Analyzing Dynamics of Crop Problems by Applying Text Analysis Methods on Farm Advisory Data of eSagu^{TM}

R. Uday Kiran
Media Lab Asia Project, ICTs for Agriculture and Rural Development,
International Institute of Information Technology (IIIT-H), Hyderabad, India
E-mail: uday_rage@research.iiit.ac.in

P. Krishna Reddy*
Media Lab Asia Project, ICTs for Agriculture and Rural Development,
International Institute of Information Technology (IIIT-H), Hyderabad, India.
E-mail: pkreddy@iiit.ac.in
*Corresponding author

M. Kumara Swamy
Media Lab Asia Project, ICTs for Agriculture and Rural Development,
International Institute of Information Technology (IIIT-H), Hyderabad, India
E-mail: kumaraswamy@research.iiit.ac.in

G. Syamasundar Reddy
Media Lab Asia Project, ICTs for Agriculture and Rural Development,
International Institute of Information Technology (IIIT-H), Hyderabad, India
E-mail: shyamiiit@gmail.com

Abstract: By extending information and communication technologies, a personalized agricultural advisory system called eSagu^{TM} has been developed in which the farmers receive agricultural expert advice for each of their farms at regular intervals. The expert advice is prepared by agricultural experts based on the crop status information received in the form of both digital photographs and text. During 2004-05, the eSagu^{TM} system was operated for 1051 cotton farms covering three villages in the state of Andhra Pradesh, India. In eSagu^{TM}, the expert advice had been delivered to every cotton farm, once in a week. As a result, the data set consisting of about 20,000 such advice texts had been generated. In this paper, we have carried out the cluster/textual analysis experiments on the data set and reported interesting results concerning the dynamics of crop problems. Normally, all are cotton farms and belonging to nearby area/region should have faced similar problems. However, the cluster analysis of the advices delivered on each day shows that significant number of farms are suffering from distinct crop production problems. The results also indicate that, a cluster of farms which face the same crop problem during one week face distinct crop problems during the subsequent weeks. Based on the results, we can conclude that it is necessary to deliver agricultural expert advice to each farm by building agricultural advisory systems which deliver farm-specific agricultural advices to reduce crop failures and improve crop productivity.

Keywords: ICTs in Agriculture, eAgriculture, Computers in Agriculture, Agricultural Information Systems, Dynamics of Crop Problems, Text Analysis, Cluster Analysis, Text Mining

Reference to this paper should be made as follows: Uday Kiran, R., Krishna Reddy, P., Kumara Swamy, M., Syamasundar Reddy, G. (2008) 'Analyzing Dynamics of Crop Problems by Applying Text Analysis Methods on Farm Advisory Data of eSagu^{TM}', Int. J. of Computational Science and Engineering, Vol. , No. , pp. – .

Biographical notes: Uday Kiran, R. is a Ph.D. student at International Institute of Information Technology (IIIT-H), Hyderabad, India. He received MS (IT for Agriculture) from Dhirubhai Ambani Institute of Information and Communication Technology, India in 2004. His research interests include Data Mining, Recommended systems and ICT's for Agriculture.
P. Krishna Reddy is a professor at International Institute of Information Technology (IIIT-H), Hyderabad, India since 2007. He has received both MTech and PhD degrees in Computer Science from Jawaharlal Nehru University, New Delhi, India in 1991 and

1 Introduction

Information or knowledge dissemination to the stakeholders is an important aspect of social development. Currently, research is going on to investigate the improved information dissemination methods by exploiting the recent developments in Information and Communication Technologies (ICTs). Print media viz. news papers, magazines, books and mass communication media viz. radio, television, telephone are being used widely to disseminate information or knowledge to masses. Normally, these methods disseminate information in an ad-hoc and generalized manner. Efforts are being made to disseminate personalized information by exploiting ICTs through search engines, web sites/portals, question-answering systems, publish/subscribe systems, topic directories, discussion forums, information blogs and call centers. In the literature, it was reported that developing information systems to deliver personalized information service is one of the problem areas (Silberschatz and Zdonik, 1996). Progress in database, data warehousing, data mining (Han and Kamber, 2006), mobile and internet technologies are enabling mass customization and personalized information services (Pine, 1993).

In the field of agriculture, agriculture extension wing deals with the dissemination of advanced agriculture technologies to the farming community. Efforts are being made to reach farmers through gatherings, news papers, magazines, journals, seminars, broadcast media, call centers and Web sites. However, they are not meeting the expectations of the farmers due to several drawbacks such as irrelevance of the delivered information, inability of the system to cover all the farmers, the lack of avenues to improve the performance, and un-accountability for the advice given by the system (Rita Sharma, 2002). Additionally, these systems do not consider the cases at the individual farmer's field level as each farmer needs a distinct guidance based on his/her socio-economic conditions, soil type, irrigation, data of sowing etc., (Krishna Reddy and Ankaiah, 2005).

Since 2004, by extending the developments in ICTs to agriculture, an effort is being made to build personalized agricultural advisory system called *eSagu*^{TM}[1] (The word "Sagu" means cultivation in Telugu language) to improve the utilization and performance of agricultural technology to improve the crop productivity under Indian farming situation (eSagu, 2008; Krishna Reddy and Ankaiah, 2005; Ratnam, Krishna Reddy and Reddy, 2006; Krishna Reddy, Ramaraju and Reddy, 2007). The *eSagu*^{TM} system aims at providing agricultural expert advices to the farmers in a timely and personalized manner. The expert advice is generated by agricultural experts based on the latest information about the crop situation received in the form of both digital photographs[2] and corresponding text. The expert advice is delivered to each farm[3] on a regular basis (typically once in a week) from the sowing stage to the harvesting stage.

The system has been developed by considering farming situation in India. During 2004-05, the eSagu^{TM} prototype was designed and the expert advices were delivered to 1051 cotton farms of three villages in Warangal district of the state of Andhra Pradesh, India. The expert advice has been delivered to each farm once in a week. Thus, about 20,000 such advice texts were gathered.

