Recommendations of Agricultural Websites Based on Modified Feature Vector Algorithm

Arundati Aralimatti [1], Indira R Umarji [2], Dr S M Joshi[3]

Abstract- *In today's world number of users using the web search to search the data are increasing day by day .WebCrawler algorithms used currently does not filter unwanted data and also the ranking results obtained for the end user are less accurate less accurate because the sequence of steps traditionally performed are data collection tokenization, frequency computation competition and feature vector competition In this paper the feature vector has been modified with the position data either in the title or description of a website so that so that more relevance more relevance is obtained during ranking of websites. one more advantage that is that there is a validation performed to search only for agricultural data cultural data sets are taken or taken from Government of India websites. The ranking results of feature vector and modified feature vector for the end user query are compared and proved that modified feature vector gives results with more accuracy.*

Keywords- Web Crawler, Tokenization, Frequency computation, Feature Vector Computation, Correlation Vector and Modified feature vector

I. INTRODUCTION

Text mining is an approach which is responsible to find out short meaningful conclusions mining the data meaning the data it has it has lot of concepts like data processing data processing voice removal stop words removal tokenization frequency competition frequency computation this in this regard in this regard there is lot of work done in the literature.

II. BACKGROUND

In the paper [1] the data is collected by using a DOM explorer and then the URL are ranked based on word count known as frequency. The data collection happens by making use of Soup Parser which hits the website and runs through the entire DOM and then collects the data and stores. After that Tokenization is performed and whenever user searches for the query the websites are ranked based on token count . The advantage of this approach is that the tokenization process is done without removing the stop words hence the ranking is faster. The disadvantage is that it takes into consideration the junk data as well as the advisement during the crawler process and second disadvantage is that the accuracy is less as the data cleaning is not performed.

In the paper [2] the task scheduler based crawler is proposed which makes use of batch processing and then repeatedly downloads the DOM and then performs the preprocessing and then removes unwanted symbols and then generates the tokenization matrix. Once it generates the tokenization matrix it uses a round robin principle in order to sort the web data and then rank it. The advantage is that it makes use of a scheduler which can run repeatedly at regular intervals and collect the data which can be later used for ranking. The disadvantage of this approach is the data cleaning is not performed hence website with huge number of stop words will also come on top which is irrelevant

In the Graph Based Web Crawler [3] it can extract the data from the websites based on AJAX request unlike other web crawlers which are based on NON AJAX implementation. The method makes multiple independent IO requests in order to obtain the data from web sites and then measures the frequency or weight and generates the graph between words with weight as the label of the link. The advantage of the method is the search results are faster because it makes use of AJAX Requests which can run in parallel and It does consider the number of repetition of words while ranking the web sites. The disadvantage is that does not have the capability of doing a cross validation so that the relationship of a word with respect to other web sites can also be obtained.

III. PROPOSED SYSTEM MODULES

The following section discusses the various modules present in the proposed system. The modules are registration, login, data collection, data cleaning, tokenization, frequency computation, feature vector computation, modified feature vector computation and finally ranking of websites

III.A Registration

This Module is responsible for allowing any external customer to perform the registration by proving the details like

First Name, Last Name, User Id, Password and Email. The validation is done so that user id will be unique for each of the users. The user enters the personal information like First Name, Last Name, User Id, Password, Email All the fields are validated for Non-Empty and Regex validations for example Email Field must follow a pattern, User Id and Password cannot be same. If there are any validation failures then error message is shown to user.

If all the basic validations are successful then list of users are obtained who have registered previously. If the given User Id exist in the list of registered users then validation error is send to user.If the user Id is new then the user information is saved and user is allowed to register. The design for the registration module can be described as below

III.B Login

Login Module is responsible for allowing the user to access the user with valid credentials and deny the access for user with invalid credentials. Fig 4 shows the login module functionality which involves the following steps. The user enters the information User Id, Password. All the fields are validated for Non Empty. If there are any validation failures then error message is shown to user. If all the basic validations are successful then list of users are obtained who have registered previously. If the given User Id does not exist in the list of registered users then validation error is send to user. If the user Id exist then password is validated. If the password entered by the user is not same as the actual password then login fails and validation error is shown. If the password entered by the user is same as that of actual password then customer is allowed to login otherwise no.

