

A Proposed Matching Algorithm for the Direction and Relevance of Information in Knowledge Assets

J. Watthananon * and A. Mingkhwan**

* Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand

**Faculty of Industrial and Technology Management, King Mongkut's University of Technology North Bangkok, Thailand

Email: watthananon@hotmail.com, anirach@ieee.org

Abstract - Knowledge Assets in knowledge management is part of the process used by information technology to assist in the data classification of organizations. Good knowledge assets are an effective way to create work successfully. Therefore, to have only data is not an important issue as in the past, so in this research, we applied the emphasis on connecting relationships of represent data, and described the archives of knowledge with the framework of standard DDC (Dewey Decimals Classification) to classify data and we propose a new efficient algorithm for the multiple matching knowledge. The experiment results, format emphasis of weight and quantity of relationships can be recommended by the direction of similar documents, this can be referred to in nature and inherited knowledge (i.e., parent and child nodes). Information in knowledge assets application using the matching algorithm is described to the core of knowledge in the fourth part of this paper.

Keyword: Knowledge Assets, Knowledge Direction, Knowledge Relevance, Multiple Relations, Relationships

I. INTRODUCTION

In recent times, many organizations focus on knowledge management to provide users with access and retrieval of information in a collected database or KA (Knowledge Assets). Because in the past, everyone would focus on quantity of data and thought if they had more data they would have a advantage over competitors. But the aspect of Knowledge Management is to have better quality data, also Knowledge Assets aims not to be too large in size [9], this means it does not collect everything in the database. KA: Knowledge Assets must be well designed, appropriately structured and divided into various topics for easy access. In addition, it must be able to identify and represent the content of knowledge that has relationship or degrees of relevance required to analyze and extract content hidden within the data. Research studies [6, 18, 10, 14 and 19] found that, they used minimal amounts of knowledge,

which is especially hard for high-level users and specialized users because data in the database is basic and less useful for classification of knowledge. The problems of clear classification of information from users are different experience levels and basic understanding of the system.

Knowledge mapping is a process in knowledge management [18, 8] to define the direction of relationship and representation to link the hidden relationships within the information. In this paper, we focused on development method of calculating the weight of knowledge link relationships and describing the ability of relevant information to define the direction of relationship to link the hidden relationships within the information. Therefore, this method shows the conceptual overview that is easy to understand and shows the sub-issues of each relevant area. The classification not only represents and shows how to use the road map information but, also the key factor used in Knowledge Assets as "The standard information analysis and the technique applied of knowledge".

The rest of this paper is organized as follows. In the next section we provide a brief review of related works. Section 3 presents the method of calculating the weight of knowledge link relationships and describes ability of the relevant information to define the direction of relationship and representation to link the hidden relationships within the information. Section 4 presents experiments with some discussion on the results. In Section 5, we conclude the paper and put forward the directions of our future works.

II. RELATED WORKS

Many researchers have studied and design methods of presentation from the information retrieval format that allows users to access and easily retrieve similar information, such as in research Sriram et al. [16] presented how to find similar documents that aims to present visualization data from the Web, and used calculation techniques to find the relationship matrix similarity. Information provided includes name, URL, Keyword and elements of data. The results showed that, the format presentation should be smooth and easy to understand. Saad et al. [13] presented a concept maps-based approach for

knowledge visualization by using the Boolean value associated with centrality to support primary education. The results showed that the prototype can adjust the concept of users to present and support in-depth level understanding. Marshall et al. [7] proposed a method concept mapping system to store and compare the knowledge of organization, this research applied element anchoring mechanism similar to flooding (SF) algorithm to match nodes and substructures between pairs of simulated maps and student-drawn concept maps. The experimental results showed significant improvements over simple string matching algorithm which combined recall accuracy of 91% for conceptual nodes. Smolnik et al. [15] studied the comparison between model systems in Groupware-based organizational memories with the period storage of knowledge in organizations, they found that knowledge had relevant consistently and distributed accordingly to basic functions but, cannot support user's search queries. So, they used a technique to create knowledge structures by relationships with topic maps, the experimental results showed that topic associations can describe the relationships between topics by presenting topic keywords. Gomez et al. [2] used conceptual graphs for the representation of text contents for similar documents by Dice coefficient method with new elements introduced by conceptual graphs. The experimental results showed that the system of information retrieval by a graph approach is especially good for short texts, it shows relationships between the elements from texts since its information retrieval works by comparison of at least one of the two elements. Hlaoui et al. [4] presented a new graph-matching algorithm for the inexact matching problem in K phases. The experimental results showed that content-based image retrieval was efficient correlating randomly generated graphs and can focus on a reduced set of model graphs.

