I.J. Intelligent Systems and Applications, **2016, 5, 55-64**
Published Online May 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2016.05.08

Fuzzy-Based XML Knowledge Retrieval Methods in Edaphology

K. Naresh kumar
Andhra University / Dept. of CS & SE, Visakhapatnam, Zip Code, India
E-mail: chakri.naresh@gmail.com

Dr. Ch. Satyanand Reddy and Prof. N.V.E.S. Murthy
Andhra University / Dept. of CS & SE, Dept. of Mathematics, Visakhapatnam, Zip Code, India
E-mail: {satyanandau@yahoomail.com, drnvesmurthy@rediffmail.com}

***Abstract*—**In this paper, we propose a proficient method for knowledge management in Edaphology to assist the edaphologists and those related with agriculture in a big way. The proposed method mainly consists two sections of which the first one is to build the knowledge base using XML and the latter part deals with information retrieval by searching using fuzzy. Initially, the relational database is converted to the XML database. The paper discusses two algorithms, one is when the soil characteristics are inputted to have the plant list and in the other, plant names are inputted to have the soil characteristics suited for the plant. While retrieving the query result, the crisp numerical values are converted to fuzzy using the triangular fuzzy membership function and matched to those in database. And those which satisfy are added to the result list and subsequently the frequency is found out to rank the result list so as to obtain the final sorted list. Performance metrics used in order to evaluate the method and compare it to baseline paper are number of plants retrieved, ranking efficiency, and computation time and memory usage. Results obtained proved the validity of the method and the method obtained average computation time of 0.102 seconds and average memory usage of 2486 Kb, which all are far better than the previous method results.

***Index Terms*—**Knowledge management, XML, Knowledge Retrieval, Soil, Edaphology, Fuzzy search.

I. INTRODUCTION

Today access to information through Web data plays a significant role. Although facing a quickly growing flood of information on the World Wide Web, we observe a rising need for advanced tools that direct us to the kind of information we are looking for. [1] Retrieval results of main search engines are increasing every day. Mostly general terms searches frequently wind up with over one million results. Generally the keyword-matching mechanisms are used in IR techniques. If one topic has different syntactic representations, the information mismatching problem may occur as in this case [2]. "Data mining" and "knowledge discovery" are the examples that refer to the same topic. If data mining is used to search documents containing "knowledge discovery", it may be missed by keyword-matching mechanism. Information overloading is the problem which occurs in when one phrase having different semantic meanings. A common example is the query, "/apple", which may mean apples, the fruit, or iMac computers. This search results may be mixed by much useless information [3, 4, 5]. If we knew that a user needed information about \apples the fruit" but not \iMac computers", we can deliver the user more useful and meaningful information thus a user's information need could be better captured. In order to better satisfy user information needs the current IR models need to be enhanced [6].

For supporting the future generations of the Web the growth and evolution of the Web makes knowledge retrieval systems is a necessary, in particular, text mining, and knowledge based systems formulate the implementation of such systems practical [7]. Knowledge Management (KM) is an intelligent process by which the raw data is gathered and is transformed into information elements. These information elements are then accumulate and organized into context-relevant structures [8]. KM is intended to approve ongoing business success all the way through a formal, structured initiative to brighten the creation, distribution, or use of knowledge in an organization [9]. In information sciences to illustrate different levels of abstraction in human centered information processing the data-information-knowledge-wisdom hierarchy is used. For the management of each of them, computer systems can be designed. Data Retrieval Systems (DRS), such as database management systems, are well appropriate for the storage and retrieval of structured data [10]. Web search engines such as Information Retrieval Systems (IRS) are very helpful in searching the significant documents or web pages that include the information necessary by a user. The management at the knowledge level is what lacks in those systems. In order to extract the useful knowledge a user must read and analyze the relevant documents [11].

Significantly the way in which information on soils is acquired and managed is changed by increasing the amount of numerical data combined with fast development of new information processing tools. Tree Analysis (TA) is a modeling technique that is being used

Copyright © 2016 MECS

increasingly. TA has numerous advantages that appear to suit well soil-landscape modeling applications [12]. Non-parametric are one of the most interesting features, which means that no assumption is made regarding variable distribution. It avoids variable transformation caused by bi-modal or skewed histograms, which are frequent in soil class signatures. The field of knowledge management is both innovative and highly volatile. Even as we were capable to find many accepted articles on knowledge management and some overviews, all dealt with comparatively small subsets of the range of work we establish referred to as knowledge management. [13]. Overview of the current state and direction of knowledge management were unfortunately unable to find therefore, much of the effort was placed on understanding the status and direction of knowledge management development under the statement that knowledge-based systems will eventually need to be integrated into a larger knowledge management system. [14].

A. Edaphology

Edaphology is about the influence of soils on living things, mainly plants. It also deal with the study of how soil influences man's use of land for plant growth as well as man's overall use of the land. Agricultural soil science is the general subfields within edaphology (known by the term agrology in some regions) and environmental soil science. (Pedology deals with pedogenesis, soil morphology, and soil classification). Soil science is the technical study of soil as a natural resource on the surface of the earth together with soil formation, classification and mapping; physical, chemical, biological, and fertility properties of soils; and these properties in relation to the use and management of soils. Sometimes terms such as pedology refer to branches of soil science (formation, chemistry, morphology and classification of soil) and edaphology (influence of soil on organisms, especially plants), are used as if synonymous with soil science. The diversity of names associated with this discipline is related to the various associations concerned. In reality, engineers, agronomists, chemists, geologists, geographers, ecologists, biologists, microbiologists, sylviculturists, sanitarians, archaeologists, and specialists in regional planning, all contribute to further knowledge of soils and the development of the soil sciences. How to preserve soil and arable land in a world with a growing population, possible future water crisis, increasing per capita food consumption, and land degradation are the concerned factors raised by soil scientists.