III.C Data Collection

The real time search results are collected by providing Google API Key and then the searched query. Each of the websites are crawled based on Google results. The data collection process can be described as given in the algorithm1. Figure shows that algorithm takes the input as URL of Google API, search query and xpath of title and description. the algorithm first the web URL is hit after that count of matching Document Object Model nodes are obtained with respect xpath by making use of JSoup parser and then stored in the format of the Matrix

III.D Data Cleaning

The Data Cleaning algorithm is responsible for removal of stop words. Each of website is cleaned by removing the stop words from description. Stop words are the set of words which do not have any specific meaning. The data mining forum has defined set of keywords which do not have any meaning like *a, able, about, across, after, all, almost, also, am, among, an etc.* The data cleaning uses a set of delimiters like comma; semicolon etc along with set of stop words are used for data cleaning

III.E Tokenization

Tokenization is a process of converting the clean data into a set of words known as tokens.

III.D Frequency Computation

This is a process in which the frequency computation is performed. For each of the reviews the frequency is computed. Frequency is number of times a i^{th} token appears in j^{th} website description. The frequency computation can be done as follows

III.E Feature Vector Computation

The feature vector computation is performed by measuring the inverse document frequency which depends on frequency and number of web sites in which token is present.

The inverse document frequency is given by the formula

$$IDFT = \log(\frac{textFrequency}{NoOfWebsites})$$

The feature vector is defined as

$$v_i = tfi * idft_i$$

Where tf is the frequency of the ith word and the IDFT is the inverse document frequency of the ith word.

III.F Modified Feature Vector Algorithm

The modified feature vector algorithm computes the position of the word along with feature vector so that the better accuracy is obtained. The modified feature vector can be computed using the following algorithm

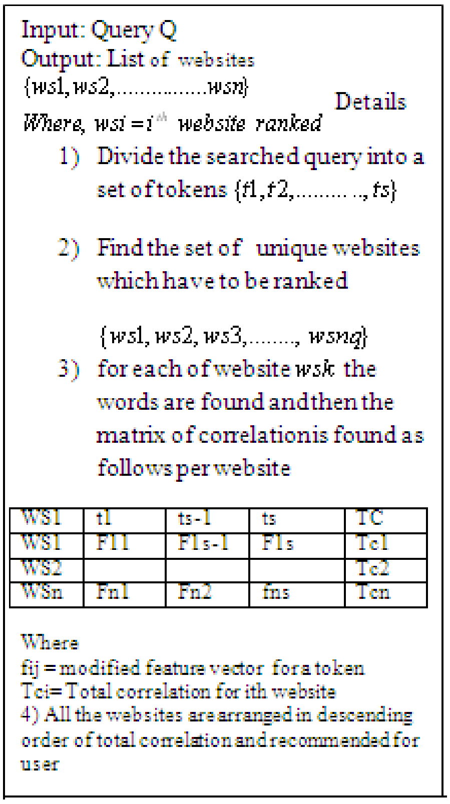
$$mfv = fv + \log(1/\ positionindex)$$
$$Where,$$
$$positionindex = position\ \ of\ \ word\ \ in\ \ title$$
$$if\ \ not\ in\ title\ \ look\ \ for\ \ desc$$
$$Note -$$
$$position - 1$$
$$mfv = fv$$

Here fv represents the feature vector And positional index is the position of the word in the web site title and if not present then in description

I.G Ra

Ranking Using Modified Feature Vector

The following snippet shows the modified
Feature vector algorithm ranking



Input: Query Q
Output: List of websites
$\{ws1, ws2, \ldots\ldots\ldots\ldots\ldots wsn\}$
Details
Where, wsi = i^{th} website ranked

