Search Engine Using Spatial Data

P.Sreedevi[1], G.Sridevi[2], B.Padmaja[3]

M.Tech. (CSE), NCET [1], Associate Professor, Department of CSE[2], Assistant Professor, DRKCET[3]

ABSTRACT:

Spatial search engines are specialized search engines primarily dedicated to retrieve geographical information through web technology. They provide capabilities to query metadata records for related spatial data, and link directly to the online content of spatial data themselves. OpenSearch-Geo extensions are developed to facilitate basic geographical data search using Open-search method.

OpenSearch-Geo extensions add new parameters of geographic filtering for querying spatial data and recommended set of simple standards responses in geographic format, such as KML, Atom and GeoRSS though spatial search engines. The communication method used in spatial search engines is based on standardized- Service-Oriented –Architecture. In this catalogue service plays a significant role. It provides a common mechanism to classify, register, describe, search, maintain and access information about resources available on a network.

In the contemporary search engine there is no mechanism to find out which of the available resources are best fit for use to users. We propose search functionality for current spatial data search engines to consider user quality requirements in addition to the geographical extent and keyword matching

KEYWORD: *Spatial, KML, Atom and GeoRSS*

I. INTRODUCTION:

Usage of spatial data resources on the web has increasingly become important in daily activates of modern society. In web technology mainstream search engines like Google, Yahoo, and ALO are used in accessing distributed information. When Internet users type a keyword or phrase into the search engine query box, they expect a list of search results which can be websites that offer information, products or services related to that keyword. However, finding proper web content is difficult due to availability of vast volume of information on the web. Therefore, searching for proper result requires specialized search engines.

BACKGROUND AND RELATED WORK

The motivation of this paper is to develop method for reasoning based on fitness for use to enable spatial search engines recommending spatial data resource for users.

Presently, there is no spatial data search engine that reason out based on consideration of fitness for use. We performed extensive study on the concepts in fitness for use and recommendation technologies. We also studied the quality of spatial data and users quality requirements to determine fitness for use.

We reviewed literature on fitness for use approach from users and producers perspective in spatial data infrastructure (SDI). We also studied the quality of spatial data and users quality requirements to determine fitness for use.

We designed the profiling algorithm and a reasoning logic to determine fitness for use of spatial data resources. We implement the model and the reasoning logic by using UML modelling language using the Enterprise Architect. For the realization of our system, we use the PostgreSQL spatial database management system with PHP: Hypertext Preprocessor for developing the front end as a web application sys- tem. The fitness for use reasoning logic is implemented by using the PostgreSQL structured language (PL/pgSQL) database programming languages.

FITNESS FOR USE AND RECOMMENDATION SYSTEMS

Fitness for use is a way of understanding the relationship between data and the users. The definition of fitness for use has been subjected to the usability of datasets. Redman suggested that for dataset to be fit for use it must be accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret. Therefore, fitness for use can be viewed as the capability of the dataset to fit stated user requirements and application specifications.

1.1 DATA QUALITY VERSUS FITNESS FOR USE

Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context and subjective to various applications. It highly depends on the need of individuals on how to use datasets. Quality can be described by object or phenomenon attributes and properties.

The term data quality is used to describe the correspondence between an object in reality and its representation in the datasets. Quality can also be expressed as a measure against a production specification or user requirements.

In GIS context a quality product is a product which is free from errors, or a product with confirmation of specifications used, or it can be a product that satisfies user’s expectations. However, widely accepted expression affirms that spatial data quality is recognized only in terms of its specific use.

Organization ISO is accepted in common to describe spatial data quality. The ISO defines quality as the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs. Therefore, for ISO quality is a result that has to be observed during use.

The standards mainly describe the spatial data quality using two main categories: quality overview elements and quantitative quality elements. The ISO provide quality elements with their sub-elements and guidelines for producers to describe the characteristics of the datasets. Spatial data quality evaluation procedure and reporting the result for quality evaluation procedure are also defined by ISO standards.

Devillers and Jeansoulin elaborates the concept of spatial quality by dividing it into two: internal and external quality. Internal quality is used to express the products with no errors. The internal quality describes characteristics that define the apparent individual nature of products. On the other hand, external quality is used to express products that meet user needs. It is associated to express the similarity between the data produced and user requirements and their needs.