From the researches above, the most effective technique to present data was calculation of the value similarities. In this paper, we also apply the DDC (Dewey Decimal Classification), base on classifying keywords and calculations to improve the classification model. Because, DDC is the standard library of hierarchical classification or family tree, it has a sufficiently wide topic for classification information and has an established relationship between class, sub-class and division by assigning notation to each call number [1]. However, we are adapting this measure to use as a concept hierarchy by the user, i.e. a direction of hierarchy, the weight of knowledge link relationships and to consider links of hidden relationships within the information. Recently, the Dewey Decimal Classification-Multiple Relations (DDC-MR) has introduced retrieval of unseen information for a set of documents [5, 18], for example, meaning and synonymy. The basic idea behind this approach is that documents are represented and linked to the hidden relevant data within the information. Each

node is represented and directed to knowledge relating to the subject.

III. PROCESS TO DETERMINE THE WEIGHT RELATIONSHIP OF THE KNOWLEDGE LINK

Process to determine the weight relationship of the knowledge link is the method of calculating and analyzing proportion value of information, to quantify the weight of knowledge link which affects the direction and relevance of knowledge from the collection data. In this paper we focused on direction of knowledge link between the information and the direction between different groups of information, this can be described in detail as follows.

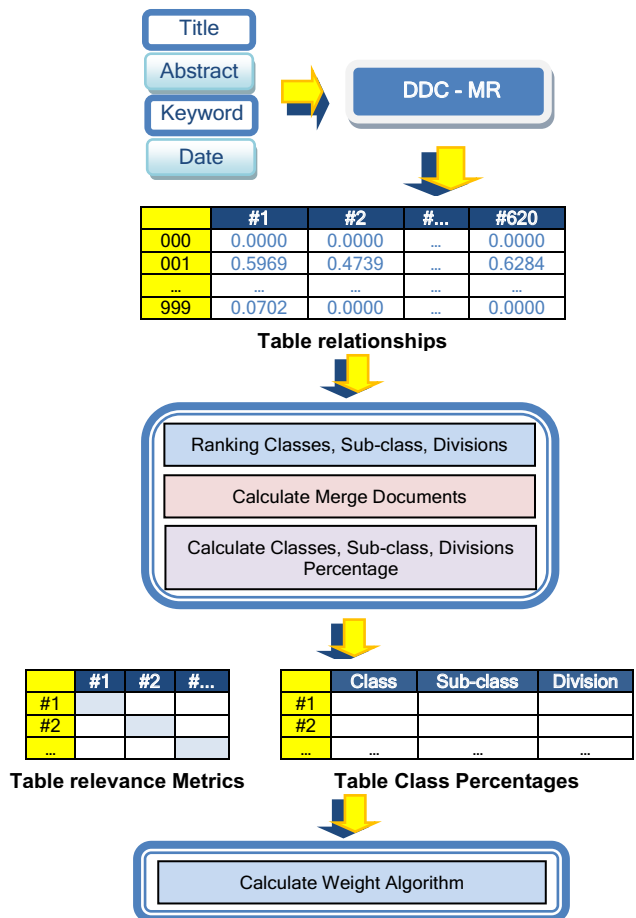


Figure 1: Procedure the weight relationship of the knowledge link

Step 1: **Calculate proportion of relationships.** From Figure 1 is a procedure of determining the weight relationship of the knowledge link, the first step, import data from research documents are divided into 4 parts: Title, Abstract, Keyword and Year published. Then data analysis with DDC-MR [1, 5] process to compute using the following equation:

Class Percentage:

$$CP = \frac{x_i * 100}{\sum_{i=1}^n x_i}$$

where x_i is value proportion of document relationship, as shown in Table 1.

Table 1: Proportional relationship of each document.

