Software Engineering And Web Engineering

Section

Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190005

PREDICTION OF WEB USAGES USING RSA ASSOCIATION RULE MINING ALGORITHM

Thu Zar Htet[(1)], Thida Lwin[(2)], Lei Lei Win[(3)]

[(1)(2)(3)]University of Computer Studies (Meiktila), Myanmar

[(1)]dawthuzarhtetisdept@gmail.com

ABSTRACT

Web mining has become a necessity to use efficient information retrieval techniques to find and order the desired information. Although there exists quite some confusion about the Web mining, the most recognized approach is to categorize Web mining into three areas: Web content mining, Web structure mining, and Web usage mining. Web usage mining is one of the most popular web mining techniques in order to generate the web usage patterns which can be further exploited in better personalization, improving navigations, recommendations, and recognition of web sites and attracting more advertisements. It focuses on the techniques that could predict user behavior while the user interacts with Web. This paper predicts the web usages using the Relative Support Apriori (RSA) algorithm which is also discovered the frequent patterns with the relative support and support measures. RSA association rule mining algorithm is applied on NASA web log data in order to predict the usage patterns. This paper accepts the NASA web log and preprocesses these data to improve data quality and produce the usages patterns. Finally, assesses the performance of RSA algorithm.

KEYWORDS: *Web Usage Mining, Web Log Data, Relative Support Apriori (RSA) Algorithm*.

1. INTRODUCTION

World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The growth of web is tremendous as approximately one million pages are added daily. Due to these huge, unstructured and scattered amounts of data available on web, it is very tough for users to get relevant information in less time. To achieve this, improvement in design of web site, personalization of contents, prefetching and caching activities are done according to user's behavior analysis. The ability to know the patterns of users' habits and interests helps the operational strategies of enterprises. Various applications like e_commerce, personalization, web site designing, recommender systems are built efficiently by knowing users navigation through web.

Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. The objects of Web mining are vast, heterogeneous and distributing documents. The logistic structure of Web is a graph structured by documents and hyperlinks, the mining results may be on Web contents or Web structures. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web usage mining is one of the applications of data mining which is used to mine of log files to discover useful patterns which can be further exploited in better personalization, improving navigations, recommendations, and recognition of web sites and attracting more advertisements etc. Every time when a server of a website receives a request from web user and the usage data captures the identity or origin of web users along with their browsing behavior at a web site. User's activities can be captured into a special file called log file. There are various types of log: Server log, Proxy server log, Client/Browser log. These log files are used by web usage mining to analyze and discover useful patterns.

Frequent pattern mining is an important knowledge discovery technique in data mining. Most frequent pattern mining has been designed with the

traditional support-confidence framework that generates more interesting kinds of patterns. This specialized framework may use different types of interestingness measures, model negative rules, or use constraint-based frameworks to determine more relevant patterns. In the basic model of frequent patterns, a pattern (or an itemset) is considered frequent if it satisfies the user-defined minimum support (min-sup) constraint. The min-sup constraint controls the minimum number of transactions a pattern must cover in a database. Since only a single min-sup constraints used for the entire dataset, the model implicitly assumes that all items in a database have uniform frequencies.

However, this is often not the case in many real-world databases. In many real-world applications, some items appear very frequently in the data, while others rarely appear. It has to be noted that considering an item in a database as either frequent or rare is a subjective issue which depends on the user and/or application requirements. To discover the useful frequent pattern, RSA association rule mining algorithm is applied. Finally, assesses the performance of RSA algorithm.

2. RELATED WORK

Due to the rapid usage of World Wide Web, websites are the information provider to the Internet users. Storing and retrieving the information from the web is always a challenging task. Web mining, the term is defined as extract needed information to the users from the Web. The information provided by the Web is not only the exact information of user needs but also suggest the information associated to the exact one. The author in this paper [1] introduces the applications and the mining process of data mining tool (open source) Rapid miner.

They proposed work analyzes the usage of web pages (i.e. Browsing behavior of user) using two different clustering algorithms such as k-means, which is incorporated in the tool and Fuzzy c means(FCM) clustering using Rapid Miner. The results showed operational background of FCM clustering and k-means clustering algorithm based on the cluster centroid.

Web Usage Mining techniques are great area of research these days. Providing users what they are looking for in websites is the ultimate aim of web usage mining. In the approach of[2], the aim is fulfilled by using association rule mining technique on clustered data i.e. data applied clustering techniques first and then applied association rule technique for frequent accessed set of link. Basic Association Rule Mining has drawback of generation of irrelevant rules, generation of too many rules leading to contradictory prediction resulting in reduction of accuracy.

Minimum support and minimum confidence parameters can be set in such a way to eliminate false discoveries. But when minimum support is too small, every rule will get a chance to be true, leading to wrong result and when minimum support is too large, for small data set, wrong prediction may occur. Clustering frequent access patterns reduce dataset for Association Rule Mining and improve result accuracy and producing results of pattern discovery of web usage mining process effective.

Analyzing the web log files through web usage mining is very important to discover the similar behavior users of particular website. The paper [3] discussed how to find useful knowledge from web log file using some data mining technique like Association rule mining and clustering. First they preprocessed the web log file then applied association rule mining and clustering algorithm on web log file to discover usage pattern and same behavioral users. The approach used in this paper [3], helps the website designers to improve their website usability.

Continued growth of user number and size of shared content on Web sites cause the necessity of automatic adjusting content to users' needs. In the literature of Web Mining, such actions are referred to personalization and recommendation which led to improve the visibility of presented content. To perform adequacy actions which correspond to the expected users' needs, [4] utilized web server log files. Mining such data with accurate constraints can lead to the discovery of web user navigation patterns.

Such knowledge is used by personalization and recommendation systems (PRS) due to performed actions against user behavior during a visit on the web portal. In this paper [4], they presented the system framework for mining web user navigation patterns in order to knowledge management and focused on constraints which are critical factors to evaluate the effectiveness of the implemented algorithm. On the other hand, these constraints can be perceived as knowledge validation criteria due to its adequacy. Thus only adequate knowledge can be added to existing in PRS knowledge base.

3. WEB MINING

Web mining is the application of data mining techniques to extract uncovers relevant, hidden information on web. Web mining can be categorized into three classes based on content, structure and usage of web pages. Three areas of web mining are: Web Content Mining, Web Structure Mining, and Web Usage Mining.

3.1 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining generally uses basic data mining algorithms such as Association rule mining, Sequential rule mining, Clustering, Classification etc. for pattern discovery phase [5]. Due to high raise in number of transactions, Association rule mining is the most basic data mining technique to be used in web usage mining to find association between web pages. It refers to the set of pages that are accessed together in a single server session. This information can be useful to restructure the web site.

3.2 Web Log Data

Web log data contains various parameters related to web server activity which are analyzed to extract useful information. Three main data sources are used to collect log data for web usage mining. Those are Server log, Proxy server log, and Client/ Browser log.

3.3 Log File Formats

Different web servers provide various format of log files such as Common log format, IIS standard/ extended log format, Combined/Extended common log format, Log markup language (LogML), because of different setting parameters. Among them common log format are commonly used. Common log format is a standard non-customized format (fixed no of attributes) suitable for http web sites. This type of log includes user's IP address/hostname, rfcname, log name, date with time zone, page access method, PATH, http version, server response code and byte received.

3.4 NASA Log File Format

NASA log file format has the parameters such asIP address/hostname, rfcname, log name, date with time zone, page access method, PATH, http version, status code, and byte received.

3.5 Prediction of Web Usages Using Relative Support Apriori Algorithm

Relative Support Apriori(RSA) algorithm which is also discovered the frequent patterns with the relative support and support measures. In this paper, RSA association rule mining algorithm is applied on NASA web log data in order to predict the usage patterns. Firstly, this system accepts the NASA web log and preprocesses these data to remove the unnecessary data and to improve data quality. In preprocessing step include the data cleaning, user identification, session identification, transaction identification and formatting. After processing these phase, this system produces the usages patterns and then calculates the performance of RSA algorithm and the system flow diagram is shown in Figure 3.1.

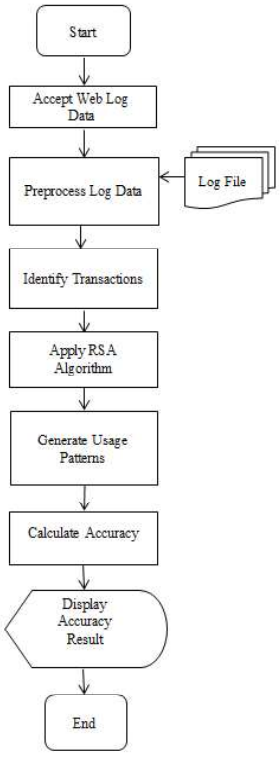

Fig 3.1. System Flow Diagram of Prediction of Web Usages Using RSA

3.6 Data Preprocessing Process

Data preprocessing process performs web data into consistent data. There is a distinction between the explicit information and implicit information. Whereas data in the explicit information is collected in a proper form of the statistical analysis, beside the data needs a pre-processing process. This process can be done in several activities: Data Cleaning, User Identification, Session Identification and Transaction Identification. These activities are described in the following section.

3.6.1 Data Cleaning Process

Some information from web log files are redundant and irrelevant. So the removing is necessary in data cleaning task. The cleaning firstly removes the unsuccessful records in which the status code field recognizes the unsuccessful http request. It removes the entries with status codes that are not 200. Removing these kinds of requests is necessary because they are just increasing the size of log file and nothing to do with analysis of user's navigational behavior.

3.6.2 User Identification Process

There are several methods to identify unique user. User identification is one of the complicated tasks due to existence of local/external proxy servers, cache systems, cooperate firewalls and shared internet. IP address is used to identify the unique. IP address is logged into log file when a user hits a page and it can be used to identify different users. But in case of proxy server when many users request a particular page then web site server logged same IP address (Proxy server IP) into the log file. Practically different users are accessing that page. Caching also creates problem to identify unique user. Whenever a user tries to access previously accessed page, browser display pages from local cache and no entry are logged into the log file.

3.6.3 Session Identification Process

To find all page references made by user, session identification process is used. These two methods are also called as "proactive" and "reactive" methods. The session considers over the duration of user request to a particular website. When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created. If the time between page requests exceeds a certain limit, it will assume that other user-session has started. This system takes 30 minutes threshold value. After preprocessing tasks, the useful log file used to identify the transaction.

3.6.4 Transaction Identification Process

According to the needs of the respective algorithms, transactions are applied to extract important information from the preprocessed data. The formatting of transactional data differs from the kind of algorithms that are used. The transactions are identified with serialization of numeric data and the related transactions are extracted.

3.7 RSA Association Rule Mining

The RSA is a computationally expensive algorithm because the frequent patterns discovered with the relative support measure. That is, although a pattern satisfies the user-defined minimum relative support threshold, all its non-empty subsets may not have satisfied the minimum relative support threshold. It also suffers from the same performance problems as the Apriori algorithm, which includes generating huge number of candidate patterns and multiple scans on the database [6].

3.7.1 RSA Algorithm

The Relative Support Apriori (RSA) algorithm is as follows [7]:Let I = {i1, i2, …, in} be a set of items, and DB be a database that consists of a set of transactions. Each transaction T contains a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let $X \subseteq I$ be a set of items, referred as an itemset or a pattern. A pattern that contains k items is a k-pattern. A transaction T is said to contain X if and only if $X \subseteq T$. The frequency (or support count) of a pattern X in DB, denoted as f(X), is the number of transactions in DB containing X. The support of X, denoted as S(X), is the ratio of its frequency to the DB size, i.e., S(X) = f(X)/|DB|. The relative support of a pattern X, denoted as Rsup(X), is the ratio of its support to the minimum support of an item (or 1-pattern) within it. That is, $Rsup(X) = S(X)/\min(S(ij)| \forall ij \in X)$. The pattern X is frequent if its support and relative support are no less than the user-defined minimum support (minsup) and minimum relative support (minRsup) thresholds. That is, X is said to be frequent if $S(X) \geq$ minsup and $Rsup(X) \geq$ minRsup.

3.8 Performance Measurement

To assess the performance of RSA algorithm, the measures of support and confidence are used [2].

3.8.1 Support

It measures the frequency of association, i.e. how many times the particular item has been occurred in a dataset.

$$Support = P(A \cap B) \quad (3.1)$$

P (A $\cap$ B) is equal to the number of transactions containing both A and B/Total number of transactions.

