Winter 12-2-2007

Intelligent agent for call center: Using data mining techniques and OLAP for automatic answering internet usage problems

Anongnart Srivihok

Narroup Rakngam

Proceedings of The Seventh International Conference on Electronic Business, Taipei, Taiwan, December 2-6, 2007, pp.54-60.

INTELLIGENT AGENT FOR CALL CENTER: USING DATA MINING TECHNIQUES AND OLAP FOR AUTOMATIC ANSWERING INTERNET USAGE PROBLEMS

Anongnart Srivihok, Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand, Email: fsciang@ku.ac.th

Narroup Raknga m, Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand, Email: g47642229@ku.ac.th

ABSTRACT

This study proposes a system for an automatic question analysis and answering using data mining techniques and OLAP. By applying information extraction techniques, relevant information is taken out from of the Internet connection questions provided by users. By using data mining, user profiles and their questions on the Internet usage are used as input for the system. Two steps clustering by SOM and K-Means algorithms are used to segment user data on their characteristics which obtained from user profile. Then output from clustering and question extraction are used for OLAP (Online Analytical Processing) to analyze the cause of the Internet usage problems. After internet usage questions have been asked to the systems, the answers for solving the Internet connections for cases are replied interactively. Results from this study reveal that it is practical to develop an automatic question answering system. The proposed automatic system had been tested with sample questions from the Internet Service Provider, Call Center in Thailand. Precision of the automatic system is about 70% which is good. This study offers useful information regarding the areas of data mining for Call Center or Customer Relationship Management Center.

Keywords: data mining, SOM, K-Mean, OLAP, Internet usage, automatic question analysis

INTRODUCTION

Nowadays, it is accepted that customers are the one who play key roles in business competition. Companies spend five times more money to acquire a new customer than to keep an old one for purchasing a new product [12]. In order to compete in the competitive local and global market, customer services are significant for all companies including small, medium and large enterprises. Therefore customer relationship is assumed as one asset of the business. Thus the collection, collation, and interpretation of customer data to attract and keep customers through business process in order to create long-lasting, customer centric and mutually beneficial relationships and customer services [5, 9].

The numbers of internet usages are increasing tremendously every year. Problems of users with the Internet connections are also increasing. Thus, the Internet Service Provider also needs to help users solve these problems. It is claimed that 80% of internet connection problems are simple and can be solved by users without the assistance of a call center. These problems include telephone cord is not plugged tightly, or misspelling password for authentications. In July 2005, the number of phone calls related to internet usage problems in True Corporation which is one of the largest internet service in Thailand are about 71,628 calls. This is huge and these calls consume a lot of service times.

The objective of this study is to offer the proposed question answering systems by using SOM and K-Means algorithms to cluster user data then applying OLAP cube to analyze cause of the internet connection problems then the answers are automatically provided. This study includes four main sections. Section 2 is the related theories and studies in the past. Section 3 is research methodology and experiment. Last, section 4 is conclusion of the study.

LITERATURE REVIEW

Collaborative filtering is the popular technique, it works by considering and comparing the feature of active users and original users who would have the similarly user database. The system which uses this technique is MovieLens [7]. Weng and Liu, 2004 proposed an approach in analysing historical data of user and cause of problems. Results show that problem characteristics and solving styles might be found in customers with similar computer usages. Clustering Feature (CF) Tree และ Agglomerative Hierarchical are two steps using for data clustering [13]. Chaimeun et al. [2] proposed the principles of data to cluster Thai handicraft customers by using a hybrid algorithm including SOM and K-Means algorithm. At the first stage, SOM is applied to calculate the optimal number of clusters. Output from the first stage had become input in the second stage, it used for K-Means algorithm. Further, two steps clustering of data by using SOM and K-Means have been applied extensively in various studies [10].

OLAP cube is applied with large and complex database resides in data warehouse. It has a powerful ability in organizing views and structuring data adapted to analysis. Messaoud et al. [11] proposed a hybrid technique which can decrease processing time of OLAP cube. This approach combines OLAP with one data mining algorithm: Agglomerative Hierarchical Clustering (AHC) for clustering complex data.

Data mining is defined as finding hidden information in a database. Otherwise, it has been named exploratory analysis, data driven discovery, and deductive learning. Many papers proposed two-stage methods of data mining techniques including SOM and K-Means algorithm in clustering. The majority used SOM to find the optimal number of clusters because of K-Means requires an input which is a predefined number of clusters, k. This two step clustering has been applied in this study to cluster user data into groups according to their user profiles and computer literacy.

