KrishiEkta: Integrated Knowledge and Information Distribution System for Indian Agriculture

Rakshit Agrawal  
CSC India Pvt. Ltd.  
Noida  
rakshit_agrawal@outlook.com

Mridu Atray  
Indian Institute of Technology  
Delhi  
mriduatrav@outlook.com

S. Krishna Sundari  
Jaypee Institute of Information Technology, Noida  
krishna.sundari@jiit.ac.in

ABSTRACT
In this publication, the authors describe KrishiEkta, an integrated knowledge distribution system for Indian Agriculture. KrishiEkta addresses the inadequate access of important information to various stakeholders in Indian agriculture. Access of rich internet platforms is not available to various farmers in small rural parts of the country. KrishiEkta system collects information from multiple sources through the internet on different categories falling under the theme of Agriculture. It then organizes this data into categorical models. This information with categorical information is the prepared for distribution in four different formats: SMS, tweet, Facebook post and digests e-mail. SMS are the preferred medium of information distribution among small farmers. Whereas many educated stakeholders use e-mails and social networking services. Summarized information with relevant links and identifiers per format are finally distributed to the subscribed users through the Distribution Engine.

1. INTRODUCTION
India is an agrarian economy with more than half of its population depending on agriculture for livelihood. Apart from scientific knowledge, farmers are also not rightly informed about weather, market and related information. It is important to note though that it is not just the farmer alone, but various other stakeholders (seed manufacturers, market workers, agriculture scientists etc.) who need to keep themselves up to date with the current happenings and new techniques being used around the world.

The motivation for KrishiEkta came from various hours of interaction with farmers from Bulandshahar district of Uttar Pradesh, researchers from Indian Agricultural Research Institute (IARI), few members from the Ministry of Agriculture, and students from engineering as well as agriculture universities.

Discussion with the farmers stressed on the fact that information access through mobile phones gives them the power of information right in their hands. The success of SMS based initiatives such as RML [2] supported the fact that SMS is a suitable medium for farmers. Since there are a number of resources on the internet carrying latest agriculture news, the scientists showed inclination towards receiving a comprehensive digest of the current happenings in the form of email. The students insisted on the usage of social media platforms such as Facebook and Twitter.

In this publication, we propose the novel concept of a platform called KrishiEkta, which collects information from different sources on the internet and distributes it to the subscribers in their preferred output media. The name of the system signifies unifying all stakeholders onto a single platform.

Copyright is held by the owner/author(s).

ACM DEV’-4, December 6–7, 2013, Cape Town, South Africa  
ACM 978-1-4503-2558-5/13/12.  
http://dx.doi.org/10.1145/2537052.2537071.

2. SYSTEM ARCHITECTURE
The KrishiEkta system architecture is shown in figure 1. It is a multiple level architecture, with four engines handling the main operations. This system is assisted by a database at the backend. First the Parsing Engine looks up through a list of sources and extracts all article content. Categorization Engine assigns relevant categories to every article. Content-generation Engine generates output strings for different formats and interacts with Translation sub-engine for language translation. Distribution Engine finally sends all information to subscribers in their desired format.

Figure 1. KrishiEkta System Architecture.

3. THE PARSING ENGINE
First stage of processing in the KrishiEkta System is the Parsing Engine. System here interacts with database table called Source table containing details of data and meta-data of every source. Objective of this engine is to extract relevant data from HTML or XML files of sources. XML can be parsed via standard library functions. In order to parse through HTML pages, the system builds an HTML code tree from the source using HTML Parser. From this tree it picks up the tag information associated with every source and accordingly extracts out the element.

This process is done in two levels. First while parsing through a general site page, it fetches compares article date with a specified date. URL and title of items beyond specific date are picked up by placed in a table named Raw Data Table. This table holds complete data that has been derived from every source. At this moment, only date, URL and title are inserted into the table. Once all the items satisfying the date criteria are retrieved, the parsing engine moves up to the second level where it identifies the title, main content of the article, location information and associated tags or keywords if available. Raw Data table now holds all the details of items to be summarized further by the system.

4. THE CATEGORIZATION ENGINE
Categorization Engine categorizes articles into specified areas. Some users need to stay updated with agriculture market news, while some might need guidance articles related to agriculture. Therefore, we categorize our data into areas like – market information, weather updates, advisory notes, event information,
etc. The Source table holds mapping with categories. Before a category is assigned to an article, it has some measure of possible categorization based on its source. For primary categorization, keywords and tags in an article are used. Using word lists per category, keywords select up to three categories and associates respective weight age with them. Then we use the Stanford Topic Modeling Toolbox [1] to assign categories selected by Topics in document. Once the categories have been assigned from all three methods, the ones with highest weights are picked in order to get final categories of an article.

