TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.5, May 2014, pp. 3297 ~ 3302
DOI: http://dx.doi.org/10.11591/telkomnika.v12i5.4949

Research on Operation-based Correlation in Personal Dataspace

Shuo Jiang, Jiajin Le, Yefeng Li
DongHua University,
No.2999 North Renmin Road of Shanghai, 0086-21-62378595-11 of DongHua University
Corresponding author, e-mail: iamprotoss@163.com

Abstract

Operation of user was defined. The weight of operation was expressed. The variable quantity of user behavior was computed by weight. 3-ary vector data definition was expanded. Data item was defined by 4-ary vector in personal dataspace. Correlation of data for user was defined by weight. Current weight of data was defined by initial weight and variable quantity of user operation. A library dataspace model was designed. The weights of data were verified by using a sample in the library dataspace for ten days. The result proved the correlation of data was very important and useful in personal dataspace.

Keywords*: personal dataspace, data item, correlation.*

1. Introduction

Information technology and internet technology with a rapid development show us a huge data amount, variety data and more closely data relationships [1]. The new features of data make people in trouble on data management and that is a great challenge for researchers about data management. People represent a new theory of data management named dataspace to face the challenge [2]. Dataspace is a set of data which relate with subject, and all the data in dataspace can be controlled by subject [3]. Dataspace is the focus in current data management technology research, there are many achievements about dataspace, such as data model [4-6], index [7], data integration [8, 9] and prototype system [10].

Dataspace is proposed for solving data integration problem, but dataspace doesn't have strict schema, the data in dataspace are heterogeneous and are saved in distributed data resources. Therefore, personal dataspace must be able to judge data correlation for user and monitor personal dataspace to catch correlation changing for keeping data or not. The step of judging correlation is shown in Figure 1.

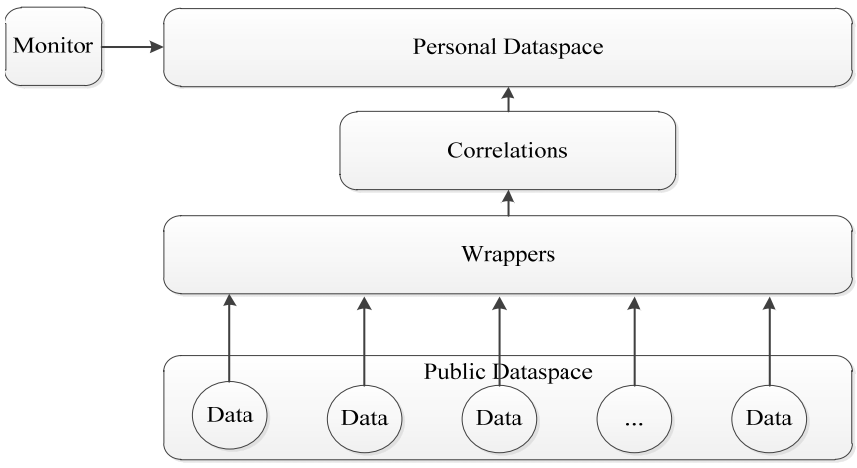

Figure 1. Judging Correlation of Personal Dataspace

Received October 16, 2013; Revised November 27, 2013; Accepted December 16, 2013

This paper researches dataspace integration which is basis of building a dataspace. User's operation and weight of each operation are defined. Operation-based variable quantity is defined by weight of operation. 3-ary vector is researched and expanded. We propose a 4-ary vector by 3-ary vector to describe data item in personal dataspace. This paper also defines correlation of data by weight of 4-ary vector. A library dataspace model is designed and an experiment base on this library dataspace model verifies the operation-based data item correlation by weight. The result from the experiment shows the correlation of data item is very important for building a personal dataspace.

2. Personal Dataspace Correlation

The personal dataspace integrates data for user and all data items in this personal dataspace have correlations for the user. This personal dataspace needs to compute correlation between data and user to ensure that the data items in personal dataspace are associated with user before saving, and avoid useless data in personal dataspace. It ensures the value of personal dataspace. Corespace model which is based on Vertical data model represents core space thought [11], Corespace gets a core space for user by threshold of data correlation, and correlations of data in this core space are very high. It proves the data in this core space are very important for user. It is shown in Figure 2.

