A DDC Multiple Relation Content Retrieval Framework

W. Lertmahakiat*, P. Kampeerapaappat*, A. Mingkhwan**

* Department of Information Technology, Faculty of Information Technology,
** Department of Information Technology, Faculty of Industrial Technology and Management
King Mongkut’s University of Technology North Bangkok (Thailand)
Email: wilaiporn.l@rmutk.ac.th, poorivat.K@rmutk.ac.th, anirach@ieee.org

Abstract - Mostly the purpose of information retrieval development is to increase the performance of searching by using keywords matching techniques between content and user’s query string based on mathematics and linguistics. However, users were inundated with thousands of results from the relevant query but irrelevant to users. In addition, the query method by using a single language query in searching data from multi language can not be implied.

In this research, Dewey Decimals Classification (DDC) scheme in library system was used to classify in order to support multiple relational DDC retrieval. Pattern of relation in DDC would be created for retrieval based on similarity relation between query string and content. Fuzzy logic was used for measuring the degree of similarity. This method can retrieve information which will enable users to meet their need while it can also retrieve content in multi language.

Keywords - Multiple Relation Retrieval, Dewey Decimals Classification Scheme

I. INTRODUCTION

In recent days, information technology provided us with more facilitation, especially information retrieval. However, human needs are unlimited as a result of sharp increase information both content and pattern of information materials. Accordingly, information retrieval processes are developed consecutively in order to be more rapid, more collective and more relevant with user needs. Thus, the process of information retrieval becomes more sophisticated than traditional, that limited by requiring a document to share some keywords with the query to be retrieved. Later, vector and probability in mathematic theory were implemented in information retrieval for calculating more accuracy [1]. Nonetheless, extreme numbers of result impact user's time in looking for their needs. Many researches used technique classification, association and clustering in data mining theory for grouping information both in query and storage databases [2],[3],[4]. In addition, knowledge scheme and classification scheme in library were used to classify subject content in order to increase access by grouping [5], [6], [7]. However, all following retrieval processes still lack of retrieval with degree of relation in content and are using query in one language for searching multi language. In this research, Dewey Decimals Classification (DDC) scheme in library system was used to develop our DDC Multiple Relational Content Retrieval Framework, based on classification keywords and to calculate degree of relation in content database and query string in order to compare degree of shape DDC relation similarity between content and query string. This technique will retrieve unseen or non representation information to present information relational in more manifest pattern that each document or query related to any subject and pattern of degree relation. Therefore, to retrieval a DDC multiple relation contents using a similarity of DDC multiple relation query string will get the more related results also enable multi language content retrieval too.

The paper’s structure is as follows: In section 2, we discuss the dominant method of information retrieval and Dewey Decimals Classification Scheme. Section 3 depicts algorithm of a DDC Multiple Relation Retrieval Framework while detailed analysis of core component is provided with pseudo code in section 4. The result of our early prototype is discussed in section 5 and finally conclusions and direction for future are provided in section 6.

II. RELATED WORKS

2.1 Information Retrieval

Traditional information retrieval process is concerned with the similarity between term in query and index in term of document. Many applications require retrieval to go beyond matching word. But the query results often do not meet users’ requirements; so many techniques of data mining theory were implemented to increase efficiency of retrieval such as classification content investigative, effort towards exploring the use of external semantic metadata available in ontology in addition to the metadata central to documents, for the task of supervised document classification. The aim is to go beyond 'word co-occurrence [8], query clustering and association which return result objects associated with a degree of being relevant to the query [9], [10]. In many cases, result of relevance determines the order in which it is presented to the user and retrieved again with user relevance feedback which
can improve the result of more accuracy relevance and meet the requirement of the user.

2.2 Dewey Decimals Classification Scheme

Dewey Decimals Classification scheme has been developed over 100 years with fine tune and revision for innovation and advancement in knowledge. With a very good hierarchical classification performance it can be used to organize a collection of hundred items up to million items, [11] and also flexibility applied to a Content Management System [12]. There are many studies that have been conducted for the purpose of automatic classifying Web pages by applying Dewey Decimals Classification schemes to define vocabularies and control vocabularies to classify document on Architecture [5] and Engineering [6]. The results had shown those vocabularies and hierarchies’ classification can be dynamically generated, but a number in some cases of many documents classes were not really set suitably and intellectually assigned. The other experiment had built set of ontology for classification Web pages linking to the Dewey Decimals Classification (DDC) and Library of Congress Classification (LLC) schemes. The results showed that strategy that was applied for classification based on the use of the domain ontology. It does result not only in a better level of accuracy, but also point to the problem of incomplete ontology in the process of automated classification [7], [13], [14].

