Cluster Analysis of Railway Directory Inquire Dialogs*

Mikhail Alexandrov,[1,2] Emilio Sanchis,[2] Paolo Rosso[2]

[1] National Polytechnic Institute, Mexico
dyner1950@mail.ru
[2] Polytechnic University of Valencia, Spain
{esanchis, prosso@dsic.upv.es}

Abstract. Cluster analysis of dialogs with transport directory service allows revealing the typical scenarios of dialogs, which is useful for designing automatic dialog systems. We show how to parameterize dialogs and how to control the process of clustering. The parameters include both data of transport service and features of passenger's behavior. Control of clustering consists in manipulating the parameter's weights and checking stability of the results. This technique resembles Makagonov's approach to the analysis of dweller's complaints to city administration. We shortly describe the MajorClust method developped by Benno Stein's group and demonstrate its work on real person-to-person dialogs provided by Spanish railway service.

1 Introduction

In the recent years, much effort has been devoted to development of automatic dialog systems. This topic is well represented at the conferences related with Dialog Processing [6]. One of the first phases in the design of such systems consists in definition of the dialog domain, as well as the different types of dialogs to be conducted. The type of dialog depends on the information requested (which is reflected in lexical and semantic restrictions of the task), the type of user (novice or expert), etc. Usually this analysis is performed manually basing on a set of real person-to-person dialogs. Then the obtained classification is used to define the behavior of the dialog manager or to acquire a corpus of dialogs by using the Wizard of Oz technique [5].

In this paper, we consider another technique consisting in:

manual parameterization of dialog set;
filtering selected parameters;
objective clustering.

These steps correspond to the approach developed by Makagonov [3] for the analysis of letters and complaints of Moscow dwellers directed to Moscow city administration.

* Work done under partial support of the Government of Valencia, Mexican Government (CONACyT), R2D2 CICYT (TIC2003-07158-C04-03), and ICT EU-India (ALA/95/23/2003/077-054). The paper is a reprint of a paper published in Proceedings of TSD-2005.

The set of parameters is supposed to reflect both the specificity of a given public service and some features of a passenger. We show how to make correct scaling for quantitative or qualitative parameters of a dialog. Analysis of the parameter's distribution allows detecting and eliminating non-informative parameters. Clustering is applied both to dialogs themselves and to their parameters.

The procedure of clustering uses the MajorClust method recently developed by Stein *et al*. [7]. This method was selected because it proved to be more adequate to data and problem settings than other methods. We briefly describe it below.

In the experiments we used examples related with Spanish railway directory inquires, obtained in the framework of Basurde projects [1].

2 Parameterization

2.1 Examples of dialogs and their parameterization

The final purpose of dialog parameterization is to present all acts of dialogs in the form of the numerical matrix "objects/parameters", where the objects are dialogs themselves and the parameters are characteristics reflecting both railway service and passenger behavior. Such a matrix allows calculating the distance (similarity) between objects, which is the input data for any method of cluster analysis.

Table 1 shows the difficulties of parameterization (the records are translated from Spanish into English). Here *US* stands for a user and *DI* for a directory inquire service. This example concerns the train departure from Barcelona to the other destinations both near the Barcelona and in other provinces of Spain. Such limited dialogs constitute approximately 25% of total number of dialogs, other dialogs being 2 to 5 times longer. This dialog gives an impression on the difficulties of dialog parameterization.

Table 1. Example of real dialog between passengers and directory inqui80

DI: Renfe, good day	two hours to Valladolid
US: Yes, good day	*US*: Are there any more?
DI: Yes, well.	*DI*: No, on Thursday only this one, eh?
US: OK, could you inform me about the trains that go from here, from Barcelona to Valladolid?	*US*: Nothing more, say me, please.
	DI: Exactly.
	US: <CONTINUALLY> before the Wednesday or Thursday.
DI: What day it will be?	*DI*: The train will be exactly at evening, on Thursday or Friday it is off.
US: The next Thursday.	*US*: Thank you, bye
DI: Let us to see. <PAUSE> on Thursday is off the one at thirteen, which come at twenty	

One can note the following three features of such dialogs: many aspects concerning the trip are not reflected in a specific dialog; many characteristics are diffuse; and much information is presented in a hidden form. To take into account these circumstances, we use nominal scales with the value "indifference" and interval scales, respectively. All parameters are normalized to the interval [0, 1]. The parameters we initially introduced are presented below:

(1) City weight (*City*). This parameter reflects the economic importance of the city. Its possible values are 0.75, 0.5, 0.25, 0, reflecting large, middle, small, and local cities, respectively. The value 1 is reserved.