In this paper, we have carried out text analysis experiments on the advice data to understand the dynamics of crop problems. The experimental results show that all are cotton farms belonging to same area are facing distinct problems. Also, a group of farms which are facing the same problem during one week are facing different problems during subsequent weeks. This indicates that the crop problems differ from farm to farm. The results can be used to re-look into the functioning of existing agricultural extension systems in India and other countries which have similar farming environment for improving their performance.

In the next Section, we briefly explain about eSagu^{TM} system and details of prototype experiment during the period 2004. In Section 3, we explain the data set and discuss the preprocessing steps. In Section 4, we present the experimental results. The discussion about the dataset is provided in Section 5. The last section consists of summary and conclusions.

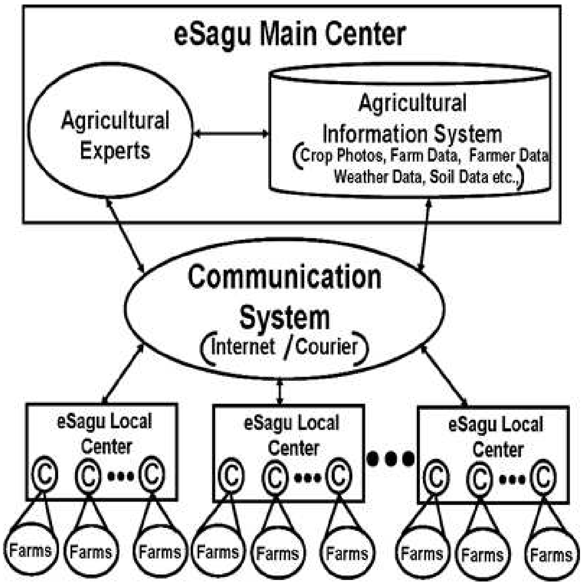
eSagu Main Center
Agricultural Experts
Agricultural Information System (Crop Photos, Farm Data, Farmer Data Weather Data, Soil Data etc.,)
Communication System (Internet /Courier)
eSagu Local Center
eSagu Local Center
eSagu Local Center
C C ••• C
Farms
Farms
Farms

Figure 1: The architecture of eSagu system. The notation © indicates coordinator.

[1] eSagu^{TM} is a trademark of IIIT-H, Hyderabad and Media Lab Asia.

[2] Here, the photograph is taken by visiting the farmer's field.

[3] The word "farm" means a piece of land in which a crop is cultivated. In this paper the words "farm" and "crop" are used interchangeably. Typically the farm size is about 1.5 hectares

2 Overview of eSagu^{TM} prototype

In this section, we briefly explain about eSagu system, its operation and the details of the prototype experiment.

2.1 The eSagu^{TM} system

In eSagu^{TM}, the agriculture scientist, rather than visiting the crop in person, delivers the expert advice by getting the crop status in the form of both digital photographs and the related information. The following are the parts of eSagu^{TM} (refer Figure 1): Farms, eSagu local center, coordinators, eSagu main center, agricultural experts, agricultural information system (AIS) and communication system. Farms are owned by the farmers, who are the end users of the systems. One eSagu local center is established for about 10 to 20 villages. It is managed by a computer operator and equipped with few computers, printer and dial-up Internet connection. The coordinator is a literate farmer having about 5 years of farming experience. In eSagu main center, a team of agricultural experts deliver the expert advice by accessing the crop status information (photos and text) from AIS. Agricultural experts possess a university degree in agriculture and are qualified to provide the expert advice. The AIS contains farm photographs along with the related information about the crop status, farmer registration data, farm registration data and farm related weather data. Communication system is a mechanism to transmit the farm observation data (photographs and text) to the agricultural experts and the corresponding expert advice from the eSagu main center to the eSagu local centers. Transmission of digital photographs from the eSagu local center to the eSagu main center requires a considerable bandwidth. If enough bandwidth is unavailable, information can be written onto compact disks and are sent by a courier or parcel service. However, the agricultural expert advice (which is a text) is transmitted from the eSagu main center to the eSagu local center through a dial-up internet facility.

2.2 Operation of eSagu^{TM}

The operation of eSagu^{TM} is as follows. A team of agriculture experts[4] work at the eSagu main center (normally in a city or a university) supported by AIS. One eSagu local center (few computers and one computer operator) is established for about 10 to 20 villages. Educated and experienced farmers (local residents) are employed as coordinators. Depending on the crop, each coordinator is assigned with a fixed number of farms. Prior to the sowing of crop, the coordinator collects the details of the farms which includes details of soil data, water resources, and capital availability. The collected information is sent to the eSagu main center. Once in a week, the coordinator collects farm observation data for each farm. The farm observation data consists of about five digital photographs, and other information like feedback from the farmer. Next, the accumulated data concerned to each farm condition and other information like weather details etc, are burnt onto a compact disk and delivered to the eSagu main center in an online manner or through a courier system. At eSagu main center, the information is stored in AIS. At the eSagu main center, the agricultural experts analyze the crop situation by considering soil, weather and other agronomic practices and prepares an expert advice for that farm, which is stored back into AIS. At the eSagu local center, the advice is printed by accessing AIS through a dial-up Internet connection. The coordinator delivers the printed advices to the concerned farmer and explains about the advice contents. In this way, the farmer gets the advice to each farm at regular intervals starting from pre-sowing operations to post-harvest actions.

Table 1: Major problems in cotton crop (Non-Bt).

Insect Pests	Aphids, Jassids, Thrips, Whitefly, Spotted boll worm, Pink boll warm, Stem borer, Mites, Mealybug, Red cotton bug, Dusky cotton bug, Red Spider mite, Leaf roller, Tobacco caterpillar, Helicoverpa, Spodoptera, Gram caterpillar.
Diseases	Damping off, Leaf spots, Wilt, Black arm, Grey mildew, Boll rot, Root rot, Anthracnose, Alteranaria leaf spot, Cercospora leaf spot, Helminthosporum leaf spot.
Deficiencies	Nitrogen, Phosphorous, Potassium, Iron, Sulphur, Manganese, Boron, Zinc, Copper, Molybdenum, Magnesium.