1) Divide the searched query into a set of tokens $\{t1, t2, \ldots\ldots\ldots.., ts\}$
2) Find the set of unique websites which have to be ranked
$\{ws1, ws2, ws3, \ldots\ldots.., wsnq\}$
3) for each of website wsk the words are found and then the matrix of correlation is found as follows per website

WS1	t1	ts-1	ts	TC
WS1	F11	F1s-1	F1s	Tc1
WS2				Tc2
WSn	Fn1	Fn2	fns	Tcn

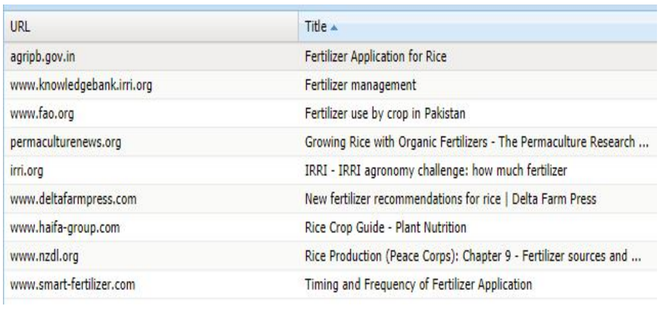
Where
fij = modified feature vector for a token
Tci= Total correlation for ith website
4) All the websites are arranged in descending order of total correlation and recommended for user

IV. EXPERIMENT RESULTS AND ANALYSIS

In the experimental results demonstrate we demonstrate data collection, data cleaning, tokenization, frequency computation, feature vector computation ,modified feature vector computation, search correlation between the user search query and feature vector , search correlation between user search query and modified feature vector and finally comparison between feature vector and modified feature vector in terms of correlation has been shown
USE

User Search Query

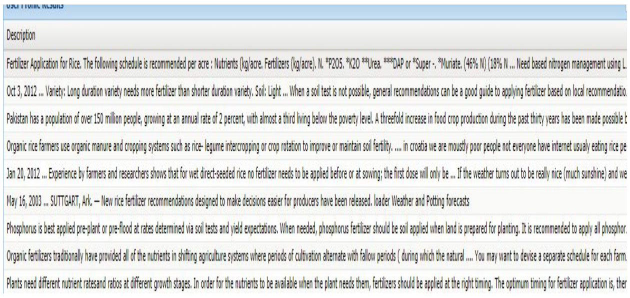
The user searches for a specific query for example "Need good fertilizers for rice production". The data collection is done for the Top 10 websites and the matrix is created in the form of a grid



URL	Title ▲
agripb.gov.in	Fertilizer Application for Rice
www.knowledgebank.irri.org	Fertilizer management
www.fao.org	Fertilizer use by crop in Pakistan
permaculturenews.org	Growing Rice with Organic Fertilizers - The Permaculture Research ...
irri.org	IRRI - IRRI agronomy challenge: how much fertilizer
www.deltafarmpress.com	New fertilizer recommendations for rice | Delta Farm Press
www.haifa-group.com	Rice Crop Guide - Plant Nutrition
www.nzdl.org	Rice Production (Peace Corps): Chapter 9 - Fertilizer sources and ...
www.smart-fertilizer.com	Timing and Frequency of Fertilizer Application

Fig2: Data Collection First 2 Columns of a matrix

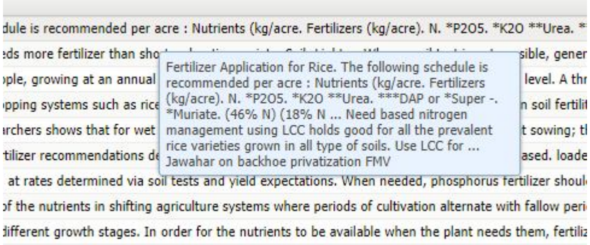
Fig 2 shows the data collecton of the searched query and two columns of the search results are shown one is the URL of the website and othe is the title of the website