1. K-means Algorithm

K-means algorithm [1, 6] is the simplest clustering algorithm and widely used in clustering or segmentation. K-means requires an input, named k, which is a predefined number of clusters. The steps of the K-means algorithm are given below.

Input:
$S = \{s_1, s_2, \ldots, s_n\}$ // set of elements
k // Number of desired clusters
Output:
K // set of clusters
K-Means algorithm:
assign initial values for means $m1,m2,m3,\ldots,mk$;
Repeat
assign each item s_i to the cluster which has the closest mean;
calculate new mean for each cluster;
Until convergence criterion is fulfilled;

2. Kohonen's Self-Organizing Map (SOM)

SOM is a neural network algorithm which is the most popular, and powerful in the unsupervised learning domains. It works effectively in unexpected and changing conditions. The basic idea is to:
1. Represent high dimension data in a low-dimensional form without loosing any of the 'essence' of data.
2. Organize data on the basis of similarity by putting entities geometrically close to each other.
Summary Steps are as follows [4]:
1. ***Initialization****:* The weight vectors and thresholds are initialized to small random values, in an interval (0,1). Then, assign small positive values to the learning rate parameter p and r values.
2. ***Activation***
Compute the neuron output at iteration p
3. ***Learning***
For each input node
- Find the shortest distance to any output node.
- Adjust selected node's weight according to current stage of p.
- Adjust neighboring node's weight according to current stage of p
- Go to next unvisited input node. If there are no unvisited input nodes left then go back to the very first on and go to Step 2.

4. ***Iteration***. Continue repeating Steps 2 and 3 until the synaptic weights reach their steady-state value.

3. OLAP (Online Analytical Processing)

OLAP system is a system which has a focus on the interactive analysis of data and actually provide more capabilities for visualizing data and generating summary statistics. OLAP provides multi dimensional of data representation (13).

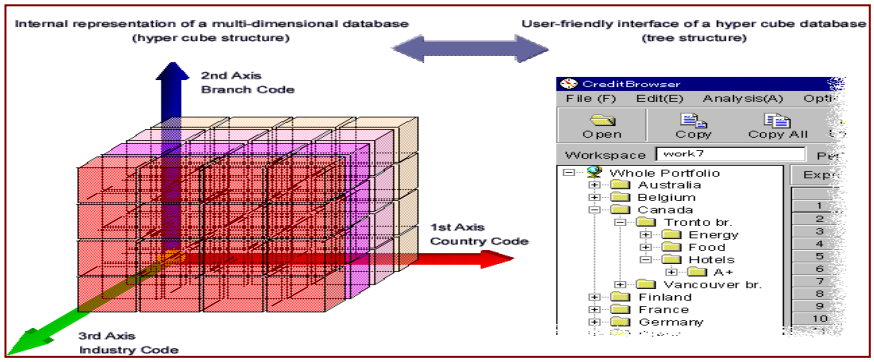

Figure 1. Multi dimensional representation for enterprise data.

4. Root Mean Square Standard Deviation (RMSSTD)

The RMSSTD [8] is the variance of the clusters; RMSSTD measures the homogeneity of the clusters to identify homogenous groups, the lower RMSSTD value means the better clustering.

$$RMSSTD = \sqrt{\frac{\sum_{i=1..n_c}\sum_{j=1..d}\sum_{k=1}^{n_{ij}}(x_k - \bar{x}_j)^2}{\sum_{i=1..n_c}\sum_{j=1..d}(n_{ij}-1)}} \quad (3)$$

Where n_c is number of cluster, d is number of dimension

$\bar{x}_j$ is expected value in the j^{th} dimension.

n_{ij} is number of element in i^{th} cluster j^{th} dimension.

5. R Squared (RS)

RS is used to measure the dissimilarity of clusters. Formally it measures the degree of homogeneity degree between groups [8]. The values of RS range for 0 to 1 where 0 means there is no difference among the clusters and 1 indicates that there are significant difference among the clusters.

$$RS = \frac{SS_t - SS_w}{SS_t}, \text{ where} \quad (4)$$

$$SS_t = \sum_{j=1}^{d}\sum_{k=1}^{n_j}(x_k - \bar{x}_j)^2, SS_w = \sum_{i=1..n_c}\sum_{j=1..d}\sum_{k=1}^{n_{ij}}(x_k - \bar{x}_j)^2$$

EXPERIMENT

This study presents methods for question analysis and answering system of call center. The system can help customers in trouble shooting of errors in the Internet connections without the assistant of call center. Framework of the study is depicted in Figure 1. There are three main steps for the systems. First, user questions on the internet connections have been captured and extracted to get the keywords which represent the questions. Then keywords are stored in the same database as the solutions of internet trouble shootings. Second, user profiles from user database are preprocessed to get the selected attributes. User data focused on the internet usages are clustered. Third, user data in each cluster, extracted questions (keywords) and trouble shooting solutions are analysed by OLAP cube.

