MINING FARMERS PROBLEMS IN WEB-BASED TEXUAL DATABASE APPLICATION

Said Mabrouk, Mahmoud Rafea
The Central Lab for Agricultural Expert Systems (CLAES), B.O.Box: 438, Dokki, Giza, 12311, Egypt

Ahmed Rafea, Samhaa El-Beltagy
Dept. of Computer Science, AUC, Cairo, Egypt
Dept. of Computer Science, Cairo University, Egypt

Keywords: Data Mining, Text Mining and Clustering Techniques.

Abstract: VERCON (Virtual Extension and Research Communication Network) is an agriculture web-based application, developed to improve communication between agriculture research institutions and extension persons for the benefit of farmers and agrarian business. Farmers' problems component is one of VERCON main components. It is used to receive farmers' problems and provide them with solutions. Over the last five years, problems and their solutions have been accumulated in a textual database. This paper presents an integrated approach for mining these problems and their solutions. The opportunity and potential of mining and extracting information from this resource was identified with several objectives in mind, such as: a) discovering patterns and relations that can be used to enhance the utilization of this valuable resource, b) analyzing solutions given for similar problems, by different experts or by the same expert at different time in terms of their similarities and differences, and c) creating patterns of problems and their solutions that can be used to classify new problems and provide solutions without the need for domain expert.

1 INTRODUCTION

VERCON: Virtual Extension and Research Communication Network (http://www.vercon.sci.eg) is a kind of help and support service. It is a web-based application, developed in Egypt by the Central Lab for Agricultural Expert Systems (CLAES) (http://www.claes.sci.eg), through a project between FAO and Egyptian Ministry of Agriculture and Land Reclamation.

This project aims to establish and improve the communication between extension and research institutions for the benefit of farmers and agrarian business at rural and village level. Improved communication incorporating newest research results and latest technologies shall ultimately improve the performance of farmers and business.

Farmers' problems component is one of VERCON main components. Farmers describe their problems to the extension officers in the villages, who in turn classify the problems according to their topics into one of four categories (Production, Administration, Marketing and Environment) and write a description for each problem, in free text. Problems are classified into other subcategories and directed to several levels of domain experts in the directorates and the specialized institutes. Domain experts study the problems and respond with recommended solutions.

Over the last five years, more than 10,000 problems and their solutions have been accumulated in a textual database. Problem text has three parts: topic of the problem (crop, weed, diagnosis, treatment, irrigation, etc.), description of the problem (facts, symptoms, findings) and questions.

The following example of the problem text, translated into English, illustrates these three parts:

"The field area is one feddan. It has been cultivated with rice variety sakha102, by baddar, ten days ago. Mild ogizza weeds are shown. What are the appropriate chemicals, concentration, and the rate? , how and when to use? ".

Topic of the Problem: Rice a-kind-of crop and ogizza a-kind-of weed.

Mabrouk S., Rafea M., Rafea A. and El-Beltagy S. (2010).
MINING FARMERS PROBLEMS IN WEB-BASED TEXUAL DATABASE APPLICATION.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 414-419
DOI: 10.5220/0002966904140419

Problem Description: Facts: area (one feddan), rice variety (sakha102), weeds type (ogizza), age of plant (ten days), and cultivation method (baddar).
Symptoms: mild ogizza weeds.

Questions: What are the appropriate chemicals, concentration and the rate? How and when to use?
The following is the answer of the above problem, given by domain expert:

"Satron with concentration 50% and 2 Litre per Feddan should be used. It should be mixed with fine sand, after 15 days of cultivation ".

Mining these problems has several objectives. First, patterns and relations can be discovered and used to enhance the utilization of this valuable resource. The discovered patterns and relations may point to certain types of widespread problems and pressing needs of people living in rural areas. Consequently, decision makers could be able to take necessary actions to tackle these pressing problems and needs of poor communities. Second, solutions given for similar problems, by different experts or by the same expert at different time can be analysed in terms of their similarities and differences. Inconsistencies can then be resolved. Third, patterns of problems and their solutions can be created and used to classify new problems and provide solutions. Fourth, outdated recommendations can be identified and removed from the database. Fifth, users using the problems database can locate problems that are similar to theirs.