Description
Fertilizer Application for Rice. The following schedule is recommended per acre : Nutrients (kg/acre. Fertilizers (kg/acre). N. *P2O5. *K2O **Urea. ***DAP or *Super -. *Muriate. (46% N) (18% N ... Need based nitrogen management using L.
Oct 3, 2012 ... Variety: Long duration variety needs more fertilizer than shorter duration variety. Soil: Light ... When a soil test is not possible, general recommendations can be a good guide to applying fertilizer based on local recommendatio.
Pakistan has a population of over 150 million people, growing at an annual rate of 2 percent, with almost a third living below the poverty level. A threefold increase in food crop production during the past thirty years has been made possible b
Organic rice farmers use organic manure and cropping systems such as rice- legume intercropping or crop rotation to improve or maintain soil fertility. .... in croatia we are moustly poor people not everyone have internet usualy eating rice pe
Jan 20, 2012 ... Experience by farmers and researchers shows that for wet direct-seeded rice no fertilizer needs to be applied before or at sowing; the first dose will only be ... If the weather turns out to be really nice (much sunshine) and we
May 16, 2003 ... SUTTGART, Ark. — New rice fertilizer recommendations designed to make decisions easier for producers have been released. loader Weather and Potting forecasts
Phosphorus is best applied pre-plant or pre-flood at rates determined via soil tests and yield expectations. When needed, phosphorus fertilizer should be soil applied when land is prepared for planting. It is recommended to apply all phosphor.
Organic fertilizers traditionally have provided all of the nutrients in shifting agriculture systems where periods of cultivation alternate with fallow periods (during which the natural .... You may want to devise a separate schedule for each farm.
Plants need different nutrient ratesand ratios at different growth stages. In order for the nutrients to be available when the plant needs them, fertilizers should be applied at the right timing. The optimum timing for fertilizer application is, ther

Fig3: Description of websites

Fig3 shows the description of each of the websutes collected using web crawler .



dule is recommended per acre : Nutrients (kg/acre. Fertilizers (kg/acre). N. *P2O5. *K2O **Urea. *
eds more fertilizer than sho sible, gener
ople, growing at an annual level. A thr
opping systems such as rice n soil fertili
archers shows that for wet t sowing; tl
tilizer recommendations de ased. loade
at rates determined via soil tests and yield expectations. When needed, phosphorus fertilizer shoul
of the nutrients in shifting agriculture systems where periods of cultivation alternate with fallow peri
different growth stages. In order for the nutrients to be available when the plant needs them, fertiliz

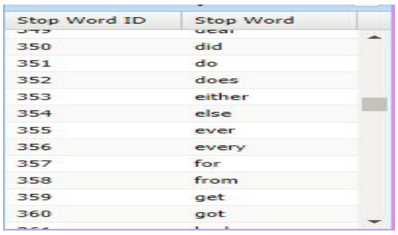
Fertilizer Application for Rice. The following schedule is recommended per acre : Nutrients (kg/acre. Fertilizers (kg/acre). N. *P2O5. *K2O **Urea. ***DAP or *Super -. *Muriate. (46% N) (18% N ... Need based nitrogen management using LCC holds good for all the prevalent rice varieties grown in all type of soils. Use LCC for ... Jawahar on backhoe privatization FMV

Fig4: Viewing Description of Specific Website

Fig 4 shows that when ever the user performs the mouse hover on the description the full description of the website is shown to the end user.

Data Cleaning Algorithm

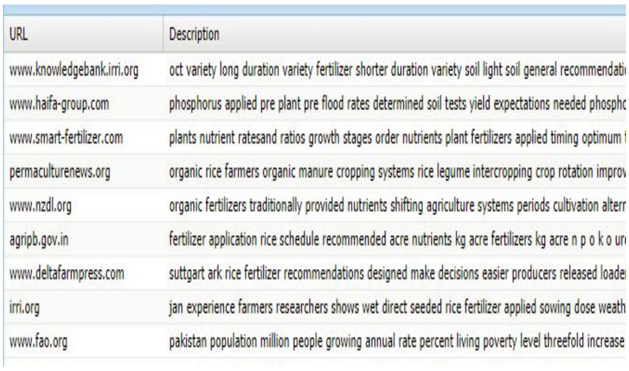
In the application the user will be able to view list of standard stopwords as well as user will be able to add customized stop words.