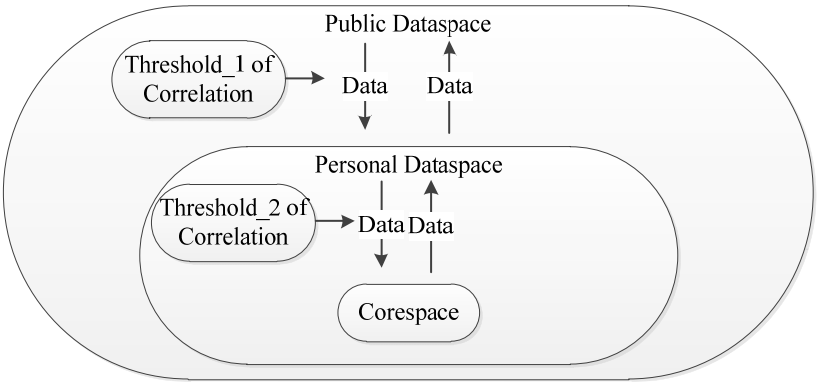

Figure 2. Corespace of Personal Dataspace

Most of correlation research for personal dataspace assumes correlation evaluation method exists like corespace and research on it without mentioning how correlation evaluates. We show a way to evaluate correlation.

Definition 1: Dataspace S = {d_1, d_2, …, d_m}, d is data item in this dataspace, m is number of data items. User operation set for dataspace S is A = {a_1, a_2, …, a_n}, a is operation of user for every time, n is number of user operations.

Definition 2: Each user's operation will produce weight set of operation, the weight set is V = {v_1, v_2, …, v_m}, v is weight of data item for operation. The number of items in V is same with the number of items in S.

Definition 2 means each operation of user will produce weight for each data item in dataspace.

Equations 1. The variable quantity after j times of operation for data item i in dataspace S is shown in Equations 1.

$$V_c = v_{i1} + v_{i2} + ... + v_{ij} = \sum_{i=1}^{j} V_i \quad (1)$$

In Equations 1, there are $1 \le i \le m$ and $1 \le j \le n$. V_i is positive when user has actions on the data, V_i is negative when user has actions on other data.

Definition 3: It takes a 3-ary vector to describe each attribute of data in dataspace. The 3-ary vector is defined as (ObjectID, AttributeName, AttributeValue). ObjectID is identification of data object, AttributeName is a set which contains the names of all attributes, and AttributeValue is a set of the values of all attributes [12].

The definition 3 expressed data in dataspace simply, but it did not show the relationship between data and user, we expand definition 3:

Definition 4: We take a 4-ary vector to describe data as (ObjectID, AttributeName, AttributeValue, Weight), ObjectID, AttributeName and AttributeValue are the same to definition 3. Weight was presented to express correlation, and the weight will change by user changing or changing of data itself.

Equations 2. The correlation of data for user is shown in Equations 2.

$$W_d = W_0 + V_c = W_0 + \sum_{i=1}^{j} V_i \quad (2)$$

W_d is current weight of data, i.e. data correlation. W_0 is initial weight which is given when the data into personal dataspace first time. V_c is the variable quantity of user actions.

We save data by threshold of weight which has set already in personal dataspace, and the data is important data for user if weight is over the threshold of correlation. Changing of user or changing of data will result in the weight changing between user and data. We will remove the data which weight is lower than threshold to keep the high correlation in personal dataspace of user. It is shown in Figure 3.