In this research we develop DDC Multiple Relational Content Retrieval, base on classification keywords and calculate degree of relation in content database and query string in all level of DDC scheme, compare similarity every classes between content and query string. This technique will retrieve unseen or non representation information to present information relational in more manifest pattern that each document or query related to any subject and pattern of degree relation.

III. PROPOSED FRAMEWORK

To represent multiple relation retrieval content, we propose A DDC Multiple Relation Content Retrieval Framework that consists of five components; Figure 1 presents an illustration of the framework.

3.1 Query String Parsing

The first part of the framework is Query String Parsing. This step will use to analyze and separate the query keyword; the user enters query string to search information from database. Parsing process will separate Query String words which have been checked on spelling and meaning from a dictionary, only word that has a meaning and correct spelling will be assigned to use as query keyword.

3.2 Query Keywords Relation Classification

In order to retrieve DDC Multiple Relational Classification content we also need to know the multiple relation of each keyword. This second process will find the relational of each query keyword in Dewey Decimals Classification (DDC) scheme. Using the library standard classification schemes one keyword will be able to classify as deep as 4 levels which assign number for notation as shown in figure 2. This scheme does divide human knowledge into 10 classes in the first level, 100 subclasses in the second level, 1000 divisions in the third level and the last level or leaf node contains more than 10000 sections.

One query keyword is organized into many hierarchies corresponding to vocabulary in concept space of DDC. A query keyword would be allowed to match more than one class, subclass, division and section in DDC scheme.
3.3 Query String Relation Integration

Usually query string contains more than one word; an average length of query is about 2.2 words [1]. In order to make the multiple relationships of each keyword more effective, we do integrate the DDC relation degree of every keyword from query string together in this process. DDC percentage in the same number notations of class, subclass, division or section had been integrated. Finally, each weight query keywords have been used to create shape of query string representing the DDC multiple relational.

3.4 Multiple Relation Retrieval

The fourth step is to retrieve multiple DDC relation content; fuzzy positive association relationship and fuzzy negative association relationship were chosen to represent degree of similarity of multiple relations between DDC query string and DDC content in database. Using SQL for access shape of multiple DDC relation content with the similarity between 0 – 0.75 (our initial criteria); degree 0 is the most similarity and degree 0.75 is the least similarity.

3.5 Retrieval Result Sets

The last process is to present the query result to users. We employ the same methods to rank the results to provide similarity degree 0.01 the “best” results first and degree 0.75 as the least similarity. The results should be shown with visualization with the purpose of increasing the accuracy in information retrieval, separating the results more clearly and navigating keywords relevant to users choosing for query string.

IV. MULTIPLE RELATION CONTENT RETRIEVAL ALGORITHMS

In this section we will present the main algorithms of A DDC Multiple Relation Content Retrieval Framework that could be separate in to 3 main algorithms.

4.1 The Parsing and Relation Classification Algorithms

The Query String Parsing and Query Keywords Relation Classification Algorithms comprise of process 3.1 Query String and Parsing and process 3.2 Query Keywords Relation Classification.

To classify query string into DDC scheme, Query String Parsing and Query Keywords Relation Classification Algorithms has been implemented for receiving query string from user, extracting keywords and analyze keywords. In details; we split keyword by keyword from query string and checked each keyword with the dictionary. Any keyword that was found in the dictionary will be kept as a query keyword, otherwise it will be abandoned.

Then, the query keywords have been classified into Dewey Decimals Classification scheme and represent degree in each number notation of class, subclass, division and section in DDC scheme that relevant to the query keyword. We also calculated that relevant degree into percentage and presented only highest percentage degree of relation in classes, subclasses, divisions and sections which could be shown the relation using radar graph as show in figure 3.
4.2 Query String Relation Integration

To integrate the relation of each keyword the process of query string relation integration we present Query String Relation Integration algorithms as show in Algorithms 1. Percentages of degree DDC relations in all keywords have been integrated in order finding lowest and highest percentage in each class and represent only highest value by starting with the first class which is also the first group of that important query.