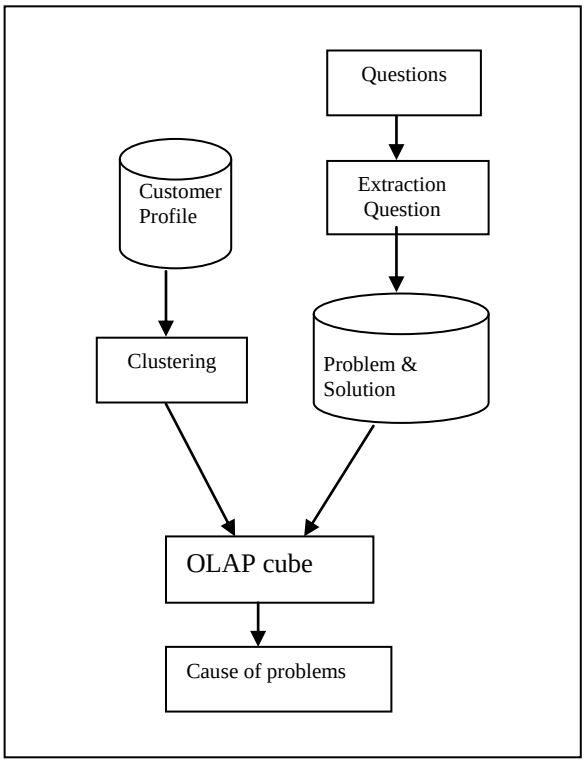

Figure 1. Framework of the study

Table1. Grouping words having the same meaning

Word synonymy	Intension
Connect, เชื่อมต่อ, ต่อเน็ต, connected	การเชื่อมต่อ / Connect
ติด.....ดับ, หลุด, เน็ตดัด, disconnected	เน็ตหลุด / Frequently disconnected
Hisp, high speed internet, ไฮสปีด, hi speed internet	Hi-speed internet
เออเรอ, Error, Errors	Massage Error
....ช้า, เร็ว...k, speed ...MB, slow	ความเร็ว / Speed

Thai documents are written continuously without stop words or spaces between words. First the questions need to be separated into single words by using Swath program developed by Charoenpornsawat [3]. Then Latent Semantic Analysis is applied for solving problems of words which have many meanings namely polynymous and the case of many words which have the same meaning as synonym. Then similar words are grouped in one concept (attribute). Table 1 depicts the word synonym and their intension.

1. Classification of problems

The problems in internet connections are divided in two main groups by the assistant of experts and engineers as: (1) internet can not be connected (Offline case) and (2) internet can be connected (ONLINE case) with some problems. Then each category is divided in subcategories. Each node of the decision tree contains clues or answers which help customers to solve their problems in the Internet connections. Figure 3 reveals the structure of decision tree used in this study. The scope of this study is only on ONLINE case.

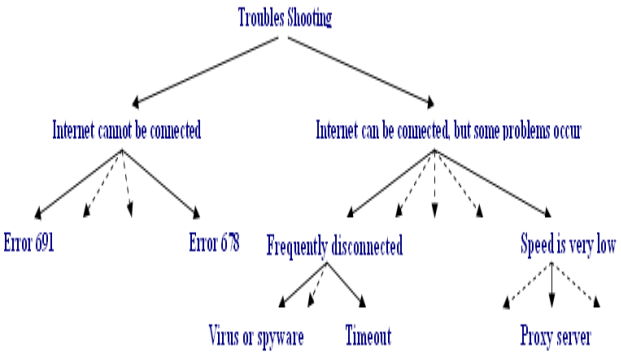

Figure 3. Hierarchical structure of trouble shooting for internet connection problems

2. Data Clustering

Sample data in this study is obtained from Call Centre at an Internet Provider Corporation in Thailand. Data was cleaned and selected for the one that related to the Internet connection problems. Then, experts in computer network answered questions from call centre. These questions-answers are stored in the knowledgebase. The 7,110 transactions are obtained from call center. Then customer profiles are used for data clustering.

