Location Specific Summarization of Climatic and Agricultural Trends

Sunandan Chakraborty, Lakshminarayanan Subramanian
Computer Science Department
New York University
New York, NY, USA
sunandan@cs.nyu.edu, lakshmi@cs.nyu.edu

ABSTRACT

Climate change can directly impact agriculture. Failure in different aspects of agriculture due to climate change and other influencing factors, are extremely rampant in several agrarian economies, most of which go unnoticed. In this paper, we describe the design of a system that mines disparate information sources on the Web to automatically summarize important climatic and agricultural trends for any specific location and construct a location-specific climatic and agricultural information portal. We have evaluated the system across 605 different districts in India. The results revealed a pan-India scenario of different problem affected areas. The key findings from this work include, around 64.58% of the districts of India suffer from soil related issues and 76.02% have water related problems. We have also manually validated the authenticity of our information sources and validated our summarized results for specific locations with findings in reputed journals and authoritative sources.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process, Retrieval models

General Terms

Design, Measurement

Keywords

Agriculture, Climate, Web service, Emerging Region

1. INTRODUCTION

Agriculture forms the backbone of several emerging economies. In India, 58.4% of the population is directly dependent on agriculture [31] for their livelihood. In the past few years, several agrarian regions have been severely affected, due to a combination of several factors including climate, lack of water availability, soil infertility etc. However, in reality, many policymakers and the general public are often unaware of the status of agricultural conditions across different localities within their countries.

In this paper, we ask the question: *Given any location, can we mine appropriate information sources and build a system that can summarize important climatic and agricultural trends in the specified location.* Such information could potentially be useful knowledge to both raise awareness about specific trends as well as for policy makers in learning about locations with problematic agricultural conditions.

For a given location, there is a wealth of information about that location on the Web from a wide range of disparate sources. There may be different news articles, blog reports, non-governmental organization reports and other information sources that can provide topic-specific information, such as, "soil erosion" or "water availability" or "rainfall", about the specified location. For example, a web search on "Vidharbha water problems" can yield several articles that may highlight specific case reports of water problems reported within Vidarbha.

In this paper, we present the design of a system that automatically constructs a location-specific climatic and agricultural information aggregation and summarization portal based on disparate information sources from the Web. Given a location, the system searches the Web for information concerning different parameters influencing agriculture and climate and presents a summary of relevant information.

Our system is built around three key ideas. First, we (manually) identify target topics of interest within climate and agriculture (such as soil, water) and construct a list of *appropriate search queries* that comprehensively describe the different aspects of the target topic. Second, for each target topic (such as soil or water), we download the top search result pages and perform information extraction on the textual content of these pages. The *information extraction* process aims to extract the critical textual snippets that can capture the key trends within the target area. Finally, we perform *information summarization* where the goal is to identify key trends corresponding to each target topic. We have tailored standard information retrieval techniques to address these problems. This summarized information on the location can be utilized to detect different problems and infer possible remedies from it. Hence, the aim is to bring to fore the important as well as lesser known facts, thereby increasing the availability of knowledge. Clearly, availability of knowledge can lead to detection and potentially prevent any catastrophes.

This work is targeted for policy makers, such as government officials, as a source for useful information. While our system is by no means perfect, we believe that this system does indeed have value for policy makers for creating general awareness of critical trends. Web information is known to be inherently noisy and we have taken several steps to both clean our data sources as well as validate our results

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

from authentic journals and articles. The system is aimed as a tool to provide an initial picture where none currently exists but is not meant to be designed as an authoritative response system; further investigation from humans is essential to corroborate the findings before any critical decision making.

As a first step towards showing the value for such a system, we are currently working with the Institute of Financial Management and Research (IFMR), India to incorporate this system as an additional information gathering tool in designing their agro-risk insurance solution. From their perspective, they currently do not have any detailed per district agricultural trend information within India to design informed risk insurance products: hence, they have expressed interest in using our system even if our inferences are not perfect.

The work has been evaluated on 605 districts of India. The results revealed an all India scenario of different problem affected areas. According to the system's findings, 64.58% and 76.02% of the districts of India suffer from soil and water related problems respectively. Around 47.94% districts have crop health, fertilizer or pesticide related issues. The figures are particularly threatening for the central India, where around 90% of the districts are affected by various serious issues, like droughts, erosion, water contamination etc. In addition, The PageRank [37] distribution of the pages showed that the average value of PageRank of the pages used was 6.4/10. It depicts the reliability and the authenticity of the pages. An analysis on the type of webpages used revealed that mostly government portals, organizational reports and news articles were used in this process. We have also validated many of our results and show that they match with inferences made in authentic scholarly articles for regions (where such articles exist).

2. AGRICULTURE IN INDIA

Historically, due to favorable geographical conditions, India's economy has always been largely dependent on agriculture. Perennial rivers have provided with abundant water, as well as formed fertile basins, both crucial for farming. Consequently, about 55.9% of total land area in India is cultivable [32] and agriculture contributes approximately one-fifth of total gross domestic product (GDP) and it accounts for about 10% of the total export earnings[31]. Also, 58.4% of the population in India is dependent on agriculture[31] for their livelihood.

Various, constantly evolving factors have drastic effects on agriculture, as well as on its influencing parameters, such as soil conditions, water availability, climate etc. Recent studies revealing that agriculture in India is under serious threat due to changing environment [33]. Changing climatic conditions, such as changes in monsoon patterns, is resulting into frequent floods in some regions and droughts in others; Rising sea levels are increasing the salinity of the soil in coastal areas. These changes are leading to low productions in farming. Apart from natural causes man-made changes, like, massive deforestation or increase in environmental pollution can also adversely affect farming. Practices such as, uncontrolled use of chemical fertilizers or pesticides can also decrease the soil fertility, resulting in lower yield in production.