When data quality description is defined by fitness for use, it should assure the user that the datasets are fit for the intended use.

1.2 FITNESS FOR USE: SPATIAL DATA PRODUCERS' PERSPECTIVE &USERS’ PERSPECTIVE

In Geographic Information Science (GIS) environments spatial datasets frequently have different origins and contain different quality levels.

Producers' perception of spatial data quality mainly depends on the dataset's internal characteristics. These intrinsic characteristics are resulted from production methods, e.g. data acquisition technologies, data models, and storages. Internal quality description of spatial dataset is independent of any task, unless it is collected and processed for a specific application. Producers of spatial data resource assume that users are able to determining a spatial dataset's fitness for use before use of the dataset. Under the fitness for use approach, producers do not make any judgment. Spatial data producers provide quality information contents is to help users to determine if spatial datasets fulfill their application's quality requirements.

Spatial data user’s quality requirements are rooted in the intended application they want the dataset to be used for. Users usually evaluate fitness for use of data sources to determine the suitability of data for problem solving and decision making and consider the datasets interoperability with other data sources. In addition, users also determine fitness for use according to their multidisciplinary information needs. Other factors, such as compliance to specific needs and availability of rules and quality control also have impact for users to determine fitness for use.

Directly or indirectly users of a dataset need to use information about spatial data quality in order to be able to assess the fitness for use of the data in their context.

1.3 APPROACHES TO DETERMINE FITNESS FOR USE

Determining fitness for use of a data resource is the only method to avoid risks caused by misuse of spatial data. Comprehensive comparison against user quality requirement and detailed quality description of dataset is the main approach to determine fitness for use. In determining fitness for use user's quality requirements, quality description of the dataset, the decision and how it will be influenced by quality are required input parameters. Given these information, evaluation of fitness for use can be implemented. For fitness for use evaluation, the user quality requirement and the dataset quality requirement should have the same base point. Otherwise, with the absence of such common agreement on quality of object, fitness for use assessment becomes much more complicated.

Each user group has certain requirements and different aspects of usability that have to be considered. The fitness for use decision can be easily determined if users quality requirement is known.

The well known approach in understanding user's quality requirements is translating subjective user's requirements into an objective technical specification.As a general approach in this research the reasoning logic design to determine fitness for use of spatial dataset is also based on comparison of user quality requirement (external quality) and the dataset quality description (internal quality).

1.4 RECOMMENDER SYSTEMS

Recommender systems are widely implemented for searching, sorting, classifying, filtering and sharing a vast amount of information available on the web to allow users to find resources that fit their need. All recommender systems take advantage of a particular set of artificial intelligence techniques. Recommender systems represent user preferences for the purpose of suggesting items to the users so that users are directed toward those items that best meet their needs and preferences. A recommender system customizes its responses to a particular user. Instead of direct response to queries, a recommender system is intended to serve as an information agent of individual users or group of users.

Recommendation techniques have a number of possible classifications. However, all recommender systems have three common fundamental components. The first component referred to as background data is the information that the system had before the recommendation process begins. The second component is the information that users must communicate to the system in order to generate a recommendation. It is referred to as input data. The third is the algorithm that combines background information and input data.

Recommendation techniques can be distinguished on the basis of their knowledge sources which can be the knowledge of other users' preferences, ontological or inferential knowledge about the domain, or added by users themselves. The main classification of recommendation techniques are:

- **Collaborative filtering**: Collaborative recommendation is probably the most familiar, most widely implemented and most mature among existing recommendation technologies. Collaborative recommender systems aggregate ratings or recommendations of objects, recognize commonalities between users on the basis of their ratings, and generate new recommendations based on inter-user comparisons. Griffith et.al conducted a survey on performance of collaborative filtering.

- **Content-based**: The system generates recommendations from two sources: the features associated with products and the ratings that a user has given them. Content-based recommender systems treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on product features. A content-based recommender learns a profile of the user's interests based on the features present in objects the user has rated. It is item-to-item or user-to-user correlation. Decision trees, neural nets, and vector-based representations have all been used. As in the collaborative case, content-based user profiles are long term models and updated as more evidence about user preferences is observed.

- Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one. Most commonly, collaborative filtering is combined with some other technique in an attempt to avoid the ramp-up problem.