B. Need for Knowledge Retrieval in Soil Database

As the plants demand varying quantities of diverse nutrients at different stages of growth, the preservation of fertility at the appropriate level in the soil and the selection of suitable vegetation type for the soil are especially vital for cropping. Therefore, in taking care of plants the knowledge of deficiency/excess of the nutrients in the soil is very significant. The large quantity of data and the multiple areas of expertise that are indispensable for soil exploration generate a massive volume of knowledge. These factors highlights the need for designing an efficient system to adjust, standardize, manage, retrieve and process soil information in order to attain improved productivity in agriculture.

The characteristics and the information about the soils collected by edaphologists are utilized to have input relational database. The input database has two tables of which one is plant description table which contains attributes that describe the plants and the other table is of the soil characteristics table, which contains the soil attributes. The tables are initially converted to XML database using plant identification number attribute in both the tables as the foreign key. The proposed method discusses two algorithms. One is to find the plants suited to the input soil characteristics and the other is to find the soil characteristics needed for the input plant name. Both the algorithm makes use of fuzzy search and ranking to have the results. In fuzzy search initially the numerical crisp values are converted to fuzzy values using the fuzzy triangular membership function and then compared with the database to have the results. After converting to fuzzy, ranking process is done by finding the frequency in order to have the final result list in response to the query.

The main contributions of our proposed technique are:

- Conversion of relational database to XML so that information retrieval happens in a faster and easier way.
- Use of fuzzy search which adds to having a greater flexibility and having better query results.
- We discuss two algorithms of which in the first one, soil characteristics are inputted to have the plants satisfying the query and in the second one, plant name is inputted to have the soil characteristics best matched to the plant.
- We compute the performance metrics having the attributes: number of plants retrieved, ranking efficiency, and computation time and memory usage in order to evaluate the method.
- We make a detailed study by comparing our proposed method to previous method [16].

The rest of the paper is organized as follows: a brief review of researches related to the proposed technique is presented in section 2. Section 3 describes proposed method for fuzzy-based knowledge retrieval in Edaphology. The detailed experimental results and discussions are given in section 4. The conclusions are summed up in section 5.

II. Review of Related Work

There have been many works in the Edaphology domain especially in the information/knowledge storing and retrieval process. In this section we make discuss some of the works related to it. Rizwana Irfan et al. [15] proposed a method that provided qualitative approach for enhancing the existing conceptual model for knowledge processing to do transformation. Modified knowledge

Copyright © 2016 MECS

management process transformed the heterogeneous data in to a uniform format and was further integrated in expert warehouse concept. Meenakshi et al. [16] presented an efficient tree-based system for knowledge management in edaphology. The system assisted edaphologists and an agricultural expert in obtaining the right crops/plants for the given soil characteristics. The characteristics and the information about the soils collected by edaphologists were utilized in the design of the presented system. The proposed system was composed of two phases namely: Knowledge Representation and Knowledge Retrieval. Firstly, a knowledge base was constructed by modeling the domain knowledge collected by edaphologists using the tree data structure. A novel algorithm was devised for effective knowledge retrieval from the modeled knowledge base. Subsequently, for the given soil characteristics, that provided with a set of plants/crops to be cultivated in that soil for better productivity from the constructed knowledge base.

Lynette L. Ralph and Timothy J. Ellis [17] investigated the use of the knowledge base of Question Point as a knowledge management tool capable of improving reference services in academic libraries. The research addressed the problem that reference librarians continually provide ineffective service to patrons. Because of the expansive exposure to resources, it is often difficult for any individual librarian to accurately recall the best resource or answer for any specific question. While individual librarians may not recall specific information, when they collaborate with their colleagues and share their collective knowledge there is usually an improvement in the quality of service they provide. It would benefit librarians therefore, if they used a knowledge management tool that could capture and store their communal knowledge for future use. This study explored the librarians' perceptions of the benefits and problems of using the Knowledge Base of Question Point and its impact of on reducing response time and duplication. The study revealed that the reference librarians did not generally use the Knowledge Base, and that there was duplication of effort and no reduction in response time.

Qiang Yang et al. [18] presented an algorithm that suggested actions to change customers from an undesired status (such as attritors) to a desired one (such as loyal) while maximizing an objective function of expected net profit. These algorithms could discover cost effective actions to transform customers from undesirable classes to desirable ones. The approach they took integrated data mining and decision making tightly by formulating the decision making problems directly on top of the data mining results in a post processing step. To improve the effectiveness of the approach, they also presented an ensemble of decision trees which was shown to be more robust when the training data changes. Rik Farenhorst and Remco C. de Boer [19] described four main views on architectural knowledge based on the results of a systematic literature review. Based on software architecture and knowledge management theory they defined four main categories of architectural knowledge, and discussed four distinct philosophies on managing architectural knowledge.

III. Proposed Methodology

In this section, we discuss about the proposed efficient technique for knowledge management in Edaphology making use of the XML and fuzzy search logic. These two features constitute to building a proficient system which gives edaphologists a solid edge when it comes to storing and retrieving informational knowledge in the concerned domain which ultimately results in having an increased productivity form the agricultural lands. This is the fact that right crop for the right soil can serve the best results. The soil is characterized by many parameters including the mineral and chemical compound content in the soil. For having the optimum outcome from the agriculture lands, the soil characteristics and the depth play a major role. In order to model and develop the relational database we make use of soil characteristics collected by edaphologists. The proposed technique mainly consist mainly two sections of which the first one is to build the knowledge base using XML and the latter part deals with information retrieval by searching using fuzzy. Fig. 1 shows the block diagram of the proposed method. The proposed technique consists of two sections:

- Creation of XML database
- Information retrieval by searching using fuzzy

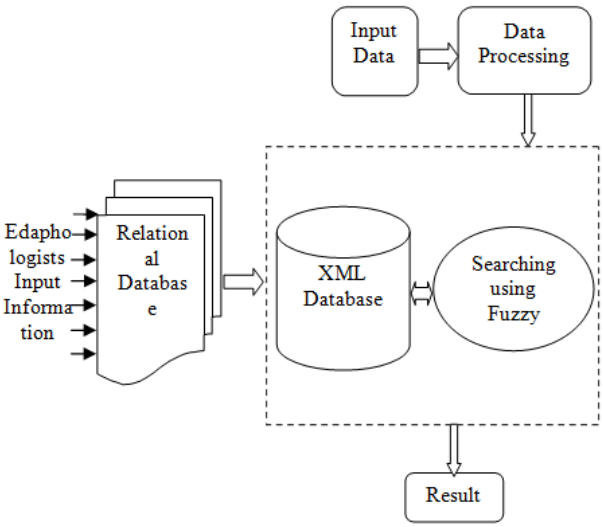

Fig.1. Block diagram of the proposed technique

A. Creation of XML Database

The primary step of the knowledge management system is to develop and model the domain knowledge or information collected from edaphologists. The optimal modeling of the information is of paramount importance as the system performance based on the effective management and retrieval of information directly depends on it. In general, proficient data structures like

K-graphs [15, 18] are chosen for knowledge modeling. In the paper [18], we make use of the tree data structure for knowledge representation which is almost like the K-graph and can be defined as an acyclic connected graph with one parent node and each node having a set of zero or more children nodes. In our proposed technique, we are improving on it and use XML which ends up in attaining better results. For the purpose, we convert relational database into XML. Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and also machine-readable. XML is widely used for the representation of arbitrary data structures. The main advantage of using the XML is the flexibility, accessibility and portability it offers. The most beneficial matter in using XML is the improved speed and performance when compared to tree structure. Also the use of XML reduces the time incurred while information retrieval.

Id	name	Geology	Taxonomy
0001	Prosophis juliflora, Cyprus sp., Hariyali,	Clay	Fine, montmorillonitic, isohyperthermic, noncalcareous, Chromic Haplusterts
0003	Palmyrah	Granite	Fine, mixed, isohyperthermic, noncalcareous, Typic Rhodustalfs
0017	Eucalyptus, Palmyrah, Neem, Tamarind	Laterite	Fine, mixed, isohyperthermic, noncalcareous, Typic Haplustepts
0019	Palmyrah, Neem	Granite	Clayey, mixed, isohyperthermic, noncalcareous, Lithic Haplustepts
0021	Palmyrah, Prosophis juliflora	Granite	Fine, mixed, isohyperthermic, noncalcareous, Typic Haplustalfs
0023	Neem, Palmyrah, Prosophis juliflora, Ta	Sand	Loamy-over-sandy, mixed, isohyperthermic, noncalcareous, Typic Ustifluvents
0024	Neem, Prosophis, Tamarind	Granite	Sandy, mixed, isohyperthermic, calcareous, Typic Ustorthents
0032	Palmyrah	Sand	Sandy, mixed, isohyperthermic, noncalcareous, Aquic Ustipsamments
0035	Neem, Palmyrah	Granite	Fine, mixed, isohyperthermic, calcareous, Calcic Haplustepts
0037	Neem, Prosophis juliflora	Western Ghats	Fine, mixed, isohyperthermic, noncalcareous, Typic Haplustepts
0039	Neem, Palmyrah, Tamarind	Granite	Fine, mixed, isohyperthermic, noncalcareous, Typic Haplustepts
0041	Palmyrah, Neem, Accacia	Granite	Clayey-skeletal, mixed, isohyperthermic, noncalcareous, Typic Haplustepts
0042	Ipomea, Thespesia populanea, Vagai	Clay	Fine, mixed, isohyperthermic, noncalcareous, Typic Rhodustalfs
0045	Palmyrah, Neem, Prosophis juliflora	Granite	Loamy, mixed, isohyperthermic, noncalcareous, Lithic Ustorthents
0047	Eucalyptus, Vagai	Laterite	Fine, mixed, isohyperthermic, noncalcareous, Fluventic Haplustepts
0050	Prosophis juliflora, Neem, Vetiver	Granite	Clayey-skeletal, mixed, isohyperthermic, noncalcareous, Lithic Ustorthents
0051	Prosophis juliflora, Palmyrah	Granite	Fine-loamy, mixed, isohyperthermic, noncalcareous, Fluventic Haplustepts
0055	Prosophis juliflora, Palmyrah, Tamarind	Granite	Fine-loamy, mixed, isohyperthermic, noncalcareous, Typic Haplustepts

Fig.2. Example of the Plant table

Initially, the knowledge is stored in the relational database with the inputs from edaphologists. Here, it comprises of two tables of which one contains first one contains the plant details and the other having the soil description. The plant details table consists of plant names, geology and taxonomy corresponding to the plant ID. Fig. 2 shows an example of plant table P having attributes plant identification number I, name Na, geology Ge and the taxonomy Ta. We can see that a plant can have multiple plant IDs and the geology and taxonomy vary accordingly. The description table contains the plant ID, depth and the description of the soil. It also has the values of various parameters like clay, silt, sand, Ph, electrical conductivity, Calcium, Magnesium, Sodium, Potassium and Phosphorus Pent-oxide, Potassium Oxide. Here we can see that the soil characteristics for the plant ID changes with the depth and because of that, each plant ID has more than one soil characteristics attached to it. Figure 3 gives an example of soil characteristics table S having attributes of plant identification number I, depth D, description G, clay Cl, silt Sl, sand Sa, hydrogen ion concentration H, electrical conductivity E, calcium Ca, magnesium M, sodium Ns, potassium Pt, phosphorous pent oxide Ph, and potassium oxide Po.