While there is a general awareness of these facts but there is a definite lack of knowledge on how these factors have an effect in agriculture or in particular locations. Hence, most of the problems are not taken care of properly and they continue to cause destruction. A probable remedy would be to acquire knowledge and deduce solutions out of it.

From the above discussion it is clear that, any decline in agriculture can bring upon devastating effects on India's economy. Hence, a system to safeguard agriculture can bring upon positive consequences.

3. SYSTEM DETAILS

The overall idea of our system is to leverage the wealth of information available on the Web to automatically construct a location-specific information portal that can summarize the key climatic and agricultural trends in a given location. Our basic system is designed around the following design steps,

- Creating a location specific repository of documents from the Web
  - Identifying target topics and appropriate search queries
  - Downloading the top results using the query set and deriving new features
  - New searches with the derived features
  - Validating the webpages to avoid inauthentic data
- Analyzing the text for identify critical patterns
  - Extracting relevant information from the text
  - Inferencing the main problem areas in the location
  - Summarizing and presenting extracted information and the key trends

For example, if we want to analyze the problems regarding agriculture and climate change in a particular region, say Jabalpur, in Madhya Pradesh, we can search the web for documents about Jabalpur and related topics, using queries, such as, "Jabalpur soil erosion" ,"Jabalpur water availability" etc. Relevant information from the downloaded webpages can be extracted and summarized to identify the critical trends regarding the major problems and threats faced in Jabalpur in terms of agriculture and climate change.

4. INFORMATION RETRIEVAL FROM THE WEB

This section discusses the process of gathering information of the web. This process has two major steps. Framing a set queries which helped to search for the best resources from the vast pool of documents the Web offers. In the subsequent step some preliminary processing was done to validate the information sources.

4.1 Base Document Set

A significant step of this work was to create a channel in the web to get relevant information. In order to obtain such information an important task was to identify some key terms associated with the topic, i.e. agriculture and its dependencies, such as climate. There are various factors on which agriculture is heavily dependent upon. A carefully chosen set of such terms can be used as queries to search

Table 1: A Segment of the Query List

Categories	Keywords
Soil	Type Quality Fertility Erosion
Water	Ground water level Rivers Dam Irrigation
Climate	Climate Weather Flood Rainfall Drought Deforestation
Agriculture, Crops, Pesticide/ Fertilizers	Crops Crop disease Pesticide Seeds Methods Harvest Fertilizers

and identify meaningful and contextual pages from the Web. Hence, the primary step of this work was to identify the important categories and frame a list of queries which can provide a good description of each category. We identified six major categories influencing agricultural production. These are: soil, water, climate, agricultural practices, crops and pesticides/fertilizers. For a more minute description, each category was associated with numerous keywords. For example, soil is an important parameter in agriculture. The main characteristics of soil which influence agriculture are soil type, soil fertility, soil moisture, etc. Additionally, there can be some important phenomenon associated with a category. For example, for soil, such a phenomenon can be "soil erosion". The location, category and a keyword are combined to form the search query. A typical search query to a search engine took the form of "Location + Category + Keyword". A simple example is "Jhabua + soil + erosion". The final list had 52 queries across all 6 categories. Table 1 shows some of the queries used across the categories.

Note that the set of keywords in this table are only representative and may not be complete. We specifically chose these keywords as one representative sample set of queries to illustrate the power of this approach. This approach can be generalized for a better choice of search queries.

For each query, we downloaded the first 64 documents featured as the top results. So, these 64 documents for each 52 queries served as the initial corpus of documents and a basis for primary analysis. Google API [35] was used as a tool to search the Web.

The process of acquiring documents is summarized below.

Let Q_c be the set of keywords related to the category $c \in C$, where C is the set of 6 categories. For any location Loc,

For all $c \in C$ do
For all $q \in Q_c$ do
$\{D_c\} \leftarrow searchWeb(Loc + c + q)$
So, $\{D_c\}$ represents a set of documents for category c
The set of documents $\{D\} = \{\{D_{soil}\} \bigcup \{D_{water}\} \bigcup$ ....} represent the preliminary set of documents after this step.

4.2 Extended Document Set

The documents obtained from firing the queries as described in the previous section represent the initial repository. This repository can potentially reveal some critical information which was not actually referred by the preliminary query set. For example, the collection of documents returned by the queries for water in Jhabua, had repeated occurrence of the phrase "fluoride contamination" but the term "fluoride" was not in the query set. This depicted that the area is suffering from fluoride contamination and this particular problem area was ignored in the primary query set. Thus, "fluoride" can be a new *feature* with respect to the topic in hand.

We used an N-gram based approach to extract these extra features from the initial set of documents. An *N-gram* refers to a collection of N consecutive words in a document. We defined an *N-gram* to represent a *critical trend* associated with a topic if the N-gram had a high rank but was not a very commonly occurring N-gram on the Web. Hence, an N-gram with high Term Frequency (TF) is an indicator of a critical trend. However, frequently occurring *N-grams* could also be just an artifact of the language in which the text is written (for example, the term *"of the"*). To differentiate those *N-grams* that are important to the topic, we eliminated the N-grams which occurred very frequently in common English texts. The list of the very common N-grams was obtained from the Linguistic Data Consortium dataset. An N-gram with significantly high frequency and not present among the frequent N-grams (from the LDC corpus) is definitely an important phrase within our text corpus. Therefore, this N-gram is one of the features identified.