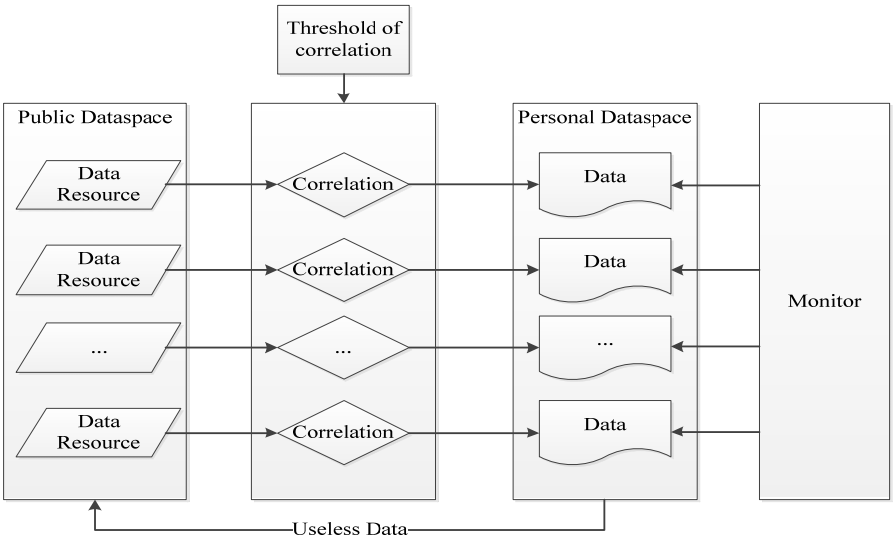

Figure 3. Data Flow for Personal Dataspace

We will give initial weight of data item when we extract information from data resources first time, and take this data item into personal dataspace by the threshold of weight when the weight of this data item is over the threshold. A monitor will monitoring the changing of all data items in personal dataspace and judge every data item in personal dataspace is useful or useless.

3. Research Method

We use some articles in the library as a test data, and use of Oracle10g database as experimental data storage environment. The library dataspace model is shown in Figure 4.

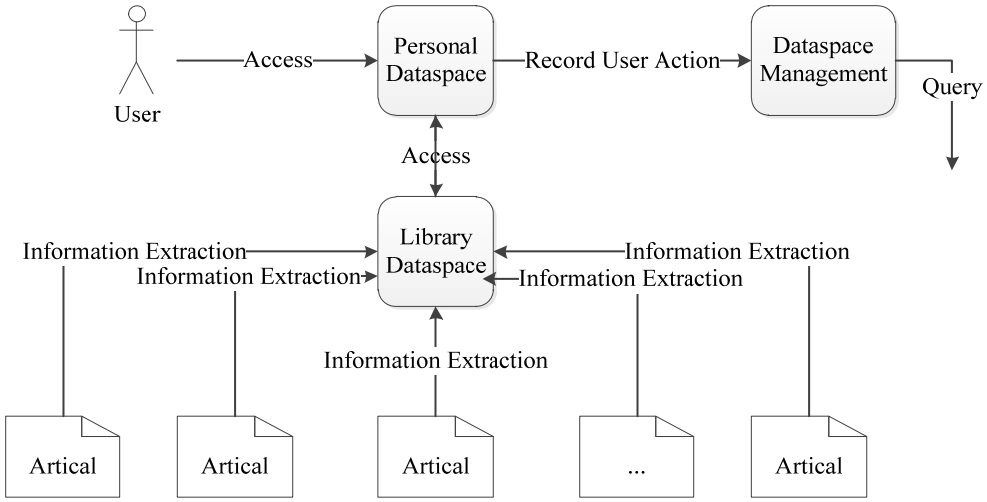

Figure 4. Library Dataspace Model

Figure 4 shows a dataspace model for library. Library dataspace extract information from articles to create a big dataspace, personal dataspace of user will access library dataspace to catch information for user when a user have operations to the personal dataspace. Dataspace management records user's actions to support user behavior analysis or compute weight of data item.

4. Results and Analysis

We select an article A which is published in 2010 in the library dataspace as a sample. We choose a reader to build personal dataspace and record this reader's behavior to compute this reader's variable quantity of operation for correlation in the first day. Let the initial weight is 1, the operation on first day is shown in Figure 5.

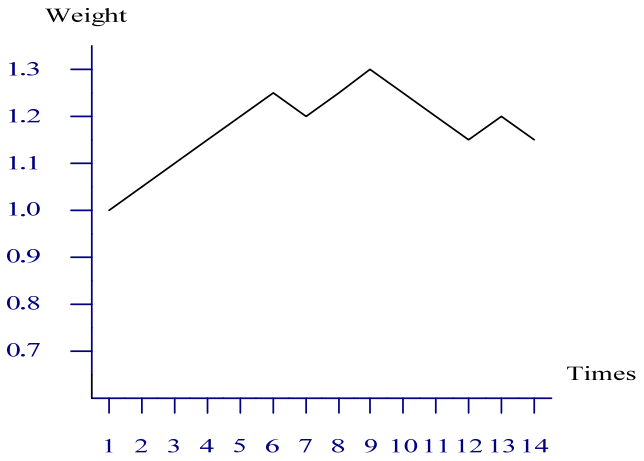

Figure 5. Variable Quantity of Operation for Article A on the First Day

The reader accesses library 14 times on the first day. Figure 5 shows reader operations on article A are 9 times and operations on other data are 5 times. We can find the Variable quantity of reader operations is 0.15 by computing with Equations 2. Therefore, the weight of reader behavior on first day is 1.15.