Stop Word ID	Stop Word
350	did
351	do
352	does
353	either
354	else
355	ever
356	every
357	for
358	from
359	get
360	got

Fig 5: Standard Stopwords

Fig 5 shows the list of stopwords of data mining. There are around 1015 stopwords of data mining amoung them few stopwords have been shown

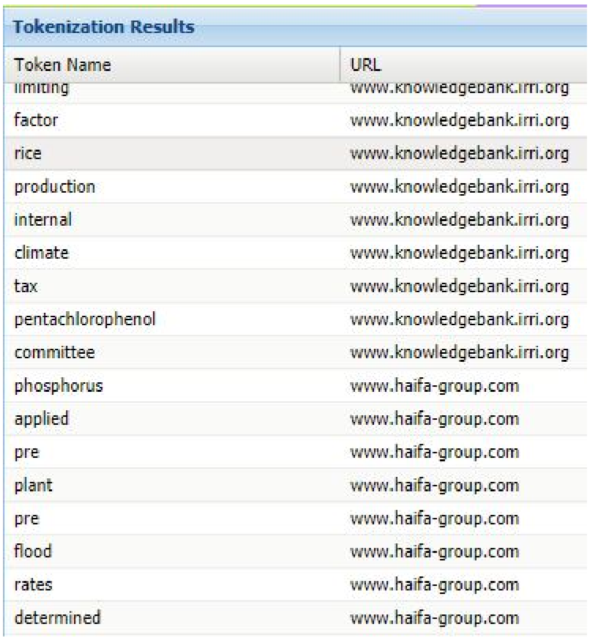


URL	Description
www.knowledgebank.irri.org	oct variety long duration variety fertilizer shorter duration variety soil light soil general recommendati
www.haifa-group.com	phosphorus applied pre plant pre flood rates determined soil tests yield expectations needed phospho
www.smart-fertilizer.com	plants nutrient ratesand ratios growth stages order nutrients plant fertilizers applied timing optimum
permaculturenews.org	organic rice farmers organic manure cropping systems rice legume intercropping crop rotation improv
www.nzdl.org	organic fertilizers traditionally provided nutrients shifting agriculture systems periods cultivation alter
agripb.gov.in	fertilizer application rice schedule recommended acre nutrients kg acre fertilizers kg acre n p o k o ur
www.deltafarmpress.com	suttgart ark rice fertilizer recommendations designed make decisions easier producers released loade
irri.org	jan experience farmers researchers shows wet direct seeded rice fertilizer applied sowing dose weath
www.fao.org	pakistan population million people growing annual rate percent living poverty level threefold increase

Fig 6: Data Cleaning Output

Fig 6 shows the output of the data cleaning that is from each of the description of the websites are the unwanted data and the stock would have been removed and only the clean description has been maintained