	#1	#2	#3	#4	#5	...	#620
000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000
001	1.0031	0.6180	0.3799	0.8069	0.3918	...	0.2633
002	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000
003	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000
...
999	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0527

Table 1 data shows proportional relationship between each of the class by DDC-MR [5] process that can classify all categories until 1000 classes. The value of the table proportion is 0 (zero), this is an arbitrary of relative zero which means relationship of the document still has relevance but is less relevant.

Step 2: **Similarity computation.** To compare and rank each document from values in Table 1, according to categories and sub-categories. The calculation has many methods but Lertmahakiat et al. (2009) presents the Pearson correlation coefficient – PCC as the best to predict similarities [6] so, we used this to compute the following equation:

Pearson correlation coefficient – PCC:

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Where $\sum X$ is sum of data measured by the variable 1 (X),

- (1) $\sum Y$ is sum of data measure by the variable 2 (Y),
- $\sum XY$ is sum of multiple data between variables 1 and 2,
- $\sum X^2$ is sum square data measured by the variable 1,
- $\sum Y^2$ is sum square data measured by the variable 2

And N is size of sample. After that, we use value from PCC to compute merge document, ranking class to sub-class and division, if the document has the same class it would calculate the number of classes and calculate degree of relationship between classes, as shown in Table 2.

Table 2: Table relevance metrics

	#1	#2	#3	...			
#1	0	147	11.7788	236	11.1373	...	
#2	147	9.5207	0	177	8.3530	...	
#3	236	15.2850	177	14.1827	0	...	
#4	223	14.4430	166	13.3013	397	18.7353	...
...	
...	

Table 2 shows the Number of class relevance and Quantity of relationships in each document, this can be described as #1 number of class relevant to #2 in class 147, #3 in class 236, #4 in class 223 and #5 in class 183 and quantity between relationships as 9.53%, 15.28%, 14.44% and 11.85%, while document #2 related to #1 in class 147, #3 in class 177, #4 in class 166 and #5 in class 149.

Step 3: **Calculate the weight of knowledge link.** This represents the size of knowledge link relationships in each node, as computed in previous methods. That proportion cannot be used to show the weight of knowledge link because it shows only data. So, to show the relationship between 3 nodes: For example with document A had relations class 100 is Philosophy, class 300 is Social sciences and node class 600 is Technology, had proportion relationship is 8%, 30% and 10% order class, as shown in Figure 2-(a).

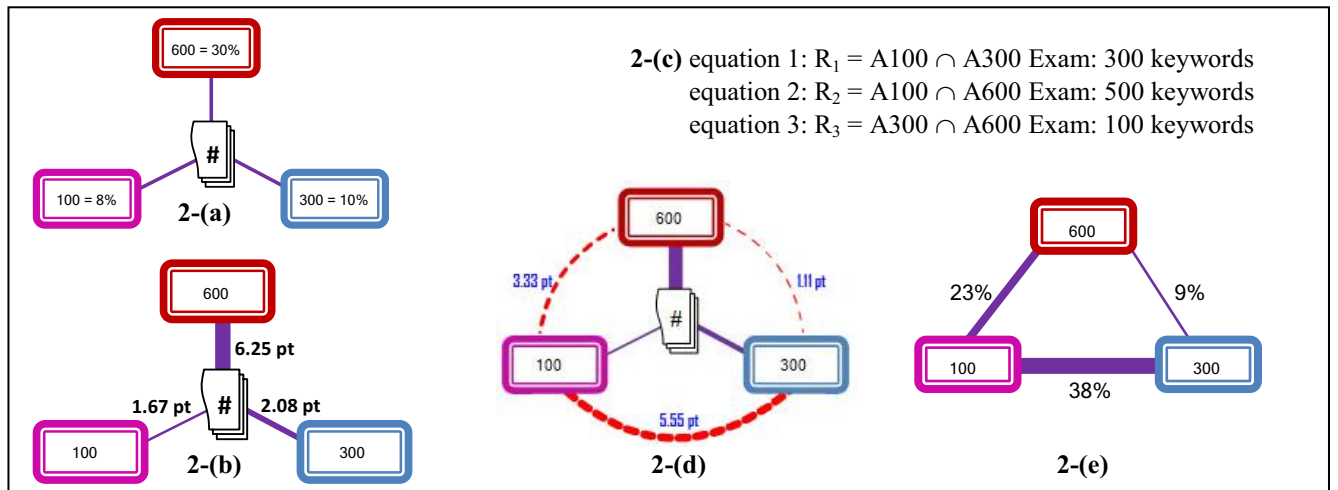


Figure 2: Preview the knowledge link relationship between 3 nodes.

