Evaluation of a Similarity Search Method for Human Behavior (Extended LDSD)

Zhang Zuo, Hung-Hsuan Huang, and Kyoji Kawagoe

Abstract—**Although large amounts of data on human behavior can be sensed, owing to rapid progress in sensor devices, the data collected are difficult to analyze and reuse. In a previous paper, we proposed a new semantic similarity measure, called Extended Linked Data Semantic Distance (Extended LDSD), for more effective mining of human behavior data. Extended LDSD can process various aspects of human behavior.**

In this paper, we conducted preliminary experiments in order to verify the previously proposed method. In the experiment, artificial data, generated by our automated test data generator, were used to simulate human behavior. T his study demonstrated that our proposed similarity search method (Extended LDSD) could provide higher similarity precision than the original LDSD and the Levenstein distance.

Index Terms—**Human Behavior Processes, Similarity Search, Linked Data, Artificial Data.**

I. Introduction

RECENTLY,the use of various kinds of sensor devices, such as motion and location sensors, has become widespread. As a result of this rapid popularization, huge amounts of data are being collected and analyzed for application-oriented purposes using advanced sensor devices. The sensor devices are currently used primarily to analyze and visualize object movements. However, efficient reuse of human behavior data is difficult, even when large amounts of data have been detected and processed in databases. In particular, when a certain behavior has been sensed, a human behavior-based service, such as E-learning, and human task support services have to be developed to search for behaviors similar to the sensed behavior.

In our previous paper, we proposed a novel similarity measure Extended LDSD [1] to obtain behaviors that are similar to a given behavior. Because a human behavior is complicated and needs to be represented from many aspects, existing similarity measures, such as graph distance and Euclid distance measures, cannot be directly used as a similarity measure of human behaviors. To overcome this issue, we previously extended an semantic distance calculation method, Linked Data Semantic Distance, LDSD, [10]. Our extended method, Extended LDSD, was developed to adapt to the semantic measurement of human behaviors from three standpoints: temporal, granularity, and content.

Although we compared our proposed extended method with other similarity methods in our previous paper, the comparison was made using a small data set. It is necessary to evaluate our method with more detailed quantitative experiments to accurately evaluate the validity of the proposed method. Therefore, in this paper, we provide a detailed evaluation of our method using larger amounts of test data. The following sections describe our experimental conditions and procedure, demonstrate our results, and draw conclusions. Because of the complexity of human behaviors, it is difficult to use actual human behavior data for our evaluation. Therefore, we generate a number of test data sets using our automated test data generator that can simulate many aspects of human behavior on the basis of our human behavior process model, MLPM [4].

The rest of our paper is organized as follows: Section II describes related work of our Extended LDSD. The Extended LDSD is briefly described in the Section III. Section IV and Section V give experiments and evaluation results of Extended LDSD. Finally, Section VI concludes the paper with future work.

II. Related Work

Work related to our proposed measure falls into two categories: behavior similarity and linked data distance. A number of definitions have been proposed for behavior similarity [2] [3]. Wang et. al. [3] proposed a similarity definition composed of two types of coupled similarities—intra-coupled and inter-coupled similarities. They define Intra-coupled Attribute Value Similarity (IaAVS) for the former, and Inter-coupled Relative Similarity based on Interaction Set (IRSI) for the latter. These similarities were primarily developed for multivariate time series and objects with multiple attributes. Therefore, these similarities cannot be used for human behaviors, even though they can be used as a component of similarity definitions for the behaviors, because they are represented by complex structures.

Neumuth et al. proposed a similarity definition exclusively for surgical process databases [2]. Their definition defines independent types of similarities, specifically, granularity, content, temporal, and transitional similarities, between two surgical processes. However, they developed the similarity metrics only for surgical process management. Further, they provide no clear instructions on how to combine these four independent metrics into a single similarity value. As regards multimedia data, much research has been conducted on similarity between both video data and audio data. One such study used Principal Component Analysis (PCA) to reduce dimensionality and extract independent aggregate dimensions [7]. However, although the PCA approach is effective for dimensionality reduction of multidimensional data, the similarity cannot be applied to structural or linked data, the predominant characteristics of human behaviors.

Zhang Zuo is a student of Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, JAPAN. (corresponding author, e-mail: gr0186rk@ed.ritsumei.ac.jp).

Hung-Hsuan Huang is an associate professor of Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, JAPAN. (e-mail: huang@fc.ritsumei.ac.jp).