Recommender systems typically determine matches via a process of identifying similar users by creating neighbor users. Determining recommendations based on selected neighbors is named as profile matching. Profile matching involves:

• Find similar users: employing standard similarity measures technique such as Nearest neighbour, Clustering and Classification

• Create a neighbour: techniques used include the creation of centroid, correlation-thresholding, and best-n-neighbours.

PROFILE BUILDING AND MAINTENANCE

The generation and maintenance of accurate user profiles is an essential component of a successful recommender system. Consequently, in analyzing how a recommendation system makes individuals recommendations or assesses a user needs, the key issue is the user profile. A recommender agent cannot begin to function until the user profile has been created.

RECOMMENDATION SYSTEM DATA MODEL DESIGN AND REASONING LOGIC

1.5 INTRODUCTION

In Section 2, we discussed the concepts of fitness for use from users and producers perspective. Both spatial data users and producers agree that fitness for use evaluation of a dataset before its usage reduce risks caused by misuse of spatial data resource. However, the two sides are not in line with the definition of fitness for use. The concept of fitness for use for spatial data users is the dataset that satisfies their need based on their quality requirement. On the other hand producers express fitness for use as the description of quality description of the dataset. Hence, the assessment and determination of fitness for use of a dataset remain users' responsibility. However fitness for use computation is not an easy task for users.

Adapting such an approach to the spatial data search engine is of great importance to search spatial data based on fitness for use. In this research work we propose a mechanize to store users spatial data search quality requirements and spatial data quality descriptions that can be used in fitness for use evaluation to recommend spatial datasets to users based on their requirements.

1.6 SPATIAL DATA RECOMMENDATION SYSTEM ARCHITECTURE

The proposed recommendation system design involves three main components as shown in figure. The figure describes the general view of spatial data recommendation system design framework.

• User interface: allows and controls user system interaction. The recommendation service obtains information about users' need through web based user interface. The user interface design considers user groups. For example, expert users group requires detailed quality information to determine if the resource is useful for their task or not. However, non GIS expert users group lacks understanding about detailed quality information. Therefore, the user interface design should support simple way of allowing these users to specify their data quality requirement. Moreover, if the users group is non human users, special web service communication facility like XML/GML standard data format should be maintained.

• Recommendation system: is the main component of the system which controls the overall interaction to provide fitness for use based spatial data recommendation.

• Profile database: is the data model of the recommendation system which store users information and spatial data quality information in a structured form. It allows automatic and active data retrieval to speed up the fitness for use evaluation, prediction and recommendation process of spatial data resources. Structured profile storage is defined by the conceptual data model of spatial data recommendation system which is discussed in the following section in detail.

FITNESS FOR USE EVALUATION FUNCTIONALITY

After users spatial data search requirements and spatial data resources information are profiled in the spatial data recommendation data model, in order to recommend the spatial data resources for users, the system should make fitness for use evaluation. In this research we discuss the fitness for use evaluation from three aspect: spatial extent matching, application matching with spatial data resources description and overview quality elements, and quantitative data quality evaluation aspect. However, since the fitness for use evaluation is performed using the system data model, the sequence does not have difference in recommending the datasets for users.

FITNESS FOR USE EVALUATION USING SPATIAL EXTENT

Fitness for use evaluation using user spatial extent requirement requires extensive spatial matching to get dataset with the best fit extent. First of all the spatial data resources that have spatial extent matching with users spatial extent requirement needs to be filtered. All the datasets which have intersection with user spatial extent requirements will be returned as a candidate dataset for further filtering.

This phase of filtering spatial datasets needs to be addressed from different aspect of spatial extent matching functions. For example, the user extent requirement may be completely inside the dataset extent or only a portion of area of user extent may intersect with the dataset extent. Therefore, spatial area difference can be known by calculating the area ratio. Hence, area ratio computation of intersection with user spatial extent requirement and area ratio computation of intersection with spatial data resources extent helps to identify the best fit spatial data resources. The value of area ratio is given in percentage.

Then by sorting datasets descending using the ratio of intersection and user extent the system can identify and return the best datasets. If there are more datasets that have similar area ratio values, again the ratio of intersection and dataset extent help us to identify the best one. Based on this logic we design algorithm 6 to rank spatial datasets using the computed area ratio.