Figure 2-(b) shows range of relation between documents A with 3 classes, this can describe the different relationship but we must calculate the new weight of knowledge link to provided meaning and appropriate proportions, which can be computed by using the following equation:

Weight value of the knowledge link: (3)

$$\frac{\text{value proportion of each link} * \text{proportion to display}}{\text{total value of proportion}}$$

where *value proportion of each knowledge link* is the sum of proportion origin to destination such as: document A → 100 = 8%, document A → 300 = 10% and document A → 600 = 30%, the *proportion displayed* is in the range of proportion relationship, users can set conditions and *total value of proportion* of raw proportion as 8 + 10 + 30 = 48.

$$W_1 = \text{document A} \rightarrow 100 = (8 * 100)/48 \cong 16.667$$

$$W_2 = \text{document A} \rightarrow 300 = (10 * 100)/48 \cong 20.833$$

$$W_3 = \text{document A} \rightarrow 600 = (30 * 100)/48 \cong 62.5$$

Step 4: Compare the size of knowledge link with criteria. To compute the appropriate size of knowledge link based on weight values calculated from 4.3 compares proportion with every knowledge link in the displayed data to show the meaning and description of the quantity of relationship. So, the criteria used in this paper is 0.1 mm per 1 proportion, this means if the knowledge link has a proportion of 40 when compared with criteria, size of knowledge link is 4 mm. In this method, the size of relation knowledge link has to change as shown in Figure 2-b, size of knowledge link in document A → 100 = 1.67 pt, document A → 300 = 2.08 pt and document A → 600 = 6.25 pt.

Step 5: Direction of the relationship. Generally relationships between variables are statistical methods to create a scatter plot which analyzes direction and proposes the relationship [12]. In this paper, the relationships are arising links between node and thickness or thinness part of related knowledge link. This consists of the number described as quantity of relationship on the knowledge link and can be applied to link similar information. In Figure 2-b we describe direction of relationship between document A and 3 knowledge's as: Technology, Social sciences and Philosophy, end of the arrow will point to the controlled knowledge. This shows the maximal direction relevant to Technology (6.25pt), while the minimal direction relevance was Philosophy (1.67 pt).

Last step: **Quantity of relationships.** To described the knowledge of one node associated with another node, and presenting quantity of the center position knowledge link relationships. Defined as a number proportion such as 18% means that the relationship between document and nodes, has values equal to 18, which if quantitative is not configured or cannot be shown, this means the quantitative relationships are uncertainty.

In Figure 2-(d) and 2-(e) shows relation weight of knowledge link between 3 classes: For example the DDC

standard can classify by keyword, so if we want to know the relation between 3 nodes in document A, we must used equation in 2-(c) to computed the same of keyword between classes. After that, calculate the weight of knowledge link in step 3, 4, 5 and 6 to represent the weight in class 100 → 300 = 5.55 pt, class 100 → 600 = 3.33 pt and class 300 → 600 = 1.11 pt and defined as a number proportion on knowledge link, as shown in Figure 2-(e).

IV. THE PROPOSE MATCHING ALGORITHM

In this section, we present a new algorithm for matching knowledge. The goal is to find the core of knowledge assets mapping between their nodes that leads to the most important class. The matching knowledge between the reach of class is a function of the similarity between each classes or nodes. It can be viewed as the distance between the knowledge's [19]. The basic idea of the new algorithm is iterative of the best possible class mapping and creates direction of the best knowledge at each iteration phase by considering both the weight of knowledge link with criteria and rank of quantity of relationship. The advantage of this algorithm is that this iterative process often allows finding the optimal core of knowledge mapping within information by classify data into 1000 classes. In the first phase, the algorithm selects the third possible rankings from 1000 ratings that minimize the multiple induced by node create knowledge link matching only. In the second phase, the algorithm examines the knowledge links that contain percentage between relationship of document and knowledge and then again computes that weight of knowledge link mappings that class to show quantity of relationships. After that, we receive that similarity measures associated with the basic relations have been defined; i.e. document number 1 have already been relation with 1000 classes of nodes and edges, selection of nodes and edges, etc. We used this to compute the following algorithm:

Algorithm and Complexity:

Input: 1000 attributed classes of each documents

Output: matching between classes or nodes in documents from the knowledge assets

1. Initialize K as follows: For each $K_{ij} = d(\mu_1(v_i), \mu_2(v_j))$.
2. Initialize R as follows: For each R_{ij} , $i = 1, n$ and $j = 1, m$, set $R_{ij} = 0$.
3. While Relation_Phase < N
 - If Relation_Phase = 1, Then
 - For $i = 1, \dots, n$
 - Set the value 1 to elements of R corresponding to the most value in i row of K ;
 - Call Matching_Classes(R).
 - Else If Relation_Phase = 2, Then
 - For $i = 2, \dots, n$
 - Set the value 1 to elements of R corresponding to the second rank value in i row of K ;
 - Call Matching_Classes(R).
 - Else If Relation_Phase = 3, Then
 - For $i = 3, \dots, n$
 - Set the value 1 to elements of R corresponding to the third rank value in i row of K ;
 - Call Matching_Classes(R).
 - Else For all $i = 1, \dots, n$

```

Set R' = R
For all j = 1, ..., m set Rij = 0
  Select the element with the most value in K that is not
  calculated 1 in R' and set it to plus 1 in R and R';
Call Matching_Classes(R);
Set R = R'.
If all the elements in R are calculated 1, Then
  Set Matching_Classes = N
Else plus 1 to Matching_Classes.

```

Matching_Classes(R):

For each valid mapping in R

1. Compute the matching relation induced by nodes.
2. Compute the weight of knowledge link.
3. Create the direction between document and nodes.
4. Show the percentage relation on each knowledge link that means Quantity of relationships.

Description with the algorithm:

$K = (K_{ij})$ is Knowledge's, where n is the numbers of classes or nodes. Each element K_{ij} in K demotes the relation between node i in K_1 and node j in K_2 and $R = (R_{ij})$ is rank. The first step is to initialize $K_{ij} = d(\mu_1(v_i), \mu_2(v_j))$ [19]. The second step consists of initializing R setting $R_{ij} = 0$. The third step contains N phases. In the $Relation_Phase = 1$, the elements of R corresponding to the most value in each row of relevance matrix. In the $Relation_Phase = 2$, the elements of R corresponding to the second rank value in each row of relevance matrix and in the $Relation_Phase = 3$, the elements of R corresponding to the third rank value in each row of relevance matrix.

V. EXPERIMENTS AND DISCUSSION.

5.1 Data used in the experiment

In our experiments, we use collection knowledge from academic articles of the participants in a national conference disciplines in the computer and information technology amounting to 620 articles, which were published during 2005 to 2009. In this paper, we provided one document for research, so the dataset used 4 sections: Title, Abstract, Keyword and Year published, this pattern can classify documents in a short period of time.

5.2 Discussions

We used DDC-MR process provided by the relationship analysis to run our experiments. DDC-MR is a hierarchy process of classify content based on DDC in level 3 (1000 classes). Thus, we used this to classify unseen data in knowledge, calculate proportion of relationships in each class, calculate the weight of knowledge link and display size of knowledge link referring to the direction and relevance of information in knowledge assets. However, the research aims to develop the method for calculating the weight of knowledge link relationships and describe ability of the relevant information to define the direction of relationship and representation link the hidden relationships within the information. This can show us the direction relation of class and what proportion is linked to another knowledge or different content, as shown in Figure 3.