Kyoji Kawagoe is a professor of Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, JAPAN. (e-mail: kawagoe@is.ritsumei.ac.jp).

Much research has also been conducted on similarity definitions of linked data and graph-based data [5] [6] [8] [9][10] throughout the past decade. The similarity definitions are classified into two types: graph matching-based [5] [9] and link-based [6] [8] [10]. In the graph matching-based approach, the similarity between two pairs of sub-graph nodes is first checked and then the similarity is calculated using the number of pairs. Conversely, in the link-based approach, the similarity between two nodes is calculated using the number of links between the nodes. The point of differentiation of the two approaches lies in whether the similarity is calculated for nodes or sub-graph nodes in a graph. Among the link-based similarity approaches, LDSD, proposed by Passant [10], is appropriate for complex and large linked data because it is simple to extend and easy to calculate, unlike other similarity approaches. we had chosen the method as our basic research.

III. EXTENDED LDSD [1]

A. Original LDSD [10]

Before introducing our proposed method called Extended LDSD, we briefly describe here the original LDSD. The objective of LDSD (Linked Data Semantic Distance) is to define a semantic distance between two nodes in Linked Data. As is well-known, Linked Data network can be abstracted into a graph, which consists of nodes and directed edges(Fig.1). Therefore, the LDSD actually calculates the semantic distance between nodes in a digraph.

Calculation of the linked data from LDSD is carried out as follows. A dataset is a graph G such as $G = (N, E, L)$, in which $N = \{N_1, N_2, ..., N_n\}$ is a set of nodes, $E = \{E_1, E_2, ..., , E_m\}$ is a set of typed links, and $L = \{L_1, L_2, ..., L_p\}$ is a set of instances of these links between data nodes, such as $E_i =< L_j, N_a, N_b >$. In this case, the semantic distance between nodes N_a and N_b, $LDSD(N_a, N_b)$ is defined as follows.

$$LDSD(N_a, N_b) = \frac{1}{1 + \alpha + \beta + \gamma + \delta} \tag{1}$$

$$\begin{cases} \alpha = \sum_i \frac{C_d(L_i, N_a, N_b)}{1+\log(C_d(L_i, N_a, N_n))} \\ \beta = \sum_i \frac{C_d(L_i, N_b, N_a)}{1+\log(C_d(L_i, N_b, N_n))} \\ \gamma = \sum_i \frac{C_{ii(L_i, N_a, N_b)}}{1+\log(C_{ii(L_i, N_a, N_n)})} \\ \delta = \sum_i \frac{C_{io(L_i, N_a, N_b)}}{1+\log(C_{io(L_i, N_a, N_n)})} \end{cases}$$

C_d is a function that computes the number of direct and distinct links between nodes in a graph G. $C_d(L_i, N_a, N_b)$ is equal to one if there is an instance of L_i from a node N_a to a node N_b, otherwise, it is zero. C_d can be used to compute the total number of direct and distinct links from N_a to N_b, which is defined as $C_d(L_n, N_a, N_b)$. Further, the total number of distinct instances of link L_i from N_a to any node $(C_d(L_i, N_a, N_n))$ can be defined and calculated.

C_{io} and C_{ii} are functions that compute the number of indirect and direct links, both outgoing and incoming, between nodes in a graph, respectively. $C_{io}(L_i, N_a, N_n)$ equals of 1 if there is a node N_n that satisfy both $< L_i, N_a, N_n >$ and $< L_i, N_b, N_n >$, 0 if not. $C_{ii}(L_i, N_a, N_b)$ equal 1 if there is a node N_n that satisfy both $< L_i, N_n, N_a >$ and $< L_i, N_n, N_b >$, 0 if not.

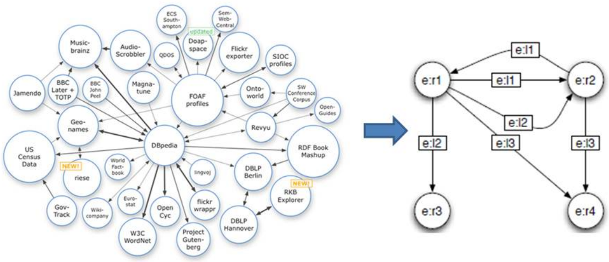

Fig. 1. Linked Data Semantic Distance. [10]

B. Similarity of Human Behaviors

In order to describe our Extended LDSD, we need to clear the characteristics of human behaviors. Human behavior processes can be compared from various aspects, such as content, granularity, and temporal relationship. content relationship refers to the situation of the behavioral components, including processes, tasks, and activities. Granularity relationship refers to the level of the hierarchical structure representing a specific process. Temporal relationship is the relationship among the behavioral components in the temporal dimension. The similarity of human behaviors can be calculated using these aspects.