(2) Complexity (*Cx*). In our case, this binary parameter reflects the necessity of transfer.

(3) Urgency and definiteness (*U/D*). This numerical parameter is introduced to reflect the profile of passenger rather then the railway service. Its possible values are 1, 0.5, and 0: urgent departure at the same day, departure at a certain day during a week or month, and the indifference to the day of departure.

(4) Round trip (*T/F*). It is a binary parameter with obvious values 1 and 0.

(5) Time of departure (*T*). This parameter is presented in the form of three nominal scales with two binary values 1 and 0 for each of them: indifference to time (*Ti*), leaving in the morning or in the day (*Tm*), leaving in the evening or at night (*Te*).

(6) Time of departure on return (*F*). This parameter is similar to the previous one.

(7) Sleeping car (*Car*). It is a binary parameter.

(8) Discounts (*Ds*). It is a binary parameter meaning whether or not a passenger discussed his/her possible discount with directory inquire service.

(9) Knowledge (*Kn*). This parameter reflects any a priori information a passenger possesses about railway service. Values 1 and 0.5 mean the passenger wants to check up or refers to any previous information, respectively; otherwise, we use 0.

(10) Length of talking (*Tk*). This parameter can serve as an indicator of question complexity or the passenger's competence. It has five numerical values from 1 to 0 with step 0.25, which correspond to non-uniform scale of question-answer numbers.

(11) Politeness (*Pl*). We introduced formal rules for evaluation of this characteristic. Value 1 means that the passenger uses "you" in the polite form (Spanish distinguishes two degrees of polite treatment, reflected in different forms of "you" and verbal conjugation), "please", and apologetic style (reflected in the use of subjunctive forms in Spanish). Value 0.5 means a passenger uses "you" in polite form or in normal familiar form together with subjunctive forms; otherwise, we use 0.

Given these parameters, our example can be presented in a parameterized form, as shown in Table 2. Here *Ti-Tm-Te* and *Fi-Fm-Fe* are nominal scales for qualitative parameters (5) and (6).

Table 2. Parameterized example

City	*Cx*	*U/D*	*T/F*	*Ti-Tm-Te*	*Fi-Fm-Fe*	*Car*	*Kn*	*Ds*	*Tk*	*Pl*
0.25	0	0.5	0	0 0 1	1 0 0	0	0	0	0	0.5

2.2 Parameter filtering

The clustering procedure requires that the introduced parameters be filtered in order to reduce the influence of any strong dominant processes and any sources of possible noise. The former can hide the real structure we want to reveal and the latter can disfigure it. For this, all parameters are divided into the following three groups [3]:

Parameters from the first group have a significant value for almost all objects;
Parameters from the second group have a significant value for a small fraction of the total number of objects;
Parameters from the third group have a significant value for more than, say, 30% of the total number of objects.

By a significant value of a parameter, we generally mean a value larger than 50% of the maximum value for this parameter. By almost all objects or a small fraction of all objects, we mean 90%–95% and 5%–10% of the total number of objects, respectively.

From the point of view of the system, the parameters in the first group reflect the processes in a system of higher level in comparison with the one under consideration. The parameters in second group reflect the processes in a subsystem [3]. On the current level of data consideration, the mentioned groups of parameters should be eliminated.

This conclusion is supported by cluster analysis. From the point of view of cluster analysis, the first group of parameters is oriented to the uniform object set, i.e. to one cluster, whereas the second group of parameters is oriented to very granulated object set, at least 10 clusters or more [4]. The first situation is not interesting at all, and the second one is too detailed for reflecting the structure as a whole. Therefore, cluster analysis approach also confirms the necessity of eliminating of both groups of parameters.

To apply these results to our data set, we calculated the average value for each parameter; see Table 3.

Table 3. Average value of each parameter for 100 dialogs, in percents.