Tokenization Algorithm Output



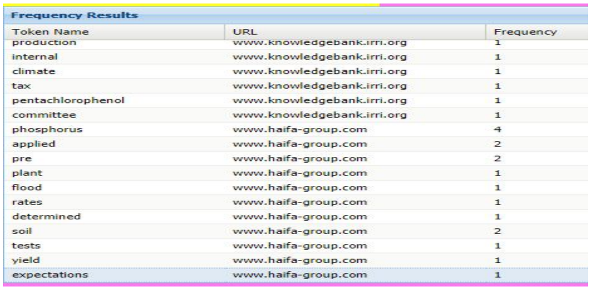
Tokenization Results

Token Name	URL
limiting	www.knowledgebank.irri.org
factor	www.knowledgebank.irri.org
rice	www.knowledgebank.irri.org
production	www.knowledgebank.irri.org
internal	www.knowledgebank.irri.org
climate	www.knowledgebank.irri.org
tax	www.knowledgebank.irri.org
pentachlorophenol	www.knowledgebank.irri.org
committee	www.knowledgebank.irri.org
phosphorus	www.haifa-group.com
applied	www.haifa-group.com
pre	www.haifa-group.com
plant	www.haifa-group.com
pre	www.haifa-group.com
flood	www.haifa-group.com
rates	www.haifa-group.com
determined	www.haifa-group.com

Fig7: Tokenization Output

Fig 7 shows the tokenization process in which each of the clean description is converted into a set of words each word is associated with website

Frequency Computation

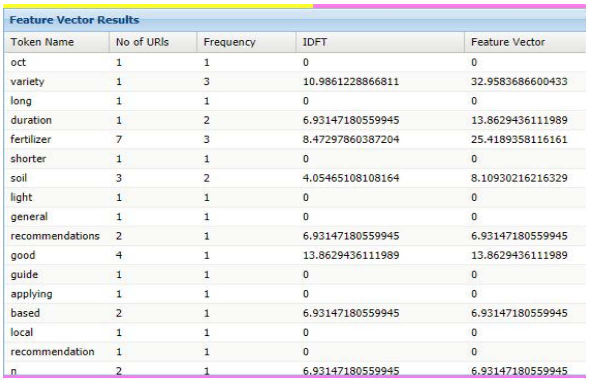


Frequency Results

Token Name	URL	Frequency
production	www.knowledgebank.irri.org	1
internal	www.knowledgebank.irri.org	1
climate	www.knowledgebank.irri.org	1
tax	www.knowledgebank.irri.org	1
pentachlorophenol	www.knowledgebank.irri.org	1
committee	www.knowledgebank.irri.org	1
phosphorus	www.haifa-group.com	4
applied	www.haifa-group.com	2
pre	www.haifa-group.com	2
plant	www.haifa-group.com	1
flood	www.haifa-group.com	1
rates	www.haifa-group.com	1
determined	www.haifa-group.com	1
soil	www.haifa-group.com	2
tests	www.haifa-group.com	1
yield	www.haifa-group.com	1
expectations	www.haifa-group.com	1

Fig8: Frequency Computation

Fig8 it shows the frequency competition is nothing but the number of times the word is repeated in the given website the frequency Matrix consists of token name URL and frequency from the figure the frequency of 2 if you have a frequency of 4 and if you have a frequency of 1 depends upon number of times that words are present in the websites



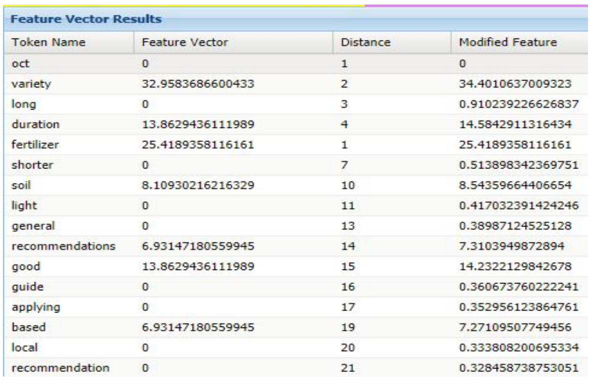
Feature Vector Results

Token Name	No of URls	Frequency	IDFT	Feature Vector
oct	1	1	0	0
variety	1	3	10.9861228866811	32.9583686600433
long	1	1	0	0
duration	1	2	6.93147180559945	13.8629436111989
fertilizer	7	3	8.47297860387204	25.4189358116161
shorter	1	1	0	0
soil	3	2	4.05465108108164	8.10930216216329
light	1	1	0	0
general	1	1	0	0
recommendations	2	1	6.93147180559945	6.93147180559945
good	4	1	13.8629436111989	13.8629436111989
guide	1	1	0	0
applying	1	1	0	0
based	2	1	6.93147180559945	6.93147180559945
local	1	1	0	0
recommendation	1	1	0	0
n	2	1	6.93147180559945	6.93147180559945

Fig9: Feature Vector Computation

Fig 9 shows the feature vecctor for each of the token number of urls in which word is present is shown in No Of Urls, Frequency of the word, IDFT is the inverse document frequency which is 10*log(NoOfUrls/Frequency) and Finally Feature Vector which is multiple of Frequency with IDFT.