These extra set of features were included as new keywords and were used to perform an extended search in the web to get additional documents. The new pages downloaded from this second phase of searches were added to the original repository. The process of identifying new features and getting additional documents is summarized below:

For all $c \in C$ do
$\{T\} \leftarrow computeNgrams(D_c, n)$ [where $1 \leq n \leq 5$]
For all $t \in T$ do
If $freq(t) \geq freq_{th}$ and $t \notin LDC$ [where LDC represents the frequent N-grams in LDC corpus]
$\{NewFeatures\} \leftarrow t$
$\{D_c\} = \{D_c\} \bigcup searchWeb(Loc + t)$
End If
End For
End For

At the end of this process, D_c will be an extended set of documents for the category c.

4.3 Page Validation

Validation of the information obtained was an essential step as some websites can produce false data. Moreover, analyzing data from the same source can falsely increase the importance of a fact. Hence, to make the information avail-

able from our system robust there was an additional step to check for duplicate sources and also their authenticity.

Duplicate Removal.

Web searches using different but similar query terms might return pages from the same domain. As a result, the same information will be repeated and it will be erroneously considered important for repeated occurrences. However, in reality the same information has been obtained from the same domain and it is a mere repetition of the same facts. Hence, during downloading the pages from the web, the pages were checked for duplicate domains. That is, a page was eliminated if contents from pages from the same domain were selected for a certain number of times previously. Thus, duplicate texts were avoided.

In addition, there was a duplicate checking module which explicitly checked whether duplicate text exists. Application of different semantic, lexical or statistical methods [7][8] can yield better results but we followed a simpler approach for similar text exclusion. We represented the paragraphs using feature vectors. Presence of a feature was represented as 1, and 0 otherwise. Thus, the paragraphs were represented by a string of 1s and 0s. Low distance between paragraphs depicted a similarity between them. Paragraphs whose distances were lower than a threshold were considered for duplicate removal. This step ensured that the information produced by the system was from disparate sources.

Authenticity of the Sites.

In some websites, like blogs or personal homepages, data can be presented without any authentication. Our system is vulnerable to this kind of incorrect information. One of the simplest way to avoid that was to use the PageRank [37] score of the page. This score gave an estimation of the reliability and relative importance of a webpage. In our system, we have considered only the pages whose PageRank was greater than 3 out of 10. This simple screening ensured some authenticity in the information retrieved by our system.

5. SUMMARIZING CRITICAL TRENDS

In the previous section, we discussed how information was gathered from the Web. The next step involved identifying the relevant text regions from the documents. In order to realize that, we used the concept of term frequency-inverse document frequency (tf-idf) [36] measurement. The tf-idf is a metric through which the importance of a word in a set of documents can be understood. It is a combination of two separate weights, term frequency (tf), which is a local parameter and inverse document frequency(idf), which is a global parameter. The tf is the frequency with which the word appears in the document and idf is the estimate of the number of documents it appears in. Low value of idf signifies the generality of the term and hence, the low importance in the current context. Higher value of combined tf-idf denotes greater associativity between the word and the document. We were interested in finding the relevant portions within a document. Hence, the concept of inverse document frequency was slightly changed and adapted to find the same measure, but with respect to the paragraphs in the documents. Here, the value of tf-idf signified the associativity of the word with a paragraph in a document.

We calculated the cosine similarity between the query (a combination of the name of the location and the keywords) and the paragraphs of the documents, using the tf-idf value between them. The following formula was used for this computation.

For, a query or a feature q and paragraph p_j

$$sim(q, p_j) = \frac{\Sigma_{i=1}^{T} w_{i,j} . w_{i,q}}{\sqrt{\Sigma_{i=1}^{T} w_{i,j}^2} . \sqrt{\Sigma_{i=1}^{T} w_{i,q}^2}}$$

where T is the number of terms in the query q and $w_{t,p}$[1] represents the weight of a query term (t) within the paragraph (p), defined as,

$$w_{t,p} = tf_t . log(\frac{|P|}{|\{t \in P\}|})$$

where, tf_t is the term frequency of t in paragraph p and P is number of paragraphs in the document.

The value $w_{t,p}$ for each query term t denoted the relevance of t in paragraph p. Note that this measure is *inverse paragraph frequency* which is slightly different from the traditional inverse document frequency (IDF) used in information retrieval. Inverse paragraph frequency is the IDF metric applied considering each paragraph as a separate document. This ranking mechanism allowed us to rank at a paragraph granularity to extract the most useful text snippet paragraphs for a given topic and location. The significance of computing cosine similarity ($sim(q, p_j)$) for all query terms t is that it is a quantitative estimate of the similarity between the query and each paragraph of the document. High value of similarity between a query q and a paragraph p_j denotes higher importance of p_j for q. This implies p_j is an important portion of the document for the query q. Hence, p_j should be selected as a part of the information being searched for q.

The decision, of selecting a part of the text as relevant, was based on the similarity value discussed above. If the similarity value exceeded a threshold value then it was considered to be important. The selection criteria was defined as:

$$sim(q, p_j) > Sim_{th}$$

The final value of Sim_{th} was determined after trials with different possible values. The system was evaluated for 10 locations for each experimental value of the threshold. Manually observing the outcomes from these trial runs the final value of Sim_{th} was set to 0.55, which demonstrated relatively better results compared to other values of Sim_{th}. Results were too generalized when the threshold value was less than 0.55. On the other hand, a greater value tended to eliminate important facts. Hence, any information with similarity value greater than Sim_{th} was assumed to have the right amount of details.

5.1 Inference Engine

The objective of the inference engine was to identify the main problems in a region with respect to a category from

[1]This is similar to tf-idf measure but here it is computed w.r.t paragraphs instead of documents

the extracted text of the documents. In other words, provide answers to questions like, "Are there any soil related problems in Jabalpur?" or "Does Sambalpur suffer from water scarcity?". Inferring such critical facts were based on the following heuristics.