In our previous study [4], we proposed a Multi-Layered Process Model (MLPM) that describes various people's behaviors and represents various kinds of processes in an integrated manner. In MLPM, behaviors are decomposed into three layers: process/task layer, activity layer, and action layer. Using this layered architecture, the overall processes of human behaviors, from the higher abstract level to the lower actual motion level, can be described. A simple example of MLPM is shown in Fig. 2.

As stated above, the human processes described by MLPM can be represented as hierarchical as well as linked structures. Specifically, a human process can be decomposed into activities, further decomposed into actions, and finally represented by various kinds of expressions that can express detailed contents of actions. Moreover, there are always associations between behavioral components. Although various methods to calculate the similarity of human processes exist, they cannot express the above two important features defined in our MLPM. Therefore, a more appropriate similarity search method is needed and we proposed in the previous paper.

C. Human Behavior Similarity Search

It is assumed that human behavior can be represented by a graph G such as $G = (N, E, LN, LE)$. In this graph, N is a set of typed nodes $N = \{N_1, N_2, ..., N_n\}$ and is categorized with node types LN, such as Process, Task, Activity, and Action [4]. Therefore, $N = NT_1 \cup NT_2, ..., \cup NT_{nLN}, NT_i \cap NT_j \neq 0$, where NT_k is a subset of nodes whose node type is LN_k and nLN is the number of node types. E

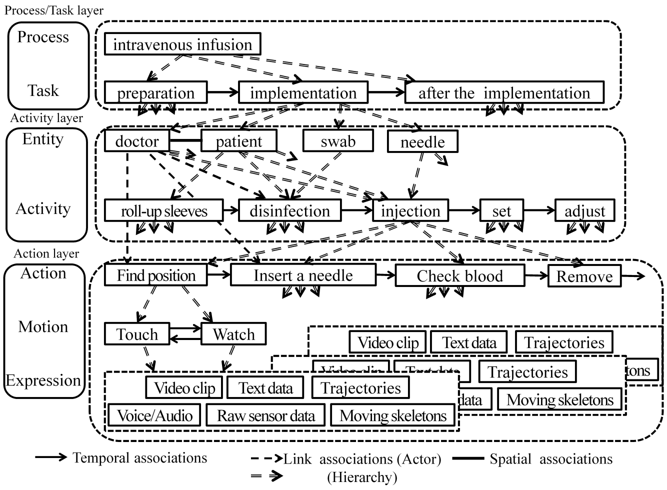

Fig. 2. MLPM example in intravenous injection process. [4]

is a set of typed links $E = \{E_1, E_2, ..., E_m\}$ and is also categorized with the abstract five link types LE. The link types are composed of the following link types: regular-link RL, hierarchical-link HL, temporal-link TL, spatial-link SL, and virtual-content-link VL. There are many concrete link types for each abstract link type. We simplify each concrete link type as $LE = \{L_{l,k}\}$, where l indicates the abstract link type and k is the k-th link type in the abstract link type L_l. An instance E_i of link type $L_{l,k}$ in G is represented by $< L_{l,k}, N_a, N_b >$, where N_a and N_b are the starting node and the end node of the link E_i , respectively.

The similarity search of human behaviors is defined and conducted to obtain the set of human behaviors similar to a given human behavior. That is, given a node N_q with node type $LN_k \in LN$, and a similarity threshold δ, find similar nodes $NS = \{N_i \in N\}$ to node type LN_k such that $HBS(N_q, N_i) \geq \delta$ or $HBD(N_q, N_i) \leq \delta$. HBS and HBD are the similarity and distance functions of two nodes with the same node type. In order to calculate HBD, we proposed extended LDSD.