In the algorithm design for the spatial computation the fallowing PostGIS built in functions are used:

- *ST _GeometryT ype*e Return the geometry type of the ST_Geometry value.
- *ST _within*: returns true if one geometry is within the geometry of the other, it takes two arguments. We used the user extent and spatial dataset extent to return true or false
- *ST _Centroid* : This function takes one argument. We used it to return the centroid of the geometry given by the user as a point. Therefore, the centre of the user extent requirement can be check within the extent of dataset.
- *ST _Intersects* : generates a boolean result after checking intersection between two geometry
- *ST _Intersection* : takes two ST_Geometry objects and returns the intersection set as an ST_Geometry object.
- *ST _GeomF romT ext* : returns a specified ST_Geometry to be enable the spatial function work

Variable definition used in Algorithm 1 - 3:

- UA - user application
- DSi Si=1…...N□ ∊ DSS - where DSS is set of selected datasets
- DSj□E=1...N□ ∊ DS S where DS is set of datasets E
- Ii E- user and dataset intersection extent E
- AI - UE and DSE intersection area
- RI_U - AI and area of UE ratio
- RI_DS - AI and area of DSE ratio
- DSiSE - selected extent dataset
- DSSS - selected and sorted dataset
- DSS - sorted DSSS

Algorithm 1: Select dataset based on user extent

Procedure:

- for all datasets, select a dataset if:
- user extent is within dataset extent
- center of user extent is within the dataset extent
- user extent and dataset extent intersection has polygon geometry
- return DS_S

Input: DS, U_E, DS_E

```
for $DS^i$ to $DS^N$ do
$DS^{iE}$□ ←extract_dataset_extent($DS^i$)
if ST_within($U_E$, $DS^{iE}$) then
$DS_S$□ ←$DS^i$
```

```
else if ST_within(ST_centroid($U_E$, $DS^{iE}$)) then
$DS_S$ □ $\leftarrow DS^i$
else
if ST_Intersect($U_E$, $DS^{iE}$) then
$I^i$ □ $\leftarrow$ST_Intersection($U_E$, $DS^{iE}$) $E$
if ST_GeometrType(ST_GeomFromText($A_I$ )) = "ST_Polygon" then
$DS_S$ □ $\leftarrow DS^i$
end if
end if
end if
end for
return $DS_S$
```

Spatial datasets which have a polygon intersection with user spatial extent requirement are returned as a result of algorithm 1. Once the spatial datasets are filtered by the spatial extent matching as given by algorithm 1, the selected datasets will be an input for algorithm 2.

Algorithm 2: Area ratio computation

Procedure:
- for all selected datasets:
- calculate user extent and dataset extent intersection
- compute intersection and user extent area ratio
- compute intersection and dataset extent area ratio
- return dataset selected, area ratios

```
Input: $DS_S$, $U_E$, $DS_{SE}$
   for $DS^{iS}$ to $DS^N$ do S
   $I^i$ □ $\leftarrow$ST_Intersection($U_E$, $DS^{iSE}$) $E$
   $R^{iI}{}_{-U}$ □ $\leftarrow^{ST}$ _Area($I_E$)) $i$
       $ST$ _Area($UE$
$R^{iI}{}_{-DS}$ □ $\leftarrow_{ST}$STArArea($I_E$) $^i$)
  end for _ $ea$($DS_{SE}$
  return $DS_S$, $R_{I\_U}$ , $R_{I\_DS}$
```

Algorithm 2 returns the same dataset that has been returned by algorithm 1 with newly computed area ratio extent information. This area ratio helps to order the dataset in order to identify the best one. The ranking procedure using spatial extent ratio value for every candidate spatial dataset is given in algorithm 3. However, for simplicity purpose we use sort function supported in PostGIS for implementation as given in algorithm 4.