2.3 Details of eSagu Prototype

Cotton is the important commercial and problematic crop in the state of Andhra Pradesh, India. Normally, the crop is grown between June to January. Based on variety, the crop duration varies between 150 and 180 days. Cotton crop is affected by several insects, diseases and nutrient deficiencies (Cotton Doctor, 2008) which are given in Table 1.

The implementation and operational information of the eSagu prototype are shown in Table 2. The system was implemented from June 2004 to March 2005 for 1051 cotton farms of 984 farmers (some farmers have multiple farms) belonging to three villages nearby Warangal district in the state of Andhra Pradesh, India. The locations of eSagu main center and eSagu local center are shown in the map (Figure 2). In one of the three villages, eSagu local center was established. The distance between the eSagu main

[4] The agricultural experts are from diverse backgrounds in agriculture like agronomy, entomology, pathology etc.

Table 2: Details of the prototype (2004-05)

Variable	**Value**
Duration	June 2004 to March 2005
Location of eSagu main center	Hyderabad, India
Number of villages covered	3
Location of villages	Near by Warangal town, Andhra Pradesh, India.
Distance between the eSagu main lab and villages	About 200 kilometers
Name of the crop	Cotton
Number of farmers	984
Number of farms	1,051
Number of agricultural experts	5
Number of coordinators	14
Periodic visit to each farm	Once in a week
Number of farm observations	20,035
Number of advices delivered	20,035
Number of Photographs	1,11,515

center and the eSagu local center is about 200 kilometers. At the eSagu local center, fourteen coordinators were identified to take farm observation photographs. Each coordinator was given a digital camera. Every day, one coordinator covered about 10 to 15 farms. Each farm was visited once in a week. So, during the week, each coordinator covered about 80 to 100 farms. In eSagu main center, five agricultural experts have delivered agro-advisory by accessing crop photos and related data from AIS. In total 20,035 farm observations consisting of 1,11, 515 digital photographs were received and the advices for the respective observations had been delivered.

2.4 Results of the Prototype

The main results of eSagu prototype implementation during 2004-05 are summarized as follows.

- It was shown that it is possible for the agricultural expert to provide the expert advice based on the crop photographs and other information available in the AIS.
- It was also found that the expert advice helped the farmers to improve the agricultural efficiency by encouraging integrated pest management, judicious use of pesticides and fertilizers.
- The impact study shows that the farmers have realized considerable monetary benefits by reducing the fertilizers and pesticide sprays, and in getting the additional yield (Krishna Reddy et al., 2005)(Ratnam, Krishna Reddy and Reddy, 2006).

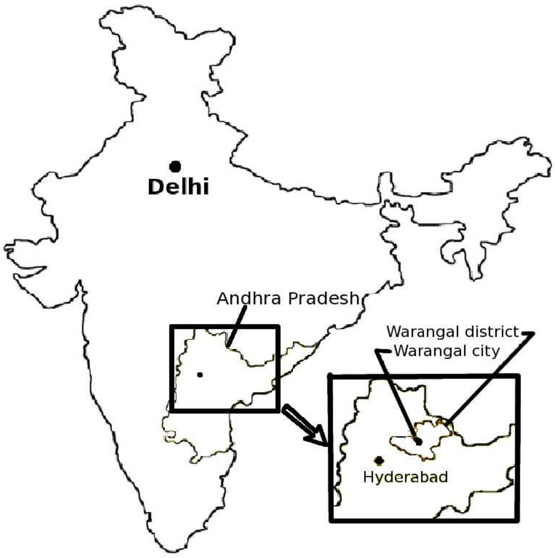

Figure 2: The Geo-Location of eSagu Implementation in 2004-2005. The eSagu main center is located at Hyderabad and eSagu local center is located nearby Warangal town.

3 Data set and preprocessing

In this section, we explain the details of the data set collected as a result of eSagu^{TM} prototype implementation. Next, we explain the preprocessing steps applied on the data set for conducting text analysis experiments.

3.1 Details of the data set

Different types of data have been collected during the execution of eSagu prototype. The details are given below.

- **Location data:** It contains the location details of villages, eSagu local center and the eSagu main center.
- **Personnel data:** It contains the profiles of agricultural scientists, coordinators and farmers.
- **Farm data:** It contains the details of soil, irrigation, sowing, and other useful information.
- **Farm observations:** It contains farm observation data and crop photographs. About five photographs are collected for each farm once in a week.
- **Weather details:** It contains the details of the daily temperature, humidity, and rainfall information of the eSagu local center.
- **Advice data:** The advice data is a collection of advice texts prepared by agricultural experts. The agriculture experts prepare the expert advice based on

the crop photographs. Each advice is a piece of English text which contains the list of steps that the farmer should take to improve the efficiency of the farm. The number of sentences in the advice text and nature of words depends on the crop problem of that farm[5]. An advice typically includes pest names, latest agricultural practices and integrated pest management methods to be followed, fertilizers to be applied and so on.

Table 3 shows the sample advices. The first column is unique identifier given to each advice, second column is an advice text prepared by agricultural expert. In the "Remarks" column, description is given regarding transliterate words, pesticide and fertilizer names, and acronyms in the advice. The advice with identifier "426" contains the corrective measures (or the practices) suggested by the agricultural expert for a farm that is being affected with sucking pests and grass hoppers. Similarly, the advice with identifier "5043" contains the corrective steps for a farm that is affected with Helicoverpa pest, and the advice with identifier "10509" contains the corrective steps suggested for a farm affected with magnesium deficiency, helicoverpa, and pink boll worm pests and a pro-active measures to manage.