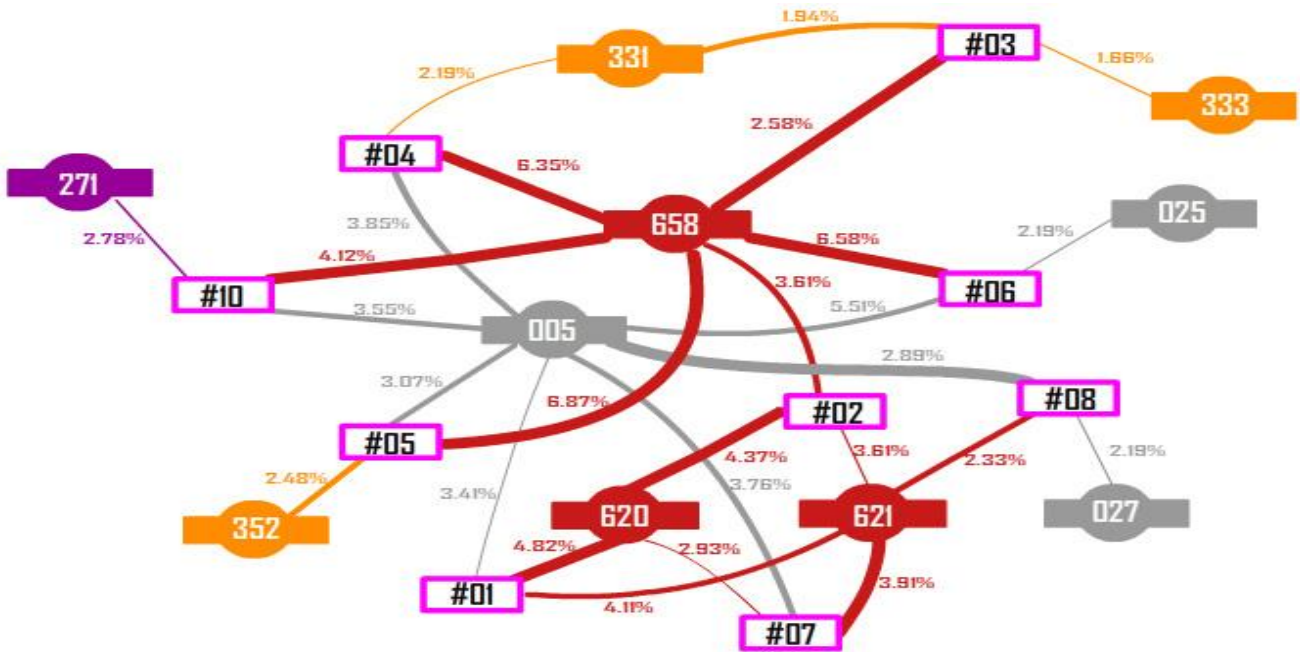


Figure 3: Shown direction and proportion relevance of each document.

Figure 3 shows an example to link each document amounting to 9 articles that assigns documents represented by a square symbol and class represented by planet symbol.

The aim is to link the size of knowledge link which was computed to the defined details of the proportion relationships, then represent it simply to help everyone

understand the data. Each document shows the direction and proportion relevance by 3 sequences because we cannot displays any class, as it would not be able to read the relationship. Shown in Table 3, descriptions of document #1 is associated with class 620 = 4.81%, class 621 = 4.12% and class 005 = 3.41%, while the document #2 is associated with class 620 = 4.37%, class 658 = 3.61% and class 621 = 3.61%. The weight of knowledge link used to display the direction of relation had different size depend on the computed value of proportion.

Table 3: Weight value of relevance between class of each document and sorted by rank.

Name	#01		#02		#03		#...
	Class	%	Class	%	Class	%	..
Rank1	'620'	4.824	'620'	4.372	'658'	2.58	...
Rank2	'621'	4.118	'658'	3.607	'331'	1.944	...
Rank3	'005'	3.412	'621'	3.607	'333'	1.661	...
Rank4	'658'	2.941	'005'	3.388	'371'	1.626	...
Rank5	'515'	2.824	'676'	3.169	'725'	1.626	...
...

The experimental results show the direction and quantity of relationship between documents, which is relevant in Computer Programming (005) and the second, is class General Management (658) by analyzing the amount of links and the weight of knowledge link relationships. However, we can refer to the relevant documents in-depth when searching for content such as document #3 and #4 which have relevance to class General Management (658) and class Labor economics (331) according to the computed value of proportion. The document #4 can be referred by document #8 in class Computer Programming (005). Thus, if we only focused on direction, the displayed data would not go to the core of knowledge or the vice of knowledge.