Algorithm 3: Rank dataset based on Extent ratio

Procedure:
- sort selected dataset by R_{I_U} (A_I and area of U_E ratio)
- for all selected and sorted datasets, if two or more consecutive datasets have equal R_{I_U} , sort these rows with R_{I_D} (A_I and area of DS_E ratio) else update the index to indicate to the next group

```
Input: $DS_S$, $R_{I\_U}$ , $R_{I\_D}$
$DS_{SS}$ □ $\leftarrow$sort $DS_S$ desc $R_I\_U$
for $i$ to $M$ do
      //where $M$ is the number of selected and sorted datasets
if $R^{iI}{}_{-U}$ = $R^i I^{+1}$ then $\_U$
for $j$ = $i$ to $M$ do
if $R^{iI}{}_{-D}$ = $R^{jI}{}_{-D}$ then
```

```
$temp$□ $\leftarrow DS^{j}$ $SS$
{temporarily save current record $DS^{j}$ } $SS$
//next two lines swap current record with the next
$DS^{j}$ □$_{SS}$ $\leftarrow DS^{j+1}$$_{SS}$
$DS^{j+1}$□ $\leftarrow temp$ $SS$
end if
//to check the next group having the same $R_{I\_U}$
if $R^{j}{}_{I}{}^{+1} = R^{j}{}_{I}{}^{+2}$ then
//if the next two records $A_I$ and area of $U_E$ ratio are not equal,
//set position for the next comparison to this group and break the
//inner loop
$i=j+2$
break
end if
end for
end if
end for
return $DS^{S}$ $SS$
```

Algorithm 4: Rank dataset based on Extent ratio

Procedure:

- sort selected dataset by R_{I_U} (A_I and area of U_E ratio)

  - for all selected and sorted datasets, if two or more consecutive datasets have equal R_{I_U} , sort these rows with R_{I_D} (A_I and area of DS_E ratio)

- return DS^{S} SS

```
$DS^{S}$□ $\leftarrow$sort $DS_S$ desc $R_{I\_U}$, desc $R_{I\_D}$ $SS$
return $DS^{S}$ $SS$
```

FITNESS FOR USE EVALUATION USING APPLICATION

To design the fitness for use evaluation based on user application requirement, it is required to define theme_keyword that represent the application by referring thematic classification of dataset. The concept of theme_keyword definition is mainly required to search different spatial datasets which can be useful for the intended application. The theme_keyword definition for the application gives wide range of possibly to search various resources for the intended application. The ISO standard metadata representation topic categories is one of metadata elements required to identify a dataset, that is used to group keywords and to learn more about main themes of the dataset to understand topics exist in the dataset description. It is high-level geographic data thematic classification to assist in the grouping and search of available geographic datasets. The topic category is also used for topic-based search of available spatial data resources. It is one of a handful element that describes the type of features that are included in a dataset.

Variable definition used in Algorithm 5-9:

- U_A - user application
- DS^{iS}□$_{i=1...N}$□ $\in DS_S$ - where DS_S is set of selected datasets based on U_A or TKW
- TKW_j□$_{j=1...M}$□ $\in TKW$ - where TKW is set of theme_keywords
- $\Omega^{jT}KW$□ $\in$□ □0, 1□ - weight assigned for theme_keyword j
- Ω_A - weight assigned for U_A
- DS – dataset
- O_Q - overview quality description of DS^{i}
- $N =$□ □ D_S□ - number of datasets D_S in database
- M - number of theme_keyword of an application
- S_i□ $\in S$ - where S is the sum of theme_keywords in DS_S

- w total number theme_keywords defined for application
- R_i percentage value of relevance based on application and TKW similarity found

Algorithm 5: Select datasets using user application and theme_keywords

Procedure:
- for all datasets select a dataset if:
- user application and the theme_keyword is similar to overview quality description of the dataset or

- the theme_keyword is similar to overview quality description of the dataset even if user application is not similar to overview quality description of the dataset.

```
Input: $U_A$, $T\,KW$, $DS$
for $i = 1$ to $N$ do
    if $U_A$□ $\sim O_Q$ then
        if $T\,KW$□ $\sim O_Q$ then
            $DS_S$□ $\leftarrow DS^i$
        end if
    else
        if $T\,KW$□ $\sim O_Q$ then
            $DS_S$□ $\leftarrow DS^i$
        end if
    end if
    $i \leftarrow i+1$
end for
return $DS_S$
```

The result of algorithm 5 is the set of datasets that the overview quality or description has matching with user application or the corresponding theme_keywords. This datasets used in the process to quantify the application and the theme_keywords matching found in the dataset as shown in algorithm 9:

Algorithm 6: Quantify application name and theme_keyword in DS_S

Procedure: - for all datasets, compare user application with each overview quality description of the dataset. When ever they are similar, set weight of the application as the sum of all the theme_keywords, otherwise set the weight to 0 - for all datasets and for all theme_keyword, if a theme_keyword is similar to overview quality description of the dataset, set the theme_keyword weight to 1, else 0 - return weight assigned for user application and weight assigned for theme keyword

```
Input: $DS_S$, $U_A$, $T\,KW$
for $DS^i{}_S$to $DS^N{}_S$ do
    $O^{iQ}$□ $\leftarrow$get_overview_quality($DS^{iS}$)
    if $U_A = O^{iQ}$ then
        $\Omega_A$□ $\leftarrow \omega$□ $/ \omega >^M \Omega^{jT} KW j=1$
    else
        $\Omega_A$□ $\leftarrow 0$
    end if
    for $T\,KW_j$ to $T\,KW_M$ do
        if $T\,KW_j = O^{iQ}$ then
            $\Omega^{jT} KW$□ $\leftarrow 1$
        else
            $\Omega^{jT} KW$□ $\leftarrow 0$
        end if
    end for
end for
return $\Omega_A$, $\Omega_T\,KW$
```

The output of algorithm 6 is conversion of subjective matching into quantitative values. As explained before, the weight given for the user application matching should be greater than sum of all the theme_keywords defined for that specific application. This enables to identify datasets that match user application in prior than other datasets selected as candidate datasets based on theme_keywords. The theme_keywords are assigned boolean values as weight to indicate match- ing is found or not in the dataset where a weight equal to 1 means that matching has been found.

This values again summed up using algorithm 7 as shown below:

Algorithm 7: Compute sum of theme_keywords in dataset

Procedure:

- compute sum of theme_keywords in dataset DS^{iS} theme_keyword

```
Input: $DS_S$, $\Omega_T KW$
for $DS^i{}_S$ to $DS^N{}_S$ do
    $S_i \square \leftarrow M \Omega^{jT} KW j=1$
end for
return $S$
```

When the process of selecting spatial dataset based on user application, searching datasets by theme_keywords and assigning weight value completed, these values are used to rank spatial datasets that best fit user application. We said that application similarity found in the spatial dataset overview quality have more theme_keyword matching has second priority. Based on this assumption we devise a relevance indicator to inform users how much percent a dataset fits their application. In order to elaborate our approach we assume that there is an application with five theme_keywords and designed algorithm 8 as shown below:

Algorithm 8: Display relevance of DS_S based on application and theme_keywords weight

Procedure:

- For each datasets DS_S filtered by U_A and $T\,KW$
- Extract the weight Ω_A and S
- If exact matching for user application and the dataset overview quality, then indicates dataset is 100% relevant
- Otherwise calculate the difference between w and s^i to indicate the corresponding relevance

```
Input: $DS_S$, $w$
for $DS^i{}_S$ to $DS^N{}_S$ do
    $\Omega_A \square \leftarrow$ get_weight_by_$UA(DS_S)$
    $S \square \leftarrow$ get_weight_by_$TKW(DS_S)$
    if $\Omega^{iA} = w$ then
        $relevance \square \leftarrow Ri\%$
    else if $w \square - s^i = 1$ then
        $relevance \square \leftarrow Ri\%$
    else if $w \square - s^i = 2$ then
        $relevance \square \leftarrow Ri\%$
    else if $w \square - s^i = 3$ then
        $relevance \square \leftarrow Ri\%$
    else if $w \square - s^i = 4$ then
        $relevance \square \leftarrow Ri\%$
    else
        $relevance \square \leftarrow Ri\%$
    end if
end for
```

Once the final result for application matching is returned from algorithm 8 the relevance indicator can be used to order the datasets as shown in algorithm 9.

Algorithm 9: Rank Selected Data Sets DS $_S$ based on application and theme_keywords

Procedure:
- sort the dataset according to application and sum of theme_keywords

```
Input: $DS_S$
$\Omega_A$ □ $\leftarrow$ get_app_weight($DS_S$ )
$S$ □ $\leftarrow$ get_sum_theme_keyword($DS_S$ )
$DS_{SS}$ □ $\leftarrow$ sort $DS_{SS}$ desc $\Omega_A$, desc $S$
return $DS_{SS}$
```

FITNESS FOR USE EVALUATION USING QUALITY ELEMENT

To design the fitness for use evaluation of spatial datasets based on quality elements, the spatial datasets need to be extracted based on user extent and application requirement and populated in the system profile. Once the datasets quality description and necessary information are populated in the system, fitness for use evaluation can be performed. In order to compute the fitness of a dataset by comparing the quantitative data quality elements, the measurement unit of each quality element should be adjusted into the same measurement unit.