3.2 Data preprocessing

We have selected only advice data set for the text analysis experiments. Each advice is a tuple consisting of <advice identifier, advice text>. The "advice identifier" uniquely identifies the advise and "advice text" contains the list of steps which have to be followed by the farmer to improve the farm productivity. These sentences consists of words of both English and agricultural literature typed in English. In addition, the advice text also includes "transliteration words" representing the phonological words in Telugu (a local language) literature. The agricultural experts use transliteration words to help the coordinators for an easy interpretation and translation of the advice to the farmers.

In the experiments, we have considered each advice as a text document and thus generated a data set of about 20,035 documents.

The data set is noisy containing typographical errors, numerical and special characters. We refer this dataset as raw advice dataset. The following steps (Figure 3) are performed on the raw advice data set to generate advice vector matrix.

i. **Removal of special and numerical characters:** From each raw advice, the special characters (, , +, -, &, etc.) and numerical characters are removed.

ii. **Correction of typographical errors:** The mistakes of English and technical terms are corrected. The words pertaining to English literature are spell corrected using Roget's Thesaurus (Roget, 1911). For technical terms, a word repository was built by selecting the words related to the names of pests, diseases, symptoms, agricultural practices, fertilizer and pesticide names etc. By comparing the words in the advice text with the words in the repository incorrect and misspelled agricultural words are corrected manually.

iii. **Removal of stopwords:** The stop words (Stopwords, 2008) are words like a, an, the, in is etc., which appear in every advice text. These stop words are removed from each advice text.

iv. **Identification of keywords:** In this step, key words are identified for each advice text. In information retrieval literature, keywords are extracted for each document and similarity between any two documents is compared by considering the keywords. A keyword represents a word of special significance in a document. The following procedure is followed to identify the keywords. First, all the advice texts are merged. A list consists of each word and corresponding frequency is prepared. The list is sorted based on the frequency. Both high (greater than 8000) as well as low (less than 500) frequency words are deleted from the list. The remaining words are considered as keywords.

Let m and n indicate the number of advices and total number of key words respectively. For each advice text, the corresponding advice vector is prepared by keeping only key words (along with their frequency in the advice text) and removing other words. The advice vector matrix $AVM[m, n]$ is prepared which contains all the advice vectors. If jth word belongs to ith advice vector, the value of $AVM[i, j]$ is equal to the frequency of jth keyword in ith advice vector. Otherwise, the value of $AVM[i, j]$ is 0.

4 Experiments Results

We have carried out different text analysis experiments using AVM[m,n]. We have analyzed advice vectors by employing document clustering techniques (Steinbach et al., 2000). We now explain the similarity measure and clustering approach employed for the analysis.

Similarity measure and similarity threshold: To compare two advice vectors, we employ the *cosine* similarity measure (Steinbach et al., 2000) which is widely employed in information retrieval literature to compare two text documents. Let $\overrightarrow{p}$ and $\overrightarrow{q}$ be two vectors with n dimensions. The cosine similarity between $\overrightarrow{p}$ and $\overrightarrow{q}$ is computed using the following formula: $cos(\overrightarrow{p}, \overrightarrow{q}) = \left(\frac{p \cdot q}{\sqrt{||p||^2||q||^2}} \right)$.

The similarity threshold value is fixed based on the following procedure. A random sample of 100 advices

[5] The word farm represents a piece of land of about two hectors. A particular crop is cultivated in the farm.

Table 3: Sample of agricultural advices. It can be noted that in the "advice text" column, the advice text which was typed by agriculture scientist is given without carrying out any corrections. So, advice text may contain both **spelling and grammatical mistakes**.

Advice identifier	**Advice text**	**Remarks**
426	1. Advise the farmer to go for stem application with either Monochrotophos (1:4) or Imidacloprid (1:20) at 20, 40 and 60 days after sowing. 2. Ask the farmer to grow Castor plants (at least 50 plants per acre) all over the field and along the borders. But do not forget to pluck and destroy the infested Castor leaves periodically. 3. Tell the farmer grow Maize crop along the borders to stop thrips. 4. Take care that the farmer should not spray any powerful chemical like Sythetic Pyrethroids.	The words 'Monochrotophos' and 'Imidacloprid' are the names of pesticides. The word 'thrips' is the name of sucking pest.
5043	Helicoverpa (PACCHA PURUGU) infestation was identified, follow the following practices. 1. Setup pheromone traps (4 traps/acre) for pest intensity identification as well as to trap the male months. 2. Setup light traps (1 light trap/5 acre) to know the range of pest incidence as well to kill moth population. 3. Arrange atleast 10 bird perches/acre. 4. In this stage there may be eggs and larvae, so it is better to spray ovicidal (GUDDU MEEDA PANICHESEDI) and larvicidal (purugu meeda (panichesedi) chemical i.e. Thiodicarb (Larvin) @ 10 grams or 15lit. OR Endosulfon+Sesamum oil in 2:1 ratio.	The words 'PACCHA PURUGU' are Telugu translation words means 'Helicoverpa'. The words 'GUDDU MEEDA PANICHESEDI' are Telugu literary words means 'A chemical agent that works (kills) on eggs'.
10509	1. Magnesium deficiency is observed (Drying of leaves) and it is very severe so suggest the farmer to spray Magnesium sulphate @ 100 g/tank. Repeat the spray agter one week. 2. Immediately put traps of Hel. and pInk boll worm. 3. It is better to spray 5% NSKE solution every week to kill the eggs of boll worms and to control sucking pests like Whitefly, Mealy bug etc.	The characters 'NSKE' is acronym of 'Neem Seed Kernel Extract'. The word 'Hel.' is an acronym of the word 'Helicoverpa'.

