Advanced Science and Technology Letters
Vol.58 (Software 2014), pp.38-43
http://dx.doi.org/10.14257/astl.2014.58.08

Similarity Analysis for the Prediction of Agent Behavior in Demographic Simulation

Eunjeong Choi, Changwon Ahn
Electronics and Telecommunications Research Institute,
128 Gajeongno, Yuseong-gu, Daejeon, Korea
{ejchoi, ahn}@etri.re.kr

Abstract. This paper describes similarity analysis for the prediction of agent behavior in demographic simulation. We use modified jaccard distance to calculate the similarity of agent. A Census Data of 407,359 people is analyzed and the similarity is calculated on the Census. The calculation time is 71 seconds and the top similarity is 75%.

Keywords: Demographic Micro Simulation, ABMS, Similarity Analysis

1 Introduction

Recently developed countries have studied to predict the composition and size of the population. They have been developed agent-based micro-simulation techniques. The state changes of birth, marriage, education, employment status have been determined using the mathematical function formulation or state transition tables. However, this is not considering individual characteristics of each agent. Also, this is a macro level rather than a micro-level simulation.

The purpose of this paper is to find the agent who has lived in the most similar environment by considering the properties of each agent based on Census Data and to use the state of the agent for the state transition of an agent. For this purpose, Census Data on 1990 year is analyzed and similarity of each agent is calculated. The results of experiments for similarity measure will be shown and analyzed.

2 Similarity Measures

There are the Euclidean distance, Cosine distance, Jaccard distance to obtain the similarity measure.

Euclidean distance[1]. measures the distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. In general, for an n-dimensional space, the distance is as follows.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_i - q_i)^2 + ... + (p_n - q_n)^2}.$$

ISSN: 2287-1233 ASTL

Cosine distance[2]. measures the cosine of the angle between two vectors of an inner product space.

$$Cosine\ Distance\ (A,B) = 1 - \frac{A \cdot B}{\|A\|_2 \cdot \|B\|_2}$$

Jaccard distance[3]. measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$Jaccard\ Distance\ (A,\ B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|}$$

3 Experiments

We use Census Micro Data on 1990 year. Because the raw data of Census Micro Data is hard to use for calculation of similarity, the data should be preprocessed. Then, similarity for every single agents is measured.

3.1 Preprocessing Census Micro Data

Table 1. shows the raw data of Census Micro Data on 1990 year. Each row in the table includes properties of an agent. We will calculated similarity with respect to the agent. Columns of raw Census are 32 but some columns are missing because of the size of a document.

Table 1. The raw data of Census Micro Data on 1990 year

C1	C2	C3	C4	C5	C6	C7	C8	C22	C23	C24	C28	C29	C30	C31
100001	1	01	2	45	3	4	1	1	1	1	23	3	0	3
100002	1	03	1	11	2	2	2	1	1			0	0	0
100002	2	03	2	17	3	4	1	1	1	9		0	0	0
100002	3	02	2	45	3	3	1	1	1	9	26	2	0	2
100002	4	01	1	44	3	4	1	1	1	9	25	0	0	0
100002	5	05	2	79	3	1		1	1	0	17	5	0	5
100003	1	21	1	28	2	4	1	1	2	1	26	0	0	0
100003	2	02	2	54	3	3	1	1	2	7	20	3	0	3
100003	3	22	2	23	3	4	1	1	3	7	21	1	0	1
100003	4	01	1	62	3	4	1	1	2	6	29	0	0	0

Table 2. shows The differences for properties between a sample agent and testing agents. Every single value is ‘0’ or ‘1’. ‘1’ means that two agents have same property. ‘0’ means that the properties between two agents are different.

Table 2. The differences for properties between a sample agent and testing agents

C29	C9	C31	C8	C6	C5	C24	C7	C4	C3	C23	C30	C22	C21
0	1	0	1	1	0	0	1	1	0	1	1	1	0
0	1	0	1	1	1	0	0	1	0	1	1	1	1
0	1	0	1	1	0	0	1	0	1	1	1	1	1
0	1	0	0	1	0	0	0	1	0	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	0	1	1	0	0	1	0	1	0	1	1	1
0	1	0	1	1	0	0	1	1	0	0	1	1	1
1	1	1	1	1	0	0	0	1	0	0	1	1	1

3.2 DMS Similarity Measure Algorithm

DMS Similarity Measure Algorithm is modified Jaccard distance. The formula is as follows. Intersection between agent A and agent B is divided by union of agent A and agent B, then multiply hundred for percentage.

$$DMS\ \ Similarity\ \ (A,B)\ =\ \frac{|A \bigcap B|}{|A \bigcup B|} \times\ 100$$

Program code to calcuating DMS similarity is as follows.

```
calculteDMSSimilarity()
for $i (0.. $#matrix)
      foreach $key (keys (%{$matrix[$i]}) )
            if($matrix[$i]{$key} ne "")
                  $num++;
                  if ($agent{$key} eq  $matrix[$i]{$key})
                        $sum += 1;
                        $similarity   =   100*($sum/$num);
                        if($top < $sum)
                              $top = $sum;
                              $top_i = $i;
                              $top_sim    =    $similarity;
                              $sum=0;
                              $num=0;
```


3.3 Environments

The program code for calcuating DMS similarity is developed by Perl 5 language. The experiments is performed on Intel Pentium 2 with Windows 7 operating system.