The first process in the paper is to store the data from two tables in the XML format. For the same we select plant ID I as the foreign key to join both the tables. Here the data is converted to the XML format and then the data is retrieved accordingly to the search query. During the conversion of the relational database to the XML structure, a tree like structure is built with the use of tags. Here, first the plant ID is taken and it acts like the parent tag. In each pant ID complete details are added in pattern having the details from both the tables corresponding to the plant ID. First the attributes form the plant table is added to the XML. Here first the name, then geology and taxonomy are given tags and are added to the structure. After that soil descriptions are added to structure corresponding to the plant ID. A single plant may have more than one plant id associated with it and also many soil characteristics attached to it as the soil characteristics vary with the depth. In each soil characteristics the depth, description, clay, silt, sand, pH and the chemical element contents are given. A separate description tag is created for each soil characteristics column in the characteristics table and a plant ID will have more than one of these description tags. After creating the complete structure for a plant ID, the structure for the next plant ID is made. Likewise for all the plant IDs in the table, the procedure is followed to get the final XML structure. In the Xml every details related to a single id is stored first and after completing it, it will move to the other plant ids. N is the total number of plant identification numbers in the tables.

For each I_j, where $0 < j \leq N$,
Find *Na*, *Ge* and *Ta* from *P* where $I = I_j$.
Store in XML.
Find *D*, *G*, *Cl*, *Sl*, *Sa*, *H*, *E*, *Ca*, *M*, *Ns*, *Pt*, *Ph* and *Po*

Copyright © 2016 MECS

from S where $I = I_j$.
Store in XML.

It can be noted that there will be only one row in the plant table corresponding to the plant id whereas there will many rows corresponding to the plant id in the soil characteristics table as with the depth the soil characteristics required by the plant changes. Fig. 3 shows the example of the XML structure for Edaphology.

```
<?xml version="1.0" encoding="UTF-8"?>
<Plant>
<Id>0001</Id>
<Name>Prosophis juliflora, Cyprus sp, Hariyali, Indigo plant</Name>
<Geology>Clay</Geology>
<Taxonomy> Fine, Montmorillonitic, Isohyperthemic, Noncalcareous, Chromic Hasplusterts </Taxonomy>
<Description>
        <Depth>0-13</Depth>
        <description> Dark brown (10 Yr 4/3); Sandy clay; Moderate, medius, sub-angular blocky: hard, Slightly firm, sticky and plastic; craks of 2-3 cm width; common, fine and very fine roots; few, very fine roots; few, very fine and fine pores; moderate permeability; clear smooth boundary
        </description>
                <Clay>38.46</Clay>
                <Slit>15.84</Slit>
                <Sand>45.70</Sand>
                <PH>8.30</PH>
                <EC>1.30</EC>
                <Ca>11.80</Ca>
                <Mg>4.10</Mg>
                <Na>3.59</Na>
</Description>
<Description>
        <Depth>65-184</Depth>
        <description> Very dark grayish brown(10 YR 3/2); clay loam; Strong, medium, angular blocky prismatic; very firm, sticky and plastic; many distinct slickesided; few, very fine and fine pores; slow permeability; few sandy streaks; clear smooth boundary; many stratified layers
        </description>
                <Clay>40.84</Clay>
                <Slit>21.90</Slit>
                <Sand>37.26</Sand>
                <PH>8.390</PH>
                <EC>1.40</EC>
                <Ca>14.20</Ca>
                <Mg>3.90</Mg>
                <Na>4.09</Na>
</Description>
</plant>
```

Fig.3. Structure of the XML Code

B. Information Retrieval using Fuzzy Search

From the knowledge base which is stored in XML format, we need to extract information in the best possible manner in-order to aid the edaphologists in the best way. For this extraction of knowledge we make use of the fuzzy search by which we can retrieve the information in a more flexible manner compared to the conventional methods and also results in having less time incurred. The advantage with the fuzzy search is based on minimization of the marginal values and the flexibility which results in faster and better execution. The paper discusses of two search scenarios, one is where the soil characteristics for the input plant name and the other part is having the soil characteristics for the plant input. In both the cases, we make of the fuzzy search. Fuzzy search deals with having fuzzy description instead of crisp values and in here mostly description crisp values are converted into fuzzy sets based on certain parameters. The fuzzy sets count to three which proves ideal in easy searching and also obtaining in results with a faster timing which is of vital importance. The fuzzy sets are designed considering the highest and lowest values in the discrete crisp values and are based on the triangular fuzzy membership function. The retrieval of information is done accordingly from the XML based on the input query, be it the plant name or the soil description.

Fuzzy search incorporates flexibility to the search which is important considering the Edaphology domain. It is because a plant survives a range of values for the attributes rather than a precise single value. For example, a particular plant A is said to grow in nine meters depth with particular soil characteristics. When the query is given for the plant having the same soil characteristics but with a depth of eight meters, it will miss out on this plant A. But in reality, soil characteristics for a depth eight meters and soil characteristics for the same plant at nine meters will be similar and can be treated as one. So incorporating fuzzy adds more flexibility to the search and matches with real life scenario.

The information retrieval has mainly three steps:

- Converting attributes to the fuzzy
- Searching in the corresponding node and retrieval of plants
- Ranking based on frequency

The three steps are explained in a detailed manner in the later part. The results are taken from the ranked results to obtain the plant or the soil characteristics required. As discussed in the earlier part the searching happens in two cases.

3.2.1.Getting the Plant based on the Soil Description

Getting the ideal plant for the available soil description is of vital importance as the plant growth and plant output directly depend on the soil characteristics. Having the right soil characteristics for the right plant will provide the best results and this can be made possible having the right answers to the search queries seeking the best plant that can be planted on the soil having the said attributes. One or more soil characteristics can be given as inputs to have the results having the list of plants suitable for the said conditions. As mentioned above, information retrieval to have the plant list based on the input soil characteristics is a three step procedure which includes a) converting attributes to fuzzy, b) searching the plants and getting the result list and c) ranking based on frequency.

3.2.1.1 Converting Attributes to the Fuzzy

First of all the crisp values of the input soil characteristic attribute are converted to the fuzzy set

based on the value. Normally, the fuzzy sets are three in number where the first one-third will come in the first fuzzy set and the second one third is in the second fuzzy set and the last one-third is in the last fuzzy set. Here the first fuzzy set is termed low, second fuzzy set is termed low and last fuzzy set is termed high.