Modified Feature Vector



Feature Vector Results

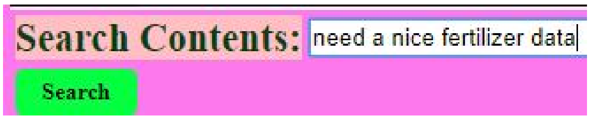
Token Name	Feature Vector	Distance	Modified Feature
oct	0	1	0
variety	32.9583686600433	2	34.4010637009323
long	0	3	0.910239226626837
duration	13.8629436111989	4	14.5842911316434
fertilizer	25.4189358116161	1	25.4189358116161
shorter	0	7	0.513898342369751
soil	8.10930216216329	10	8.54359664406654
light	0	11	0.417032391424246
general	0	13	0.38987124525128
recommendations	6.93147180559945	14	7.3103949872894
good	13.8629436111989	15	14.2322129842678
guide	0	16	0.360673760222241
applying	0	17	0.352956123864761
based	6.93147180559945	19	7.27109507749456
local	0	20	0.333808200695334
recommendation	0	21	0.328458738753051

Fig 10: Modified Feature Vector

Fig 10 shows the modified feature vector Matrix in which the feature vector is computed along with that the position of the word is computed in terms of the distance if the word is present in the title than the position is taken from the title of the word is not present in the title than the position is taken from the description then the modified feature vector is obtained by the combination of Regular Feature vector and the position of the word so that the accuracy is improved

Ranking of Agricultural Data

The ranking for the search query

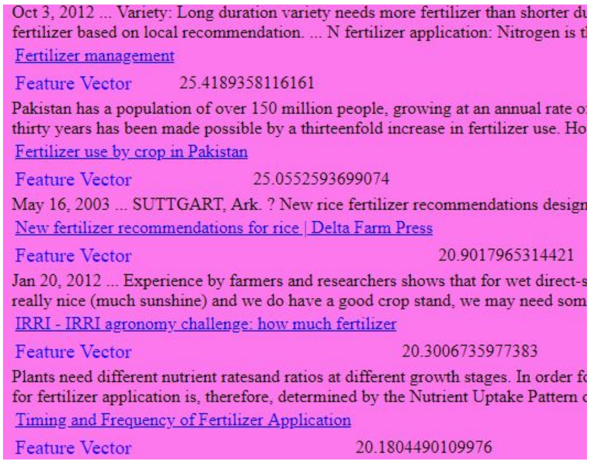


Search Contents: need a nice fertilizer data

Fig11: Search Query

Fig 11 shows the search query performed by the user.

The search results are ordered by the decending order of modified feature vector



Oct 3, 2012 ... Variety: Long duration variety needs more fertilizer than shorter du
fertilizer based on local recommendation. ... N fertilizer application: Nitrogen is t

Fertilizer management

Feature Vector 25.4189358116161

Pakistan has a population of over 150 million people, growing at an annual rate o
thirty years has been made possible by a thirteenfold increase in fertilizer use. Ho

Fertilizer use by crop in Pakistan

Feature Vector 25.0552593699074

May 16, 2003 ... SUTTGART, Ark. ? New rice fertilizer recommendations design

New fertilizer recommendations for rice | Delta Farm Press

Feature Vector 20.9017965314421

Jan 20, 2012 ... Experience by farmers and researchers shows that for wet direct-s
really nice (much sunshine) and we do have a good crop stand, we may need som