3.8.2 Confidence

Confidence basically measures the strength of the association rules. It is defined as the fraction of the transactions that include both A and B to the total number of records that contain A. It determines how frequently item B occurs in the transaction that contains A. Confidence expresses the conditional probability of an item.

$$\text{Confidence} = P(A \mid B)$$
$$= P(A \cap B) / P(A) \quad (3.2)$$

3.8.3 Predictive Accuracy

Predictive accuracy is also another way to measure interestingness of an association rule. The definition of predictive accuracy is: Let D be a data file with n number of records. If [a $\rightarrow$ b] is an Association Rule which is generated by a static process P then the predictive accuracy of [a $\rightarrow$ b] is c([a $\rightarrow$ b]) = Pn[n satisfies b |n satisfies a] where distribution of n is govern by the static process P and the Predictive Accuracy is the conditional probability of a $\rightarrow$ n and b $\rightarrow$ n.

4. IMPLEMENTATION

Step 1: This process accepts one hour NASA log data which has 1396 transactions with 150KB.

Table 4.1. NASA log file format for an hour

No	Log
1	www-c3.proxy.aol.com - - [01/Aug/1995:00:00:01 -0400] "GET / HTTP/1.0" 200 7280
2	picard.cistron.nl - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
3	mccaughey-thoma.eigenmann.indiana.edu - - [01/Aug/1995:00:00:07 -0400] "GET /ksc.html HTTP/1.0" 200 7280
	. .

Step 2: The log file format is a text-based format, the NASA log file also text- based. So, this file is changed into the useful format that is shown in Table 4.2.

Table 4.2. Parsing the data with relevant format

IP	rfc Name	Log Name	Time	Port	Method	address	http	Status code
www-c3.proxy.aol.com	-	-	01/Aug/1995:00:00:01	-0400	GET	/	HTTP/1.0	200
picard.cistron.nl	-	-	01/Aug/1995:00:00:07	-0400	GET	/	HTTP/1.0	304
mccaughey-thoma.eigenmann.indiana.edu	-	-	01/Aug/1995:00:00:07	-0400	GET	/ksc.html	HTTP/1.0	200
. .	. .	. .	. .	. .	. .	. .	. .	. .

Step 3: Some information from web log files are redundant and irrelevant. So the removing is necessary in data cleaning task. The cleaning firstly removes the unsuccessful records in which the status code field recognizes the unsuccessful http request. It removes the entries with status codes that are not 200. Removing these kinds of requests is necessary because they are just increasing the size of log file and nothing to do with analysis of user's navigational behavior. Apply the preprocessing on web log files and store them into the database. The preprocess log file is shown in Table 4.3 which has 1257 transactions.

Table 4.3. Preprocessed log file

IP	rfc Name	Log Name	Time	Port	Method	address	http	Status code	Byte
www-c3.proxy.aol.com	-	-	01/Aug/1995:00:00:01	-0400	GET	/	HTTP/1.0	200	7280
mccaughey-thomas.cigsmann.indiana.edu	-	-	01/Aug/1995:00:00:07	-0400	GET	/ksc.html	HTTP/1.0	200	7280
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Step 4: To find all page references made by user, session identification process is used.The session considers over the duration of user request to a particular website. When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created. If the time between page requests exceeds a certain limit, it will assume that other user-session has started. This system takes 30 minutes threshold value. After preprocessing tasks, the useful log file used to identify the transaction, 146 sessions transactions are shown in Figure 4.1.

/history/apollo/apollo.html, /history/apollo/images/footprint-small.gif, /images/KSC-logosmall.gif, /history/apollo/images/apollo-logo1.gif, /history/history.html, /history/apollo/images/apollo-small.gif, /images/NASA-logosmall.gif, /ksc.html, /images/ksclogo-medium.gif, /images/MOSAIC-logosmall.gif, /images/USA-logosmall.gif, /images/WORLD-logosmall.gif, /images/ksclogo.gif, /history/apollo/apollo-13/apollo-13.html, /history/apollo/apollo-13/apollo-13-patch-small.gif, /images/ksclogosmall.gif, /history/apollo/images/footprint-logo.gif, /facilities/lc39a.html, /images/kscmap-tiny.gif, /images/lc39a-logo.gif
/, /images/MOSAIC-logosmall.gif, /images/USA-logosmall.gif, /images/ksclogo-medium.gif, /images/WORLD-logosmall.gif, /whats-new.html, /images/whatsnew.gif
/software/winvn/winvn.gif, /software/winvn/wvsmall.gif, /images/USA-logosmall.gif, /images/WORLD-logosmall.gif
. .

Fig 4.1. Transactions of Sessions

Step 5: This process involves determining frequent patterns. Association rules are used for prediction of next event or discovery of associated event. In the web data set, the transaction consists of the number of URL visits by the client, to the web site. To find the associated pattern RSA Algorithm is used.

Count1

/images/ksclogo-medium.gif

/images/MOSAIC-logosmall.gif

/images/USA-logosmall.gif

/images/NASA-logosmall.gif

/images/WORLD-logosmall.gif

/images/KSC-logosmall.gif

Count2

/images/USA-logosmall.gif<==>/images/MOSAIC-logosmall.gif

/images/WORLD-logosmall.gif<==>/images/MOSAIC-logosmall.gif

/images/WORLD-logosmall.gif<==>/images/USA-logosmall.gif

Count3

/images/WORLD-logosmall.gif<==>/images/USA-logosmall.gif<==>/images/MOSAIC-logosmall.gif

Fig 4.2. Rules produced by implementing with RSA algorithm

Step 6: Finally, this access the performance of RSA algorithm.

Table 4.4. Correct rules with confidence

Frequent pattern			Confidence
{/images/USA-logosmall.gif}	→	{/images/MOSAIC-logosmall.gif}	92.86
{/images/MOSAIC-logosmall.gif}	→	{/images/USA-logosmall.gif}	100.0
{/images/WORLD-logosmall.gif}	→	{/images/MOSAIC-logosmall.gif}	92.86
{/images/MOSAIC-logosmall.gif}	→	{/images/WORLD-logosmall.gif}	100.0
{/images/WORLD-logosmall.gif}	→	{/images/USA-logosmall.gif}	100.0
{/images/USA-logosmall.gif}	→	{/images/WORLD-logosmall.gif}	100.0
{/images/WORLD-logosmall.gif}	→	{/images/USA-logosmall.gif, /images/MOSAIC-logosmall.gif}	92.86
{/images/USA-logosmall.gif, /images/WORLD-logosmall.gif}	→	{/images/MOSAIC-logosmall.gif}	92.86
{/images/USA-logosmall.gif}	→	{/images/WORLD-logosmall.gif, /images/MOSAIC-logosmall.gif}	92.86
{/images/MOSAIC-logosmall.gif, /images/WORLD-logosmall.gif}	→	{/images/USA-logosmall.gif}	100.0
{/images/MOSAIC-logosmall.gif, /images/USA-logosmall.gif}	→	{/images/WORLD-logosmall.gif}	100.0
{/images/MOSAIC-logosmall.gif}	→	{/images/WORLD-logosmall.gif, /images/USA-logosmall.gif}	100.0

This algorithm implemented an hour NASA data, there are 1396 transactions. The accuracy is tested on 12 corrected rules, the minimum support count is 0.28 and relative minimum support count is 0.65. The values of confidence are illustrated in Table 4.4 and the average of confidence is 97.02%.

5. CONCLUSION

Web Usage Mining is a great research area in discovering the interested patterns of user's usage data on the web. In this paper, implementation of a system is pattern discovery using association rules which introduce the process of web log mining, and show how to find frequent pattern from the web log data in order to obtain useful information about the user's navigation behavior when the user browses or makes transactions on the web site. Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on the threshold value called support, identifies the frequent item sets. In RSA algorithm, the frequent patterns discovered with the relative support measure. So, it is a computationally expensive. That is, although a pattern satisfies the user-defined minimum relative support threshold, all its non-empty subsets may not have satisfied the minimum relative support threshold. This algorithm implemented with one day of August of 59588 transactional NASA data with 6.09MB and the average confidence is 92.52% based on 180 rules. Moreover, 80976 transactions with 8.27MB for two days NASA data are also implemented and 92.33% confidence over 180 rules corrected. Using these web usages mining not only knows the knowledge pattern and usefulness but also understands of web usage mining processing. This approach of Association rule mining produces the interesting and frequent pattern. These include the crucial information which helps the website designers to improve their website usability.

ACKNOWLEDGEMENT

I would like to express my special thanks to **all my teachers** who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank **my parents** for their continuous support and encouragement throughout my whole life.

REFERANCES

[1] M. Santhanakumar and C. C. Columbus, "Web Usage Based Analysis of Web Pages Using RapidMiner," vol. 14, p. 10, 2015.

[2] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, "Mining association rules on significant rare

data using relative support", J. Syst. Softw., 67:181–191.

[3] Aarti Parekh, Anjali Patel, Sonal Parmar, Prof. Vaishali Patel, and Shri S'ad vidhya mandal institute of technology/Bharuch, "Web usage Mining:Frequent Pattern Generation using Association Rule Mining and Clustering," *Int. J. Eng. Res.*, vol. V4, no. 04, p. IJERTV4IS041467, Apr. 2015.

[4] P. Weichbroth, M. Owoc, and M. Pleszkun, "Web user navigation patterns discovery from WWW server log files," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 1171–1176.

[5] M. Srivastava, R. Garg, and P. K. Mishra, "Preprocessing Techniques in Web Usage Mining: A Survey," *Int. J. Comput. Appl.*, vol. 97, no. 18, pp. 1–9, Jul. 2014.

[6] R. U. Kiran and M. Kitsuregawa, "Towards Efûcient Discovery of Frequent Patterns with Relative Support," p. 8.

[7] H. Yun, D. Ha, B. Hwang, and K. Ho Ryu, "Mining association rules on significant rare data using relative support," *J. Syst. Softw.*, vol. 67, no. 3, pp. 181–191, Sep. 2003.

Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190035

CLASSIFYING AND ASSIGNMENT PROJECT USING MULTI CRITERIA DECISION AID APPROACH

Pa Pa Win [(1)], Thin Thin San [(2)], Seint Wint Thu [(3)]

[(1)(2)(3)]University of Computer Studies (Meiktila), Myanmar

[(1)]*papawin.pale@gmail.com*

ABSTRACT

Project manager selection is one of the most important decisions in project management. Competent project manager is one of the key factors for the success of project. Project manager selection due to special requirements is significantly important. A project manager must have the ability of managing costs, time and resources through the optimistic way. Furthermore, project manager has to own general management skills and benefits from adequate information about the project context. This paper presents an integrated model to support the process of classifying projects and selecting project managers for these projects in accordance with their characteristics and skills using a multiple criteria decision aid (MCDA) approach. Such criteria are often conflicting. The model also supports the process of allocating project managers to project by evaluating the characteristics/types of projects. As a result, it was possible to classify the projects and project managers into definable categories, thus enabling more effective management as different projects require different levels of skills and abilities.

KEYWORDS: *project manager, integrated model, multiple criteria decision aid*

1. INTRODUCTION

Project manager selection is one of the most important problems that organizations have to deal with. Project manager choices may cause failure of a routine project or conversely may cause an unbelievable success in projects with many unforeseen obstacles and problems. As organizations increasingly focus on human assets as a competitive advantage, they expect higher levels of performance from their employees. Schoonover et al. anticipate the use of competencies as a strategic intervention to continue, and even to accelerate firms' success. Competencies are behaviors that encompass the knowledge, skills, and attributes required for successful performance. In addition to intelligence and aptitude, the underlying characteristics of a person, such as traits, habits, motives, social roles, and self-image, as well as the environment around them, enable a person to deliver superior performance in a given job, role, or situation[1].

The project manager has specific accountability for achieving the entire defined project objectives within the time and resources allocated. The project manager performs the day-to-day management of the project. One or more assistant project managers with the same responsibilities over specific portions of the project may support the overall project manager, without diluting his or her responsibility. Project managers must demonstrate knowledge, skills and experience commensurate with the size, complexity and risk of the project. Since different levels of competency are required for various levels of project management and project size, the project manager role is divided into three proficiency levels. Depending on the size, complexity and risk of the project, more than one level of project manager may share responsibility for managing the project. The selection of project manager considers the concepts of the project in relation that roles characteristic. This concept contains the typical role of the project manager and links it to the skills that are required by an effective project manager. Interviewing related candidates is one of the techniques concerning human resource selection. There are many studies fulfilled in the literature, which are based on interviews, work samples, tests, assessment centers, job knowledge and personality tests in human resource management

and in the special case of project manager selection, also we could consider project management (PM) knowledge, social awareness, leadership abilities and stakeholder management as important criteria. But multi criteria decision making (MCDM) techniques were used by only few of them. Traditional personnel selection method used an experimental and statistical techniques approach. Searching for MCDM, fuzzy logic, and human resource, selection separately has a few results in research databases, but searching for the keywords together results in more researches. In this paper, we consider a number of criteria and related sub-criteria in order to match of project managers of petroleum and gas projects, and because of the potential importance of this industry in countries with huge resources of fossil fuels, these managers should have the essential competencies. The proposed criteria and sub-criteria were identified based on associated references and literature of project management involved gas and petroleum project management.