Algorithms 1: Query String Relation Integration Pseudo Code

Move Zeros to LowPercentageofClass.
Move Zeros to UpperPercentageofClass.
Integration DDC Percentage of all keywords.
Move 1 to Class.
Repeat.
If Percentage of Class Greater than zeroes then
  If Percentage of Class is single value then
    Move Percentage of that class to LowPercentageofClass[Class].
    Move Percentage of that class to UpperPercentageofClass[Class].
  Else
    Move Low Percentage of that class to LowPercentageofClass[Class].
    Move Low Percentage of that class to UpperPercentageofClass[Class].
End if
End if
Add 1 to Class.
Until end of class in keyword class.
Display DDC Percentage of all keywords with radar graph.

In this algorithm, we set conditions, if any class gets percentage of relation more than zero; it means that there are data of that class. If percentage value of relation ratio has only a single value in that class, the percentage of the lowest and highest level will be the same value. While many values limit percentage of relation ratio as the lowest value and limit the highest value in return with increasing each class gradually till the end of the accumulated important class.

4.3 DDC Multiple Relation Retrieval

To compare the degree of DDC multiple relation more complex, we implied fuzzy as show in algorithms 2.

Algorithms 2: Multiple Relation Retrieval and Retrieval Result Sets Pseudo Code

Move zeroes to NegativeFuzzy.
Move zeroes to PositiveFuzzy.
Move zeroes to RankingBook.
Repeat
  Move 1 to Class.
  Move null string to SQL.
  Repeat
    If Percentage of Class Greater than zeroes then
      Add NegativeFuzzy to LowerPercentage[Class].
      Add PositiveFuzzy to UpperPercentage[Class].
      Addition Condition of Range between lowerPercentage, UpperPercentage to SQL.
    End if
  End if
  Add 1 to Class.
Until end of class in keyword class.
Use SQL for Searching in DDC Relative Index DB.
For each read DDC Relative Index DB of Found
  Add 1 to Ranking Book.
Display DDC Percentage of that Book with radar graph.
Next.
If PositiveFuzzy*10 < 0.75 Then
  Subtract 0.01 from NegativeFuzzy.
  Add 0.01 to PositiveFuzzy.
End if
Until PositiveFuzzy*10 >=0.75.

This algorithm is used to compare percentage of relation ratio as per DDC standard for all important queries with the content in our sample database of percentage ration by setting value of Negative Fuzzy, Positive Fuzzy and Ranking Book as zero.

Starting from the first class which is the first mode of query and setting value of SQL as Null String. If any class has percentage of relation ratio more than zero, increasing the percentage of the lowest value with Negative Fuzzy, increasing the percentage of the highest value with Positive Fuzzy in to build condition to SQL. This condition will start from DDC class 000 and repeat again on a new loop until the end of DDC class 900. Then, SQL was implied for searching the book relevant in Book DDC Multiple relations database. Sets of result matching to query had been ranked by percentage of relation, and show degree of DDC relation in each book with radar graph. If not matching to SQL condition, it will be searching by degree of relevant. In this process, we will decrease Negative Fuzzy and increase Positive Fuzzy value 0.01 to 0.75 in order to compare similarly.

V. IMPLEMENTATION RESULTS

We had tested our DDC Multiple Relation Content Retrieval framework prototype by having “Human Machine Interface” as the query string for learning and checking the correctness of our DDC Multiple Relation Content Retrieval Framework. Results are as follows:

Figure 4 shows the Radar Graph chart after having “Human Machine Interface” as a query string, keyword for analysis, result of Query Keywords Relation Classification, classification and percentage of relation according to DDC concept domain.

V. IMPLEMENTATION RESULTS

We had tested our DDC Multiple Relation Content Retrieval framework prototype by having “Human Machine Interface” as the query string for learning and checking the correctness of our DDC Multiple Relation Content Retrieval Framework. Results are as follows:

Figure 4 shows the Radar Graph chart after having “Human Machine Interface” as a query string, keyword for analysis, result of Query Keywords Relation Classification, classification and percentage of relation according to DDC concept domain.
As shown in figure 4, query keyword “Human” in level 1 has been related to DDC class 300, class 600 and class 200 as follow, and have been related to DDC subclass 390, subclass 610 and subclass 340 in level 2, and also related to division 398, 344 and division 611, 612 as followed in level 3.

![Fig 4: Query Keywords Relation Classification of “Human” “Machine” “Interface” in Level 2](image)

In figure 5, we have shown Query Keywords Relation Classification of 3 query keywords in level 2. Query keyword “Human” has been related to DDC subclass 390, subclass 610 and subclass 340 respectively. Query keyword “Machine” has been related to DDC subclass 620, subclass 680 and subclass 670 accordingly. But query keyword “Interface” has been related to DDC only subclass 000.