Section 2 is a review of related work. In section 3, a methodology for mining the problem parts is given. Three parts can be extracted from the problem's text. They are topic, description and questions parts. Similar problems are clustered. Solutions associated with each cluster are retrieved and analysed.

Section 4 illustrates the difficulties encountered when the clustering techniques was used as a means for identifying similar problems. An alternative more structured approach, based on transforming the problems data base into structured data base using extracted data set of features for each set of problems before applying the data mining, is presented. Result of experimentation with weed control problems is discussed. Section 5 is conclusion and future work.

2 RELATED WORK

Mining problems and their solutions, accumulated in textual databases of help and support services is a novel application of web mining. Previous mining works focused in dealing with one type of documents. For example, in opinion mining systems, documents or reviews of customers are considered. All opinion holders are of one type which is the customer (Nauskawa, Yi, Bunescu, R., 2003. Popescu, A., and Etzioni, O., 2005, Bo Pang and Lillian Lee, 2008). In our work mining will be in two different types of documents. Farmers' problems documents and domain experts' solutions documents. Furthermore, there is an association between these two types of documents.

Data mining and text mining techniques can be used in this application in an integrated manner. In problem part, feature extraction, text clustering, and text analysis techniques (Salton, G., 1989. Ayed, H., and K. M, 2002) are used to cluster similar problems and to analyse the problems in terms of their dominant features and the asked questions. Data mining techniques (Margaret, H., 2003) are used to discover patterns and relations among these problems. In solution part, feature extraction, and text analysis are used to analyse the solutions and data mining techniques are used to discover patterns and relations among solutions. In clusters of problem-solution pairs, data mining techniques are used to discover association rules (Jean Marc Adamo, 2000) and text analysis techniques are used to find the similarities and differences among solutions of similar problems.

3 METHODOLOGY

Two modes of operation are considered, training mode and test mode. In training mode, grouping similar problems, extracting patterns/relations, forming exemplars of similar problems, retrieving solutions associated with each cluster of problems, summarizing solutions and forming pairs of problem and solution are done. In test mode, discovered problem-solution exemplars are used to classify new problems.

3.1 Problem Analysis

Figure 1, summarises the main steps of the methodology as follows:

1. Pre-processing: using Arabic language stemmer to remove affixes and stop words from problem text.

2. Feature Extraction: two approaches are considered, simple approach that uses terms of text as features and more sophisticated one that identifies specific features to be extracted using compiled lists of words from agricultural ontology [http://www.fao.org/agrovoc].

3. Indexing: using term frequency and inverse document frequency schema (TF-IDF).

4. Clustering: grouping similar problems using different clustering algorithms such as partitioning and agglomerative ones.

5. Summarization: problems in the same cluster are summarized in terms of their extracted dominant features, focusing on the three parts of the problem text, i.e., topic, symptoms, and questions.

6. Generalization: features of texts, in one cluster, are generalized using different generalization rules (John Anderson, and Stanislaw, 1983) to obtain an exemplar text that represent all texts in that cluster.