D. Basic Concepts on Human Behavior Similarity

Using the concept of Linked Data network in [10], we can build a human behavior linked data network comprising nodes and the links between them. Based on MLPM, a human behavior linked data network is defined with seven kinds of nodes: process, task, entity, activity, action, motion, and expression. Human behaviors are first defined into processes–the abstract view of human behaviors. The processes are then decomposed in order to define the tasks comprising each process. Then, entities appearing in the behaviors are defined, followed by definition of activities. Further, each activity is decomposed into a set of actions, and each action decomposed into a set of motions. Finally, nodes of expressions are constructed to represent concrete motions.

There are also various kinds of links between the above nodes in the network. For example, for each constructed MLPM node, there are links between the process and its activities, which can be classified hierarchically. Further, each activity is composed of actions and can be described by various kinds of expressions. Thus, it is clear that there are many hierarchical links between the various types of nodes, such as activity nodes and action expression nodes. In addition, a kind of virtual-content-link is defined in our extended LDSD. When two nodes both have a hierarchical-link with one or some of the same nodes, they are linked by a virtual-content-link that represents the deep association between them in the content. In addition to these two kinds of links, there are other kinds of links, such as temporal-link, granularity-link, content-link, representing associations between nodes.

E. Extensions to LDSD

As stated in the previous section, human behaviors can be represented by a group $G = (N, E, LN, LE)$. In our extended LDSD, N is a set of network nodes $N = \{N_1, N_2, ..., N_n\}$, LN is a set of node types, specially in our method $LN = \{process, activity, expression\}$, E is a set of links $E = \{E_1, E_2, ..., E_m\}$, and in our method $LE = \{hierarchical - link, temporal - link, granularity - link, content-link, virtual-content-link\}$. An instance E_i of link type $L_{l,k}$ in G is represented by $< L_{l,k}, N_a, N_b, >$, where N_a and N_b are the starting node and the end node for link E_i, respectively. Further, T is defined as a set of possible times or time intervals. For each node N_i an effective time or time interval $T_i \in T$ is assigned.

The purpose of our extensions is to find the similarity between entities such as processes, activities, actions. There are three types of extended distance definitions: temporal extension, granularity extension, and content extension, which all are extensions to the original LDSD. In original LDSD, whether there is a link between two specific nodes or not is concerned. But in our Extended LDSD, in each extension we also concern the distance of each kind of links mentioned above.

$ExtendedLDSD(N_a, N_b)$is defined as follows.

$$ExtendedLDSD(N_a, N_b) = \frac{1}{1 + \alpha + \beta + \gamma + \delta + \tau + \chi + \psi} \tag{2}$$

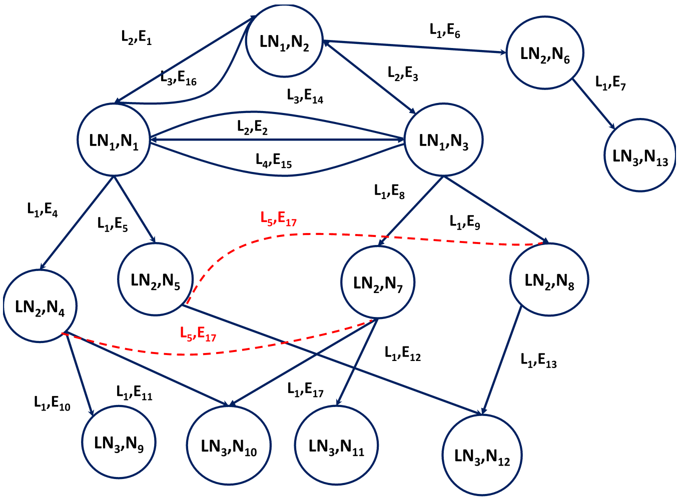

Fig. 3. Extended LDSD example. [1]

whereα, β, γ, and δ are defined in the previous section. τ, χ, and ψ are defined below.