We evaluate correlation of article A for personal dataspace of a reader by weight of article A and record of the reader who access to article A in ten days. Let the initial weight is 1, the threshold of weight is 0.8, we compute weight of every day by Equations 2, the result is shown in Figure 6.

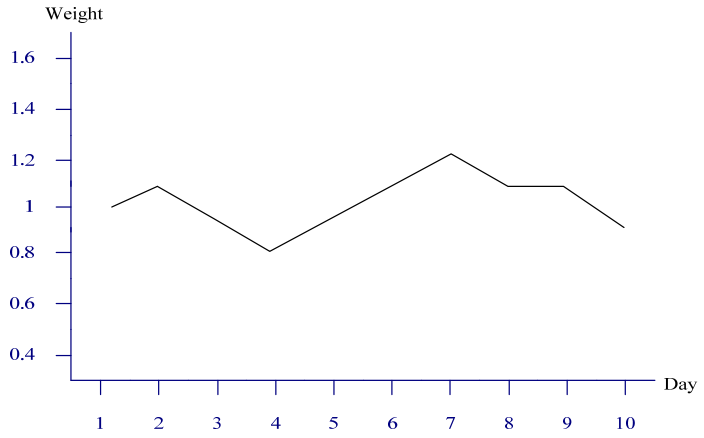

Figure 6. Weights of Article A for a Reader in Ten Days

Figure 6 shows this reader accessed article A many times from the first day to the second day, but article A was not read by this reader from the second day to the fourth day instead of researching other data. The reader accessed article A from the fourth day to the seventh day, and never touched article A after the eighth day, even without accessing the library from the eighth day to the ninth day. It lead the weights of article A for this reader have no change in this period. Fig.5 shows the article A is very important for this reader in the ten days, and the article A must be the data in personal dataspace of this user in the ten days.

5. Conclusion

The new features of data make people research on dataspace, and there have been many achievements in several aspects of dataspace already. In this paper, we researched dataspace integration and pointed out the necessity of correlation. Behavior of user was defined and operation-based variable quantity was expressed by weight of operation. 3-ary vector was expanded and 4-ary vector was proposed by 3-ary vector to describe data item and correlation of data. A library dataspace model was designed and we verified the correlation by this model. The result shows the correlation of data is very important and useful in personal dataspace.

References

[1] Meng XF. *From Database to Dataspace, From Enterprise to People.* School of Information, Renmin University of China. 2006.
[2] Franklin M, Halevy A, Maier D. From databases to dataspaces: A new abstraction for information management. *ACM SIGMOD Record*. 2005; 34(4): 27-33.
[3] Jones W, Bruce H. *A report on the NSF-sponsored workshop on personal information management*. Seattle. 2005.
[4] Dong X, Halevy A. *A platform for personal information management and integration.* Proceedings of the 2nd Conference on Innovative Data Systems Research. Asilomar. 2005: 119−130.
[5] Dittrich JP, Antonio M. *iDM: A unified and versatile data model for personal dataspace management*. Proceedings of the 32nd Int'l conference on Very Large Data Bases. New York. 2006: 367−378.
[6] Karger DR, Bakshi K, Huynh D, Quan D, Sinha V. *Haystack: A customizable general-purpose information management tool for end users of semistructured data*. Proceedings of the 2nd Conference on Innovative Data Systems Research. Asilomar. 2005: 13−26.
[7] Dong X, Halevy A. *Indexing dataspaces*. Proceedings of the 27th Int'l Conference on Management of Data. New York. 2007: 43−54.

[8] Dong X, Halevy A, Yu C. Data Integration with Uncertainties. *The International Journal on Very Large Data Bases*. 2007; 18(2): 469-500.
[9] Sarma AD, Dong X, Halevy A. *Bootstrapping Pay-as-you-go Data Integration Systems.* Proceedings of the 2008 ACM SIGMOD international conference on Management of data. New York. 2008: 861-874.
[10] Blunschi L, Dittrich JP, Girard OR, Karakashian SK, Salles AV. *A dataspace odyssey: The iMeMex personal dataspace management system.* Proceedings of the 3rd Conference on Innovative Data Systems Research. Asilomar. 2007: 114−119.
[11] Li YK, Meng XF. *Exploring Personal CoreSpace for DataSpace Management.* Proceedings of the 2009 Fifth International Conference on Semantics, Knowledge and Grid. Zhuhai. 2009: 168-175.
[12] Li YK, Meng XF. *Research on personal dataspace management*. Proceedings of the 2nd SIGMOD PhD workshop on Innovative database research. Vancouver. 2008: 7−12.