City	*Cx*	*U/D*	*T/F*	*To-Tm-Te*	*Fi-Fm-Fe*	*Car*	*Kn*	*Ds*	*Tk*	*Pl*
37	7	44.5	35	32 32 36	80 9 11	18	4	9	31	40

Since the maximum value of each parameter is equal to 1, we can easily select the parameters of the second group: *Cx*, *Fm*, *Fe*, *Kn*, *Ds*. For all these parameters, the number of significant values is less or equal to 10%. As for the first group, we decided to eliminate the parameter *Fi*, because its value is very close to the boundary value of 90% and from the other hand, this parameter is not interesting for interpretation: it means the indifference to the time of departure from the point of destination.

3 Clustering

3.1 Method of clustering

For a moment, there are dozens of methods and their modifications in cluster analysis, which can satisfy practically all necessities of users. The most popular ones are *K*-means, oriented on the structures of spherical form, and Nearest Neighbor (NN), oriented on the extended structures of chain form [2]. In our work, we use the MajorClust method, recently developed by Stein *et al*. [7], which has the following advantages over the mentioned two:

MajorClust distributes objects to clusters in such a way that the similarity of an object to the assigned cluster exceeds its similarity to any other cluster. This natural criterion provides the grouping of objects, which better corresponds to the users' intuitive representation. Neither *K*-means nor NN methods possess such optimization property: they do not evaluate the similarity between clusters.

MajorClust determines the number of clusters automatically and in all cases tends to reduce this number. *K*-means requires the number of cluster to be given, and NN does not determine this number at all: cutting of the dendrite is performed by the user.

MajorClust has been successfully used with various data sets and demonstrated very good results [8]. The main disadvantage of MajorClust is its runtime. However, in case of sparse matrix of relations this disadvantage is not essential. In our problem, we are faced just with this case because of a weak similarity between the majority of objects. These weak connections are eliminated, which gives the mentioned matrix.

3.2 Distance matrixes and manipulations with them

In our clustering procedure, we used two distance matrices: objects/objects (cities/cities) and parameters/parameters. To construct such matrices, we define the distance measure and apply it to the source objects/parameters matrix. It is well known that:

Cosine measure is used if the proportion between object's coordinates is important, but not their specific values. This is the case when the coordinates have the same meaning.

Euclidean measure is used when the contribution of each coordinate to object's properties is important. This is the case when the coordinates have different meaning.

Therefore, we used the cosine measure to evaluate the distance between parameters whose coordinates were cities, and Euclidean measure to evaluate the distance between objects (cities) whose coordinates were parameters.

During clustering procedure we changed the distance matrix:

To emphasize the role of certain objects (cities) while clustering parameters or certain parameters while clustering objects;

To reveal stronger but less numerous clusters;
To determine the stable number of clusters.

The first goal is reached by weighting the coordinates of objects or parameters, respectively. The second goal is achieved by eliminating weak connections between objects. At in the last case we vary the connections between objects and observe the changes of number of clusters.

4 Experiments

4.1 Experimental data

The data we used in the experiments were a corpus of 100 person-to-person dialogs of Spanish railway information service. The short characteristic of the corpus (length of talking, volume of lexis) is described in [1]. The data were analyzed in detail in [5] for constructing artificial dialogs.

4.2 Clustering parameters

Here in all experiments we used the cosine measure with the admissible level of connections not less than 0.7. In the first experiment all objects (cities) had no any privileges. In the second one the more important cities, that is the large and middle cities (see above) obtained the weight 5. It was the minimum weight, which allowed revealing new result.

Experiment 1. Two parameters *City Weight* and *Length of Talking* were joined to one cluster and the others remained the independent ones.

Experiment 2. Three parameters *City Weight, Urgency and Definiteness* and *Length of Talking* were joined to one cluster and the other parameters remained independent.

These results can be easily explained: the larger the city, the more possibilities to get it, the longer discussion a passenger needs. The urgency of trip is usually related with large cities: usually the trip to the small cities is completed without any hurry.

4.3 Clustering objects (dialogs)

Here in all experiments, we used Euclidean distance measure with the admissible level of connections not greater than 0.5 of the maximum. Then the distances were re-calculated to the similarity measure. In the first experiment, all parameters were equal. In the second experiment, we wanted to emphasize the properties of passengers. For this, we assigned the weight 2 to the parameters *Urgency and Definiteness, Length of Talking* and *Politeness.* This weight was the minimum one to obtain the significant differences with the first experiment. Parameters presented in

nominal scales were weighted by the coefficient 0.33 that is inverse value to the number of scales. Cluster descriptions are presented below.