- **Presence of Keywords:** Each topic is associated with a set of keywords. Some of them were framed manually and the rest were automatically extracted from the repository. Presence of such keywords in a text establishes the importance of the sense portrayed by the query terms. For example, from high frequency of the terms "soil erosion" or "fluoride contamination", the presence of such problems in that location can be deduced. As, we considered N-grams ($N \leq 5$) during feature extraction, repeated mention of "no soil erosion" or "soil erosion was not observed" cannot be mistakenly inferred as soil erosion problems. In such a case the frequency of "no soil erosion" or "soil erosion was not observed" would exceed "soil erosion".
- **Number of occurrences from disparate sources:** This metric provided an estimate of the popularity of the information. The severity of the problem can be interpreted from the number of disparate sources which reported the problem.
- **PageRank [37] of the pages:** This metric is a quantitative estimation of the reliability of the information source. Verification of the authenticity of the source can help to avoid inclusion of trends incorrectly termed as critical.

The following algorithm was proposed for the inference engine,

Procedure: SummProblems

1. $Keywords_c$ is the set of keywords for category c
[$Keywords_c$ is the collection of both manually constructed and automatically extracted keywords]
2. For all i where $terms_i \in Keywords_c$ do
3. $S \leftarrow \sum_{k=1}^{D_c}[tfidf_k(term_i) + (A_{ik} * PR_k)]$
4. If $S > T_{th}$ then
5. The presence of the problem described by $term_i$ is $true$
6. End if
7. End for
8. Repeat steps 1-7 for all categories $c \in C$

D_c represents the total number of documents gathered for a category c (e.g. soil, water). For example, for soil, D_{soil} is the collection of the documents obtained after firing all soil related queries (such, as soil erosion, soil fertility) etc. $tfidf_k(term_i)$ represents the tfidf of $term_i$ with respect to the document D_k. A_{ik} is a binary term and $A_{ik} = 1$ if $term_i$ is present in the document D_k and 0 otherwise. Finally,PR_k is the PageRank value of the webpage from where D_k is obtained. The $\sum(A_{ik} * PR_k)$ represents the sum of PageRank of all the different sources which had these terms (combining PageRank with the number of disparate sources).

The N-gram analysis also helped in avoiding misleading inferences. For example, repeated occurrence of "water shortage" or "soil erosion" can mean presence of these problems in that region. However, the phrases could have been "no water shortage" or "little soil erosion". An N-gram analysis is resistant to such ambiguities as it involves a larger portion of the text.

5.2 Summarization

The final step was to aggregate the whole information gathered from disparate sources and prepare a summarized version of all the relevant information and critical trends and present them to the users. Text summarization involves extracting only the meaningful portions from larger text and prepare a concise form. There are various methods for automatic text summarization [10][11][12]. We adopted a simple *N-gram* based approach.

As discussed in the previous section, we obtained a set of keywords (N-grams) which identified the critical parts of the text. For each location and categories we have set of query keywords (manually framed) and a set of features extracted automatically from the gathered text. Using these key terms the important text regions within the documents were identified and retained them as part of the summarized text. The paragraphs in the documents were ranked on the basis of the occurrences of these terms. The ranking values were computed as,

$$R_p = \sum_{i=1}^{n}(f_p^i)$$

where, R_p denotes the rank of the paragraph p and f_p^i denotes the frequency of the i^{th} key term in the paragraph p for all the n key terms. The n terms include the predetermined query set as well as the extracted features. Paragraphs with high R_p values are included in the summarized text.

The resulting text is a concise form of the all the text gathered but includes most of the crucial parts. Moreover, this summarized form can be a useful tool for the users to get a quick reference of the important facts.

6. EVALUATION

In this section, we discuss about the performance of the system. First we discuss the methodology used in this evaluation process. Then we elaborate on some of the key findings from the results returned by the system. Then we talk about the variety and the authenticity of the sources used. Finally, we try to perform a qualitative validation on the information obtained by comparing the findings with established theory and facts published in reputed journals.

6.1 Results and Observation

6.1.1 Methodology

The system was applied on 605 districts of India [34], which covers almost whole of the country. For each district, the documents were downloaded and analyzed as discussed in the previous section. The performance of the system was evaluated on the basis of the extracted and summarized information from the results of these 605 districts in India.

Inferencing was based on the method discussed in Section 5.1. A location is said to be suffering from a problem (e.g. soil erosion) if the inference module reported the presence of that problem within the document set returned for the location. However, the current version of the inference module is incapable of reporting absence of a problem. If in a location there was no mention of a certain problem the module will cannot assert that the problem is not present there. In such cases the report can be stated as "do not know" or "cannot

Table 2: Zone-wise Soil affected areas (%)

	Erosion	Infertility
North	52.70	31.65
East	65.14	38.13
North-East	70.33	27.78
Central	87.37	85.35
South	53.61	22.84
West	43.44	56.83

Table 3: Zone-wise water affected areas (%)

	Scarcity	Drought	Flood
North	45.32	53.29	33.16
East	78.56	50.26	69.65
North-East	72.81	23.52	39.17
Central	93.49	89.32	20.06
South	30.35	40.21	35.12
West	64.14	70.56	30.78

decide". For example, if the system reports 40% of the locations suffer from water scarcity it does not mean the rest 60% is free from water scarcity. It suggests that those 60% of the locations may or may not suffer from water scarcity, the system was unable to infer that.

6.1.2 Results

The summarized text returned by the system revealed some general facts for each location. Like, primary soil type, average temperature, average annual rainfall, normal water availability, major crops, coverage of irrigation etc. In addition, the inference engine reported various problems related to climate and agriculture in these locations.