VI. CONCLUSIONS AND FUTURE WORKS

The direction and relevance of information are applied to link knowledge that is extracted from Knowledge Assets in a manner known as Knowledge mapping. This is supported and represent by contributing information from 620 articles. This method can represent the conceptual map of relation and open dimensional thinking of the total system by linking the correct path. In addition, when considering the benefits, quantity of knowledge or the thickness weight of the knowledge link can correlate to continuous data. We can analyze all relevant data and display it, so the problem is not the searching of knowledge but the issue understands the knowledge clearly.

In our future work, we plan to develop and apply our proposed method with a variety of information retrieval such as star relationship or triangle relationship etc.

REFERENCES

- Dewey, M., *Dewey decimal classification and relative index*. 21st ed: ed, ed. J.P. Comaromi. 2003, Albany, New York OCLC Forest Press.
- Gomez, M. M., Lopez, A.: Gelbukh, A., *Information Retrieval with Conceptual Graph Matching*, Mexico, 2000.
- Golub, K., *Automated subject classification of textual web documents*, Journal of Documentation., 2006(62).
- Hlaoui, A., Wang, S., *A New Algorithm for Graph Matching with Application to Content-based Image Retrieval*, Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2002.
- Lertmahakiat, W., Mingkwan, A., *Information Retrieval by Multiple DDC Relational Classification*: Proceedings NCCIT 2008. The 4th National Conference on Computing and Information Technology, 22 – 23 May 2008, King Mongkut's University of Technology North Bangkok, Thailand.
- Lertmahakiat, W., Mingkwan, A., *A Study of Retrieval Algorithm for DDC-Multiple Relation Classification in Content*: Proceedings PGNet 2000. The 10th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting. 23 – 24 June 2009, Liverpool, U.K.
- Marshall, B., Chen, H.: and Madhusudan, T., *Matching Knowledge Elements in Concept Maps using a Similarity Flooding Algorithm*, October, 2005.
- Mostafa, J., Peyman, A., Atieh, B., Roozbeh, H. A., *A Framework for the Selection of Knowledge Mapping Techniques*, Journal of Knowledge Management Practice, Vol. 10, No. 1, March 2009.
- Nonaka, I., *Knowledge Based Management Synthesizing East and West*, Ministry of Education with the Collaboration of JICA, 27 June 2006.
- Oded Nov., *Creativity, Knowledge and IS: A Critical View*, Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
- Peery., C., A.M. Wei Wang., T.D. Nguyen., *Multi-dimensional search for personal information management systems*: in EDBT'08. 2008, Nantes, France.
- Rucker, G., Schumacher, M., Simpson's paradox visualized. *The example of the Rosiglitazone meta-analysis*: BMC Medical Research Methodology, 30 May 2008.
- Saad, A., *A Concept Maps-based Approach for Knowledge Visualizations and Autonomous Self-Assessment in Distance Learning Environments*: International Conference on Engineering Education, 21 – 23 July 2003, Valencia, Spain.
- Sajjad M. Jasimuddin., *Storage of Transferred Knowledge or Transfer of Stored Knowledge, Which Direction? If both, then how?*, Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
- Smolnik, S., Nastansky, L., K-Discovery: *Using Topic Maps to Identify Distributed Knowledge Structures in Groupware-based Organizational Memories*, Proceedings of the 35th Hawaii International Conference on System Sciences, 2002.
- Sriram, V., *Term and Document Correlation and Visualization for a set of Documents*, 2008, Stanford University.
- Wattananon, J., Mingkwan, A. *A Relation of Distributed Knowledge Classification Using DDC and Time Domain*, Proceedings PGNet 2008. The 9th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting. 23 – 24 June 2008, Liverpool, U.K.
- Wattananon, J., Mingkwan, A., *A connection relationship group of knowledge technique with DDC-Multiple Relations*, Proceedings WUNCA 21st and CIT2010. National Conference on Computer Information Technologies 2010. 13 – 16 January 2010, Burapa University, Thailand.
- Wattananon, J., Mingkwan, A., *A Magnify Classification technique for group of Knowledge using DDC-MR*, Proceedings NCCIT 2009. The 5th National Conference on Computing and Information Technology, 22 – 23 May 2009, King Mongkut's University of Technology North Bangkok, Thailand.