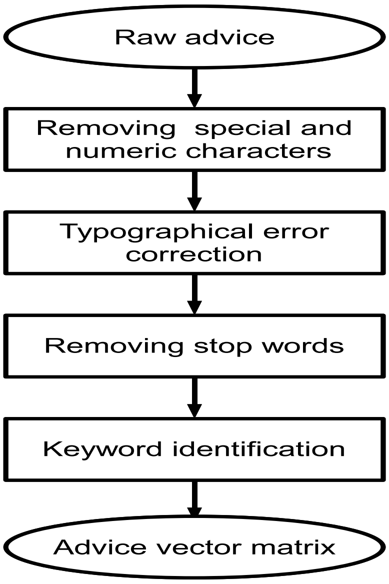

Figure 3: Identifying keywords from the advice dataset.

were given to five agricultural experts and requested them to find the similarity among the advices. The agricultural experts have grouped these advices into different clusters. With different similarity threshold values, the advice clusters are generated through a program using cosine similarity. By comparing the results generated from various threshold values with the results obtained from manual process, the similarity threshold value was fixed at 80%.

Clustering algorithm: For clustering the advices, we have followed the two step procedure. The pseudo code is given in Algorithm 1.

Step 1: In this step, the clusters are computed by assigning the similar advice vectors into one cluster. By considering the first advice vector in AVM, all the other advice vectors which are near to it are assigned to one cluster. The advice vectors which have been assigned to a particular cluster are ignored during further steps. The next cluster is obtained by repeating the same procedure by picking the next unclustered advice. The procedure is repeated until all the advice vectors are clustered.

Step 2: The clusters are refined in this step. After computing the centroids of the clusters, each advice vector is pushed to the nearest centroid. The step is repeated until the values of the cluster centroids do not change or the change is less than the given threshold value.

Algorithm 1 Clustering algorithm

```
Input: AV[m]. AV[m] is the array of advice vectors which corresponds to the rows of AVM[m × n].
ST: Similarity threshold;
Output: Clusters.
Variables: i,j, nc, nctemp: integers; bav[m]: array of boolean values; cluster[m]: An array of m cluster identifiers. Each cluster identifier represents a cluster of advice vectors.

1 Finding the clusters
    1.1 i=0; nc=0; for i=0 to i=(m-1) {bav[i]=tr ue;}
    1.2 if ((bav[i]= true) and (i ≤ (m-1)) { bav[i]=false; Assign AV[i] to cluster[nc]; nctemp=nc; nc=(nc+1); for j=(i+1) to (m-1) do
        { if ((bav[j] ≠ false) and (cosine(AV[i],AV[j]) ≥ ST)), then { AV[j] is assigned to cluster[nctemp]; bav[j]=false} } } else {if i=(m-1) go to the step (2); else { i=(i+1); Go to step (1.2); } }
2 Refining the clusters.
    2.1 For i=0 to (nc-1) { Compute the centroid[i] } /* The centroid[i] is of the type advice vector and represents the centroid of the cluster[i]. */
    2.2 for j=0 to (m-1) { for i=0 to i=(nc-1) { Compute the cosine similarity between AV[j] and centroid[i]. Assign AV[j] to the nearest cluster. } }
    2.3 Repeat the step (2) until the values of the cluster centroids do not change or the change is less than the given threshold value.
```

4.1 Cluster analysis of advices delivered on each day

In this experiment, we have analyzed the clusters obtained from the advices delivered on each day. Figure 4 shows several graphs; each graph shows the variation of cluster size verses number of clusters for the advices delivered on one day which has resulted into an exponential curve. Figure 5 shows a curve of cluster size versus the number of clusters on a log-log scale by considering the mean values of all days. It resulted into a straight line with slope of -1.6. The equation of corresponding exponential curve comes to $\left(\frac{1}{2e^{1.6i}}\right)$. A sample of clustering results for four days is shown in Table 4.

It can be observed that, on any day, significant number of singleton clusters (clusters having one one farm) are formed. Also, it can be observed that as cluster size increases the number of clusters decrease sharply. It indicates that significant number of farms are facing district crop problems. It can be noted that, as cluster size increases, a few large clusters are being formed. It indicates that a reasonable number of farms are facing the same problem.

For the four days cluster data shown in Table 4, the farms are divided into three different cluster sizes of 1, 2 to 5, and above 5. The total number of farms in three different cluster size is shown in Table 5. Figure 6 shows the bar graph for the data in Table 5. From the analysis of four days data, it can be observed that significant number of clusters are having only one farm. There are also significant number of small clusters having number of farms varying from 2 to 5. The remaining significant number of farms constitute few large clusters above 5.

We have carried out similar analysis for the entire dataset of all days. On an average, the results are as follows. The number of clusters having only one farm comes to 20%. It means that about 20 percent of farms are facing distinct problems. Also, the number of clusters having farms varying from 2 to 5 comes to 40%; each small cluster of farms is facing a distinct problem. The remaining 40 percent of farm advices have formed a few clusters of large size above 5.

The results show an interesting phenomena. Overall significant number of distinct advices are delivered on each day which shows that significant number of farms is facing distinct problem.

4.2 Cluster analysis of advices delivered during each week

During eSagu prototype implementation of 2004-05, agricultural expert advice was delivered to each farm once in a week in a regular manner. In this experiment, we have analyzed the number of distinct problems faced by the farms for each week by carrying out the cluster analysis for the advices delivered during each week.