7. Extracting Patterns and Relations: association rules technique is used to extract useful patterns and relations.

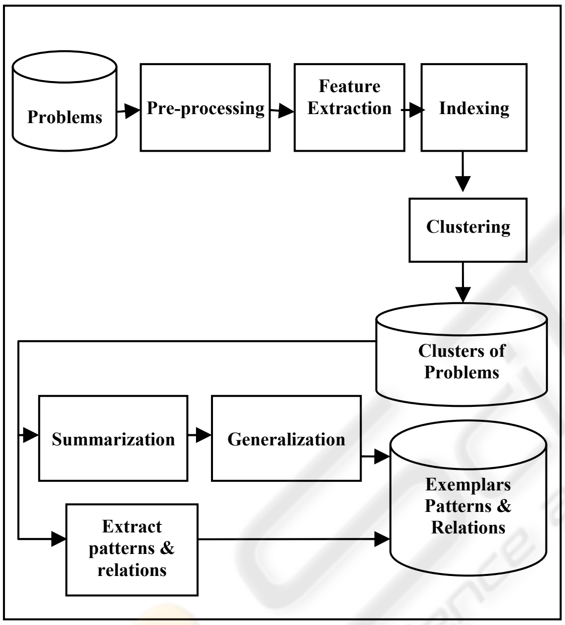


Figure 1: Framework of Problem Parts Analysis.

3.2 Solution Analysis

Figure 2, illustrates the methodology used to analyse solution parts. Clusters of problems are used to retrieve their associated solutions from the textual database. Solution texts are pre-processed by removing stop words and affixes using Arabic language stemmer. Features are extracted using the same approaches used with the features of the problems parts. Texts are summarized in terms of their similarities and differences. Pairs of problem and solution summaries are stored.

4 EXPERIMENTS

Several experiments were conducted to investigate the use of clustering techniques as a means for identifying similar problems. GCLUTO [http://glaros.dtc.umn. edu], which is a clustering tool kit, was used. Different clustering methods such as bisection, K means, and agglomerative clustering with various selected cluster sizes were tried. Terms in problem description were used as features and their weight were calculated based on the TF * IDF model.

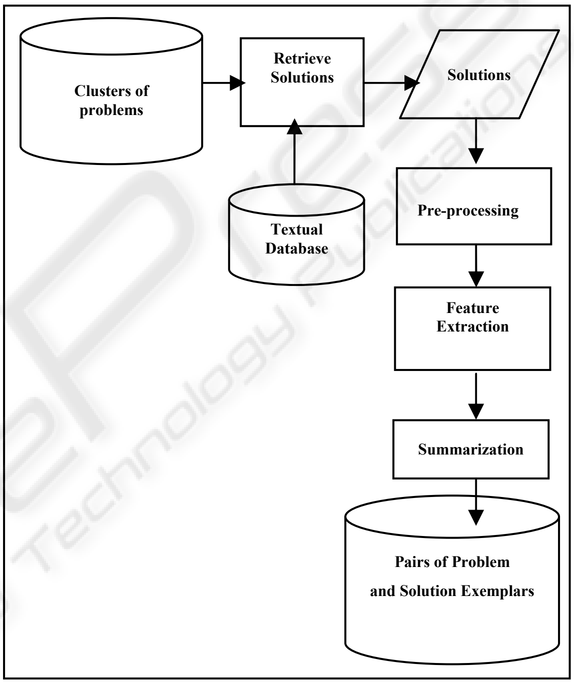
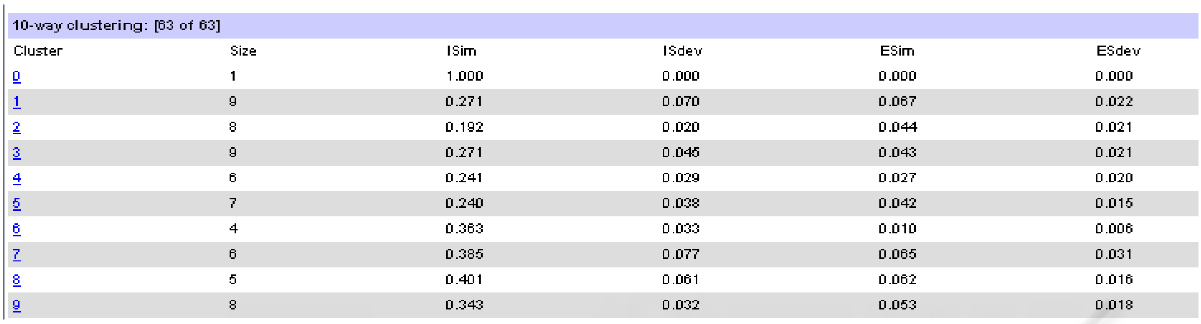


Figure 2: Framework of Solutions Parts Analysis.