The most important competencies for a project manager are different in fields but these fields are common in many categories, these categories could contain even a manager's social behaviors or decision making in uncertainty situation. But in practice the essential feature that a project manager should own it genetically is to understand the actual weight of activities and make appropriate decision when several parameters combined each other simultaneously, regarded to the weights and impotency of each activates in the shortest time. Actually modeling and solving it with computer is not only impossible and there is not enough time and resource for everybody. Thus, this paper aims to address this gap by putting forward an integrated model to support the process of sorting projects into different categories and classifying project managers, taking their competences into account, by using a structured process with the support of Multi criteria Decision Aid - MCDA methods.

A MCDA methodology seems very appropriate for this study since it is by using this methodology that various aspects of a project can be evaluated simultaneously: the size, complexity, resources, and so on, of the project itself and then what experience, training and knowledge of the tools and techniques of project management, etc., potential project managers have. This paper contributes towards understanding and developing a classification of projects and project managers in the energy sector. In the review of the literature, no study was found on using multiple criteria to sort projects and classify project managers.

2. Multi-Criteria Decision Aid – MCDA

MCDA aims to assist decision-making against multiple criteria, which are often conflicting, by applying a set of structured techniques and methods. When choosing a multi-criteria method, consideration needs to be given to the context of the problem, the actors of the process, decision makers' preference structures and rationality. Moreover, MCDA methods can be distinguished from each other, according to Roy (1996), as there are four types of basic problematic: choice, ranking, sorting and description. This study aims to categorize projects and project managers. Thus it is a problem of classification which means it is a sorting problematic. This consists of identifying what aspects of decision making are causing problems, generating alternative solutions and subsequently distributing each alternative to a predefined category. These categories have some ordering implicit to the categories, relative to each other [2].

Classification methods can be distinguished into two categories. The first uses techniques based on questioning the decision maker (DM) directly and the second uses preference disaggregation classification methods. Several methods have been developed for this type of problem, such as ELECTRE TRI, PROMETHEE TRI and PROMSORT, these being characterized by relying on questioning the DM directly; and UTADIS and PAIRCLASS, which are about preference disaggregation. The PROMSORT method, a procedure based on Preference Ranking Organization Method for Enrichment Evaluations – PROMETHEE, is the one chosen to be used in this study and is described in the following section.

2.1. PROMETHEE sorting – PROMSORT

According to Araz & Ozkarahan the PROMSORT is an effective tool to assign the alternatives to the ordered categories. It provides reliable classification in terms of the preference relation between alternatives and valuable information to the decision maker about the weaknesses and strength of the alternatives and features of the categories[3]. It was chosen for this study since it is possible to guarantee the ordering of the alternatives also within classes unlike what happens with PROMETHE TRI and ELECTRE TRI [8][9].

According to Araz & Ozkarahan in sorting problems there are two ways to define 'a priori' categories: using alternative references or using the profile limits of the categories. There are also two ways to categorize the alternatives: in a nominal or ordinal way. In its nominal form, there is no sorting of the classes and this is called a nominal sorting problem. For the other form, the classes are sorted in the order of from best to worst and this is called an ordinal sorting problem. This study will focus on the problem of pre-sorted categorization.

PROMSORT allocates alternatives to predefined sorted categories. To designate an alternative **a** to a certain category, results are taken from a comparison of **a** with the profiles that define the limits of categories and with reference to the alternatives in different phases[3].

Araz & Ozkarahan note that: G is a set of criteria 1*g*, 2 *g*,..., *n g*{}() 1,2,, *G n*= …and B a set of profiles that distinguish limits K +1 categories B (B = {1,2, ..., n}) wherein *h b*represents the upper limit of category, *h C*the lower limit of category *hC*+1, h = 1,2, ..., k. Assume 2 1 *CC*> means that category 2 outranks 1, and the set of profiles { }( ) 12 ,,, *k B bbb*= … should have the property: [ ] [ ] [ ] 1 12 21 , ,, . *kk k k bP bbP bbP b*---…This property says that the categories should be ordered and distinguishable. Assuming this ranking is given from the most preferred to the least preferred, the following condition helps in obtaining orderly and distinct categories: ( ) ( ) 1 , 1,, 1, *j hj h jjh k g bg bp* + ∀∀ =…- ≥ + . Comparison between two profile limits 1 *hb* - and *h b*which distinguish categories 11, *hh hCCand C* - +, is defined using the PROMETHEE methodology. PROMSORT allocates alternative categories by following the three steps recommended by Araz & Ozkarahan: (1) determining an outranking relation using PROMETHEE I; (2) using the outranking relation to describe the alternatives in the categories, except in situations of incomparability and indifference; (3) finally designating alternatives based on a one to one comparison.

2.2. Classification and evaluation of projects

In the literature on project management, classifying projects has mainly been used to develop capability and has focused on (1) tailoring the management style to suit the project type or (2) prioritizing and selecting projects. This study focuses on the first of these with a view to providing a classification so as to choose the best management approach given that the projects to be evaluated have been previously selected.

In this area, an approach widely used is the NTCP Model that evaluates the **N**ovelty, **T**echnology, **C**omplexity and **P**ace of projects in order to classify them as set out in Shenhar & Dvir. Other studies have used this model, for example, Dvir et al. and Howell et al. However this approach is very focused on research and development projects (R&D) besides which the model does not use a multiple criteria method [4][5].

2.3. The fundamental skills and abilities of the project manager

A project manager must be familiar with, fully understand and apply the tools and techniques regarded as good practice in project management (Project Management Institute). According to the International Project Management Association "Competence is a collection of knowledge, attitudes, skills and relevant experience required for the successful exercise of a given function." Further, according to IPMA (International Project Management Association), the necessary skills for project management consist of technical skills, behavioral skills, and contextual skills.

Darrell et al. identified that in addition to technical skills, project managers must possess general management skills (namely, they must know how to delegate, how to lead, and how to draw up procedures), interpersonal skills (i.e. those to do with communication, conflict management, motivation) and skills in project management (specific knowledge in the project area about how to use specific tools and techniques). Also, the project manager must first have decision-making skills, communication, leadership and motivation and problem solving skills. The selection of a project manager is also a multiple criteria problem. Also the project manager must be evaluated in terms of experience and personal skills. In the context of project allocation the management skills and project assessment must be considered. As to skills in project management, other studies can be consulted include (Ahadzie et al), (Zhang et al) and (Bredin & Soderlund).

3. Integrated system to classify projects and project managers: a model proposed

Thus with a view to filling this gap in the literature, the model proposed sets out to enable projects and project managers to be classified

according to their characteristics and abilities. It is hoped that this proposal will be regarded as making a major contribution to the management of projects, both in academic terms and in organizations that can make use of a systematic process such as this [6]. The issue of classification seeks to assign alternatives (in this case, projects and project managers) to pre-existing categories in accordance with the evaluation of these alternatives on a set of criteria, which are determined by classes which have specified limits[4][5]. The methodology proposed in this research which ends with the classification of projects and project managers is. This model comprises two main phases from the time of launch of the project. The following steps will generate the information needed for classifying projects and project managers and can be performed in parallel. In multi criteria modeling, the first stage is to identify the decision maker (DM). He/she will interact with the analyst to establish the parameters of modeling and evaluate alternatives. The DM is essential to establish the company´s values and goals and doing so enables consistent decisions to be made. The stages of the model may then be followed always with the participation of the DM and his/her interaction with the analyst. The model is described in Fig 1.

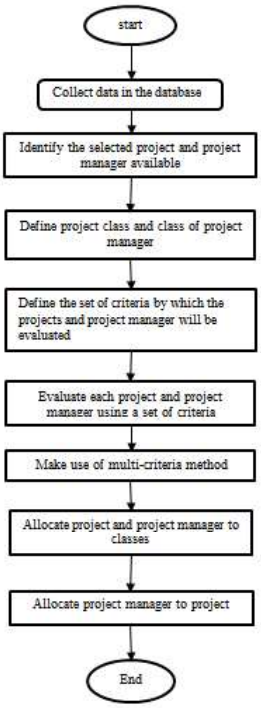

Fig 1. Overview of System Flow

4. Criteria for classifying projects and project manager

Table 1. Criteria for classifying projects

ID	Criteria	Weighting	Verbal Scale	Numerical Scale
g1	Complexity	0.2	High Medium Low	3 2 1
g2	Resources (Man-hours)	0.2	-	Man-hours
g3	Rate of growth	0.2	Urgent Critical Competitive Regular	4 3 2 1
g4	Contribution to achieving organizational strategy	0.3	-	%
g5	Technology	0.1	High Medium Low	3 2 1

Table 2. Criteria for classifying project manager

ID	Criteria	Weighting	Verbal Scale	Numerical Scale
g1	Experience as project manager	0.1	-	Years
g2	Level of training	0.5	Master Specialization Graduate	1 2 3

g3	Knows how to use techniques and tools of project management	0.5	-	Years
g4	Focus on needs of the client	0.1	-	%
g5	Ability to resolve problems	0.1	-	Ratio
g6	Maturity of project manager	0.15	-	n°
g7	Commitment	0.05	-	n°
g8	Size of previous projects	0.2	Large Medium Small	1 2 3
g9	Interoperability of previous projects	0.1	-	n°

The DM evaluated the projects in accordance with the criteria of Table 1 and Table 2 and this evaluation is presented in Table 3.

Table 3. Evaluation of project

ID	Application Area	g^1	g^2	g^3	g^4	g^5
1	Engineering and Construction	2	640	2	3	2
2	Organization and Business	3	480	2	2	1
3	Information System	3	640	1	3	2
4	Information System	1	720	2	2	2
5	Organization and Business	1	160	3	3	2
6	Organization and Business	1	160	2	2	1
7	Information System	2	640	3	3	2
8	Organization and Business	1	120	2	3	1
9	Engineering and Construction	2	3200	1	2	1
10	Information System	3	960	2	3	2
11	Organization and Business	1	160	2	2	2
12	Engineering and Construction	3	7680	4	3	3
13	Engineering and Construction	2	2880	1	2	1
14	Organization and Business	1	160	2	2	3
15	Organization and Business	2	320	2	2	1
16	Engineering and Construction	3	5760	4	3	1
17	Information System	2	160	1	2	1
18	Engineering and Construction	1	160	2	2	1
19	Information System	2	240	2	3	2
20	Engineering and Construction	3	2880	4	3	1

Table 4. Evaluation of the qualities required of the project managers.

ID	g^1	g^2	g^3	g^4	g^5	g^6	g^7	g^8	g^9
PMr 1	3years	2	3years	60.0 %	8	5	7	3	8
PMr 2	5years	3	5years	80.0 %	12	15	8	2	8
PMr 3	3years	2	3years	62.5 %	10	8	7	1	8
PMr 4	1years	1	1years	66.7 %	5	3	7	1	4
PMr 5	2years	3	2years	70.0 %	6	10	8	2	6
PMr 6	3years	2	3years	60.0 %	15	10	8	2	6
PMr 7	4years	2	4years	66.7 %	8	12	8	2	6
PMr 8	4years	2	4years	80.0 %	16	10	8	1	6

PMr9	4year s	2	4year s	100.000	3	2	9	1	3
PMr10	5year s	2	5year s	87.500	12	3	3	1	4
PMr11	2year s	2	2year s	100.000	4	2	3	1	4
PMr12	3year s	3	3year s	66.700	10	6	3	1	5
PMr13	1year s	1	1year s	66.700	5	3	7	1	5
PMr14	2year s	2	2year s	50.000	7	2	7	1	5
PMr15	1year s	1	1year s	66.700	14	6	9	1	5
PMr16	3year s	3	3year s	62.500	15	3	7	2	3

Table 5. Profile limits of the of the classes of the projects

Profile	g1	g2	g3	g4	g5
b1	2	1200	2	2	2
b2	1	600	1	1	1

Table 6. Profile limits of the project managers

Profile	g1	g2	g3	g4	g5	g6	g7	g8	g9
b1	2	2	2	80 %	10	10	12	2	8
b2	1	1	1	70 %	7	8	10	1	6

The profile limits of classes were defined as the parameters to be used for PROMSORT, and are presented in Table 5 for the projects and Table 6 for PMrs. It is important to emphasize that nay decision made is dependent on a given issue and the DM's view, both of which change from case to case.