4 Result

The agent with top similarity has almost same properties with the sample agent. It seems two people have parallel life in the movie "Parallel Life".
Table 3. shows the comparison of calculation time for similarity measure, top similarities, and agent index with top similarity according to the number of agents. A Census Data of 407,359 people is analyzed and the similarity is calculated on the Census. The calculation time is 71 seconds and the top similarity is 75% as shown in Table 3.

Table 3. The comparison of calculation time for similarity measure, top similarities, and agent index with top similarity according to the number of agents

Agents	Time(sec)	Top Similarity	Index
10	1	52.94117647	3
100	1	64.70588235	93
200	1	70	150
300	2	70	150
500	7	70	150
1000	10	70	150
10000	23	75	2288
407359	71	75	2285

Fig.1. shows Similarity calculation time measures according to the number of agents as shown in Table 4. Similarity calculation time increases as the number of agents increase.

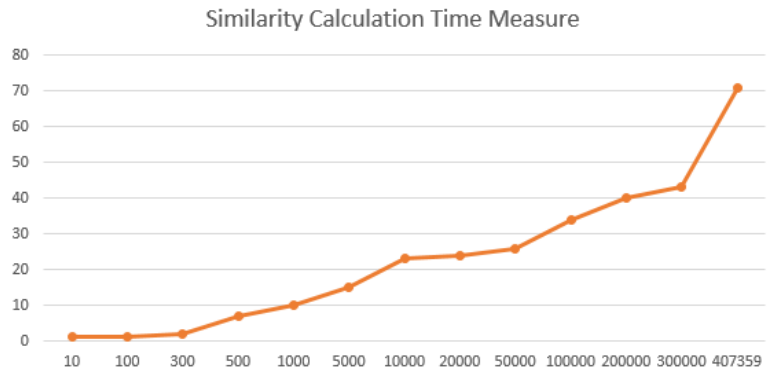

Fig. 1. Similarity calculation time measures according to the number of agents

Fig.2. shows top similarity values according to the number of agents. Top similarity increases as the number of agents increase. However, the top similarity values are not changed after the top similarity agent is found.

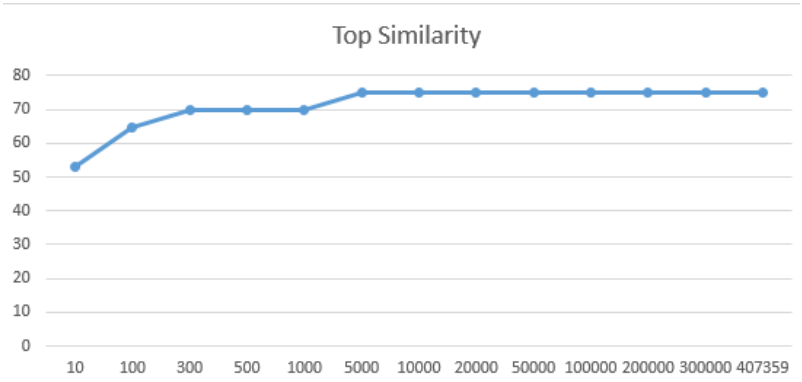

Fig. 2. Top similarity values according to the number of agents

5 Discussions

A Census Data of 407,359 people is analyzed and the similarity is calculated on the Census. The calculation time is 71 seconds and the top similarity is 75%. The agent with top similarity has almost same properties with the sample agent. It seems two people have parallel life in the movie “Parallel Life”.

In the future, we will develop more elaborated similarity measure for specific agent behaviors and apply to the demographic simulation system.

Acknowledgments. This work was supported by the ICT R&D program of MSIP/IITP. [10047117, Development of Distributed/Parallel Multi-Dimensional Demographic Micro Simulation Technologies for Population Dynamics and Socio-Economic Experimentation]

References

1. Elena Deza & Michel Marie Deza (2009) Encyclopedia of Distances, page 94, Springer.
2. P.-N. Tan, M. Steinbach & V. Kumar, "Introduction to Data Mining", , Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500
3. Simone Santini, Ramesh Jain, Similarity Measures, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 21, NO. 9, SEPTEMBER 1999