Experiment 1.

Cluster 1 (10 objects). The large and middle cities, no urgent trips (only 10%), round trips, night trips (70%-90%), sleeping cars, enough long talking.

Cluster 2 (25 objects). No urgent trips (only 8%), round trips, a few number of night trips (25%), no sleeping cars.

Cluster 3 (8 objects). Small cities (75%), undefined day of departure, one-way trips, night trips (90%), sleeping cars.

Cluster 4 (57 objects). Small or local cities (75%), one-way trips, no sleeping cars, short talking (80%).

Experiment 2.

Cluster 1 (31 objects). No urgent trips, no night trips (only 20%), only ordinary politeness.

Cluster 2 (44 objects). Urgent trips or defined days of trips (95%), advanced level of politeness (85%).

Cluster 3 (12 objects). Only small and middle cities, no urgent trips, one-way trips (75%), short talking (85%), the highest level of politeness.

Cluster 4 (13 objects). Only small and local cities, undefined days of trip, one-way trips (75%), no night trips (only 15%), short talking (75%), advanced level of politeness.

Some of the clusters were expected (e.g. the cluster 4 in both experiments) and the others need to be analyzed more closely. In all cases in comparison with manual classification where only costs and time-table were considered, our experiments gave the additional information [5]. Table 4 presents some examples of clustered objects.

Table 4. Examples of objects from cluster 3 in the experiment 2

City	*U/D*	*T/F*	*Ti*	*Tm*	*Te*	*Car*	*Tk*	*Pl*	*Name of city*
0.25	0.5	0	1	0	0	0	0.25	1	Girona
0.5	0.5	0	0	1	0	0	0.25	1	Alicante

5 Conclusions

Results The quality of automatic dialog systems used in public transport service crucially depends on the scenarios of dialogs. These scenarios may be determined by means of clustering in the space of parameters defined by an expert. We have shown (a) how to parameterize the records of dialog and to solve the problems of incompleteness and diffuseness of the source data; (b) how to accomplish the

clustering procedure providing stability of results and their usefulness for a user. We have tested the MajorClust method for this and recommend using it for such type of problems. The obtained results were judged by experts as interesting and useful for determining the main themes of dialogs and the profile of passengers related with these themes. This information can be used to the design of scenarios for an acquisition of dialogs person-to-machine by means of the Wizard of Oz technique.

Future work In the future, we plan to consider more extensively the problems of Knowledge Discovery and to use both geographic information and the other parameters related with transport service.

Acknowledgement. The authors thank Sven Meyer zu Eissen for valuable discussion of application and tuning MajorClust, which allowed us to significantly improve our results.

References

1. Bonafonte, A., et. al.: Desarrollo de un sistema de dialogo oral en dominios restringidos. In: *I Jornadas en Tecnologia de Habla, Sevilla, Spain, 2000*
2. Hartigan, J.: Clustering Algorithms. Wiley, 1975.
3. Makagonov, P.: Evaluating the performance of city government: an analysis of letters by citizens to the Mayor by means of the Expert Assistent System. *Automatic Control and Computer Sciences*, Allerton Press, N-Y, vol. 31, N_3, 1997, pp. 11-19
4. Makagonov, P., Alexandrov, M., Sboychakov, K.: A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: *Data Analysis, Classification, and Related Method*, Springer, series "Studies in classification, data analysis, and knowledge organization", 2000, pp. 83-88
5. Sanchis, E., Garcia, F., Galiano, I., Segarra E.: Applying dialog constraints to the understanding process in a dialog system. In: *Proc. of TSD-02* (*"Text, Speech, Dialog"*), Springer, LNAI, N 2248, 2002, pp. 389-395
6. Sojka, P., et. al. (Eds.): Proceedings of Conf. Text, Speech, Dialog. Springer, LNAI, N_3206, 2004
7. Stein, B., Eissen, S. M. Document Categorization with MajorClust. In: *Proc. 12th Workshop on Information Technology and Systems*, Tech. Univ. of Barcelona, Spain, 2002, 6 pp.
8. Stein, B., Eissen, S. M. Automatic Document Categorization: Interpreting the Performance of Clustering Algorithms. In: *Proc. 26th German Conference on Artificial Intelligence (KI-2003)*, Springer, LNCS, N 2821, 2003, pp. 254-266