The system reported that about 64.58% of the districts suffer from soil related problems, mainly soil erosion and infertility and 47.94% of the districts has issues related to crop health, pesticides and fertilizers. The system also reported that 76.02% districts suffer from water related issues, such as water unavailability, droughts, floods etc. This findings are summarized in Figure 1. Table 2, 3, 4 shows zone-wise data of different problems regarding soil water and agriculture.

The pan-India data reveal the graveness of the situation and can act as a serious warning. Almost every region of the country has reported problems regarding soil, water or agricultural issues. This data also signals the necessity of taking immediate steps to prevent further deterioration because any worsening may prove catastrophic.

Particularly threatening picture is obtained from the central part of the country, including the states of Rajasthan, Madhya Pradesh, Chattisgarh, north-west Orissa, Jharkhand and western part of West Bengal. Detailed results

Table 4: Zone-wise crops/pesticide/fertilizer related problems affected areas (%)

	Crops	Pesticide	Fertilizer
North	38.46	48.71	30.0
East	23.52	48.55	65.23
North-East	56.33	27.78	34.72
Central	61.23	53.07	40.78
South	53.61	22.84	29.89
West	43.44	32.18	42.18

Table 5: Problem affected districts from central India. The districts are from the states of Rajasthan, Madhya Pradesh, Chattisgarh, north-west Orissa, Jharkhand and western part of West Bengal

Problem Areas	No. of districts (Total: 115)
Soil erosion	103
Soil infertility	95
Drought	108
Absence of irrigation	62
Water contamination	98
Crop production	106

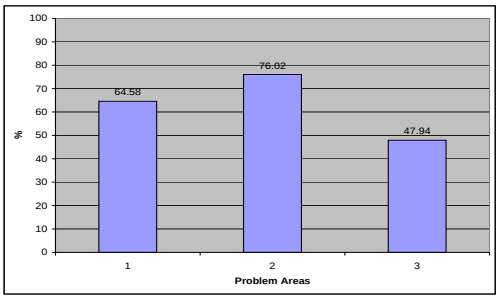

Figure 1: Percentage of districts in India suffering from various problems. 1: Percentage of districts suffering from soil related problems; 2: Percentage of districts suffering from water related problems; 3: Percentage of districts suffering from pesticide/fertilizer/crop health and quality related problems

from this region is enlisted in Table 5. The situation here is far worse than the average national level data. Out **115** districts in this region **103** reported soil erosion, **95** has soil infertility issues, **108** suffer from droughts and only **53** districts has irrigation facilities. Moreover, **86** districts has reported contamination (fluoride or arsenic) in drinking water. These facts vindicate the need for giving special attention to this region. This region is by far the least developed and the most poverty-stricken part of the country. The data revealed from this region sends a strong message to take urgent precautionary steps and to stop further negligence.

A different kind of fact was also reported by the system, how some locations were potentially prone to certain kind of natural calamities. For example, Kutch in Gujrat falls in an earthquake prone zone, proximity to sea makes it vulnerable to cyclones and regular inflows from the sea leads to frequent flooding. Some other such instances include, regular floods in Brahmaputra valley in Assam (district: Nalbari) or devastating cyclones in the coastal districts of Orissa. These calamities cause destruction by damaging crops as well as affects soil by erosion or increasing salinity (in case of cyclone). Moreover, irrigation is also affected by natural calamities and outbreak of diseases, like malaria are also common.

We also learnt about some location specific factors affecting various parameters. In Faridabad, Haryana, paddy-

Table 6: Keywords extracted from 4 locations

Location	Keywords
Amreli	fluoride levels in drinking water fluoride content in thriving chemical industry
Bharuch	the sardar sarovar dam flood dams have drowned cause disastrous flood dams
Kurukshetra	with ground water pollution which affect ground water
Jhabua	of fluoride exposure levels of fluoride effects of fluoride

wheat rotation has caused degradation in soil fertility. Moreover, mine debris blocked feeding channels to drying up lakes, leading to depletion of ground water level. In the regions around Chhatrapur, Madhya Pradesh, numerous ravines created by gully erosion, are heavily under soil loss. Government of Madhya Pradesh has tried to check this soil erosion and expansion of ravines by the means of watershed development and by aerial-seeding for plants like Prosopis, Acacia, and Jatropha in the ravines. In another case, burning is the normal method of rice stubble management in mechanically harvested rice-wheat growing areas of North-West India. This causes air pollution and loss of soil health as well as impacting on animal health.

Similarly, some local success stories also came up within this obtained information. For example, construction of watersheds in Ahemdnagar, helped to fight erosion and drought, both of which are quite prevalent there. Again, Government of Bihar took a positive step when they enacted the Bihar Ground Water act, 2006, which provides mandatory provision of roof top rain water harvesting structures in the building plan in an area of 1000 sq. m. or more. This could potentially resist ground water depletion.

6.1.3 Extraction of Additional Keywords

In Section 4.2, we discussed that additional features were extracted from the preliminary set of documents, based on N-grams. Here, in Table 6 we present some of the relevant features identified from 4 locations.

From this list it can be inferred that Amreli District in Gujrat and Jhabua in Madhya Pradesh might suffer from fluoride contamination. Further searches based on these terms on the location can reveal more facts about this problem. Again, for Kurukshetra district in Haryana, the major N-grams based keywords identified were related to "ground water". Hence, we can interpret that the area is suffering from some issues with ground water. For Bharuch district the main keywords were related to Sardar Sarovar dam and floods. This entails that in that area Sardar Sarovar dam is one of the highlighting features there may be some problems related to the dam and frequent floods. This demands further investigations.

The keyword identification module also reported some features like, "Goverment of Haryana" (for Kurukshetra, Haryana) or "the gulf of Cambay" (for Bharuch, Gujrat). These are contextual features, as Kurukshetra is in the state of Haryana and Bharuch is close to the gulf of Cambay. However, this kind of features are not very relevant in this context and can add noise to the inference mechanism. To avoid such issues, these kinds of keywords need to excluded. Hence, a future improvement would be to devise a method which would identify and omit such keywords.