Features used for customer data segmentation included: (1) customer type, (2) time, (3) operating system, (4) IT Literacy, (5) internet connection speed and (6) modem.

Clustering data by combining algorithm between SOM and K-Means from Clustering Technique
Step 1: Use Kohonen's self-organizing maps neural network (SOM) to group data into 2-10 groups. For choosing the best number of clusters, RMSSTD (Root Mean Square Standard Deviation) and RS (R square) are used to measure similarities and differences among different clusters. The results from both measurements suggested that the optimal number of groups is 8. Then, the output from this step (k value=8) is used as input for the next Step.
Step 2: K-Means Algorithm is used for segmenting data. The number of group or k value is derived from step 1. In this step customer data are segmented based on customer transactions and their behaviors. Results are depicted in Table 2. Characteristics of each cluster are as follows:

Table 2. Clustering user data into 8 groups by using SOM and K-Mean Algorithms.

Features		Clusters							
		1	2	3	4	5	6	7	8
Period of working	0:00 –9:59 am.	5	13	0	24	283	0	3	0
	10:00 am.–17.59 pm.	70	36	0	165	888	1	10	0
	18:00 pm.-23:59 pm.	476	177	3,078	453	0	8	31	678
Operating System	Windows XP	551	226	3,068	639	1,162	0	0	677
	Windows 2000	0	0	4	1	3	0	0	0
	Windows ME	0	0	0	0	0	3	27	0
	Windows 98	0	0	0	0	0	6	17	0
	Others	0	0	6	2	6	0	0	1
Modem	Zyxel	0	0	0	202	334	0	10	678
	Billion	0	0	3,078	440	805	0	34	0
	Huawei	418	95	0	0	32	7	0	0
	Other	133	131	0	0	0	2	0	0
IT literacy	High	57	41	275	78	123	0	3	70
	Medium	444	174	2,555	529	938	7	36	543
	Low	50	11	248	35	110	2	5	65
Customer	New	6	3	216	71	319	0	3	27
	Old	545	223	2,862	571	852	9	41	651
Speed of internet	<256 Kb	127	0	377	0	308	5	9	133
	512 – 768 Kb	424	0	2,701	0	845	3	33	545
	1 – 2 Mb	0	123	0	333	17	1	2	0
	> 2 Mb	0	103	0	309	0	0	0	0
Total		**551**	**226**	**3,078**	**642**	**1,171**	**9**	**44**	**678**
percent		8.6	3.5	48.1	10.0	18.3	0.1	0.69	10.8

Cluster 3 is the largest of all, it is about 48.1%, the second biggest is **Cluster 5** (18.3%) and the smallest is Cluster 6 (0.1%). **Cluster 3**, data characteristics is old customers who use only modem brand: Billion and Window XP to access the Internet in the evening with broad band (512-768 KB). **Cluster 5**, the second largest (18%), includes the old customers who use Window XP, modem: Billion to access the Internet in the evening with broad band (512-768 KB). **Cluster 4** is about 10%, include the old customers who use Window XP, modem: Billion to access the Internet in the evening with high speed (> 1 Mb). **Cluster 8** is about 10%, include the old customers who use Window XP, only modem: Zyxel to access the Internet in the evening with broad band (512-768 KB). **Cluster 2** is about 3.5%, include the old customers who use Window XP with modem: Huawei to access the Internet in the evening with broad band (1-2 Mb). **Cluster 6**, the smallest, is about 0.1%, include the old customers who use Window 98 modem: Huawei to access the Internet in the evening with band width <256 Kb. **Cluster 7** is about 0.69%, include the old customers who use Window ME, modem: Billion to access the Internet in the evening with broad band (512-768 KB). Majority of customers are old customers.

3. Using OLAP cube

After the data were segmented to 8 clusters, then data in each cluster together with keywords from questions extractions were analyzed by OLAP Cube. Table 3 shows lists of factors and their probabilities which relate to the internet disconnections. These factors include (1) splitter not installed, (2) line deteriorated, (3) many connecting points, and (4) virus. In cluster 1 (Table 3), the frequencies of disconnect is related to using too many connection points since this factor has the highest probabilities (64.41%).