IRRI - IRRI agronomy challenge: how much fertilizer

Feature Vector 20.3006735977383

Plants need different nutrient ratesand ratios at different growth stages. In order fo
for fertilizer application is, therefore, determined by the Nutrient Uptake Pattern c

Timing and Frequency of Fertilizer Application

Feature Vector 20.1804490109976

Fig12: Search Results

Fig 12 shows the search results for the search query and each of the websites are rank based on the descending order of the modified feature vector first website has a feature vector of around 25.41 second website with the feature vector of 25.055 III website is having a feature vector of 20.901

Comparison of Algorithms

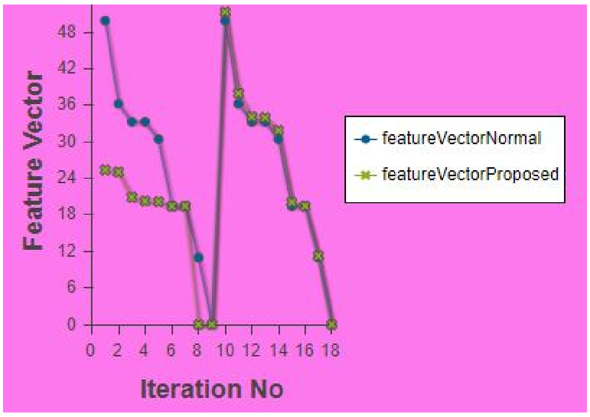


Fig13: Comparision of Algorithms

Fig 13 shows the comparison between the Regular Feature vector competition and modified feature vector competition from the comparison figure one can know that the modified feature vector is always higher as compared to the normal feature vector and hence the accuracy of modified feature vector method is better as compared to the normal method

V. CONCLUSION

In this paper we have collect your we have shown the data collection using web crawler and paper home data cleaning on the collected data so that the stop words are removed and special symbols are removed from the description okay frequency and modified feature vector are computed and prove that the modified feature vector is the best for providing more accurate results for the search query the search results will be obtained if and only if the agricultural words have been found out in the search query otherwise no search results forgiven because there is a validation performed on the agricultural keywords

REFERENCES

[1] Yang Y, Du Y, Sun J, et al. A Topic-Specific Web Crawler with Concept Similarity Context Graph Based on FCA”.Advanced Intelligent Computing Theories and Applications. with Aspects of Artificial Intelligence, International Conference on Intelligent Computing, Icic 2008, Shanghai, China, September 15-18, 2008,Proceedings. 2008:840-847.

[2] Dong H, Hussain F K. Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery[J]. IEEE Transactions on Industrial Informatics, 2014, 10(2):1616-1626.

[3] Álvarez M, Raposo J, Pan A, et al. DeepBot: a focused crawler for accessing hidden web content” Data Engineering Issues in E-Commerce and Services, Third International Workshop, Deecs 2007, in Conjunction with ACM Conference on Electronic Commerce. 2007:18-25.

[4] Dongdong Z, Pengpeng Z, Zhiming C. On the research and design of deep web crawler .Journal of Tsinghua University(Science and Technology)

[5] Cope, N. Craswell and D. Hawking, "Automated discovery of search interfaces on the web," ADC '03 Proceedings of the 14th Australasian database conference, Volume 17, pp. 181-189, 2003.

[6] T. Dong and W. Shang, "Identification of Sensitive Information Based on Improved Naïve Bayesian Classifier", IEEE Computational Sciences and Optimization (IEEE CSO 2011), pp. 816-820, Kunming&Lijiang, Yunnan, China, Apr., 2011.

[7] K. C.-C. Chang, B. He, C. Li, and Z. Zhang, "Structured databases on the web:Observations and implications", Technical report, UIUC.

[8] D. Florescu, A. Y. Levy, and A. O. Mendelzon, "Database techniques for the world-wide web: A survey", SIGMOD Record, 27(3), pp. 59-74,1998.