Table 7. Results of classifying the projects

Classes	Projects
CLASS1– PROJECTS VERY CRITICAL	P44,P46,P47
CLASS 2– PROJECTS CRITICAL	P1,P2,P3,P4,P5,P6,P7,P8,P9,P10,P11,P12,P13,
	P14,P16,P17,P18,P19,P21,P23,P24,P25,P27,
	P28,P29,P30,P32,P33,P35,P36,P37,P38,P39,
	P40,P41,P42,P43,P47,P48,P49,
CLASS3– PROJECTS NON-CRITICAL	P12,P15,P20,P22,P26,P31,P32

Table 8. Results of classifying project managers

Classes	Project Managers
CLASS 1 – SENIOR MANAGER	PMr2, PMr7, PMr8
CLASS 2 – MIDDLE MANAGER	PMr1, PMr3, PMr5, PMr6, PMr9, PMr10, PMr12, PMr16,
CLASS 3 – JUNIOR MANAGER	PMr4, PMr11, PMr13, PMr14, PMr15

Reviewing the results obtained indicates that the PROMSORT projects marked P12, P16 and P20 is "very critical",

P1, P2, P3, P4, P5, P7, P8, P9, P10, P13, P14, P19, which were assigned to class 2 and marked "critical"

Projects P6, P11, P14, P15, P17, P18 were allocated to Class 3 project "non-critical".

According to the class of project, the projects assigned to suitable class of project manager. This show in Table 9.

Table 9. Assign to Project Manager based on class of project

ID	Class of Project	Assign to Project Manager
P12,P16,P20	Class1:Very Critical	Class1:Senior Manager (PMr2, PMr7,PMr8)
P1,P2,P3,P4,P5, P7,P8,P9,P10, P13, P19	Class2: Critical	Class2:Middle Manager (PMr1, PMr3, PMr5, PMr6, PMr9, PMr10, PMr12, PMr16)
P6,P11,P14, P15,P17,P18	Class3:Non-Critical	Class3:Junior Manager (PMr4,PMr11,PMr13, PMr14, PMr15)

5. CONCLUSIONS

This study put forward a model that offers systematic integrated support to the process of classifying projects and project managers and allocating PMrs and includes a multi-criteria analysis and a flexible process. The results of application showed the efficacy of the model and that the PMO is satisfied. By applying the model proposed, it was possible to classify projects and project managers into distinguishable categories, thus enabling them to be managed more effectively, as different projects require different levels of skills and abilities. Using the model put forward in this study also enables PMrs to be chosen more carefully, thus assisting an organization to allocate its most critical projects to the best prepared professionals, especially when the organization is developing multiple projects simultaneously. Future studies should be undertaken in order to ensure a formal approach to allocating PMrs is being followed, as well as to explore other different methods for assessment, for example, by considering how a group decision might best be taken.

ACKNOWLEDMENT

I would like to express my special thanks to **all my teachers** who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank **my parents** for their continuous support and encouragement throughout my whole life.

REFERENCES

[1] D. Ahadzie, D. Proverbs, and I. Sarkodie-Poku, "Competencies required of project managers at the design phase of mass house building projects," *Int. J. Proj. Manag.*, vol. 32, Jan. 2013.

[2] L. H. Alencar and A. T. de Almeida, "A model for selecting project team members using multicriteria group decision making," *Pesqui. Oper.*, vol. 30, no. 1, pp. 221–236, Apr. 2010.

[3] E. C. B. de Oliveira, L. H. Alencar, and A. P. C. S. Costa, "A decision model for energy companies that sorts projects, classifies the project manager and recommends the final match between project and project manager," *Production*, vol. 26, no. 1, pp. 91–104, Nov. 2015.

[4] "Projects and Project Managers: The Relationship between Project Managers' Personality, Project Types, and Project Success - Dov Dvir, Arik Sadeh, Ayala Malach-Pines, 2006." [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/875697280603700505. [Accessed: 10-Dec-2019].

[5] "Overzicht IPMA C en IPMA D cursussen," *ICM opleidingen & trainingen*. [Online]. Available: https://www.icm.nl/opleidingen-en-trainingen/proces-projectmanagement/overzicht-ipma/. [Accessed: 10-Dec-2019].

[6] "An automated procedure for selecting project managers in construction firms: Journal of Civil Engineering and Management: Vol 19, No 1." [Online]. Available: https://www.tandfonline.com/doi/abs/10.3846/13923730.2012.738707. [Accessed: 10-Dec-2019].

[7] "'A Decision Support Model for Project Manager Assignments' by Peerasit Patanakul, Dragan Miloševi et al." [Online]. Available:

https://pdx scholar.library.pdx.edu/etm_fac/22/. [Accessed: 10-Dec-2019].

[8] "Multicriteria Methodology for Decision Aiding | Bernard Roy | Springer." [Online]. Available: https://www.springer.com/gp/book/9780792341666. [Accessed: 10-Dec-2019].

[9] "Linguistic Extension for Group Multicriteria Project Manager Selection." [Online]. Available: https://www.hindawi.com/journals/jam/2014/570398/. [Accessed: 10-Dec-2019].

Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190068

COMPARATIVE ANALYSIS OF FCFS, SJF SCHEDULING ALGORITHMS BASED ON SIMILAR PRIORITY JOBS

Yin Lin Thu [(1)], May Zin Htun[(2)], Su Myat Sandar Win[(3)]

[(1)(2)(3)]University of Computer Studies (Meiktila), Myanmar

[(1)]*yinlinthu.ucsmtla@gmail.com*

ABSTRACT

One of the primary roles of the operating system is job scheduling. Scheduling is one of the most important activities of the process which takes decision to choose which of the process in the ready queue will be assigned to the CPU. There are different types of scheduling algorithms available for taking decision. One of them is priority scheduling algorithm, which is based on the priority assigned to each process. In priority scheduling, the processes are executed on the basic of priority and the process having highest priority is executed first. In this paper, the priority scheduling algorithm is used in such a way and comparison of Shortest Job First (SJF) scheduling algorithm and First Come First Served (FCFS) scheduling algorithm based on similar priority is calculated and then the average waiting time and average turnaround time is also calculated. The comparison analysis is performed on the SJF based priority scheduling and FCFS based priority scheduling to compare the average waiting time and average turnaround time.

KEYWORDS: *SJF, FCFS, Scheduling, Priority Scheduling and Similar Priority Jobs*

1. INTRODUCTION

In a single-processor system, only one process can run at a time; any others must wait until the CPU is free and can be rescheduled. The objective of multiprogramming is to have some process running at all times, to maximize CPU utilization. Scheduling is a fundamental operating-system function. For the system of single processor, multiple process come and then one process can be executed at a time and other process remain in waiting state until the CPU becomes idle or can be scheduled again. To expand the CPU usage, the goal of multiprogramming is to have some procedures running at all times. CPU scheduling manages the issue of choosing which of the procedures in the ready queue is to be assigned the CPU [1].

Different number of algorithms is used to schedule process such as First Come First Served (FCFS), Shortest Job First (SJF), round robin and priority scheduling algorithm. FCFS selects the process that has been waiting the longest for service. The SJF selects the process with the shortest expected processing time, and do not preempt the process. In round robin scheduling, uses time slicing to limit any running process to a burst short of processor time, and rotate among all ready processes. The CPU is preempted, if a process does not complete before its CPU time-expires and given to the next process waiting in a queue [2].

2. Review of related Literature

In [3] *Nazleeni SamihaHaron ,et.al*. has analyzed in any distributed systems, process scheduling plays a vital role in determining the efficiency of the system. The comparative study was done based on the Average Waiting Time and Average Turnaround Time of the processes involved. The general need of CPU usage builds the interest to enhance the CPU time. An overview has been made to analyze the different scheduling algorithm and to enhance them.

E.O. Oyetynji and A.EOluleye, "Performance Assessment of Some CPU Scheduling Algorithm", 2009 [4], attempted to compare the different scheduling algorithms on the basis of waiting time and turnaround time. This paper gives a brief overview to the problem of scheduling jobs/processes on the central processing unit (CPU) of the computer system. *Jyotirmay Patel1 and A.K. Solanki*, "CPU Scheduling: A Comparative Study",[5], discuss about scheduling policies of Central processing unit (CPU) for computer system. A number of problems were solved to find the appropriate among them. Therefore, based on performance, the shortest job

first (SJF) algorithm is suggested for the CPU scheduling problems to decrease either the average waiting time or average turnaround time. Also, the first-come-first served (FCFS) algorithm is suggested for the CPU scheduling problems to reduce either the average CPU utilization or average throughput.

TaqwaFlayyihHasan, "CPU scheduling Visualization",[6], evaluated the different number of algorithm to analyze the average waiting time and average turnaround time. The results show that, in FCFS recommended for the CPU scheduling problems of minimizing either the average CPU utilization or average throughput. In RR algorithm, selecting of time quantum is the major problem. In round robin scheduling algorithm average waiting time is often quite long.

3. SCHEDULING CRITERIA

Different CPU scheduling algorithms have different properties, and the choice of a particular algorithm may favor one class of processes over another. In choosing which algorithm to use in a particular situation, we must consider the properties of the various algorithms. Many criteria have been suggested for comparing CPU scheduling algorithms. The main goal of CPU scheduling algorithms is to utilize the resources effectively and efficiently. It can be accomplished by CPU busy as much as possible. And the number of processes in the job queue must be maximized. It is called the throughput**.** It is the task of operating system is to provide the fair time of CPU to the each process in the ready queue. By this, each process participates in the execution of the CPU time. The characteristics that are used for comparison can make a substantial difference in which algorithm is judged to be best. The criteria include the following:

1. CPU Utilization: It keeps the CPU as busy as possible. It must have maximum value.

2. Throughput: The number of processes that complete their execution per time unit It must have maximum value.

3. Turnaround time: The amount of time to execute a particular process. It must have minimum value.

4. Waiting time: The amount of time a process has been waiting in the ready queue. It must have minimum value.

5. Response Time: The amount of time it takes from when a request was submitted until the first response is produced, not output (for time-sharing environment). It must have minimum value.

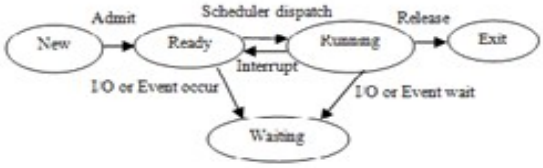

Fig 1. Life cycle of Process

3.1 First Come First Served (FCFS)1: This algorithm allocates the CPU to the process that requests the CPU first. This algorithm is easily managed with a FIFO queue. New process enters the queue through the tail of the queue and leaves through the head of the queue. A process does not give up CPU until it either terminates or performs s I/O.

3.2 Shortest Job First (SJF): The SJF algorithm may be implemented as either a preemptive non-preemptive algorithms. When the execution of a process that is currently running is interrupted in order to give the CPU to a new process with a shorter next CPU burst, it is called a preemptive SJF. SJF will allow the currently running process to finish its CPU burst before a new process is allocated to the CPU.

3.3 Round Robin (RR): It is often used in time sharing system. RR is similar to FCFS except that preemption is added to processes. In this algorithm, a time slice of 3ms has been taken. After the time slice is expired, executing process will leave the CPU free and allocate the CPU to the next process in the ready queue.

1.4 Priority Algorithm: A priority number is associated with each process. The CPU is allocated to the process with the highest priority .the smaller number is generally used for the highest priority. It runs the highest priority algorithms first. The disadvantage of the Priority Based Scheduling is that it may cause low-priority processes to starve.

2. FIRST COME FIRST SERVED (FCFS) BASED PRIORITY SCHEDULING ALGORITHM

In the FCFS algorithm the processes are executed according to priority, such that the process having highest priority will execute first. On the basis of execution of each process, the waiting time and

turnaround time is calculated but in the case of similar priority FCFS in which two or more processes have similar priority then the process which comes first executed first, algorithm for priority scheduling are as follow:

First Come First Served (FCFS) Algorithm:

Step 1: Assign the process the ready queue.