5. THE CONTENT-GENERATION ENGINE
After the categorization engine has updated the Raw Data table with relevant categories for all the articles, the content generation engine summarizes articles for four different output media – SMS, Twitter, Facebook and E-mail.

5.1 SMS
SMS being 160 7-bit character sized requires summarization. In our application, it essentially means reduction of article information to few keywords with little or no grammatical attachment. SMS string therefore keeps article title length to a maximum of 60 characters. Next it adds location kept to a maximum of 10 characters. Third known item is the source short name. This short name is placed at the end with a hyphen mark as prefix. Now the engine calculates possible length for remaining text and puts words associated with the article’s category specific vocabulary based on their frequency in the article and completes SMS string.

5.2 Twitter
Twitter, a social networking enables its users to share messages of up to 140 characters, known as "tweets". Tweet generation begins with placing complete title of article first. KrishiEkta system uses the URL of article and places it in tweet string as a short link of 20 character length. Content in tweet will be based on keywords used as hash-tags by placing "#" in front. All the words used as hash-tags are picked up entirely from the article text.

5.3 Facebook
Facebook is the most used social networking service in the world. Technically there is no character limit in a Facebook post but size has to be short for comfortable reading to. Title of the post is used exactly along with URL for article source. Article summarization will be done using standard genetic algorithm-based attribute selection technique described in [3]. The engine uses standard algorithm for unconstrained length format.

5.4 E-mail
E-mails are the most common methods of communication over internet. A large number of stakeholders in agriculture are not social networking enthusiasts but active users of e-mail. E-mails are not generated by the content generation engine for individual articles but as digest mail. All the items belonging to selected category beyond a specified date are combined together in the form of a mail consisting title, URL and summarized text.

5.5 Translation Sub-Engine
India is a land of multiple languages. English, the common language of digital media, is not the common language in Indian rural areas. While serving the educated stakeholders, English is perfect. But to reach farmers, we need language translation of our content. At present, KrishiEkta performs only basic translation of content using Microsoft Bing Translation API into Hindi language.

<table>
<thead>
<tr>
<th>Distribution Media</th>
<th>Source</th>
<th>Content</th>
<th>Date</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMS</td>
<td>Short name</td>
<td>Keyword + headings - max 60 chars</td>
<td>Not needed</td>
<td>Needed (10 chars max), cut name to 10 if bigger</td>
</tr>
<tr>
<td>Tweet</td>
<td>Short link (20 chars)</td>
<td>Keyword summary within minimum 85 chars</td>
<td>Publishing date + Tweet date</td>
<td>Needed</td>
</tr>
<tr>
<td>Facebook post</td>
<td>Link</td>
<td>Algorithmically Summarized</td>
<td>Publishing date + Post date</td>
<td>Full</td>
</tr>
<tr>
<td>E-mail</td>
<td>Full Name and Link</td>
<td>Algorithmically Summarized</td>
<td>Date to be placed with every mail</td>
<td>Full details</td>
</tr>
</tbody>
</table>

6. THE DISTRIBUTION ENGINE
Distribution engine sends specific format outputs to subscribers based on their selected categories. For SMS, A GSM modem attached to the system sends SMS to user with the help of an internal tool created using AT-commands and connected with system via PySerial. For e-mail, based on the selected categories and location, digest mail is created as explained in the previous section for a user. System then pushes that mail to respective users. Facebook post is a non-subscriber based platform and the system posts all items on its own page itself. Twitter is primarily a non-subscriber method but can be used as subscriber method also. Generally every tweet when generated is pushed to Twitter through a script similar to Facebook post. But for subscriber with Twitter handle (@subscriber), tweets are sent using subscriber’s handle in front. Both Facebook and Twitter have private message options also which might be used by this system in future.

7. CONCLUSION AND FUTURE WORK
KrishiEkta as a working model is in Alpha Testing mode now. We’ve been running tests with some farmers and gardeners for the SMS model. E-mails are tested with some experts from universities and few members from the Agriculture Marketing Department. This work has been motivated by hope to provide the inaccessible information to users as even heading of a news item or an advisory method can make help a farmer in many ways.

KrishiEkta system needs to help the inner core of Indian farmers and therefore regional language support is needed in the system. Current version of translation sub-engine needs to be highly optimized for practical success. Also a lot more categorization will be created in the system allowing finer filtering of data.

8. REFERENCES