Table 1. Showing the conversion to fuzzy

Crisp Values	Fuzzy Value
Miminum-33.33% of maximum	Low
33.33% -66.66% of maximum	Medium
66.66% - maximum	High

The method is improved having overlapping functions by having fuzzy triangular member in-order to improve flexibility. The depth, clay, silt, sand, Ph, electrical conductivity, Calcium, Magnesium, Sodium, Potassium and Phosphorus Pent-oxide, Potassium Oxide values (*D, Cl, Sl, Sa, H, E, Ca, M, Ns, Pt, Ph* and *Po*) have the crisp values are converted to the fuzzy set. The other text inputs like name, geology, taxonomy and the description forms the text inputs (*G, Na, Ge* and *Ta*) which are not changed and is compared in the text format during the search operation.

For each I_j, where $0 < j \leq N$,
For every attribute *D, Cl, Sl, Sa, H, E, Ca, M, Ns, Pt, Ph* and *Po* where $I = I_j$.
Convert to fuzzy $F_D, F_{Cl}, F_{Sl}, F_{Sa}, F_H, F_E, F_{Ca}, F_M, F_{Ns}, F_{Pt}, F_{Ph}$ and F_{Po}
For other attributes *G, Na, Ge,* and *Ta* No change

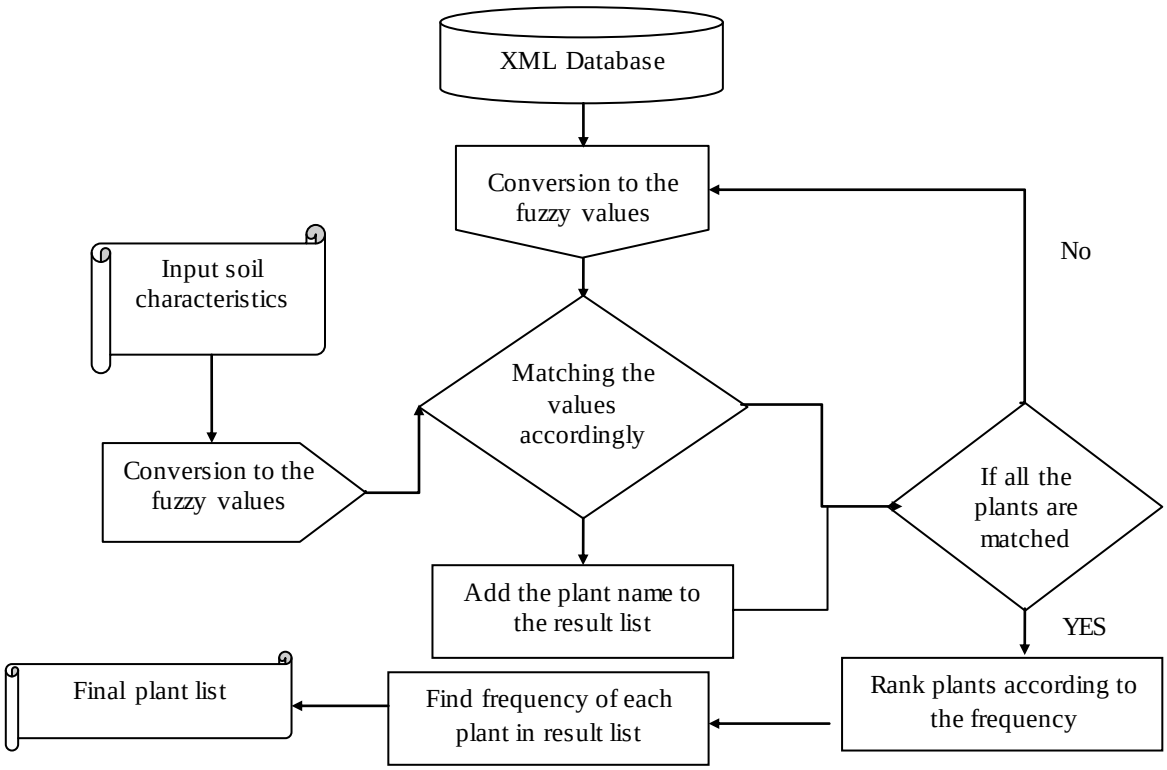

Fig.4. Block Diagram of Algorithm 1 (Getting the plant list for the given soil conditions)

The conversion to the fuzzy is based on the fuzzy triangular membership values discussed in the previous section. Here the conversion of the values is into three fuzzy sets HIGH, MEDIUM and LOW.

For each I_j, where $0 < j \leq N$,
For every element E_j
Where E = {*D, Cl, Sl, Sa, H, E, Ca, M, Ns, Pt, Ph* and *Po*},
Convert to LOW, MEDIUM or HIGH fuzzy set

Fuzzy Triangular Membership Function

The attributes having numerical values in the XML database is transformed into the fuzzy using the triangular membership function. Membership functions can either be chosen by the user arbitrarily or be designed using machine learning methods like artificial neural networks, genetic algorithms and others. There are different shapes of membership functions; triangular, trapezoidal, piecewise-linear, Gaussian, bell-shaped, etc. Here, we have chosen the Triangular membership function in which a, b and c represent the x coordinates of the three vertices of $f(x)$ in a fuzzy set A (a: lower boundary and c: upper boundary where membership degree is zero, b: the centre where membership degree is 1). One of the key issues in all fuzzy sets is how to determine fuzzy membership functions,

- The membership function fully defines the fuzzy set.
- A membership function provides a measure of the degree of similarity of an element to a fuzzy set.

Copyright © 2016 MECS

- Membership functions can take any form, but there are some common examples that appear in real applications.