Step 2: Assign the process to the CPU according to the priority, higher priority process will get the CPU first than lower priority process

Step 3: If two processes have similar priority then the FCFS is used to break the tie.

Step 4: Repeat the step 1 to step 3 until the ready queue is empty.

Step 5: Calculate waiting time and turnaround time of individual process.

Step 6: Calculate the average waiting time and average turnaround time

3. PROBLEM STATEMENTS

One of the most important problems in operating systems designing is CPU scheduling and challenge in this field is to build a program to achieve proper scheduling. In case of priority scheduling algorithm, FCFS that arrives the similar priority jobs is used and the average waiting and turnaround time relatively higher. The process that arrives first is executed first, no matter how long it takes the CPU. So, if long burst time processes execute earlier then other process will remain in waiting queue for a long time. This type of arrangement of processes in the ready queue results in the higher average waiting time and turnaround time. The objective of this paper is:

To analyze of SJF priority based and FCFS priority based scheduling algorithm.

- To reduce the average waiting time and average turnaround time of CPU.

4. SHORTEST JOB FIRST (SJF) BASED PRIORITY SCHEDULING ALGORITHM

In SJF scheduling algorithm, SJF based priority based scheduling algorithm is used in which each process that have similar priority is executed on the basis of burst time, i.e. the process which have least burst time will execute first. The SJF based priority algorithm results in reduced average waiting time and turnaround time.

Shortest Job First (SJF) Algorithm:

Step 1: Assign the job to ready queue.

Step 2: Assign the job to the CPU according to the priority, higher priority job will get the CPU first than lower priority job.

Step 3: If two jobs have similar priority then SJF is used to break the tie.

Step 4: Repeat the step 1 to 3 until ready queue is empty.

Step 5: Calculate the waiting time and turnaround time of individual process.

Step 6: Calculate the average waiting time and average turnaround time.

5. IMPLEMENTATION

Different cases are used to implement the algorithm. The cases have number of processes along with the burst time and priority.

To calculate the waiting time and turnaround time, we use the following formula [7]:

Turnaround time = Process completion time - Arrival Time

Waiting time = Turnaround time - Burst Time

Case1: For 10 processes

Table 1 . INPUT VALUE

Process Name	Burst Time	Priorities
P1	40	2
P2	120	5
P3	180	7
P4	26	1
P5	150	6
P6	70	3
P7	15	1
P8	50	2
P9	90	4
P10	100	5

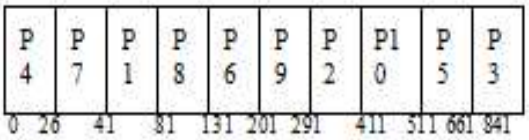

Fig 2. Expected Gantt chart for the above job (For Similar Priorities)

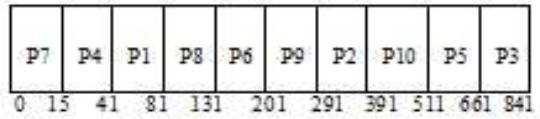

Fig 3. Expected Gantt chart for the above job (For Similar Priorities)

8. RESULT AND ANALYSIS

In this paper, an algorithm is proposed in which SJF is used to schedule the similar priority jobs. For this different cases were used through which we can easily compare the result of FCFS based priority scheduling algorithm and SJF based priority scheduling algorithm. In table 2, On the basis of the above outcome for FCFS and SJF based priority scheduling algorithm, the average waiting time and average turnaround time is calculated which is shown in graph. Turnaround time is the time that includes the actual execution time plus time spent waiting for resources, including the processor.

Table 2. COMPARISON OF WAITING TIME AND TURNAROUND TIME FOR 10 PROCESSES

Process	Waiting Time (ms)		Turnaround Time (ms)	
	FCFS based Priority Scheduling Algorithm	SJF based Priority Scheduling Algorithm	FCFS based Priority Scheduling Algorithm	SJF based Priority Scheduling Algorithm
P1	41	41	291	81
P2	291	391	411	511
P3	661	661	841	841
P4	0	15	26	41
P5	511	511	661	661
P6	131	131	201	201
P7	26	0	41	15
P8	81	81	131	131
P9	201	201	291	291
P10	411	291	511	391
Avera-ge	235.4	232.4	319.5	316.4

In First Come First Served (FCFS) algorithm, the process that has been in the ready queue the longest is selected for running for the similar priority jobs. In Shortest Job First (SJF) scheduling algorithm, the process with the shortest expected burst time is selected next for the similar priority jobs. In SJF, P4 and P7 have the same priority. P7 is selected because P7 is shortest burst time than P4. In FCFS, P4 is already waiting in the ready queue. So, the average turnaround time of SJF scheduling algorithm is the minimum value than the FCFS because SJF always choose the shortest job for the same priority.

8.1 Comparison of average waiting time and average turnaround time for 10 processes

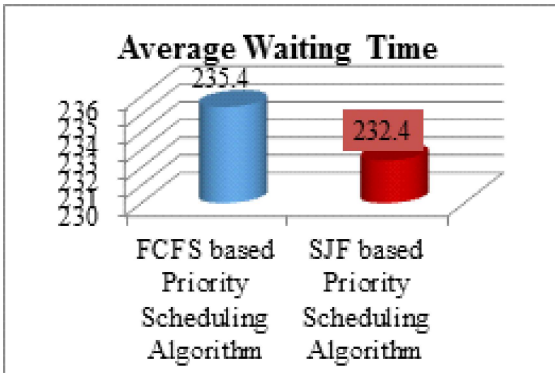

Fig 4. Average Waiting Time of FCFS based and SJF based Priority Scheduling Algorithm

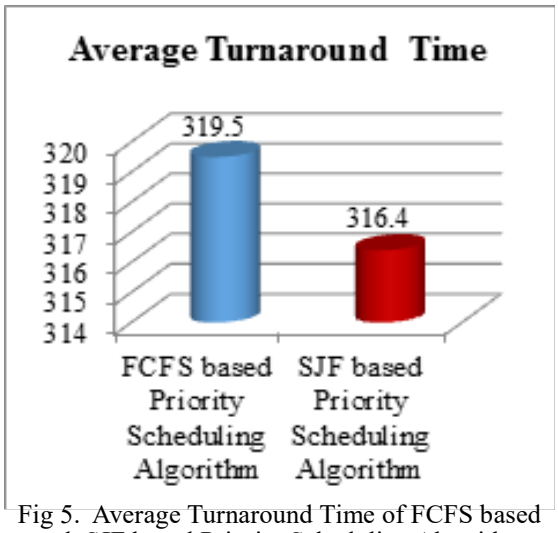

Fig 5. Average Turnaround Time of FCFS based and SJF based Priority Scheduling Algorithm

9. CONCLUSIONS

The operating system is crucial for ensuring the smooth and efficient operation of any computing device. There are many scheduling algorithms having their own benefits and drawbacks. Scheduling can also be done on the basis of priority. Each process assigned a priority and the process which has highest priority will be executed first. In case of similar priority generally FCFS is used to select the next process. If SJF based priority scheduling priority is used when two or more processes having similar priority, instead of FCFS, then the average waiting time and average turnaround time is reduced. We SJF based priority scheduling algorithm in which, the process that having lowest burst time will execute first. The FCFS based priority scheduling algorithm and SJF based priority scheduling algorithm is analyzed and the result shows that the average waiting time and average turnaround time is reduced.

ACKNOWLEDGEMENT

I wish like to express my cordial thanks to my supervisor , Prof. Dr. Thida Win, Faculty of Computer Science, University of Computer Studies (Meiktila), for her valuable advices and suggestion so that I could complete my work comfortable.

REFERENCES

[1] "Operating System Concepts, 8th Edition [Book]." [Online]. Available: https:// www.oreilly.com/library/ view/operating-system-concepts/9780470128725/. [Accessed: 10-Dec-2019].

[2] F. Sabrina, C. D. Nguyen, S. Jha, D. Platt, and F. Safaei, "Processing Resource Scheduling in Programmable Networks," *Comput Commun*, vol. 28, no. 6, pp. 676–687, Apr. 2005.

[3] T. Gabba and V. Singh, "COMPARATIVE STUDY OF PROCESSES SCHEDULING ALGORITHMS USING SIMULATOR," *Int. J. Comput. Bus. Res.*, vol. 4, May 2013.

[4] M. N. Sarisakal, "DESIGN OF A SCHEDULER: COMPARISON OF DIFFERENT SCHEDULING ALGORITHMS," p. 19.

[5] "Single-chip microprocessor that communicates directly using light | Nature." [Online]. Available: https://www.nature.com/articles/ nature16454. [Accessed: 10-Dec-2019].

[6] "Full text of 'Modern Operating System (3th Edition) Tanenbaum.'" [Online]. Available: https:// archive.org/stream/ ModernOperatingSystem3thEditionTanenbaum/ Modern%20Operating%20System%20%283th%20Edition%29%20-%20Tanenbaum_djvu.txt. [Accessed: 10-Dec-2019].

[7] "(PDF) A Comparative Study of CPU Scheduling Algorithms." [Online]. Available: https://www. researchgate.net/publication/ 249645533_A_ Comparative_Study_ of_CPU_Scheduling_Algorithms. [Accessed: 10-Dec-2019].

[8] "A CPU scheduling algorithm simulator - IEEE Conference Publication." [Online]. Available: https://ieeexplore.ieee.org/document/4417885. [Accessed: 10-Dec-2019].

[9] W. Stallings, *Operating Systems (3rd Ed.): Internals and Design Principles*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1998.

Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190075

THE ROLE OF DECISION SUPPORT SYSTEM (DSS) IN DECISION MAKING PROCESS FOR REAL-WORLD MANAGERS

Seint Wint Thu[(1)], Pa Pa Win[(2)], Thin Thin San[(3)]

[(1)(2)(3)]University of Computer Studies (Meiktila), Myanmar

[(1)]*seintwint241@gmail.com*

ABSTRACT

In today's business world, an organization coordinates run through its business processes that square logically connected tasks and behaviors for accomplishing work. The aim of the system is to develop decision support system (DSS) in creating a decision for a manager to boost the choice skills within the real-world. Managers play key roles in organizations. Their responsibilities vary from creating decision and selections, to writing reports, to attending conferences. Management's job is to form sense out of the various things featured by organizations. Managers built decisions, and formulate action plans to resolve structure issues. A decision support system (DSS) is interactive software-based system. A decision support system (DSS) helps managers in decision-making simply by accessing massive volume of data collected in structure business processes. DSSs serve the management, operations and designing levels of a corporation. DSS may be a tool to facilitate structure business higher cognitive processes. DSS mirror the hopes, dreams, and realities of real-world managers.

KEYWORDS: *Decision support system, Software-based system, Business process, Decision making process*

1. INTRODUCTION

A decision support system or DSS could be a computer based mostly system. A DSS serves to facilitate the answer of unstructured issues by a specific manager or typically a gaggle of managers operation along within the same location or in several locations. DSS uses the outline information, exceptions, patterns, and trends victimization analytical model. DSS provides tools and technology expressly toward deciding. A DSS provides the particular want of the individual and cluster managers. Therefore, the DSS will extend this support through the remaining steps (in objective and criteria setting, various search, various evaluation, creating call the choice and decision review) of the choice creating. Finally, DSS has additional roles in decision-making processes. A manager has to have data management. Data management is that the set of business processes to form, store, transfer, and apply data. The management's job will increase the flexibility of organization by incorporating data into its business processes. All the data from any organization is drawn within the style of charts and graphs. So, DSSs facilitate the managers for taking strategic deciding process.

This paper is organized as follows: Section 2 describes about some related works. Section 3 describes about Decision Making Process. Section 4 describes about Decision Support System (DSS). Section 5 introduces a DSS application for the proposed system. Section 6 describes the role of the DSS in the process of decision making. Finally, it concludes the paper in Section 7 and then describes acknowledgement.

2. RELATED WORKS

Uma (2009) has expressed that a choice network is associate integrated set of computer tools permitting a decision maker to act directly with computer to retrieve information helpful in creating

semi structured and unstructured choice [5]. This paper illustrates the process of evaluating the performance of the employees. A manager manages the employees of a company or organization by adjusting to the standards set by the company. Yoon and Hwang [15] first introduced the multiple criteria decision making methods for the best employee selection. This paper presented Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method in decision making to improve the work quality of employees.

3. DECISION MAKING PROCESS

A manager plans, organizes, staffs, leads, and controls her/his team by execution selection. Creating a choice may be a multistep method [2]. There are totally different stages in call making: intelligence, design, and choice. The intelligence stage discovers the matter. The design stage discovers the potential answer and the choice stage chooses the most effective answer. The stages within the decision making process is shown in Figure 1.