Clustering was applied on three classes of rice crop problems: weed control, seeding rate and land preparation. The aim of the experiments was to investigate whether the simple approach to cluster similar complaints, based on their wordings would work or not. Our assumption is that wording the problems may be too different from such an approach to work, but we've decided to pursue this approach to validate this assumption. Clustering based on the bag of word features can also serve as a tool for analysing a sample of input complaints. Identifying and extracting features then constitutes the next step, followed by formalizing similarity function.

GCLUTO clustering tool was used for experimentation. GCLUTO is capable of taking vectors and clustering them based on their similarity

Table 1: Experiment 1 with target number of clusters = 10.



10-way clustering: [63 of 63]

Cluster	Size	ISim	ISdev	ESim	ESdev
0	1	1.000	0.000	0.000	0.000
1	9	0.271	0.070	0.067	0.022
2	8	0.192	0.020	0.044	0.021
3	9	0.271	0.045	0.043	0.021
4	6	0.241	0.029	0.027	0.020
5	7	0.240	0.038	0.042	0.015
6	4	0.363	0.033	0.010	0.006
7	6	0.385	0.077	0.065	0.031
8	5	0.401	0.061	0.062	0.016
9	8	0.343	0.032	0.053	0.018

(different similarity measures are supported, but the one used is the cosine similarity) and the number of desired clusters (k) which the user specifies beforehand. Experimentation has been carried out with various values of k. The goal of any cluster task is to maximize the similarity of the contents of each cluster and at the same time maximize the distance between all other clusters. Two metrics: intra and inter cluster distance are used to evaluate these criteria respectively.

4.1 Clustering Results

In seeding rate and land preparation classes, problem parts were used in clustering while in weed control clustering was carried out using both the problem as well as solution parts.

Table 1 shows the result of experiment with seeding rate data set, using number of cluster = 10 (Where: Size = number of problems in the cluster, ISim = average intra cluster similarity, ISdev = standard deviation of ISim, ESim = average inter cluster similarity, and ESdev = deviation of ESim).

This experiment was repeated with different number of clusters (15, 20, 25). Figures 3 summarizes the intra and inter cluster distances. The graphs indicate that both intra and inter cluster distances grow when the desired number of clusters is increased. This might indicate high degree of overlap in the created clusters.

Similar experiments were done with land preparation problem parts while in weed control, clustering was carried out using both the problem as well as solution parts.

Analysing the contents of the various clusters obtained with weed control data set, revealed that the actual distribution of similar problems amongst various clusters is more scattered than in the seeding rate and land preparation data sets. This can be attributed to the fact that solutions were included in the clustering process, and that there is a lot of overlap in the solutions' text, which means that clustering of problems pertaining to the same weed is not achieved.

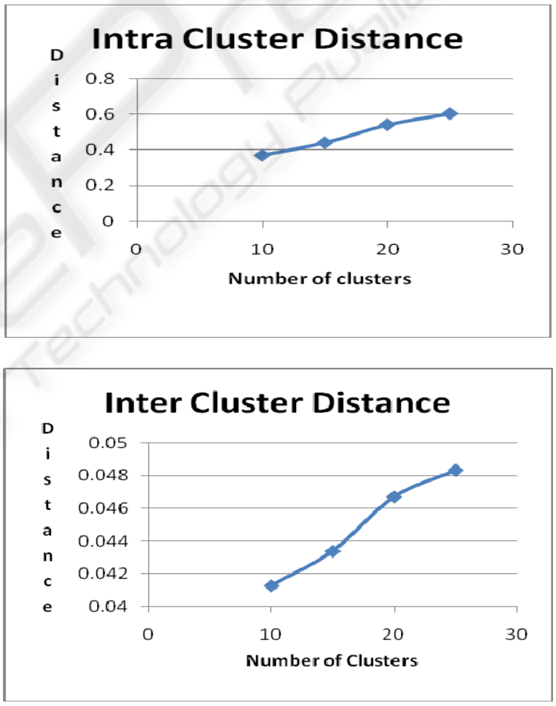


Figure 3: Intra and Inter cluster Distances.