The formula used to compute the membership values is depicted as below,

$$f(x)=\begin{cases}0 & if\ x\le a\\ \dfrac{x-a}{b-a} & if\ a\le x\le b\\ \dfrac{c-x}{c-b} & if\ b\le x\le c\\ 0 & if\ x\ge c\end{cases}$$

Fig. 5 shows a triangular membership function for a single fuzzy set. Here we can see that at a and c the value is zero and it reaches steadily to a maximum of value one at the centre point b between a and c.

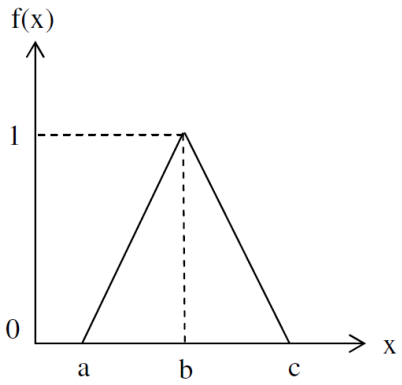

Fig.5. Triangular membership function

Fig. 6 shows the plot considering all the three membership functions having overlapping values. Here the curves for low, medium and high are shown for the attribute, say depth.

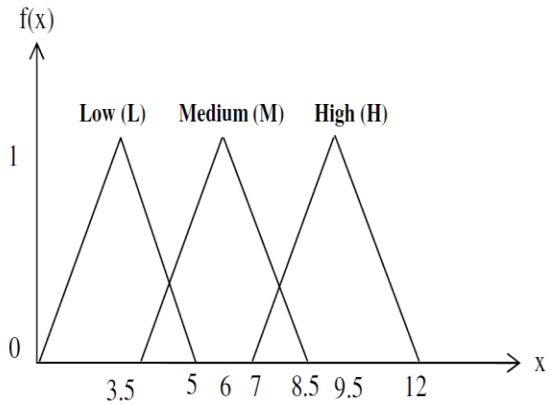

Fig.6. Triangular membership function with defined parameters and their values

By using the fuzzy membership formula, we have transformed the numerical attributes into the fuzzy.

3.2.1.2 Searching in the Corresponding Node and Retrieval of Plant Lists

After converting to the fuzzy, the searching process happens where the information is retrieved according to the input query and the searching happens in the node of the XML corresponding to the input query attributes. For example, when a depth of eight meters is given as the input, first it is converted fuzzy and then all the plants that is having the same fuzzy is found out by searching in the depth node. For the searching, we compare using the string compare function comparing the input attribute fuzzy word to others in the database under the same root node. If a range is given instead a single value as the word, it too is converted to the fuzzy. The plants that satisfy the input condition are found out and listed. The searching happens inside the XML database with the use of fuzzy search where initially the values are converted to the fuzzy values. For a description of depth giving arbitrary value D_i, we have to convert it to fuzzy value and do the search in the database under the fuzzy values for the node depth.

For an input D_i, convert to Fuzzy F_{Di},
For each I_j, where $0 < j \le N$,
Search in root node depth if $F_{Di} = F_D$, then select the corresponding *Na*
Add *Na* to the result list *R*.

For those having the same fuzzy depth values in the database, the corresponding plant names are added to the result list. The same process happens for all cases {*D, Cl, Sl, Sa, H, E, Ca, M, Ns, Pt, Ph* and *Po}* where some soil characteristic is given as input X_i where the values are converted to the fuzzy values F_{Xi} and compared with the fuzzy root nodes in the XML database {F_D, F_{Cl}, F_{Sl}, F_{Sa}, F_H, F_E, F_{Ca}, F_M, F_{Ns}, F_{Pt}, F_{Ph} and F_{Po}}. Those which satisfy the conditions are noted and are added to the result list *R*.

R = {Na_1, Na_2, … , Na_k}, where *k* is the total number of results in the list which contains the names of the plant *Na* which satisfies the condition. When there are multiple input conditions then, names of the plants which satisfy all the input conditions are only added to the list.

For an input X_i and Y_i convert to Fuzzy F_{Xi} and F_{Yi}
For each I_j, where $0 < j \le N$,
Search in root node depth if $F_{Xi} = F_X$ and $F_{Yi} = F_Y$ then select the corresponding *Na*
Add *Na* to the result list *R*.

X_i and Y_i are the input conditions, F_X and F_Y are the fuzzy values from the database corresponding to the X and Y nodes.

3.2.1.3 Ranking based on the Frequency and Fuzzy Value

After the search, we get the plant list having the plant names which satisfy the conditions. In the list plant names will appear in many places and will look random. In order to have a better understanding and also to know the best plant that is suitable for the given conditions we

Copyright © 2016 MECS

have to arrange it in the best possible way. For the purpose we find out the number of times the plant appears in the list or rather the frequency of the plant in the list. The frequency of the plant directly gives the direct knowledge how good that plant can grow in the said conditions. Better the frequency, the better the chance of the plant growing well under the conditions. Hence, we rank the plants based on the frequency of the plant and its fuzzy value to get the final list.

Form the result list *R;* we have to find the most appropriate answers for the input conditions. For the purpose, we find the frequency of each plant in the list. *K* is the total number of results in the list.

$$\text{For each } Na_i \text{ in } R,\ 0 < j \le k$$
$$\text{If } Na_j = Na_i, \text{ for } 0 < j \le k$$
$$C_i = C_i + 1,$$
$$\text{Then, } S_i = \frac{1}{C_i}\sum_{j=1}^{C_i} F(C_j)$$

Here C_i is the frequency of i^{th} name in the result list R and S_i is the final fuzzy score of the i^{th} plant name. After the finding out the fuzzy score of each plant, the list is sorted accordingly so that the plant with maximum fuzzy score comes first. Let m be the number of unique plant names in the list.

$$\text{For } Na_i \text{ in } R,\ 1 < i \le m$$
Sort in descending order with respect to S_i,

For given input soil conditions, the plants in the top of the list will yield good results and this knowledge will prove beneficial for the edaphologists. Hence the plant fit for the given conditions are obtained.