Fig 1. Stages in decision making process

The different levels in a company (strategic, management, operational) have totally different deciding needs requirements. Deciding may be performed by individual or teams of managers. Deciding includes workers further as operational, middle, and senior managers. The effectiveness and quality of this selection verify however winning a manager is going to be. Managers are perpetually known as upon to create selection so as to unravel issues [3]. Decision making and downside resolution are current processes of evaluating things or issues. Then, decision making process considers alternatives, makes choices, and follows them up with the required actions. In other situations, the decision process will drag on for weeks or perhaps months [5]. The entire decision making process is dependent upon the right information being available to the right people at the right times. The decision making process begins when a manager identifies the real problem [7].

The accurate definition of the problem affects all the steps that follow. If the problem is inaccurately defined, every step in the decision making process will be based on an incorrect starting point [10]. One way that a manager can help to determine the true problem in a situation is by identifying the problem.

4. DECISION SUPPORT SYSTEM (DSS)

A decision support system or DSS may be a computer based mostly system. DSS is employed by a specific manager or typically a bunch of managers at any structural level in decision making process [6]. Multiple decision makers are working together as a group to reach solutions. In this particular situation, the term GDSS, or a group decision support system is used. GDSS helps folk operating along in very reach choices additional expeditiously. The decision makers represent a committee or a project team [9].

The cluster members communicate with each another both, directly and by means that of the group ware. Because the DSS idea was broadened to supply support to two or addition decision maker working together as a team or committee, the idea of special group oriented software or groupware, became a reality [10]. A Decision Support System (DSS) is associated in nursing an interactive, flexible, and adaptable computer based mostly information system. Using DSSs in structured decision-making tasks allows users to know an outside rang of parameters and relationships [8].

DSS utilizes decision rules, models, and model base as well as a comprehensive database and therefore the decision maker's own insights. A DSS is resulting in specific, implementable decisions in determination issues that will not be amenable to management science models. Decision support system to support decision making do not always produce better manager. Management decisions improve firm performance. Thus, a DSS supports advanced higher cognitive process and will increase its effectiveness [11].

5. A DSS APPLICATION

In the proposed system, a manager uses DSS application in creating a decision. The decision making process for a manager is illustrated in Figure 2. A manager gathers data and information and stores these data in the database. DSS application interacts with the database, the user interface and other computer-based systems. A DSS application can be composed of following subsystems [13]:

1. Data Management subsystem: The database management subsystem includes a database that contains relevant information for the matters and is managed by software system referred to as the database management system (DBMS). The database management subsystem may be interconnected with the company information warehouse, a repository for company relevant decision-making information.

2. Model Management subsystem: The model base provides decision makers access to a spread of models and assist them in higher cognitive process. The model base will embrace the model base management software (MBMS). MBMS coordinates the use of models in an exceedingly DSS. External storage of data connected to the current part.

3. Knowledge-based Management subsystem: This subsystem will support any of the other subsystem or act as an independent component. It provides intelligence to enhance the decision maker's own. It is interconnected with the organization's knowledge repository that is named the organizational knowledge base.

4. User Interface subsystem: The user interface referred to as the dialog management facility. It permit users to move with the DSS to get data. The user interface requires two capabilities:

(1) The action language that tells the DSS what is needed and passes the information to the DSS and

(2) The presentation language that transfers and presents the user results.

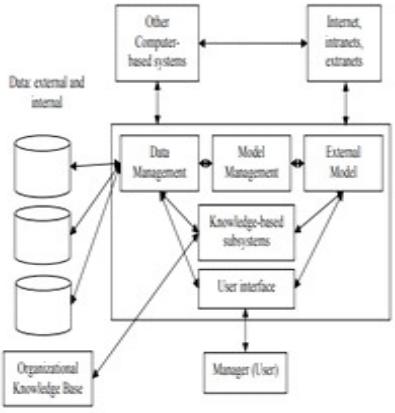

Fig 2. The decision making process for a manager

5.1 DSS characteristics

The key purpose of decision support system is to gather the information when [5]. The decision support system has a range of characteristics [8], that embody following:

1. Data collection-The system should collect information. A well-known example of a decision support system is daily company report. This is a system that collects information about the market in ecommerce. They use a spread of devices to gather sale measurements, like the sale [5]. If a decision support system is to be helpful, it should initial collect information that is relevant.

2. Data Management-Data management provides access to hold on information by the users or managers. Information is holding on anyplace. For instance, the information of sales of a corporation must be stored. It is stored by analyzing the information. There are some kinds of database or data log. A decision support system provides decision making process for a manager to manage the data that it collects.

3. Data Analysis-Raw data is rarely useful. The data may make big decisions. Therefore, analysis makes a world of difference when it comes to decision support.

4. Data Presentation-Data is presented and distributed to people. It is the interface and interaction between data and user. This is the user interface, look and feel, column graph or pie chart.

5.2 DSS capabilities

A decision support system (DSS) has the capabilities to create a decision for a manager. The DSS capabilities area unit as follows:

1. DSS will support for problem-solving phases as well as the intelligence, design, choice, implementation and observance.

2. DSS will support for different decision or choice frequencies.

3. DSS will handle one-of-a-kind decisions or choices.

4. Institutional DSS will handle repetitive decisions or choices.

5. DSS will support for various downside structures starting from high structured and programmed to unstructured and non-programmed.

6. DSS will support for numerous decision-making levels as well as operational-level decisions, tactical-level decisions and strategic decisions [8].

6. THE ROLE OF THE DSS IN THE PROCESS OF DECISION MAKING

A properly designed DSS is associate interactive software-based system. DSS is meant to assist decision makers compile helpful information from a mix of data, documents, and private data, or business model to spot and solve issues and create decisions [10]. DSS is extensively employed in business and management processes. A DSS will handle great deal of knowledge for database management package. DSS permit decision makers to go looking database for information. A DSS may solve issues wherever a little quantity of needed information. DSSs will facilitate managers to make attractive, informative graphical presentations displays on computer screens and on written documents. DSSs perform elect cognitive decision making functions. DSSs are based on artificial intelligence [13]. Information technology provides new tools for managers to hold out both traditional and newer roles to reply earlier to the dynamical business environment. Information systems facilitate to managers by supporting their roles in distributive data. Managers use decision support systems with powerful analytics and modeling tools. Analytic and modeling tools embody spreadsheets and pivot tables [12].Managers altogether levels of organization hierarchy want precise and appropriate data and information to create decisions that increase structure performance.

6.1 The managerial roles of a manager in decision making

Managers are able to influence on the things like business process, customer satisfaction, and employee training. There are five classical functions of a manager area unit coming up with, planning, organizing, coordinating, deciding and controlling. Most of managers are intensively used information systems in business processes. Managers act because the nerve centers of their organizations. They receive the foremost concrete, and up-to-date information [11]. Managers create in decisions. In their decisional role, they discuss conflicts and mediate between conflicting teams in organization. They analyze regular behaviors. So, information systems additionally facilitate for managers in their decisional roles [14]. The managerial behavior of a manager has five attributes [10]. These attributes are as follows:

First, managers perform a good deal of labor at associate degree unrelenting pace.

Second, managerial activities are fragmented; most activities last for fewer than nine minutes, and only 10 percent of the activities exceed one hour in duration.

Third, managers like current, specific, and ad hoc information.

Fourth, they like oral sorts of communication to written forms as a result of oral media give larger flexibility, require less effort, and bring a quicker response.

Fifth, managers provide high priority to maintaining a various and sophisticated web of contacts that acts as an information system.

There are many different kinds of employees within the company or the organization. Employees are very dominated to increase productivity and to achieve the company's target. The companies and economies are successful when employees are respected by the organization. Employees should have a level of responsibility and reward that reflects their skills. So, the managerial role of a manager is to select the best employee. A decision support system can help to facilitate the manager in decision making to determine the best choice based on standard criteria. A manager finds the best employee

based on predetermined criteria. Making a choice or decision in the proposed system uses Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method to perform the calculation on selection of best employee [15]. This method can choose the best alternative from a number of alternatives based on the criteria specified. The criteria can change because its weight value is dynamic with the desired user. Then, the ranking process will determine the best employee.

6.2 Methodology

TOPSIS is one of the methods of decision making system. TOPSIS uses the principle of the selected alternatives that have the shorted distance from the positive ideal solution and the farthest from the negative ideal solution from a geometrical point by using the Euclidean distance. It determines the relative proximity of an alternative to the optimal solution. This method takes the relative proximity to the positive ideal solution. Alternative priority order can be achieved based on the comparison of the relative distance [14]. This method can measure the relative performance of the alternatives decision. This method can facilitate to select the best employee in making a decision or choice [12]. The steps to calculate the TOPSIS method are as follows:

1. Make a decision matrix is normalized.

$$r_{ij} = X_{ij} / \sum_{i=1}^{m} X_{ij}2 \qquad (1)$$

2. Normalized weighted.

With the weight w j = (w1, w2, w3,..., wn), where w j is the weight of the criteria for all j and $\sum j = 1$ wj= 1,

The normalization of weight matrix V, is

$$v_{ij} = w_j * r_{ij} \qquad (2)$$

3. Determining the ideal solution matrix of positive and negative ideal solution by using this formula: v_{ij}

$$A+ = \{(\max v_{ij} \mid j \in J), (\min v_{ij} \mid j \in J) \qquad (3)$$

i = 1,2,3,…
m = { V1+,V2 +,V3 +,…,Vn +}

$$A- = \{(\min vij \mid j \in J), (\max vij \mid j \in J) \qquad (4)$$

i = 1,2,3,…
m = { V1-,V2-,V3-,…,Vn-}

4. Calculating separation

(a) S+ is an alternative distance from the positive ideal solution is defined as:

$$S_i^+ = \sqrt{\sum_{j-1}^{n} (v_{ij} - v_j)^2} \qquad (5)$$

Where i= 1, 2, 3, ..., m

(b) S- is an alternative distance from the negative ideal solution is defined as:

$$S_i^- = \sqrt{\sum_{j-1}^{n} (v_{ij} - v_j)^2} \qquad (6)$$

5. Calculating positive ideal solution with this function: Ci+ = si/(si – +si+) (7)

6. Alternative rank.

Alternative 5Ø6Ü+sorted from largest value to the smallest value. Alternative with the largest value of C+ is the best solution.

Experiment is performed by using a few criteria. The data of criteria is shown in Table 1.

Table 1. Criteria TOPSIS

Criteria
Job Responsibilites
Work Discipline
Work Quality
Behavior

The criteria are ranged with 1: very bad, 2: bad, 3: very good, 4: good ranges. The criteria in Table 1 are determined to gain the value of weight by using weight metrix. The alternative data is used. The value of each alternative is determined by using the TOPSIS formula, as shown in Table 2.

Table 2. Values of Each Alternative

No	Name	Alter native	Job Responsi bilities	Work Discip line	Work Quality	Be havior
1.	Sanny	A1	4	3	4	4
2.	Danial	A2	5	4	3	3
3.	Amy	A3	3	4	5	4
4.	Elisa	A4	4	4	3	3
5.	Hybris	A5	5	4	5	4

Finally, the values of each alternative are determined by applying the TOPSIS method of decision support system to select the best employee. The TOPSIS result is shown in Table 3.

Table 3. TOPSIS Result

No	Code	Name	Ranking
1.	0005	Hybris	1
2.	0002	Danial	2
3.	0001	Sanny	3
4.	0003	Amy	4
5.	0004	Elisa	5

According to the TOPSIS calculation process, the name of Hybris is the best employee. The given result is not a final decision but only gives recommendation to the manager to make better decision.

7. CONCLUSIONS

Nowadays, most firms and economies are successful due to the great management of a manager. Except kind of system in business world, decision support system (DSS) is either totally computerized or human-powered, or a mix of each. Therefore, DSS users see that DSS could be a tool to facilitate decision making and structure business processes. DSSs are being employed to assist managers in simply creating a decision with customized information. So it is over that DSS will extend its support to the identical steps of decision making process. DSS has additional roles in decision-making and drawback resolution. DSS tends to be aimed toward the less well structured, underspecified drawback that higher level managers generally face. DSS makes an attempt to mix the employment of models or analytic techniques with ancient knowledge access and retrieval functions. The use of TOPSIS method on DSS can assist the managerial role of a manager in making a decision to gain a high and best skills employees. Within the future, DSS emphasizes flexibility and adaptability to accommodate changes within the environment and the decision making approach of the managers in the real world.