The total data set is collected for 20 weeks starting from 22-07-2004 to 23-12-2004. A week constitutes days between Sunday and Friday (Saturday was weekly off) in eSagu operations. The 1^{st} week, 2^{nd} week and so on refers to the advices delivered during the first week, second week, and so on respectively. We have not compared clusters from 1^{st} week to 3^{rd} week and from 18^{th} week to 20^{th} week, because the number of advices delivered in those

Table 4: Details of clusters found for the advices delivered on four days

Cluster Size	27-August		20-September		20-October		23-November	
	Clusters	Farms	Clusters	Farms	Clusters	Farms	Clusters	Farms
(a)	(b)	(c=a∗b)	(d)	(e=a∗d)	(f)	(g=a∗f)	(h)	(i =a∗h)
1	22	22	33	33	49	49	40	40
2	5	10	8	16	10	20	8	16
3	6	18	6	18	15	45	6	18
4	4	16	6	24	3	12	4	16
5	3	15	5	25	1	5	2	10
6	1	6	4	24	2	12	1	6
7	1	7	0	0	1	7	0	0
8	0	0	1	8	1	8	1	8
9	0	0	0	0	2	18	1	9
10	2	20	1	10	0	0	1	10
11	1	11	1	11	1	11	0	0
12	0	0	1	12	1	12	0	0
13	0	0	2	26	0	0	0	0
14	0	0	0	0	0	0	0	0
15	0	0	1	15	0	0	0	0
16	1	16	0	0	0	0	0	0
17	1	17	0	0	0	0	0	0
18	0	0	1	18	0	0	0	0
19	0	0	1	19	0	0	0	0
20	0	0	0	0	0	0	0	0
21	0	0	1	21	0	0	2	24

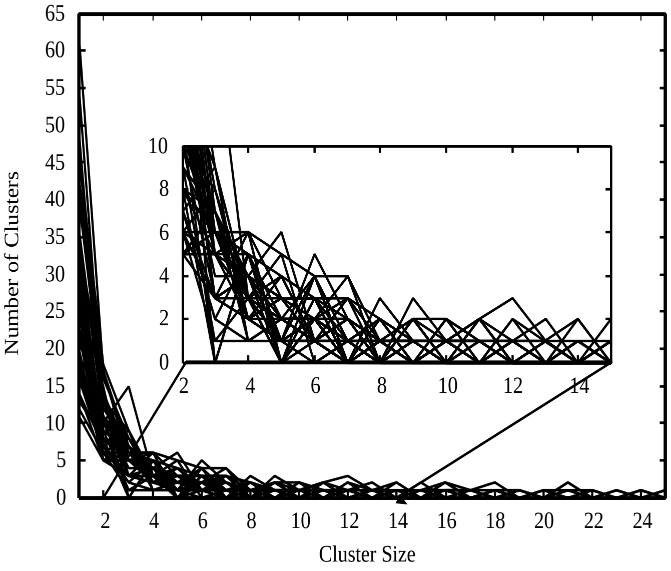

Figure 4: Cluster size versus number of clusters.

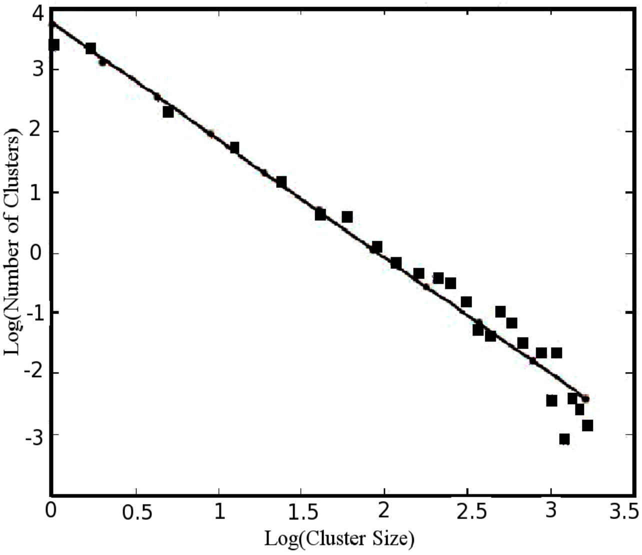

Figure 5: Cluster size versus number of clusters on a log-log scale.

weeks were less than 30 percent of registered farms (about 300 farms out of 1051). (From 1^{st} week number to 3^{rd} week very few farms had been registered and advices were delivered to those farms. After 18^{th} week most of the farms were at harvesting stage).

Table 6 shows the number of clusters found in different weeks. First column, “Week”, refers to the week number and the second column refers to the number of clusters during that week. The graph plotted between number of clusters and number of weeks (ranging from 3^{rd} week to 20^{th} week) is shown in Figure 7. It can be noted that from 3^{rd} week to 8^{th} week, the number of clusters is increased. This reflects that the number of distinct problems also has increased. The basic reason is that the cotton crop is subjected to different kinds of pest problems along with its growth and development. Some problems persist more or less the whole crop duration while others confine to a particular growth stage of cotton crop. The other important reason is that the pest dynamics change with the timing and effectiveness of management interventions initiated by the farmer. The cotton crop was being affected by a few insect pests like aphids, thrips and green jassids in the initial stages of its development besides some seedling blights and zinc deficiency which resulted in 32 different clusters during the 3rd week. Later as the crop grows, pests like spodoptera, helicoverpa, whiteflies along with magnesium deficiency lead to an increase in the number of problem clusters from 108 in the 4th week to 190 in the 8th week. The number of problem clusters was reduced from 190 in the 8th week to 95 in the 10th week owing to intensive pesticide sprays by the farmers. Again a new set of pests has emerged from 9th week onwards viz. pink bollworm, mealy bug, grey mildew, black arm etc besides the resurgence of some of the old problems. As a result, there was an increase in the number of problem clusters from 95 in the 10th week to 190 and 183 in the 12th and 13th weeks respectively. There after, the crop reaches the maturity phase and the farmers gradually recede from the management interventions. Hence the fluctuations in the later stages largely depend on the interaction between the crop, pests and the climate.

About 50 types of common problems occur on cotton crop which includes insect pests and diseases along with macro- as well as micro-nutrient deficiencies. However, the text analysis results show that the number of problem clusters are increased to 190 in the 12^{th} week. On analyzing agricultural advices, it was observed that multiple problems are faced by each crop. Each farm has faced about 3 different problems on an average. So, due to different combinations, the number of distinct clusters increased to 190.

Table 5: Clusters statistics for different sizes for the data of Table 4.