These Results revealed that clustering using the vector space model where terms in the problem represent its features is inappropriate for this kind of task, in this domain. The main reason for this is that similarity is primarily determined through matching of the problem wording which means that two similar problems with different formulation may not be considered as similar and two different problems with a high degree of overlap in terminology and a difference of major term may match. These results have driven us to design a more structured approach for extracting features, to store those in database, and then carry out mining on the database.

4.2 Structured Database

Analysis of weed control problems was done during clustering experiments to populate weed problems structured database. Features to be extracted have been determined. Table 2 shows the dominant features of these problems and their sources.

For simplicity, extraction of weed names and herbicides is carried out through a list of known weeds and herbicides. This is particularly applicable since these are not long lists and can be easily obtained from agricultural resources. However, it must be stated that exact matching between weed names or herbicide names in problems and entries in the lists will not be always possible because of the fact that experts offering the solutions very often misspell both. A rapid intelligent string matching utility thus has been built in order to determine whether an entry in the text matches an entry in the used lists or not.

4.3 Discovered Patterns and Relations

Association rules were applied on subset of the structured weed problems and their solutions database. Multiple frequency item set method was used to find useful patterns and relations among selected features. The strength and confidence of features association is computed. The minimum strength and confidence thresholds were set to different levels. Several interesting patterns and relations were found. The following are some of the discovered patterns and relations:

Pattern1: The most Frequently occurring Weeds and their occurrence Frequency. This is obtained by applying selected thresholds on the "weed name" one item set. Geographical distribution of the problem can also be detailed alongside weed names.

Pattern2: The distribution Pattern of Weed Problems among Planting Methods. This is obtained by using the "Planting type" one item set.

Pattern3: The most Commonly used Herbicides and their occurrence Frequency. This is obtained by applying the selected thresholds on the "herbicide name" one item set.

Relation1: Relationship between a Certain Weed and a Specific Herbicide. This relationship is obtained using two item set that includes the "weed name" and "herbicide". Herbicide related attributes

Table 2.

Feature	Source
Weed Name	problem or solution text
Weed age in days	problem or solution text
Field type (Nursery/production field)	Problem text or deduction rule.
Planting method (Seedlings/Seeds)	Problem text or deduction rule.
Control method (chemical, manual)	Solution text
In case of chemical control, the following are possible additional features:	
Herbicide name	Solution text
Herbicide concentration (percentage)	Solution text
Rate of application	Solution text
Unit for Rate of Application (kg/feddan or litres/Feddan)	Solution text
Application Method (free text representing the solution)	Solution text
Application Time	Solution text
Application Reference (After transplantation/ After planting seeds)	Solution text
Problem Metadata	
Problem ID	From VERCON's Database
Crop Name	From VERCON's Database
Problem's solution Date	From VERCON's Database
Originating Governorate	From VERCON's Database

Table 3.

Doniba (22 %)				
Herbicide name /other	Concentra-tion	Rate per feddan	Application Reference	Application Range (days)
Satron (54%)	50%	2 Litre	Since "Shatel" (50%)	3-4 (64%)
				1-7 (36%)
			Since cultivation (32%)	7-10 (57%)
				8-9 (29%)
				8-8 (14 %)
			Since seeding (14%)	8-9 (100%)
			Unspecified (4%)	
Cafrosatron (22%)	50%	2 Litre	Since cultivation (44.4%)	7-10 (75%)
				8-9 (25%)
			Since "Shatel" (33.3%)	1-7 (100%)
			Since seeding (22.3%)	8–9 (100%)
Aniloguard (7.3%)	30%	7503 Cm^3	Since "Shatel"	5-10 (100%)
Nomini (4.9%)	20%	800 Cm^3	Since seeding	14-18 (100%)
Machit (2.4 %)	60%	1.5 Litre	Unspecified	(100%)
Bazgran (2.4 %)	50%	1.5 Litre	Since "Shatel"	12-15 (100%)

such as (concentration, rate, reference, etc), can also be presented to the user. Table 3 clarify an example of this relation for the weed name "doniba" with strength 22%.