ACKNOWLEDGEMENT

I would like to express my special thanks to **all my teachers** who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank **my parents** for their continuous support and encouragement throughout my whole life.

REFERENCES

[1] S. Alonso, E. Herrera-Viedma, F. Chiclana, and F. Herrera, "A Web Based Consensus Support System for Group Decision Making Problems and Incomplete Preferences," *Inf Sci*, vol. 180, no. 23, pp. 4477–4495, Dec. 2010.

[2] D. J. Power, *Decision support systems: concepts and resources for managers*. Westport, Conn: Quorum Books, 2002.

[3] J. Simonsen, "Herbert A. Simon: Administrative Behavior How organizations can be understood in terms of decision processes," p. 12.

[4] "Role of Decision Support System For Decision-Making Process in Global Business Environment." [Online]. Available: https://ezinearticles.com/?Role-of-Decision-Support-System-For-Decision-Making Processin Global Business Environment&id=2315787.[Accessed: 12-Dec-2019].

[5] S. Nowduri, "Management information systems and business decision making: review, analysis, and recommendations," p. 8.

[6] R. J. Chambers, "The Role of Information Systems in Decision Making," *Manag. Technol.*, vol. 4, no. 1, pp. 15–25, 1964.

[7] "(PDF) An Integrated Framework Of Decision Support System In Crime Prevention." [Online]. Available:https://www.researchgate.net/publication/303940849_An_Integrated_Framework_Of_Decision_Support_System_In_Crime_Prevention. [Accessed: 12-Dec-2019].

[8] G. B. Davis and M. H. Olson, *Management Information Systems: Conceptual Foundations, Structure, and Development (2Nd Ed.)*. New York, NY, USA: McGraw-Hill, Inc., 1985.

[9] Al-Zhrani, "Management Information Systems Role in Decision-Making During Crises: Case Study," *J. Comput. Sci.*, vol. 6, no. 11, pp. 1247–1251, Nov. 2010.

[10]"The Role of Management Information Systems in Decision-Making | Chron.com." [Online]. Available: https://smallbusiness.chron.com/role-management-information-systems-decisionmaking-63454.html. [Accessed: 12-Dec-2019].

[11]R. McLeod and G. P. Schell, *Management information systems*. Upper Saddle River, N.J.: Pearson/Prentice Hall, 2007.

[12]"Managerial Decision Making Process (5 Steps)." [Online].Available:http://www.economicsdiscussion.net/decision-making/managerial-decision-making-process-5-steps/6099. [Accessed: 12-Dec-2019].

[13]"Decision Support Systems: An Organizational Perspective."[Online].Available:https://dssresources.com/books/contents/keen78.html. [Accessed: 12-Dec-2019].

[14]E. Turban, J. E. Aronson, and T.-P. Liang, *Decision Support Systems and Intelligent Systems*, 7 edition. Upper Saddle River: Prentice Hall, 2004.

[15]"Multiple Attribute Decision Making Based on Cross-Evaluation with Uncertain Decision Parameters."[Online].Available:https://www.hindawi.com/journals/mpe/2016/4313247/. [Accessed: 12-Dec-2019].

Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190078

QUERY PROCESSING FOR XML DOCUMENT USING RDF REPOSITORY

Win Lai Hnin [(1)], Su Myat Sandar Win[(2)], Myat Thet Nyo[(3)]

[(1)(2)(3)]University of Computer Studies (Meiktila), Myanmar

[(1)]*winlaihnin.84@gmail.com*

ABSTRACT

XML and its schema language are becoming a primary data exchange format on the current web. In next generation of the Semantic Web, the drawbacks of XML and its schema will appear. XML adds tags to a text stream to provide some structure and additional information in the same way that HTML does, however, XML tags do not provide any explicit meaning. Semantic web would give more structure and computer-understandable meaning to the data on the WWW. The semantic web is not a separate web but an extension of the current one, in which information is given a well-defined meaning, better enabling computers and people to work in cooperation. Therefore, finding a way to utilize the available XML documents for the Semantic Web is a current challenge research. To harvest such power requires robust and scalable data repositories that can store RDF data. This paper proposes a set of rules to map DTD to RDFS/OWL and provide algorithm to interpret XML documents as RDF. This approach mainly considers the integrity constraints of XML document to map RDF data. So, this approach is to ensure the integrity of the structure and to provide more meaning for the original XML document while transforming them into RDF and then performing RDF query processing for the performance of RDF query evaluation.

KEYWORDS: *XML, DTD, RDFS/OWL, Semantic Web, Integrity Constraints*

1. INTRODUCTION

XML (Extensible Markup Language) is an important language for data interchange, and also provides a serialization format for Semantic Web languages. Anyone who has knowledge of this predefined format can read and interpret a given document correctly. XML is used to create human and machine-readable documents. To specify the exact names of element nodes and attributes, a Document Type Definition (DTD) or XML Schema is required. The information contained in these definitions should be known to anyone using the XML document. This method targets on DTD, utilizes its declarations to produce suitable mapping rules and more compact and higher readable than XML Schema. XML is a meta-language, which is used to describe the structure of documents. Although XML plays an important role in structuring the document, it has disadvantages to use in the semantic interoperability. So, we can't directly use XML data for the Semantic Web, and need another language to interpret this data.

The meaning in the Semantic Web is mostly represented by Resource Description Framework (RDF). RDF, based on the concept of semantic networks, provides a simple yet powerful data model for semantic assertions. This model is based on so called triples. Using these triples a semantic of information is formulated like an elementary sentence consisting of a subject, predicate and object. There are various serialization syntaxes for RDF, but the most common one for a RDF document exchange is RDF/XML (based on XML). These are recognized by the Universal Resource Identifiers (URIs) which tie meanings to a unique definition so that users can easily find them and their relationships on the web. Most of the existing RDF storage techniques rely on relation model and relational database technologies for these tasks. The mis-match between the graph model of the RDF data and the rigid 2D tables of relational model jeopardizes the scalability of such repositories and frequently renders a repository inefficient for some types of data and queries. RDF is a directed, labeled graph data format for

representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunction and disjunction. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results set or RDF graphs.

The remainder of this paper is organized as follows. In section 2, we briefly introduce the related works. In. section 3 defines the overview of the proposed system. Section 4, describes a set of rules for mapping DTD to RDFS/OWL and XML mapping algorithm. Section 5 defines the SPARQL query language. In section 6 describes the experimental result. Finally, section 7 concludes this paper.

2. RELATED WORKS

Several approaches related to schema mapping and XML transforming have been proposed. This section summarizes and analyzes the strength and weakness of these approaches. Based on such examining, this method proposed a more comprehensive and efficient solutions for DTD mapping and XML transforming.

P.T.T. Thuy et al. [5] proposed a procedure for transforming valid XML documents into RDF via RDF Schema. This procedure derived classes and properties from DTD, then matched them with elements in XML documents and interpreted all XML data as RDF statements. This approach is closed to the proposed method. But they didn't consider the constraints in the transformation process. On the contrary, the proposed method expands RDF Schema by defining OWL ontology property to describe the constraints of DTD in the RDF Schema.

P.T.T. Thuy et al. [11] proposed a procedure for transforming valid XML documents into RDF via RDF Schema. This procedure derived classes and properties from XSD, then matched them with elements in XML documents and interpreted all XML data as RDF statements. However, in order to describe the relationship between parent class and child class, the authors defined new RDF vocabulary, rdfx:contain. This definition is not recognized by the RDF evaluation tools or Semantic Web applications. Therefore, the proposed method use existing RDF vocabularies by using rdfs:Container. So, the result of this approach is used directly on the Web without any changes.

This paper proposes a strategy to map DTD to RDF Schema and transformation algorithm for interpreting valid XML document as RDF data which can be loaded immediately by RDF editors and other Semantic Web applications.

3. OVERVIEW OF THE PROPOSED SYSTEM

The transforming framework of DTD2RDFS/ OWL is shown in Figure 1. Having DTD as input, a mapping process converts all DTD components to RDFS/OWL which captures the semantic and maintains the constraints of the element names, attribute names, data types and other declaration of DTD. Moreover, RDFS/OWL is better than the DTD by adding definition of the meaning and relationship between elements in DTD..

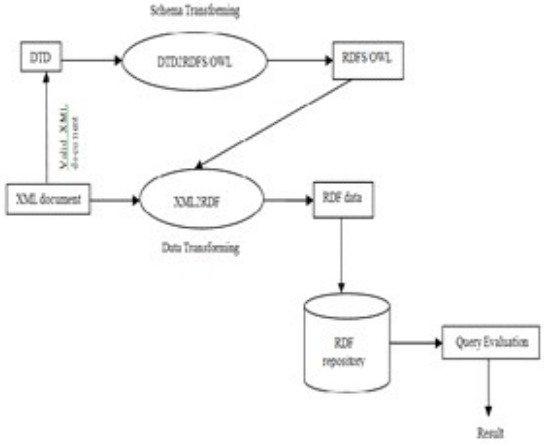

Fig 1 A framework for transforming XML into RDF

During this stage, the proposed method also checks and solves the problem whether the next element has the same name with the previous one, if it does, these elements are renamed. Specifically this system collect the element from XML and transform XML document into RDF data and store RDF repository to extract information by SPARQL queries. Figure 1 shows the architecture of proposed system.

4. MAPPING AND XML TRANSFORMING

The proposed method has two main steps. The first one presents the set of rules for mapping of DTD

to RDFS/OWL. The Second uses RDFS/OWL to transform XML document into RDF/XML.

4.1 Mapping DTD to RDFS/OWL

In this section, the proposed method presents the rules for the mapping of the DTD to RDFS/OWL. This method tends to convert every DTD elements and attributes to class and property in the RDFS/OWL. The result of this mapping is an RDFS/OWL that maintains the structure and captures the semantics of the DTD. The idea of this step is as follows:

Root element: Element defined by <! DOCTYPE> in DTD is mapped to the root-class of RDF schema, which is the first class declared by rdfs:Class.

Class (rdfs:Class): A DTD is made up of three main building blocks: ELEMENT, ATTLIST and ENTITY. ELEMENT is the main building block of XML documents. In the DTD, XML elements are declared with an ELEMENT. An element definition has the following syntax:

<!ELEMENT *element-name* (*element-content*)>

element-content may be EMPTY, or data type, or sequences of children. As ELEMENT is used to describe elements of a document and each element can contain children elements, the function of these elements is like a class in a structure program, therefore this element will be considered as RDF class. Each rdfs:Class is represented by a unique identifier, rdf:ID.

```
<!DOCTYPE catalog[
<!ELEMENT catalog (journal+)>
<!ELEMENT journal (article □ name)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT article (title, author)>
<!ELEMENT title (#PCDATA)>
<!ATTLIST catalog publisher #FIXED "O'Reilly">
<!ATTLIST catalog title CDATA #REQUIRED>
<!ATTLIST catalog photo ENTITY #REQUIRED>
<!ENTITY mt-catalog-1 "mt-catalog1.jpg">
<!ATTLIST journal date CDATA #REQUIRED>
<!ATTLIST author gender (Male/Female)
#REQUIRED>
<!ATTLIST article aid ID #REQUIRED>
<!ATTLIST author id IDREF #REQUIRED>
]>
```

Fig 2 Definition of complex classes in DTD

Moreover, DTD attributes normally contain constraints. Table 1 shows the mapping of DTD constraints to RDFS/OWL concepts.