Table 3. Data analysis of internet connection problems and their causes in Cluster 1 by using OLAP

OLAP Cubes (Group 1)			
Problems	Causes	N	Probability
Frequently disconnect	Splitter not install	2	3.39 %
	Many connecting points	38	64.41%
	Virus	4	6.78 %
	Line deteriorated	15	25.42 %
Total		59	100.0%

4. Evaluation

7,110 transactions from customer questions of the internet connections are used for evaluation. The data are divided into two sets: 6,399 transactions (90%) are used for training the system and 711 transactions (10%) are used for testing the system. The precision for this experiment is 70.46% which is good result for prediction.

CONCLUSION

This study proposes a question analysis and answering system for call center. We have proposed the new approach by combining data mining techniques and OLAP. Data of internet customers are clustered by using SOM and K-Means algorithm. First, SOM is used to find the appropriate number of clusters (k). Then, K-Means algorithm is applied to segment data in k clusters. Data in each cluster together with key words extracted from questions and causes of problems provided by domain experts are analyzed by OLAP cube. Output is the probability of each item to be the cause of problems. The system has been evaluated by measuring precision power in predicting the cause of internet failure. The performance of system is about 70% which is good. The proposed system will be useful for building a prototype of help desk or customer relationship. Further, the processing time for OLAP cube is improved since data have been segmented into clusters before using OLAP processing. It further benefits in improvement the effectiveness since clustering data will decrease the data complexity.

Future research should be done on the development of intelligent systems/agent for customer relationship management; this system can be an automatic system that helps customers and company to work more efficiently and effectively.

REFERENCES

[1] Bradley, P. and Fayyad, U. (1998) "Refining Initial Points for K-Means Clustering", *Proceeding of the 15th International Conf. on Machine Learning*.

[2] Chaimeun, O. and Srivihok, A. (2005) "Clustering of Thai handcraft customers using combined SOM and K-means algorithm", A Scientific and Technical Publish Company, *OACTA Press*.

[3] Charoenpornsawat, P., Swath, (2007) http://www.links.nectec.or.th /~yai/.

[4] Dunham, M. H. (2003) *Data Mining, Introductory and Advanced Topics*, Prentice Hall.

[5] Eechambadi, N. and Quaero, M. O. (2004) Creating a CRM Business Case, www.quaero.com.

[6] Fayyad, U., Piatetsky, S. G. and Smyth, P. (1996) "Knowledge Discovery and Data Mining: Towards a Unifying Framework", *The AAAI press*.

[7] Goldberg, D., Nichols , D., Oki , B. M. and Terry, D. (1992) "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM*, Vol. 35, No. 12, pp. 61-70.

[8] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002) "Cluster validity methods: part II", *SIGMOD Rec.*, Vol. 31, No. 3, pp. 19-27.

[9] Karakostas, B., Kardarasb, D. and Papathanassiou, E. (2005) "The state of CRM adoption by the financial services in the UK: an empirical investigation", *Information and Management*, Vol. 42, pp. 853–863.

[10] Mongkolsripattana, S. and Srivihok, A. (2006) "Clustering of technology transfer centers by using SOM algorithm and project performances", *The 3rd Asia-Pacific International Conference on Knowledge management*, Hong Kong, China.

[11] Messaoud, R. B., Boussaid, O. and Rabas´eda, S. (2004) "A New OLAP Aggregation Based on the AHC Technique", *DOLAP'04*, Washington, DC, USA.

[12] Payne, A. (2002) *Customer Relationship Management*. http://www.crm2day.com.

[13] Tan, P.N., Steinbach, M. and Kumar, M. (2006) *Introduction to data mining*. Pearson Education, Inc., USA.

[14] Weng, S.S. and Liu, M. (2004) "Feature-based recommendations for one-to-one marketing", *Expert systems with Applications*, Vol. 26, pp. 493-508.

APPENDIXES

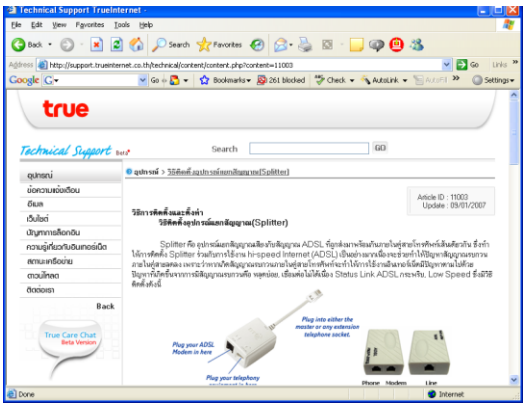

Figure 4. Internet disconnections due to splitters, system provides splitter installation procedures.

Figure 5. System providing assistance to user on trouble shootings for the internet connections with error codes.