Cluster Size	27^{th} August	20^{th} September	20^{th} October	23^{th} November
1	22	33	49	40
2 to 5	59	83	82	60
above 5	77	143	110	58

4.3 Dynamics of farm problems over weeks

We have analyzed the dynamics of farm problems over weeks by conducting two types of analysis.

4.3.1 Analysis 1

In this experiment, we have considered the farms in a largest cluster in a week and analyzed as to how the problems of these farms vary during the subsequent week. The analysis is made by considering the largest cluster for each week. Table 7 shows the results extracted from the advice data set. The first column "Week", refers to the week number in which advices were delivered. The second column refers to the number of farms in the largest cluster during that week. The last column indicates the number of clusters during the subsequent week in which the advices in the preceding week's largest cluster have become members. In 4^{th} week, the largest cluster contains 67 farms; these farms have become members of 14 different clusters in the next week. In 5^{th} week, the largest cluster contains 15 farms, and these farms have become members of 2 clusters in the subsequent week. Similarly, in 7^{th} week, the 34 farms which have formed as one cluster have become members of 18 clusters in the subsequent week. Overall, it can be observed that for all weeks, the farms in one cluster are shifted to different clusters during the subsequent week.

It can be noted that the farms in one cluster are facing similar problem. So, the experiment results indicate that, though a group of farms have the same farm condition during one week, each sub-group of farms face different farm conditions or problems during the subsequent weeks.

4.3.2 Analysis 2

To analyze how farm problems vary, we have selected one sample cluster of farms which received the same advice in a particular week and examined the advices received by these farms in the subsequent and the preceding weeks. (Note that, in eSagu prototype, every farm has received the advice once in a week.) Table 8 shows the sample results. First column "Farm ID" indicates farm identifiers and the other columns indicate the cluster numbers in different weeks. For example, one can observe that how the farm with identifier "10761" changes clusters from 15^{th} to 20^{th} week in the following sequence $< 13, 18, 18, 58, 92, 48 >$. It means that the problems faced by the farm "10761" are changing in that manner. Similar is the case for the other farms.

It can also be noted that all four farms are belong to the same cluster number 18 and have received the similar

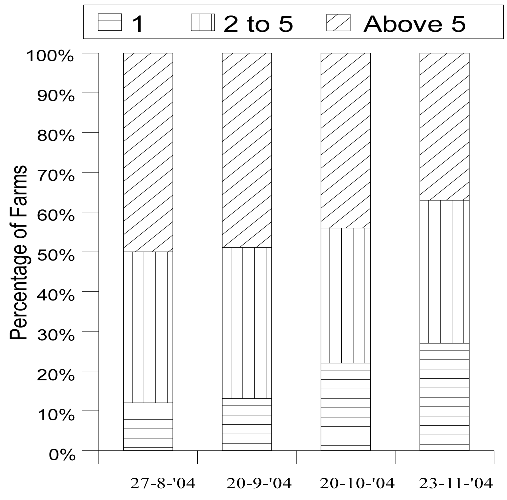

Figure 6: Percentage of farms in different cluster sizes for the data of Table 5.

advice in 17^{th} week. Even though these four farms belong to one cluster in 17^{th} week, they fall in different clusters in the subsequent weeks (18^{th}, 19^{th} and 20^{th} weeks) and also in the preceding weeks (15^{th} and 16^{th} weeks). We found similar phenomenon for the farms of other clusters.

The results show an interesting phenomena. Normally, as all are cotton farms/crops and belong to nearby area, the farms should face the same kinds of problems. However, on the contrary, the results show that that farm conditions are dynamic and several farms are facing a distinct sequence of problems during the crop period.

5 Discussion about the data set

Table 6: Number of clusters in different weeks

Week	Number of different clusters
3	32
4	108
5	137
6	141
7	148
8	190
9	102
10	95
11	124
12	190
13	183
14	110
15	140

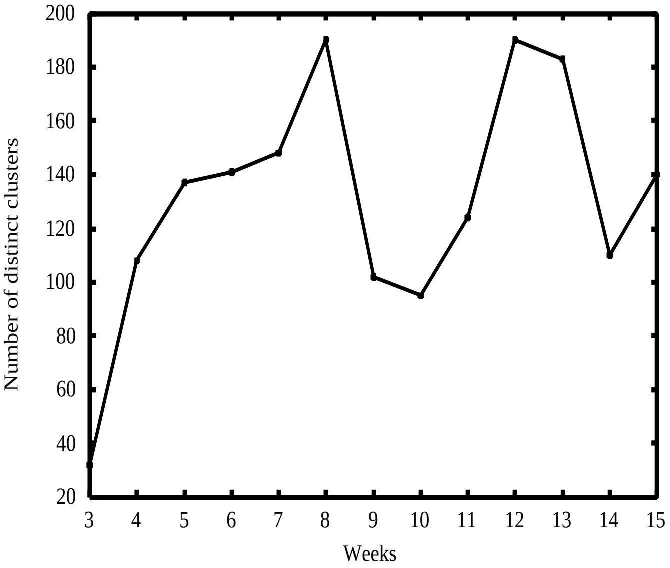

Figure 7: Number of clusters in different weeks.

Table 7: Movement of farms from one cluster to other clusters in subsequent week

Week	Number of farms in a large cluster	Number of clusters in subsequent week
4	67	14
5	15	2
6	4	3
7	34	18
8	17	7
9	12	3
10	32	13
11	15	11
12	10	7
13	10	5
14	12	3
15	22	3
16	21	8
17	7	6

Table 8: Dynamics of farm problems in different weeks. First column indicates farm identifier. The entries in the other columns indicate the cluster numbers in different weeks.