1) Temporal extension to LDSD
In this extension, the distances τ of temporal-link are calculated using the time interval assigned to nodes.
$$\tau = |T_{N_a} - T_{N_b}|$$
2) Granularity extension to LDSD
In this extension, the distances χ of granularity-link are calculated by comparing the number of hierarchical-link linked to nodes.
$$\chi = |C_d(E_i, N_a, N_n) - C_d(E_j, N_b, N_m)|$$
$C_d(E_i, N_a, N_n)$ is the total number of direct instances of all hierarchical-link from N_a to any node N_n.
3) Content extension to LDSD
According to the hierarchical structure of the human behavior linked data network, we can build the virtual-content-link on the network. For example, in Fig. 3, node N_4 and node N_7 are not directly linked, but they linked to the same node N_{10}, thus it is considered that there is an association between them.
Firstly, We define a value ω to describe the tightness of this kind of association between N_c and N_d as follow.
$$\omega = C_{io}(E_i, N_c, N_d)/C_{io}(E_i, N_c, N_n)$$
Secondly, we set a threshold value Θ to check whether a virtual-content-link can be constructed or not. That is, if the ω of two nodes is greater than Θ which means the indirectly association between them is strong , then, a virtual-content-link will be constructed.
Finally, due to the increase of virtual-content-link, the human behavior network can be rebuilt. The virtual content distance between two upper lever nodes N_e and N_f in this situation is defined as follow:
$$\psi = \sum_i \frac{C_{io}(E_i, N_e, N_f)}{1+log(C_{io}(E_i, N_e, N_n))}$$

IV. EXPERIMENTS

A. Experiment conditions

In our experiments, artificial data instead of real human data is used, because of the complexity of human behaviors. In order to generate sufficient artificial data, we developed an automated test data generator. As stated in the previous section, human behaviors can be represented by MLPM process model, which is consisted of nodes and links between nodes. The automated test data generator can automatically generate many artificial test data sets to simulate human behavior according to the MLPM model.

We defined human behavior into tetrad groups, e.g., G-tetrads, based on our MLPM. Many nodes and edges between two nodes are generated according to these definitions. The process nodes consist of an activity node, a content node, and links between them. Processes in our data generator are all stored in the form of independent text documents. Characters or strings in the document represent activity and context nodes. As the MLPM model is a graphical structure, links between nodes are stored in the form of an adjacent matrix, according to the documents. Our automated test data generator consists of a series of function classes, such as deformation, I/O, and link matrix refresh.

We implemented a method of deforming processes in order to generate a large amount of data. Our data generator inputs a collection of manually constructed processes. The outputs are a series of varying data. In the deformation, some predefined variation functions, such as add nodes, cut nodes, and change nodes , are used. Moreover, the deformation process can automatically and reproducibly be performed for quantitative and subjective evaluations.

B. Experiment procedure

Our experimental procedure is composed of two stages: the data selection stage and the evaluation stage. In the first stage, our data generator creates multiple data sets, called patterns. In the second stage, the similarities between two processes are calculated for each pattern. The details of these stages are described below.

C. Data selection experiment stage

In this stage, we prepare NP process nodes as the original processes in each experiment. NP is the number of the original processes, and it is set to four, through our experiments. For each original process, 12 variations of processes are generated using our data generator. Therefore, 12*NP (=48) processes are totally generated as a test data set, called a pattern. The number of nodes in one original process is set to 260 (52 activity nodes, 208 content nodes).

We introduced three parameters to control the generator, to generate various kinds of processes that are similar to the original process. These three parameters are the degree of similarity to the original process (P1), the deformation degree (P2), and the number of deformations (P3). The ranges of these three parameters are shown in Table I. P1 defines how similar a given original process is to another given original process on average. A larger P1 value increases the similarity of the processes.

Our data generator has three types of variation functions: cut, add, and change. We chose cut as the variation function in this experiment. P2 defines what percentage of the nodes should be removed from the original process: 1/2 means to cut half of the nodes, and zero means to cut nothing. P3 defines the number of times the deformation should be conducted. For example, P2 = 1/2, P3 = 4 means the deformation type ‘cut half’ is conducted four times in the deformation.

TABLE I
RANGE OF PARAMETER

	P1	P2	P3
Range	25% ∼ 75%	0% ∼ 50%	0 ∼6

TABLE II
DATA SET

	P1	P2	P2	P2	P2	P3
Pattern1	50%	0	1/32	1/16	1/8	3
Pattern2	75%	0	1/32	1/16	1/8	3
Pattern3	50%	0	1/32	1/8	1/4	3

Finally, we selected three data patterns as our evaluation data, as shown in II.

D. Evaluation stage

As stated previously, our proposed method, Extended LDSD, includes three extension types, temporal, granular, and content . In order to clearly demonstrate the effect of each extension on the results, four types of evaluation experiments were conducted using the data sets selected in the previous stage. Experiments for the three extensions were independently conducted. The final experiment used the combined extensions. We used the original LDSD and the Levenshtein Distance as existing distance measures to compare with Extended LDSD.