In this section we address the process of computing the fitness for use evaluation using quantitative data quality elements. to recommend a spatial datasets based on users quality requirements, we design algorithm 13 to compute range for all user quality requirement values. This is because it is not always possible to find spatial dataset that exactly match users requirement. We decided to subtract and add half of user quality requirement value to a user quality requirement value itself for each element to set the minimum and maximum of range.

Algorithm 10: sum of weighted X (dataset relevance to user quality requirement)

Procedure:
- for all datasets weighted boolean value, compute sum of weighted X
- return sum weighted X

```
Input: $DS$, $X^W$, $Q_{DS}$, $S_j = 0$
for $DS_j$ to $DS_M$ do
        $X^W$ □ $\leftarrow$ get_weighted_boolean($DS$)
    for $X^W{}_{JI}$ to $X^W{}_{JN}$ do
    $S_j$ += $X^W ji$
  end for
end for
return $S$
```

Algorithm 11: Calculate distance of dataset quality from user quality

Procedure:
- for all datasets:
- fetch boolean values of the datasets quality elements
- for each user quality requirements
- for a single dataset, if each boolean value is true, then compute the distance from user quality subelement to dataset quality subelement

```
Input: $DS$, $Q_R$, $Q_W$
for $DS_j$ to $DS_M$ do
        $X$ □ $\leftarrow$ get_boolean($DS$)
    for $Q^{iR}$ to $Q^N$ do $R$
    for $X_{ji}$ to $X_{jN}$ do
     /*$j$th dataset and $i$ to $N$ quality subelements*/
      $X_{ji}$ = fetch($X$) /*$X$ - boolean value*/
       if $X_{ji}$ = 1 then
      $d_{ji}$ □ $\leftarrow$ □ $/Q^{iR}$ □ $-Q^{ji}$ □ □ □$DS$
```

```
            end if
        end for
    end for
end for
return $D$
```

Finally using the computed relevance indicator from algorithm 10 and the distance value of each quality element of the dataset from algorithm 11, it is be possible to identify the best dataset that fits user quality requirements.

Algorithm 12: Rank datasets *DS* by relevance based on quality element evaluation

Procedure:

- sort the dataset according to their aggregated weighted boolean *X*

Input: *DS*

```
$S$ □ $\leftarrow$ get_sum_weight_boolean($DS$)
$DS_S$ □ $\leftarrow$ sort $DS$ desc $S$
return $DS_S$
```

Variable definition used in Algorithm 16:

- DS_S - sorted datasets
- DS_{SS} - sorted DS_S by distance
- Q_W - set of weight of quality elements provided by user
- Q^{max} - the maximum weight of quality elements provided by user *w*
- Q^{name} - the name of the maximum weighted of quality elements *w*
- *D* - distance between dataset quality and user quality
- d_{ji} - is distance between dataset quality and user quality in *i*th column

Algorithm 13: Rank dataset by identifying relevance by distance

Procedure:

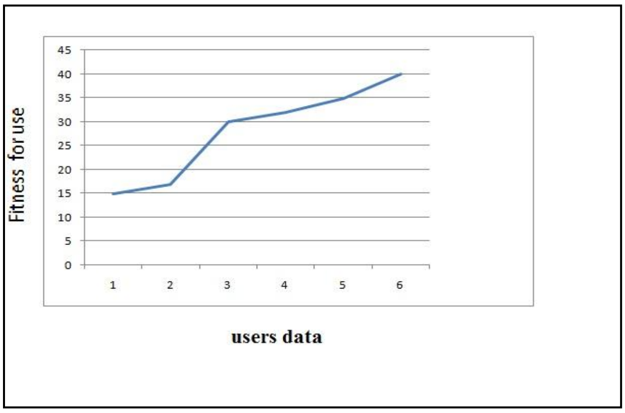
- Input datasets sorted by relevance value:
-for all user quality requirement weight
- identify the maximum user quality requirement weight assigned
- identify the name of quality element which have maximum weight
- find the dataset quality element which have the same name
- for each DS_S identify the distance value of quality element
- sort the DS_S based on the distance value