Farm ID	15^{th} Week	16^{th} Week	17^{th} Week	18^{th} Week	19^{th} Week	20^{th} Week
10761	13	18	18	58	92	48
10875	31	11	18	51	31	54
10220	31	11	18	92	88	65
10521	31	67	18	82	88	65

In the experiments, we have carried out textual analysis on the advice texts and reported the dynamics of farm problems. For the analysis, we have considered that the advice text represents the farm problem. It can be noted the agriculture scientist prepares the advice text mentioning the details of corrective steps that farm. So while preparing the advice text, the agriculture scientist also mentions the farm problem. So, the advice also contains the description of the problem. For example, the advice identifiers 5043 and 10509 (Table 3) have a mention about the problems viz. "Helicoverpa infestation was identified" and "Magnesium deficiency is observed" respectively. Hence, it is to state that the advice contains the terms which reflects the problem. So, we have considered that the analysis of advice texts is equivalent to the analysis of the corresponding crop/farm problems. However, there are some limitations. A single pesticide may control multiple pest problems which is considered as a single problem. For the same problem, each scientist could have prepared the distinct advice. Some pro-active measure is suggested in anticipation of the pest problem which has been counted as if the problem had occurred. Overall, we the analysis results reflect the dynamics of cotton farms in the state of Andhra Pradesh, India.

6 Summary and Conclusion

In this paper, we have reported the results about the dynamics of crop problems by applying text analysis methods on the advice data set generated under eSagu prototype. The prototype was implemented for 1051 nearby cotton farms belong to three contiguous villages during 2004-05 in the state of Andhra Pradesh, India. Normally, as all are cotton farms and belong to the same area/region, one expects that all or most of the farms should face the similar production problems. On the contrary, the text analysis results show that majority of the farms are facing distinct crop production problems. The results are summarized as follows.

- At any given point of time, the majority of farms are suffering from distinct farm-specific problems.
- The farms are facing a distinct problem sequence during the crop period.

The results imply that the crop condition is influenced by multiple factors. Some of the major factors that influence the crop condition are as follows: crop variety, history of the farm, type of soil and soil nutrients, cultural practices followed in a farm and farm environment.

The crop variety is one of the factor because, the level of crop tolerance towards certain pests or diseases attacks depends on the crop variety. The history of the farm will also influence the crop condition because the current problems on the crop depends on the type of the crops that have been grown in the preceding seasons and corresponding farm management practices followed. The soil type and its nutrient levels play a major role in crop condition. The farm management practices carried out on the farm becomes a factor because the farm condition might be different from the other farms for which correct cultural practices have been followed including crop sanitation (clean cultivation). The farm environment which includes the population of pests and predators in a farm becomes one of the factor. Suppose, if the predators for certain pests are very high in number in a farm, then it is more unlikely to face the problem from those pests.

When the farm condition is influenced by such multiple factors, disseminating agricultural information in a generic manner may not be effective to reduce crop failures or to improve crop productivity. The generic agro-advisory delivered through print media like news papers and magazines or electronic media like radio and television do not take farm-specific dynamics into account. For each farm, a distinct information is needed at regular intervals. So, it is necessary to deliver agricultural expert advice to each farm by building personalized agricultural advisory systems to reduce crop failures and improve crop productivity. The recent developments in ICTs can be exploited to build such systems in a cost-effective and efficient manner.

ACKNOWLEDGMENTS

The eSagu system has been developed by IIIT-H, Hyderabad, India and Media Lab Asia. The initial part of the project was supported by Ministry of Communications & Information Technology, New Delhi.

REFERENCES

Cotton Doctor, 12 Dec. 2008, http://www.ipni.net/cottondoc.

'eSagu: An IT Based personalized agro-advisory system', 12 Dec. 2008, http://www.esagu.in.

Han, J., and Kamber, M. (2006) 'Data Mining Concepts and Techniques', *Morgan Kaufmann Publishers*, II Edn.

Krishna Reddy, P., and Ankaiah., R. (2005) 'A Framework of information technology based agricultural information dissemination system to improve crop productivity', *Current Science*, vol. 88, no. 12, pp. 1905-1913.

Krishna Reddy, P., Sudharshan Reddy, A., Venkateswar Rao, B., and Reddy, G.S., (2005) 'eSagu: Web-based Agricultural Expert Advice Dissemination System (2004-05)', *Final Completion Report of Research Project. Submitted to Ministry of Communications and Information Technology (2005).*

Krishna Reddy, P., Ramaraju, G.V., and Reddy, G.S., 2007 '*eSagu*^{TM}: A data warehouse enabled personalized agricultural advisory system', In: Proceedings of the 2007 *ACM SIGMOD Conference.*

Pine II, B.J. (1993) 'Mass Customization', *Hardward Business School Press*, Boston, Massachusetts.

Ratnam, B.V., Krishna Reddy, P., and Reddy, G.S., 2006 'eSagu: An IT based personalized agricultural extension system prototype - Analysis of 51 farmers case studies', *International Journal of Education and Development using ICT (IJEDICT)*, Vol. 2, No. 1.

Rita Sharma, (2002) 'Reforms in Agricultural extension: New policy framework', *Economic and Political Weekly*, pp. 3124-3131, (27 July 2002).

Roget, P.M., (1911) MICRA Inc. An electronic thesaurus derived from the version of Roger's Thesaurus published in 1911, 12 Dec. 2008, http://www.infomotions.com/etexts/ gutenberg/dirs/etext91/roget15a.htm.

Silberschatz, A. and Zdonik, S.B. (1996) 'Strategic directions in database systems - breaking out of the box', *ACM Computing Surveys*, pp. 764-778.

Michael Steinbach, George Karypis, and Vipin Kumar, (2000), *'A comparison of document clustering techniques', In: KDD Workshop of Text Mining (2000).*

'Stop words', 12 Dec. 2008, http://www.ranks.nl/tools/stopwords.html.

Uday Kiran, R., and Krishna Reddy, P., (2007), 'Understanding the dynamics of crop problems by analyzing farm advisory data in eSagu^{TM}', *Lecture Notes in Computer Science*, pp. 272-284.