IV. Results and Discussions

This section presents the results and discussions of our proposed method for knowledge retrieval in Edaphology. Here we evaluate both the algorithms used in the search operations where in the first, plant list fit for the input soil conditions are found out and in the other one, the soil characteristics list for the input plant name is found out from the XML database. We also compare this paper to our baseline paper with the help of the performance metrics obtained in response to various user input queries. The obtained data are analyzed with the help of bar charts which proves the validity of our proposed technique.

4.1 Expermential Setup and Dataset Description

The proposed technique is implemented in JAVA on a system having 4 GB RAM and 2.10 GHz Intel i-3 processor. Initially, the domain knowledge collected from edaphologists is modelled into a knowledge base, which acts as the input data set. The input database consists of two tables, of which one is the plant list table and the other, soil characteristic table. The two tables are linked by the foreign key plant identification number. There are 148 plant ids in the database in each plant table there are four attributes and in soil characteristics table there are 15 attributes. The plant table attributes are plant identification number, name, geology and the taxonomy. The soil characteristics table attributes are plant identification number, depth, description, clay, silt, sand, hydrogen ion concentration , electrical conductivity, calcium , magnesium, sodium, potassium, phosphorous pent oxide and potassium oxide.The input database is stored in a file and later converted to XML database, from where the results are searched in reference to the user input query.

4.2 Performance Metrics

In order to find the performance and to evaluate our proposed method, we make use of certain parameters that constitute to the performance metrics. Selection of performance metrics parameters is of high importance as it should give a clear cut idea of how well the method works when compared to other existing technologies and also should be able to validate the effectiveness of the method. Here in this paper, we make use of four parameters that form the evaluation metrics.

Number of plants retrieved: The input to the method will be a user query which will have the soil characteristics and the output will be the plant list which will have the names of plants that satisfy the input user query. The parameter, number of plants retrieved is the number of plants in the plant list. As the number of plants retrieved increases, the effectiveness of the plant retrieval method also increase.

Ranking efficiency: The plant list will have many plants that satisfy the input conditions which are subsequently ranked. Ranking is done so that the most appropriate plants for the input soil conditions come top in the plant list. So the ranking procedure is of vital importance because the best fit plants should come in the top. In our method, we rank based on the frequency count and fuzzy score. Similarly we perform the ranking for the soil characteristics list in response to the input plant name. Here the ranking is done for each individual attribute in the soil characteristic list to get the best fit soil characteristics list for the input plant.

Computation time: Computation time refers to the time incurred between the input query and the output list. The input query may be soil characteristics or a plant name and the output will be the plant list or the soil characteristics list accordingly. Reducing the computation time shows better and faster processing of the query. Our method had a great advantage in reducing the computation time as we are using the fuzzy search method.

Memory usage: The amount of memory used up while executing the query is known as the memory usage. Having a lesser memory usage will validate the effectiveness of the method.

4.3 Experimental Sample Resluts

In our method for knowledge retrieval in Edaphology, we make use of two algorithms. In the first one, we input the soil characteristics to get the plant list that satisfy the

Copyright © 2016 MECS

input condition. Sample input and corresponding output is given in table 2. The table only shows the top 10 results of the total 44 plant names retrieved by the algorithm.

In the analysis, parameters performance metrics used are 1. Number of plants retrieved, 2. Computation time and 3. Memory usage. Table 2 and table 3 shows the values obtained for different metrics attributes for different queries for the proposed method and the baseline method. Fig. 8, 9 and 10 shows the chart graph for number of plants retrieved, computation time and memory usage for various queries for the two methods.

Table 2. Showing performance metrics values for input query for our method

Performance Metrics	Q1	Q2	Q3	Q4	Q5	Q6
No of plants Retrieved	32	31	27	42	34	33
Computation Time(s)	1.102	1.130	1.102	1.105	1.098	1.104
Memory Usage (Kb)	4486	4487	4687	4823	4561	4486

Table 3. showing performance metrics values for input query for previous method

Performance Metrics	Q1	Q2	Q3	Q4	Q5	Q6
No of plants Retrieved	28	13	13	27	28	25
Computation Time(s)	1.092	1.101	1.083	1.095	1.089	1.092
Memory Usage (Kb)	4483	3697	3665	4442	4421	4824

Next we plot the diagrams our proposed method by comparing the results of our method to the baseline paper [16] with the help of evaluation metrics.

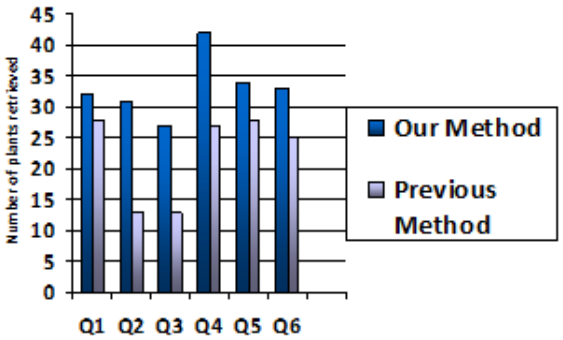

Fig.7. Chart showing the number of plants retrieved for various queries by the two methods

Fig. 7 above shows the number of plants retrieved for different queries for our method and the baseline method. Our method gives more plant options that satisfy the given soil characteristics. In this process we executed several queries against information retrieval system. However, in Fig. 8 above we show only 6 queries results for our method. Further notice that all these queries generated much finer results compared to the previous method. All these queries are independent to each other.

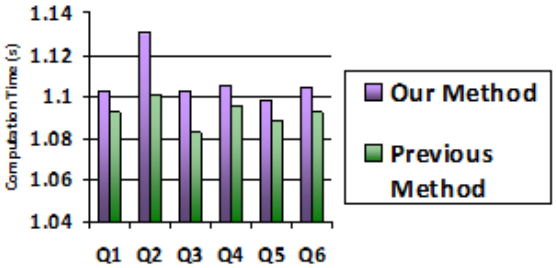

Fig.8. Chart showing the computation time for various queries by the two methods

Fig. 8 above shows the computation times taken for different queries for our method and the baseline method. As expected, evaluating the fuzzy membership functions our method took a little more execution time than other one, but it gave better results.