Using distance matrices, the data set, which included 48 processes, was classified into NP(=4) categories. The classification precision of each method was calculated as an evaluation measure. We used the classification precision to reflect the precision of the similarity. Because of the high classification precision and the simplicity of the operation, we chose K-means as the experimental classification method. In our experiment, accuracy is defined as (number of successful classifications) / (number of classification trials). The number of trials equals the number of processes in one test data set.

Specific experimental steps are shown below.

1) In the temporal extension experiment, a time interval was assigned to each process of the data set, according to the definition in the previous section. Time intervals are random data divided into four categories.
2) In the granularity extension experiment, the granular similarity between processes was evaluated. We counted the number of links between processes and activities and added them into the distance matrices.
3) In the content extension experiment, the relationships between the activity and content were considered. Using these relationships, the process constructions were rebuilt. In our experiment, the activity-content matrix was prepared beforehand in four categories.
4) In the combined experiment, we made a linear combination of all the above extensions to discover the impact of the combination precisely.

V. Evaluation

In this section, we evaluate the extensions presented in the previous section.

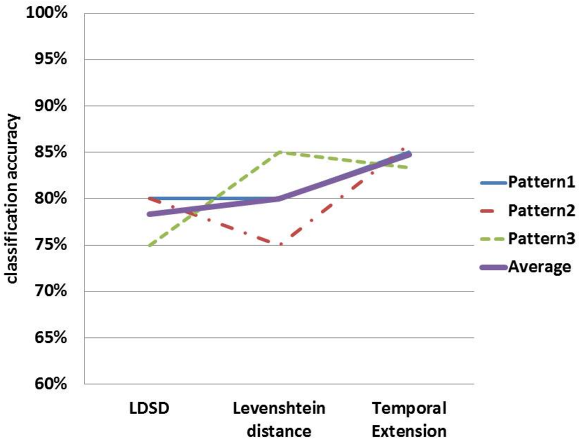

Fig. 4. Evaluation of temporal extension to LDSD

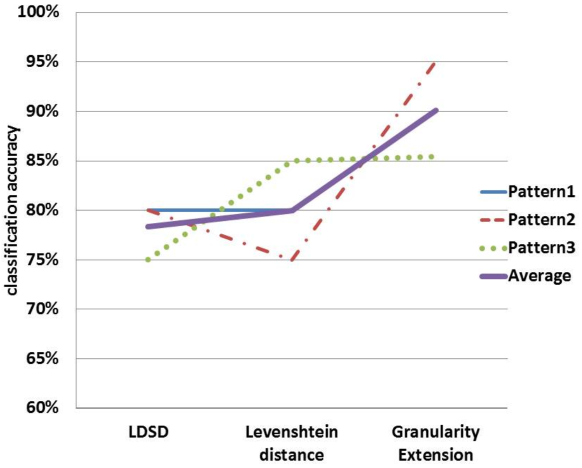

Fig. 5. Evaluation of granularity extension to LDSD

A. Evaluation of temporal extension to LDSD

The results of the temporal extension evaluation are shown in Fig. 4. The temporal extension shows a significant improvement of about 10%.

B. Evaluation of granularity extension to LDSD

The results of the granularity extension evaluation are shown in Fig. 5. The granularity extension has an 85% average classification accuracy, which is better than the other methods and very consistent.

C. Evaluation of content extension to LDSD

The results of the content extension evaluation are shown in Fig. 6. The content extension shows a high classification accuracy, particularly in Pattern 1.

D. Evaluation of extended LDSD

The results of the entire combination are shown in Fig. 7. This shows that our proposed method, Extended LDSD, has the highest classification accuracy with the combination of temporal, granularity, and content extensions.