6.2 Analysis of the Validity of Information

There are chances that false information might penetration through the system. In other words, how authentic is the data reported by the system? For example, if the system reports that a location is suffering from soil related problems, is it true or a false alarm? To ensure the authenticity of the information made available by our system, we followed two approaches. First, analyzed the nature of the websites the system got its information from. Then, we tried to analyze the information qualitatively. We compared the system's findings with scholarly articles on similar topics published in government reports, reputed journals and magazines. It is assumed that these articles carry authentic and accurate information. By finding the similarity between the accounts of these scholarly articles and our system's findings, we verified the authenticity of the facts produced by our system.

6.2.1 Quality of the Information Sources

Here we present some statistics related to the webpages used by the system for retrieving information. These data show the kind of websites used by the system for information retrieval, the authenticity of the webpages, uniqueness of the sources etc. The results are summarized in the following figures.

PageRank Distribution.

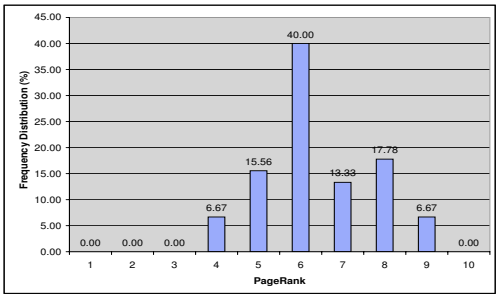

Figure 2: PageRank distribution

Figure 2 shows the PageRank [37] distribution of the webpages used by the system. The average PageRank is **6.4/10**. This is a fairly high value. Usually, pages belonging to reputed institutions which are frequently visited have PageRank value as high as that. Hence, it can be inferred that the system used quite reliable web pages for the information extraction.

Diversity of Websites.

Figure 3 depicts the various web sources used by the system and their share of contribution. From the figure it can be seen that the major share of the information were taken from sites like, governmental portals, organization sites, news articles and educational portals. These are quite

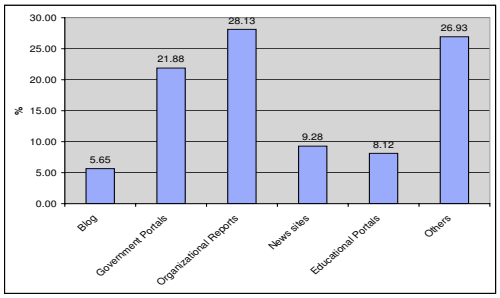

Figure 3: Classification of webpages

reliable sources of information and it can be assumed that the information obtained by the system through these sites are authentic.

Main Contributing Sites.

Apart from the data discussed above, we also tried to identify the top contributing sites. They are as follows,

www.indiaenvironmentportal.org.in, en.wikipedia.org, www.fao.org, agricoop.nic.in, www.tribuneindia.com, planningcommission.nic.in, www.indiawaterportal.org, www.thehindu.com

The list includes some of the specialized environment and agricultural portals from India and also some leading Indian news sites.

Furthermore, we tried to check the number of distinct domains the system used for the whole process. We found that about **83.94%** of the sites were distinct, i.e. only 16.06% pages were repeated during the information retrieval process. This fact proves that most of the information gathered by the system has been from very disparate sources and chances of duplication of facts are very low.

6.2.2 Manual Validation of Information Retrieved

Some of the findings of the system were validated using articles from journals on related topic. For example for many locations, the data presented in Figure 1 and other findings from the system were verified from articles published in journals like Journal of Soil and Water Conservation, Journal of Indian Academy of Science, Journal of Hydrological Sciences, Proceedings of International Union of Geophysics and Geodesy, etc . Similarities were found between these scholarly articles and the system's results.

For example, the system reported high fluoride contamination in the water of Jhabua district in Madhya Pradesh. From the system's output,

> "In tribal areas of Jhabua district, in which some of the villages were affected by Guinea worm (water borne disease) and high fluoride content in drinking water"

The same fact has been validated from a report published by Central Ground Water Board, Ministry of Water Resources, Government of India [14]. Similarly, other outputs from the system were manually validated by comparing them from different scholarly articles or organizational reports. Table 7 shows some of the location-specific problems reported by the system which were verified through this process.

Table 7: Manual verification of the problems reported in some locations

Location	Reported problem	Verified from
Jhabua	soil, water contamination	Govt. report [14]
Vellore	water scarcity	Govt. report [15]
Bidar	crop, water	Govt. report [18]
Malda	soil erosion	Journal article[13]
Haryana	soil quality	Journal article[17]
Nalbari	flood	Govt. report[19]
Bharuch	crop disease	Govt. report[23] Tech. Report[27]
Korba	air pollution, soil	Conf. article[22]
Kurnool	rainfall	Journal article[25]
Kozhikode	ground water	Govt. report[30]
Ratnagiri	drought	Tech. Report[28]
Bareilly	crop production	Tech. Report[29]
Sikkim	soil erosion	Govt. report[20] Journal article[21]
Gaya	fertilizer, soil quality	Tech. Report[26]

From the above discussion, it can be inferred that the information produced by the system is authentic and can be found mentioned in various scholarly articles or government reports. As, the system is only capable of retrieving and processing html files from the Web and the articles used for manual validation were all in pdf format, there are no chances that the system found the information from the same source. This claim further vindicates the authenticity of the system's findings.

However, there can be more rigorous validations of the system's output. This can be done in variety of ways. For example, evaluating the system's performance by domain experts and taking an opinion score, manually gathering related data from the Web and other sources and finding the similarity between the outcomes of the manual mode and the system's output etc. Such further validations will greatly enhance the system's reliability. These validations are kept as future work of this paper.