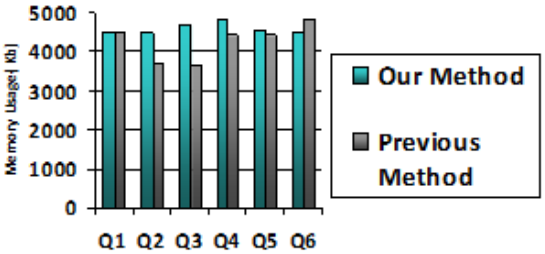

Fig.9. Chart showing memory usage for various queries by the two methods

Fig. 9 above shows the memory usage for different queries for our method and the baseline method.

REFERENCES

[1] B. Koester, "Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies". *International Conference on Data Mining*, July 2006, pp. 176-190.

[2] B.D. Newman, and K.W. Conrad, "A Framework for Characterizing Knowledge Management Methods, Practices, and Technologies". *In Proceedings of the Third International Conference on Practical Aspects of Knowledge Management*, 2000, pp. 30-32.

[3] Y. Li, and Y. Yao, "User profile model: a view from Artificial Intelligence". *In proceedings of 3rd International Conference on Rough Sets and Current Trends in Computing*, Oct 14-16 2002, pp. 493-496.

[4] Y. Li, and N. Zhong, "Web Mining Model and its Applications for Information Gathering". *Knowledge-Based Systems*, Vol.17, 2004, pp. 207-217.

[5] Y. Li, and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs, *IEEE Transactions on Knowledge and Data Engineering*, Vol.18, No.4, 2006, pp. 554-568.

[6] X. Tao, Y. Li, Y, and R. Nayak, "A Knowledge Retrieval Model Using Ontology Mining and User Profiling". *Integrated Computer-Aided Engineering*, Vol.15, No.4,

Copyright © 2016 MECS

2008, pp. 1-24.
[7] Y. Yao, Yi. Zeng, N. Zhong, and X. Huang, "Knowledge Retrieval (KR)". In proceedings of IEEE International Conference on Web Intelligence, 2007, pp. 729-735.
[8] M. Apistola, L. Mommers, and A. Lodder, "A Knowledge Management Exercise in the domain of Sentencing: towards an XML Specification". In: Proceedings of the Second International Workshop on Legal Ontologies, Amsterdam, the Netherlands: December 13, 2001, pp. 49-57.
[9] S. Denning "The role of ICT's in knowledge management for development". The Courier ACP-EU, Vol.192, 2002, pp. 58 - 61.
[10] R. Irfan, and M. Shaikh, "Enhance Knowledge Management Process for Group Decision Making". In Proceedings of World Academy of Science, Engineering and Technology, 2009.
[11] J. Whittaker, M. Burns, J.V. Beveren, "Understanding and measuring the effect of social capital on knowledge transfer within clusters of small-medium enterprises". In proceedings of the 16th Annual Conference of Small Enterprise Association of Australia and New Zealand, 2003.
[12] C. Grinand, D. Arrouays, B. Laroche, and M.P. Martin, "Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context". Geoderma, Vol. 143, Issue 1-2, Jan 2008, pp. 180–190.
[13] E.N. Bui, B.L. Henderson, and K. Viergever, "Knowledge discovery from models of soil properties". Ecol. Model, Vol.191, 2006, pp. 431–446.
[14] E.N. Bui, "Soil survey as a knowledge system". Geoderma, Vol.120, May 2004, pp.17–26.
[15] R. Irfan, and M. Uddin-Shaikh, "Enhance Knowledge Management Process for Group Decision Making". In Proceedings of World Academy of Science, Engineering and Technology, World Congress on Science, Engineering and Technology (WCSET 2009), Penang, Malaysia, February 2009.
[16] A. Meenakshi, and V. Mohan, "An Efficient Tree-Based System for Knowledge Management in Edaphology". European Journal of Scientific Research, Vol.42, No.2, 2010, pp. 253-267.
[17] L.L Ralph, and T.J. Ellis, "An Investigation of a Knowledge Management Solution for the Improvement of Reference Services". Journal of Information, Information Technology, and Organizations, Vol. 4, 2009, pp. 17-38.
[18] Q. Yang, J. Yin, Ch. Ling, R. Pan, "Extracting Actionable Knowledge from Decision Trees". IEEE Transaction on Knowledge and data Engineering, Vol.19, No.1, 2007, pp.43-56.
[19] R. Farenhorst, and R.C. de Boer, "Knowledge Management in Software Architecture: State of the Art". Software and Architecture Management, 2009, pp. 1-277.

Authors' Profiles

Naresh kumar. Kalapu was born in July 30 1984. He completed his Master's degree in Computer Science and Systems Engineering and presently pursuing PhD in Computer Science and Systems Engineering from Andhra University, Visakhapatnam, India.

Dr. Satyanand Reddy. Ch working as Sr. Assistant Professor at Computer science and systems engineering department, Andhra University, India. His research interests include Software Engineering, Software Cost Estimation, Object Oriented Analysis & Design, Web Technologies, Data Base Management Systems, and Operating Systems.

Dr. Murthy V.E.S. N working as Professor at Mathematics department, Andhra University, India. His research interests include Various Fuzzy Set Theories/Logics and Their Applications to Computer Sciences.

How to cite this paper: K. Naresh kumar, Ch. Satyanand Reddy, N.V.E.S. Murthy, "Fuzzy-Based XML Knowledge Retrieval Methods in Edaphology", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.5, pp.55-64, 2016. DOI: 10.5815/ijisa.2016.05.08