![Fig 5: Query Keywords Relation Classification of “Human” “Machine” “Interface” in Level 2](image)

The results of Query String Relation Integration collecting & gathering percentage and group as per subclass DDC show in table 1 and represents the highest percentage degree of DDC relation with radar graph in figure 6.

In this paper we will use only 2 levels, so that the result will show percentage and distinct classification to 100 subclasses. The Result of Query String “Human Machine Interface” in table 1; show only the highest percentages of degree DDC relations which have been tracking only 31 subclasses in level 2 that integrated weight more than zero and show highest percentage of each subclass. In order to show degree of relation more clearly and complexly, we transfer the data from table 1 to represent with radar graph in figure 6.

![Fig 6: Query String Relation Integration of “Human Machine Interface” in Level 2](image)

<table>
<thead>
<tr>
<th>DDC(Scheme)</th>
<th>Weight</th>
<th>Highest Percentage of Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>16</td>
<td>100.00</td>
</tr>
<tr>
<td>070</td>
<td>1</td>
<td>2.70</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>120</td>
<td>4</td>
<td>7.41</td>
</tr>
<tr>
<td>140</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>170</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>210</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>230</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>290</td>
<td>3</td>
<td>5.56</td>
</tr>
<tr>
<td>300</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>340</td>
<td>4</td>
<td>7.41</td>
</tr>
<tr>
<td>350</td>
<td>3</td>
<td>5.56</td>
</tr>
<tr>
<td>360</td>
<td>2</td>
<td>5.41</td>
</tr>
<tr>
<td>370</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>380</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>390</td>
<td>5</td>
<td>9.26</td>
</tr>
<tr>
<td>400</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>560</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>570</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>590</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>600</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>610</td>
<td>5</td>
<td>9.26</td>
</tr>
<tr>
<td>620</td>
<td>12</td>
<td>32.43</td>
</tr>
<tr>
<td>630</td>
<td>4</td>
<td>10.81</td>
</tr>
<tr>
<td>640</td>
<td>1</td>
<td>2.70</td>
</tr>
<tr>
<td>650</td>
<td>1</td>
<td>2.70</td>
</tr>
<tr>
<td>670</td>
<td>5</td>
<td>13.51</td>
</tr>
<tr>
<td>680</td>
<td>6</td>
<td>16.22</td>
</tr>
<tr>
<td>700</td>
<td>3</td>
<td>5.56</td>
</tr>
<tr>
<td>740</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>750</td>
<td>2</td>
<td>3.70</td>
</tr>
<tr>
<td>790</td>
<td>1</td>
<td>2.70</td>
</tr>
<tr>
<td>800</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>900</td>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>920</td>
<td>1</td>
<td>1.85</td>
</tr>
</tbody>
</table>

Table 1: Result of Query String Relation Integration of “Human Machine Interface” in Level 2.

As shown in Figure 6 Query string “Human Machine Interface” present a huge related to DDC subclass 000, subclass 620 and subclass 680.

By comparing a set of percentage of query string with a group of books that pre-classification and created DDC Multiple Relative, the Fuzzy Positive & Negative technique has been used to select (0.00% to 0.75%). Resembling books will be categorized as first priority. Results are as follows:
VI. CONCLUSIONS AND FUTURE WORK

In our DDC Multiple Relation Content Retrieval Framework, Dewey Decimals Classification (DDC) scheme in library systems was used to classify and calculate degree content relation in order to support multiple relational DDC retrieval. In this paper we used query string “human machine interface” for Multiple DDC Relational Classification experiment. We found that it can analyze content into more complex classes. Query keyword “human” has been related to DDC class 300, class 600 and class 200 respectively. Query keyword “machine” has been related to DDC class 600, class 000 and class 300 respectively. On the other hand, query keyword “interface” has been related to DDC only class 000. After we integrated degree relation of three query keywords, the results show that “human machine interface” has been related to DDC class 000, class 600 and class 300 respectively. Multiple DDC Relational Classification technique will retrieve unseen or non representation information to present information relational in more manifest pattern that each document or query was related to any subject and pattern of degree relation. Therefore, retrieval of multiple relation contents that has similarity with query string are possible and also enable users to use one query language for retrieval of multi language content.

In the future, we plan to develop the process of query string relation integration, implement fuzzy logic in multi relation retrieval process and representation retrieval result sets with visualization.

REFERENCES