7. DISCUSSION

In this paper, we have presented a system that mines information from the Web and tries to identify critical trends related to agriculture and climate, for different locations. This system provides a channel for getting localized information on the topics. It is mainly targeted for different decision-making bodies such as government departments, various organizations working on agriculture etc. The main idea is to make these bodies aware of the critical information of a location so that they can act upon the main problem areas and consequently avoid major catastrophes that may arise due to unawareness.

However, the system is incapable of providing the complete picture. There can be various events and incidents which the system might not capture. The goal of the system is to procure the maximum possible information from the web and reproduce it in an organized and precise form.

Hence, the system can still be relevant in producing a preliminary platform to obtain such information. The targeted users can be greatly benefitted with an exploratory data source, which the system can provide them with.

The system has been evaluated for almost all the districts in India and has shown to produce important data related to the topics. The results reveal a very serious picture of the current state. It has reported that a major part of India suffer from water related (76.02% of the districts), crop related(47.94% of the districts) or soil related(64.58% of the districts) issues.

Although this system is using the Web as the source of information, the technology is portable. The same technique can be applied to different other information sources. Already collected data sources such as, various governmental and non-governmental organization reports, news articles can be used for this purpose. These kinds of sources can yield better results too.

The system has been evaluated based on it's performance on retrieving information for 605 districts of India. However, it is yet to be deployed and be actually used by the targeted users. This can help to evaluate the system qualitatively and also in terms of usability. Furthermore, repeated usage by the potential users, domain experts can validate the system's output. This can provide scopes for a better evaluation of the system through experts' opinion and potentially increase the system's reliability. Based on such user feedbacks the system can be improved and come closer in achieving its goals.

8. LIMITATIONS

While our system achieves its design goals, there are scopes for improvement. The present version of the system has some limitations, overcoming which can greatly improve the performance. For example, one way to improve the performance and relevance of the results obtained by our system is to also consider the date. It might happen that the page downloaded is quite old and it contains stale information. As a consequence, irrelevant and unimportant information might be given focus or problems already solved may be put to the front. A significant improvement of the system would be to filter out older pages and accept only current articles and reports.

Again, the query set used in the current version is static, i.e. the same set of queries are fired for all locations. However, there are certain issues typical to certain locations. Like, hilly regions, particularly the eastern Himalayan regions are vulnerable to frequent landslides. Depending upon the nature of the location some extra queries further improve the performance. Moreover, the query set can be further enriched by consulting domain experts. This expert induced improved query set can result into a better information retrieval from the Web. Another way, the query set can be improved is by automatically extracting frequently occurring terms from literatures related to agriculture and its dependent factors. This method can potentially extract more authentic keywords.

The system, currently cannot perform any predictions. Based on the past facts if it could predict trends in the future that would immensely help the target users. This feature could enable early detection of probable disasters and lead to timely action on those, thus preventing major devastations.

Another important drawback of the system is, although it is capable of reporting major problems in a location, not reporting any problem in a location does not mean that the area is devoid of such problems. This can happen if the system could not find relevant sources for that location and that problem. Hence, even for those areas, for which major problems were not reported, a separate study might reveal different facts. This is an example of incompleteness of the system.

9. RELATED WORK

Previously, there have been various attempts to apply Information and Communication Technology (ICT) based solutions for agriculture. The project Digital Green [1] has investigated the option of using videos to disseminate information about agricultural practices and methods among the community. Avaaj Otalo, is an example of using mobile phones and audio interface to provide information to the farmers [3]. The system aAQUA [2], provides a multilingual web interface for farmers. With the help of aAQUA, farmers can get answers to their queries from remote experts, through rural kiosks. The system AgrIDS [5] is another instance of using information technology based agricultural information dissemination system. AgrIDS is a web-based system, which helps to send agriculture experts' opinions to farmers with the help of human coordinators. They have demonstrated how experts' help through AgrIDS has helped some cotton farmers in India, with better yields and reduced expenditure.

Some similar systems developed or used outside India, include, the web-based application developed by Xin and Hu [4], to assist in agricultural emergency against natural disasters. In this work, the authors used Google Earth application to obtain necessary geographical information to asses disasters and provide a remedy, thereafter. Also, the web decision support system built by Tambour *et al* [6], which focuses on helping farmers with strategic and operational assistance to overcome uncertain environment. Some works have focused on applying ICT on environment related issued such as water resource management [16]. Data Observation Network for Earth (DataONE) [9] is an initiative to provide, distribute and share information about environment for better understanding and awareness.

10. CONCLUSIONS

This paper is based on the premise that information dissemination and awareness can help in preventing major devastation in agriculture due to environment related issues. Our system incorporates a novel scheme where localized information on various agriculture and environment related topics are retrieved from the Web and processed to provide concise information from which critical trends on the topic for a location can be interpreted.

The work has been evaluated by obtaining such information for 605 districts in India. The results demonstrated the some effectiveness of the system in achieving the desired goals. Fetched information revealed various facts about the regions, major areas of concern and their vulnerability to different factors.

The current work is still in a developmental stage and there are scopes of improvement within the system. For example, an improved query list or enhanced information

retrieval and summarization strategies can greatly improve the performance. Finally, another major objective of the work will be to deploy this system at various points and be in actual use. Incorporating feedback from actual users of our system will be invaluable in modifying its design to further improve performance, scope and applicability.