Relation2: Relationship between the Control Method and Control Time. The relationship is obtained using two item set that includes the features: "Control method" and "Application time".

Relation3: Relationship between Herbicides and Control Times. This relationship is obtained using two item set that includes the features: "herbicide" and "application time".

Relation4: Breakdown of weeds into (wide and narrow weeds) and their occurrence frequency as well as relationship between generalized weeds and herbicides.

5 CONCLUSIONS

Mining Textual databases accumulated through the use of Help and Support services is new web mining application. Farmers' problems module contained in the Virtual Extension and Research Communication Network (VERCON) is one of these services.

A methodology for mining both the farmers' problems and their solutions was presented. Clustering experiments were carried out using subsets of the complaints database. The result of these experiments revealed that clustering using the vector space model where terms in the problem represent its features is inappropriate.

A more structured approach for extracting features and transforming the concerned subsets of the database into structured database before applying the mining was developed and applied to the weed control problem. Result of the experiments and examples of the discovered patterns and relations were discussed.

The following activities are under investigation:

1. Developing an automatic approach to extract dominant problem features.
2. Devising method to generalize similar problems and their solutions into pair of problem-solution exemplar and using the created exemplars to classify new problems and automatically find their solutions without the need for human experts.
3. Investigating the use of opinion mining techniques to analyse the expert solution of a problem as his opinion for solving it, where the expert is the opinion holder, the problem is the object of the opinion and the solution is the opinion or the view of the expert.

ACKNOWLEDGEMENTS

This work has been supported by the Egyptian Science and Technology fund # 79/2009

REFERENCES

Nauskawa, Yi, Bunescu, T., and Niblack, R., 2003. Sentiment analyser: Extracting sentiments about a given topic using natural language processing techniques. In *Proceeding of the third IEEE Conf. on Data Mining*. Melbourne, Florida, USA.

Popescu, A., and Etzioni, O., 2005. Extracting product features and opinion from reviews, 2005. Inproceeding of HLT-EMNLP, Vancouver, Canada.

Salton, G., 1989. Automatic Text Processing AddisonWesley.

Ayed, H., and K. M, 2002. Topic discovery from text using aggregation of different clustering methods. In *15th Conference of the Canadian society for Computational Studies of Intelligence.*

DY, J., and Brodley, C., 2004. Feature Selection for Unsupervised Learning. *The Journal of Machine Learning Research*, Vol. 5, pp. 845-889, 2004.

Margaret, H., 2003. Data Mining: Introductory and Advanced Topics, *Pearson Education*, Inc.

Jean Marc Adamo, 2000. Data Mining for Association Rules and Sequential Patterns. *New York, Springer-Verlag.*

Bo Pang and Lillian Lee, 2008. Opinion Mining and Sentiment Analysis," Foundations and Trends in information Retrieval, Vol 2, No 1-2.

Bing Liu, 2008. Opinion Mining and Summarization-Sentiment Analysis, tutorial given at WWW-2008, Beijing, April 21, 2008.

Dave, K., Lawrence, S. and Pennock, D., 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Review.

http:// www.fao.org/agrovoc.

John Robert Anderson, and Ryszard Stanislaw, 1983. Machine Learning: An Artificial Intelligence Approach. http://glaros.dtc.umn.edu/gkhome/views/cluto/

VERCON, 2006. http://www.vercon.sci.eg.