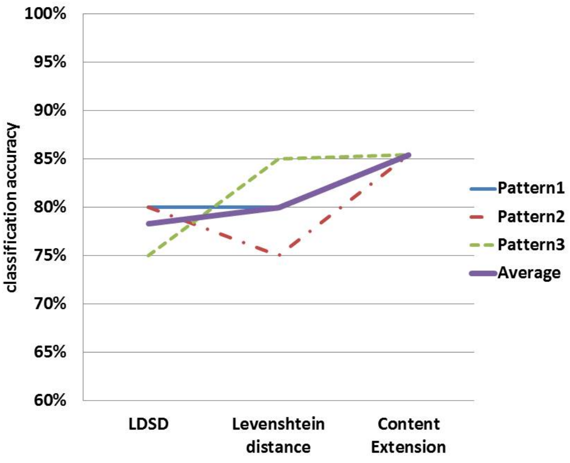

Fig. 6. Evaluation of content extension to LDSD

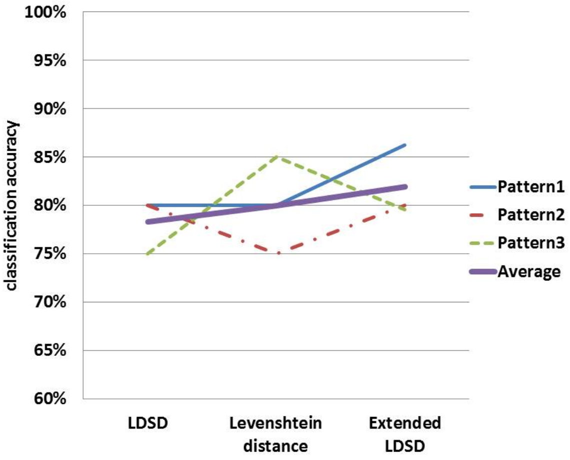

Fig. 7. Evaluation of extended LDSD

E. Discussion

Through the above figures, we found that our proposed extension has higher similarity precision in varying degrees, and is superior to the other semantic similarity methods. Because we joined the human behavior elements of time, granularity, and content into the semantic similarity calculations, we also confirmed that our Extended LDSD shows promise in its measurement of human behavior.

However, some issues were not clarified in our experiment. Although we verified the superiority of the three extensions, we found that the weights used to combine the extension values in the combined experiment can have a significant influence on the similarity precision. Further investigation is required to clarify the association between the weights and the similarity precision.

VI. Conclusion

In this paper, we presented the evaluation of our previously proposed human behavior similarity search method, Extended LDSD. For the experiment, we developed an automated test data generator to generate artificial data to simulate human behavior processes. In addition, using data sets created by the generator, we evaluated the validity of each extension and their combination, and we compared the results with other similarity methods. Finally, the evaluation experiments showed the superiority of our method in terms of similarity precision. Our future work will include more experiments to improve our method and investigate its application to real human activities.

Acknowledgment

This work was partially supported by MEXT-Supported Program for the Strategic Research Foundation at Private Universities, 2013-2017.

References

[1] Z. Zuo, H. H. Huang and K. Kawagoe. Similarity Search of Human Behavior Processes Using Extended Linked Data Semantic Distance. ISSASiM2014, the 4th DEXA Workshop on Information Systems for Situation Awareness and Situation Management,2014. p.178-182.

[2] N. Thomas, L. Frank and J. Pierre. Similarity metrics for surgical process models. Artificial intelligence in medicine, 2012. 54.1: 15-27.

[3] C. Wang, L. B. Cao and M. C. Wang. Coupled nominal similarity in unsupervised learning. Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011. p. 973-978.

[4] Z. Zuo, H. H. Huang and K. Kawagoe. MLPM: A Multi-Layered Process Model Toward Complete Descriptions of People's Behaviors. eKNOW 2014, The Sixth International Conference on Information, Process, and Knowledge Management, 2014. p. 167-172.

[5] G. Brian. Matching structure and semantics: A survey on graph-based pattern matching. AAAI FS, 2006, 6. p. 45-53.

[6] R. John, H. Kathleen and D.V. Denise. A unifying semantic distance model for determining the similarity of attribute values. In: Proceedings of the 26th Australasian computer science conference-Volume 16. Australian Computer Society, Inc., 2003. p. 111-118.

[7] K. Y. Yang. A PCA-based similarity measure for multivariate time series. In: Proceedings of the 2nd ACM international workshop on Multimedia databases. ACM, 2004. p. 65-74.

[8] H. Michael, K. Yannis and T. Yannis. Similarity-based Browsing over Linked Open Data. arXiv preprint arXiv,2011. 1106.4176.

[9] J. W. Zhong et al. Conceptual graph matching for semantic search. Conceptual structures: Integration and interfaces. Springer Berlin Heidelberg, 2002. p. 92-106.

[10] P. Alexandre. Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010.

[11] A. Tatsuya, F. Daiji and T. Atsuhiro. Approximating Tree Edit Distance Through String Edit Distance. Fundamenta Informaticae, 2010. p. 157-171.