11. REFERENCES

[1] Gandhi, R., Veeraraghavan, R., Toyama, K. and Ramprasad V. 2007. Digital Green: Participatory Video for Agricultural Extension, Information and Communication Technology for Development. 2007

[2] Ramamritham, K., Bahuman, A., Duttagupta, S. 2006. Innovative ICT Tools for Information Provision in Agricultural Extension, Information and Communication Technology for Development. 2006

[3] Patel, N., Agarwal, S., Rajput, N., Kumar, A., Nanavati, A., Dave, P., Parikh, T. S. 2008. Experiences Designing a Voice Interface for Rural India, IEEE Workshop on Spoken Language Technology for Development. 2008

[4] Xin, J., Hu, P. 2008. Web Services, Portals and Internet Applciations Managing Agricultural Emergency Resources through Information Mashups. J.l of Information Technology in Agriculture, Vol 3(1), 2008

[5] Reddy, P. K. and Ankaiah, R. 2005. A framework of information technology-based agriculture information dissemination system to improve crop productivity, Current Science, vol. 88, Num.12 2005

[6] Tambour, L., Houles, V., Cohen-Jonathan, L., Auffray, V., Escande P., and Jallas, E. 2009. Design of a Model-Driven Web Decision Support System in Agriculture: Scientific Models to the Final Software, Advances in Modeling Agricultural Systems, 2009

[7] Metzler, D., Dumais, S. T. and Meek, C. 2007. Similarity Measures for Short Segments of Text. European Conference on Information Retrieval, 2007

[8] Sahami, M. and Heilman, T. A web-based kernel function for measuring the similarity of short text snippets. WWW 2006, pp 377-386, 2006.

[9] https://dataone.org/ [Last accessed: July, 2010]

[10] Fattah, M. A. and Ren, F. 2008. Automatic Text Summarization. World Academy of Science, Engineering and Technology 37 2008

[11] Kan, M., Klavans, J. 2002. Using librarian techniques in automatic text summarization for information retrieval. 2nd ACM/IEEE-CS joint conference on Digital libraries. 2002. pp 36-45.

[12] Neto, J. L., Freitas, A. A., and Kaestner, C. 2002. Automatic Text Summarization Using a Machine Learning Approach . Advances in Artificial Intelligence, Lecture Notes in Computer Science, Springer. Vol 2507.

[13] Iqbal, S. 2010. Flood and Erosion Induced Population Displacements: A Socio-economic Case Study in the Gangetic Riverine Tract at Malda District, West Bengal, India. Journal of Human Ecology 30(3).

[14] Ground Water Quality in Shallow Aquifers of India. Central Ground Water Board, Ministry of Water Resources, Government of India. Faridabad, 2010

[15] Balakrishnan , T. 2009. District Groundwater Brochure Vellore District, Tamil Nadu. Central Ground Water Board, Min. of Water Resources, Govt of India. 2009

[16] Pereira, A. G., Rinaudo, J., Jeffrey, P., Blasques, J., Quintana, S. C., Courtois, N., Funtowicz, S., and Petit, V. 2003. Ict Tools To Support Public Participation In Water Resources Governance and Planning: Experiences From The Design And Testing Of A Multi-Media Platform. J. of Environmental Assessment Policy and Management. Vol. 5(3). 2003. pp. 395-420

[17] Datta, K. K., Jong, C. 2002. Adverse effect of waterlogging and soil salinity on crop and land productivity in northwest region of Haryana, India. Agricultural Water Management Volume 57(3), 2002.

[18] Gore, P. G., Prasad, T., Hatwar, H. R. 2010. Mapping of Drought Areas over India. National Climate Centre Research Report No. 12/2010

[19] Nalbari District Disaster Management Database, 2010

[20] Soils of Sikkim. Forest Environment & Wildlife Management Department, Government of Sikkim, 2010.

[21] Pathak, A. Cultivation of Large Cardamom in Sikkim. Journal of Ishani, Vol 2(6), 2008

[22] Singh, G., Pal, A. K., Tiwari, A. 2007. Air Pollution and its Impact on Social Spectrum with Special Reference to Korba Coalfield of Chattisgarh. Intl. Conf. on MSECCMI, New Delhi, 2007

[23] Final District Agriculture Plan: Bharuch, Agriculture and Co-operation Dept, Govt of Gujrat, 2007

[24] Disaster Management in Maharashtra. Envis Newsletter, Environment Department, Govt. of Maharashtra. Vol 1, 2006.

[25] Systematic Evaluation of IWDP and DPAP Programmes: A Study of Kurnool District, AP. Journal of LNRMI, 2010.

[26] Jha, T.N., Viswanathan, K.U. 1999. Probtems and Prospects of Agricultural Development in Bihar. NABARD, 1999.

[27] Srinivasa, N., Mallik, B., Gowda, C.C. 2004. Aceria cajani (Acari : Eriophyidae) Transmitted Pigeonpea Sterility Mosaic Disease. All India Network Project on Agricultural Acarology. Department of Entomology, University of Agricultural Sciences, Bangalore, 2004

[28] Gadekar, M. Tracking the Drought-II Ratnagiri: Water scarcity amidst plenty. InfoChange News & Features, 2003. [http://infochangeindia.org]

[29] Singh, R.P., Islam. Z. 2010. Land Use Planning In Western Uttar Pradesh: Issues And Challenges. Recent Research in Science and Technology 2010, 2(9)

[30] Joji, V.S. 2009. Ground Water Information Booklet Of Kozhiode District, Kerala. Central Ground Water Board, Min. of Water Resources, Govt. of India, 2009

[31] http://india.gov.in/sectors/agriculture/index.php

[32] http://www.nih.ernet.in/water.htm

[33] http://www.teriin.org/coping/index.htm

[34] http://en.wikipedia.org/wiki/List_of_districts_of_India

[35] http://code.google.com/apis/ajaxsearch/

[36] http://en.wikipedia.org/wiki/Tf%E2%80%93idf

[37] http://en.wikipedia.org/wiki/PageRank