly focused on continuing to verify and improve the current growable model with more cases study.

References

[1] N.Y. Yen, J. Ma, R. Huang, Q. Jin, and T.K. Shih, "Shift to Cyber-I: Reexamining personalized pervasive learning," Proc. of the 3rd IEEE/ACM Int'l Conf. on Cyber. Physical and Social Computing, December 2010, pp. 685-690.

[2] J. Wei, B. Huang, and J. Ma, "Cyber-I: Vision of the individual's counterpart on cyberspace," IEEE International Conference on Dependable. Autonomic and Secure Computing, 2009, pp. 295-302.

[3] J. Ma, J. Wen, R. Huang and B. Huang, "Cyber-Individual meets brain informatics," IEEE Intelligent Systems, vol. 26, pp. 30-37, October 2011.

[4] V. Marco, B. Nadia, and E.Z. Elod, "A survey on user modeling in multi-application environments," The 3rd International Conference on Personalized Mechanisms, Technologies and Services, 2010, pp.111-116.

[5] B. Peter, S. Sosnovsky, and O. Shcherbinina. "User modeling in a distributed e-learning architecture," The 10th International Conference on User Modeling, Springer, 2005, pp. 387-391.

[6] Yaoxue Zhang, Yuezhi Zhou. Transparent Computing: A New Paradigm for Pervasive Computing, in Proc. of the 3rd Int'l Conf. on Ubiquitous Intelligence and Computing (UIC06), LNCS 4159, pp.1-11, Sept. 2006.

[7] T. Ke, A. Fabian, G. Qi and H. Geertjan, "TUMS: Twitter-based user modeling service," ESWC 2011 Workshops, Springer, 2012, pp.269-283.

[8] H. Dominik, S. Tim and B. Boris, "GUMO - The general user model ontology," The 10th International Conference on User Modeling, Springer, 2005, pp. 428-432.

[9] J. Key, and B. Kummerfield, "Lifelong user modeling goals, issues and challenges," Proceedings of the Lifelong User Modelling Workshop at UMAP, 2009, pp. 27-34.

[10] S.L. Daniel, Q. Yang, and L. Li. "Lifelong machine learning systems: Beyond learning algorithms," AAAI Spring Symposium on Lifelong Machine Learning, 2013.

About Authors:

Mr. Song Zhang Graduated in 2012 from Northeast Normal University, China and received his MSc from the graduate school of Computer and Information Sciences, Hosei University, Japan. His research is focused on Cyber-I growable model.

Dr. Jianhua Ma is a professor in the Faculty of Computer and Information Sciences, Hosei University, Japan. His research fields include Cyber Intelligence, Human Modeling, Ubiquitous Computing, Social Computing,

Dr. Runhe Huang is a professor in the Faculty of Computer and Information Sciences, Hosei University, Japan. Her research interests include Artificial Intelligence, Data Mining, and Human Modeling.

Dr. Dongming Chen is an associate professor in the Software College in Northeastern University, China. His research areas include complex networks, social network analysis and information security.