Input: *D*, Q_W, DS_S

```
for $Q^{iw}$ to $Q^N$ do $w$
    $Q^{max}$ □ $\leftarrow$ fetch_max($Q_W$) $w$
    $Q^{name}$ □ $\leftarrow$ get_name($Q^{max}$)
    if $DS_S \sim Q_W{}^{name}$ then
        for $DS^j{}_S$ to $DS^M{}_S$ do
            $d_{ji}$ □ $\leftarrow$ fetch($DS_S$)
        end for
    end if
    $DS_{SS}$ □ $\leftarrow$ sort($DS_S$) desc $d_{ji}$
end for
return $DS_{SS}$
```

II. EXPERMINETAL RESULTS

The recommendation service result page allows users to access recommended datasets as a response to their requirements. The recommendation process starts first by filtering spatial datasets based on user extent requirement. The datasets that satisfy user spatial extent requirement returned as a candidate datasets. Then application based filtering process starts specifically by comparing the user application to the overview data quality information usage and purpose and continues matching the theme_keywords of the application to the description of datasets



Finally the system recommend spatial datasets with corresponding values that indicates level of relevance

As computed by the fitness for use evaluation logic. The values returned with the recommended datasets enables users to easily observe which dataset best fits which requirements. In order to help users on picking the datasets based on extent, the computed spatial extent will also be visualized on the map.

CONCLUSION

The fundamental approach to determine fitness for use is comparison of users quality requirements and quality of data resources. In order to use fitness for use as a searching criteria in GIS, comprehensive comparison against user quality requirement and detailed quality description of spatial dataset is required. Thus understanding users' view towards spatial data quality and quality description of spatial data resources, gave us an idea on how to design a concptual model of recommendation system.GIS spatial data quality is a perception or an assessment of data fitness to serve its purpose in a given context and subjective to various applications. Widely accepted expression affirms that spatial data quality is recognized in terms of its specific use and the quality definition given by ISO is accepted in common to describe spatial data quality.

Therefore, we followed the spatial data quality according to ISO standard to represent the spatial data quality and users spatial data search quality requirements in our system. This simplification is required because the standard gives common ground on spatial data quality to evaluate its fitness for use. We also consider OGC catalogue service as a source to extract required datasets quality description. However, data quality description to determine fitness for use should not be limited to the ISO standard. Other factors such as: currency, cost, accessibility of the dataset, dataset granularity, popularity and users opinion about the dataset need to be included.

REFERENCES

[1] E. and B. Vaßeur. How to select the best dataset for a task. In *Proceedings of 3rd International Symposium on Spatial Data Quality (ISSDQ'04)*, pages 197-206, 2004.

[2] ISO/TC 211. Text of 19113 geographic information - quality principles, as sent to the iso central secretariat for registration as fdis, 2002.

[3] A. Agumya and G.J. Hunter. A risk-based approach to assessing the fitness for use of spatial data. *URISA Journal*, 11(1):33-44, 1999.

[4] Victorian Spatial Council. Spatial Information Data Quality Guidelines. Victorian Spatial Council, 2009.

[5] M.A. Gebresilassie, I. Ivánová, and J. Morales. User profiles for data quality models. Master's thesis, University of Twente Faculty of Geo-Information and Earth Observation ITC, 2011.

[6] ISOTC211. Revised text of 19115 Geographic information - Metadata, as sent to the ISO Central Secretariat for registration as FDIS , 2003.

[7] ISOTC211. Text of 19114 geographic information - quality evaluation procedures as sent to the iso central secretariat for publication, 2003.

[8] A.U. Frank, E. Grum, and B. Vaßeur. Procedure to select the best dataset for a task. *Geo- graphic Information Science*, pages 81-93, 2004.

[9] K.T. Huang, Y.W. Lee, and R.Y. Wang. *Quality information and knowledge*, volume 141. Prentice Hall PTR, 1999.

[10] A. Zargar and R. Devillers. An operation-based communication of spatial data quality. In *2009 International Conference on Advanced Geographic Information Systems & Web Services*, pages 140-145. IEEE, 2009.