TABLE 1 THE MAPPING OF DTD CONSTRAINTS INTO RDFS/OWL

DTD	RDFS/OWL
NOTATION	rdfs:comment
#REQUIRED	owl:minCardinality (=1)
#IMPLIED	owl:Cardinality (=0)
+	owl:minCardinality (=1)
?	owl:minCardinality (=0)
*	owl:minCardinality(=0) owl:maxCardinality (=unbounded)

For nested elements, this procedure does not use the rdfs:subclassOf, which is available in RDF syntaxes. The reason is because some nested elements in DTD are not actually the sub-class of their parent element. Therefore, rdfs:Container is defined to establish the relationship between child node and parent node For example, the relationship of parent element and child element between three classes, "catalog", "journal", and "article" in Fig.2 are described as RDFS/OWL concepts shown in Fig.3.

```
<rdfs:Class rdf:ID= "catalog">
   <rdfs:comment> catalog class </rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID= "journal" owl:minCardinality=
"1">
   <rdfs:comment> journal class </rdfs:comment>
   <rdfs:Container rdf:resource= "catalog"/>
</rdfs:Class>
<rdfs:Class rdf:ID= "article">
    <rdfs:comment> article class </rdfs:comment>
    <rdfs:Container rdf:resource= "journal"/>
</rdfs:Class>
```

Fig 3 RDFS/OWL declaration in Fig 2

Content constraints are relationship of subelements. In Figure 2, article and name are subelements of journal and title and author are subelements of article. For instance, these are mapped to RDFS/OWL concepts in Fig 4.

```
<rdfs:Class>
  <owl:intersectionOf rdf:parseType="Collection">
    <rdfs:Class rdf:ID="article"/>
    <rdfs:Class rdf:ID="name"/>
  </owl:intersectionOf>
    <rdfs:comment> Class Collection
</rdfs:comment>
    <rdfs:Container rdf:resource="journal"/>
</rdfs:Class>
<rdfs:Class>
  <owl:unionOf rdf:parseType="Collection">
    <rdfs:Class rdf:ID="title"/>
    <rdfs:Class rdf:ID="author"/>
  </owl:unionOf>
    <rdfs:comment> Class Collection
</rdfs:comment>
    <rdfs:Container rdf:resource="article"/>
</rdfs:Class>
```

Fig 4 RDFS/OWL declaration in Fig 2

Property (rdf:Property) : For this case an element in DTD is described by <!ELEMENT> tag but its element-content contains data type (#PCDATA or #CDATA), this element will be considered as RDF property, rdf:Property. The property's domain is the parent class of this property, and its range is the data type of this property. On the other hand, #PCDATA and #CDATA are used for declaring character data in XML, so this procedure maps them to "String" data type in RDFS/OWL. For instance, in Fig 5, element "title" has a data type, so it is mapped to RDFS/OWL concepts in Fig.6.

```
<!ELEMENT title (#PCDATA)>
```

Fig 5 Definition of complex classes in DTD

```
<rdf:Property rdf:ID="title">
  <rdfs:domain rdf:resource="article"/>
  <rdfs:range rdf:resource=
"http://www.w3.org/1999/02/22-rdf-syntax-
ns#Literal"/>
</rdf:Property>
```

Fig 6 RDFS/OWL declaration in Fig 5

ATTLIST provides extra information about elements so its function is to describe the property of a class. The attribute definition has the following syntax:

<! ATTLIST *element-name attribute-name attribute-type default-value*>

element-name is the name of element (class) and *attribute-name* is a name of the attribute, in the propose method, it is a name of the property. *attribute-type* is a data type and *default-value* specifies default value of the attribute. For instance, one simple attribute in Fig.7 is mapped to RDFS/ OWL concepts in Fig.8.

```
<!ATTLIST journal date CDATA #REQUIRED>
```

Fig 7 Definition of complex class in DTD

```
<rdf:Property rdf:ID="date">
  <rdfs:domain rdf:resource="#journal"/>
  <rdfs:range rdf:resource=
"http://www.w3.org/1999/02/22-rdf-syntax-
ns#Literal"/>
</rdf:Property>
```

Fig 8 RDFS/OWL declaration in Fig 7

If *attribute-type* contains "FIXED" constraints, *default-value* will be considered as the range of its property. In Fig 9, attribute "publisher" has data type "FIXED", so it is mapped to RDFS/OWL concepts in Fig 10.

```
<!ATTLIST catalog publisher #FIXED
```

Fig 9 Definition of complex classes in DTD

```
<rdf:Property rdf:ID="publisher">
  <rdfs:domain rdf:resource="#catalog"/>
  <rdfs:range rdf:resource="#O'Reilly"/>
</rdf:Property>
```

Fig 10 RDFS/OWL declaration in Fig 9

There is another notice that XML syntax allows, elements with the same name in a document, but RDFS/OWL does not. RDFS/OWL requires each element has a unique identifier. Since there are two elements that have the same name, **title,** the second repeated name is renamed by adding its parent name in front of its name as in Fig 11.

```
<rdf:Property rdf:ID="article-title">
  <rdfs:domain rdf:resource="#article"/>
  <rdfs: range rdf:resource=
"http://www.w3.org/1999/02/22-rdf-syntax-
ns#Literal"/>
</rdf:Property>
```

Fig 11 RDFS/OWL declaration in Fig 2

Beside these, another constraint in DTD is enumeration constraint. Its purpose is to declare a list of possible value of its attribute and attributes in the document must be assigned a value from this list. Furthermore, in order to depict the attribute's enumerated type, this procedure borrows the OWL expressions, such as owl:oneof. For example, in Fig 12, attribute "gender" has enumeration constraint, so it is mapped to RDFS/OWL concepts in Fig 13.

```
<!ATTLIST author gender (Male/Female)
#REQUIRED>
```

Fig 12 Definition of complex classes in DTD

```
<rdf:Property rdf:ID="gender" owl:oneof
              rdf:parseType="Resource">
   <rdfs:domain rdf:resource="#author"/>
   <rdfs:range rdf:resource="#Male"/>
   <rdfs:range rdf:resource="Female"/>
</rdf:Property>
```

Fig 13 RDFS/OWL declaration in Fig 12

ENTITY is used to define a shortcut for a common text in XML. Its syntax is as follows:

<! ENTITY *name definition*>

In this case, name is the name of ENTITY and definition is its definition. Because of the function in the DTD, this procedure handles name as a variable and definition as its value. When this procedure meets this variable in the document, its value will be called. For example, one simple entity in Fig 2 is mapped to RDFS/OWL concepts in Fig 14.

```
<rdf:Property rdf:ID="photo">
    <rdfs:domain rdf:resource="#catalog"/>
    <rdfs:range rdf:resource="&mt-catalog-1;"/>
</rdf:Property>
```

Fig 14 RDFS/OWL declaration in Fig 2

In DTD, there are two kinds of keys constraints such as ID and IDREF. ID is key and IDREF is foreign key. For instance, ID/IDREF in Fig.2 is mapped to RDFS/OWL concepts in Fig.15.

```
<rdf:Property rdf:ID="aid">
    <rdfs:domain rdf:resource="article"/>
    <rdfs:range rdf:resource=
"http://www.w3.org/1999/02/22-rdf-syntax-
ns#Literal"/>
</rdf:Property>
<rdf:Property rdf:ID="id">
    <rdfs:domain rdf:resource="author"/>
```

Fig 15 RDFS/OWL declaration in Fig 2

4.2 XML TRANSFORMING ALGORITHM

In this section, the proposed method presents XML transforming algorithm for transforming valid XML document into RDF data. This method use namespace http://www.w3.org/2002/07/owl for OWL syntaxes and http://www.w3.org/1999/02/22-rdf-syntax-ns for providing RDF syntaxes. The URI of the XML document will be the URI of each class. The algorithm starts traversing from the beginning of the XML document and finishes when it meets the close tag of root element. The comments are skipped during the transforming process. When it meets an element, it will compare this element to the definition of the RDF schema to decide whether it is a class or a property. If it is a class, it creates rdf:Description and describes new resource for that class. Otherwise it is a property, it will drag this value and tag from XML document to RDF. The XML transforming algorithm is as follows:

```
1. Input: XML document
2. Read the description of root-class in the XML
   document to draw its properties if they are available.
3. For 1 to total number of child-node (of the XML
document)
4.     Create the namespace for XML document;
5.     For each complex element (class)
6.         Generate an RDF description for each
           class
7.         Create a resource for class (URI=
           baseURI+resourcename+#class+number)
8.     For each attribute in the class
9.         Create tag "namespace:attribute name";
10.        Copy attributes values;
```

Fig 16 XML document example

During the schema mapping process, this method changes some name of the DTD element. So, this data transformation step must update the changed element in the XML instances too.

This section illustrates the transforming from XML document into RDF. The XML document is also taken on the same website with DTD. The

proposed algorithm automatically generates RDF data; it does not require any human intervention so the result is independent from user. The RDF result obeys RDF syntaxes, so it does not require any changes in order to be used by the Semantic Web. This algorithm can also be applied for scalable XML documents which exist enormously on the current web. The corresponding RDF data for above XML document is in Figure 17.

Fig 17 RDF/XML document for Fig 16

5. SPARQL QUERY LANGUAGE

SPARQL is an RDF query language that is a semantic query language for databases able to retrieve and manipulate data stored in RDF format. SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions and optional patterns. It allows users to write queries against what can loosely be called "key-value" data or more specifically, data that follow the RDF specification of the W3C. Thus, the entire database is a set of "subject-predicate-object" triples. It provides a full set of analytic query operations such as JOIN, SORT, and AGGREGATE for data whose schema is intrinsically part of the data rather than requiring a separate schema definition. However, schema information (the ontology) is often provided externally to allow joining of different datasets unambiguously. In addition, SPARQL provides specific graph traversal syntax for data that can be thought of as a graph. In the case of queries that read data from the database, the SPARQL language specifies four different query variations for different purposes.

- SELECT query

Used to extract raw values from a SPARQL endpoint, the results are returned in a table format.

- CONSTRUCT query

Used to extract information from the SPARQL endpoint and transform the results into valid RDF.

- ASK query

Used to provide a simple True/False result for a query on a SPARQL endpoint.

- DESCRIBE query

Used to extract an RDF graph from the SPARQL endpoint, the content of which is left to the endpoint to decide based on what the maintainer deems as useful information.

Each of these query forms takes a WHERE block to restrict the query, although, in the case of the DESCRIBE query, the WHERE is optional. SPARQL query example of RDF/XML document for figure 16 is as follow:

```
PREFIX ex:
< http://www.recshop.fake/Catalog#>
SELECT  ? title
            ? publisher
WHERE
    {
        ?x   ex:titlename ?title;
         ?y  ex:publishername   ?publisher;
```

ex:Catalog

}

Properties relevant to SPARQL language design include support for the RDF format:

- Support for RDF data, which is a collection of triples that form the RDF graph
- Support for RDF semantics and inference that allows for entailment, the reasoning about the meaning of RDF graphs
- Support for schema data types and for desirable language features
- Expressiveness: the power of query expression that may be constructed
- Closure: data operations on an RDF graph should
- Orthogonality: data operations are independent of the context in which they are used
- Safety: every expression returns a finite set of results.

6. EXPERIMENTAL RESULTS

6.1 Validation of efficiency

Efficiency of a system quantifies the consumption of resources during the run-time of the system. Depending on the purpose of the system, different resources might be of interest. Usually, time is the main resource: the efficiency shows how quick a system completes a given task. Other kinds of resources include electricity consumption, human working hours, and cost of using leased resources such as web-services or network. In a mapping research, efficiency has mostly been ignored and most work does not report on any efficiency validation. In this thesis, however, efficiency is in the focus and it exclusively considers time consumption. In particular we are interested in expressing the efficiency improvement acquired by introducing a set of rules for transforming DTD to RDFS ontology and providing algorithm to interpret XML documents as RDF. To access the performance of the instance transformation process, we have used four datasets with HDF5.xml, mondial-3.0.xml, Sigmodrecord.xml and yahoo.xml. We have transformed four XML documents with different sizes, which validate the mapped schema. Time performance of schema mapping is not very important however in some applications such as communication in peer-to-peer networks, the time needed to find "good enough", mapping plays the key role. The performance results are shown in Table 2.

Table 2 Time Performance Evaluation

Schema name	Size in (XML file)	Size out (RDF data)	Time Processing
HDF5	4.48KB	9.25KB	0.155s
Yahoo	14.6KB	24.5KB	0.255s
Sigmodrecord	16.0KB	37.9KB	0.301s
Mondial-3.0	22.3KB	39.2KB	0.412s

6.2 Validation of the Results

In order to validate our RDF output result [29], we use the ICS-FORTH VRP validation tool. This RDF validation service is based on Another RDF Parser (ARP). It currently uses version 2-alpha-1. ARP was created and is maintained by Jeremy Carroll at HP-Labs in Bristol. This W3C service was created by Nokia's Art Barstow (a former W3C Team member). The service now supports the Last Call Working Draft specifications issued by the RDF Core Working Group, including data types. It no longer supports deprecated elements and attributes of the standard RDF Model and Syntax Specification and will issue warnings or errors when encountering them. In order to use this service, we only need to copy and paste the RDF file to the input window. When parsing large RDF files, requesting Triples only will significantly shorten the response time of this service. For testing our RDF result, we paste our RDF statements to this validator, the result is verified successfully. By pressing the button "Process", we can see the validation result as in the figure 18. This means that our RDF document can be used directly by other RDF editors or Semantic Web applications.

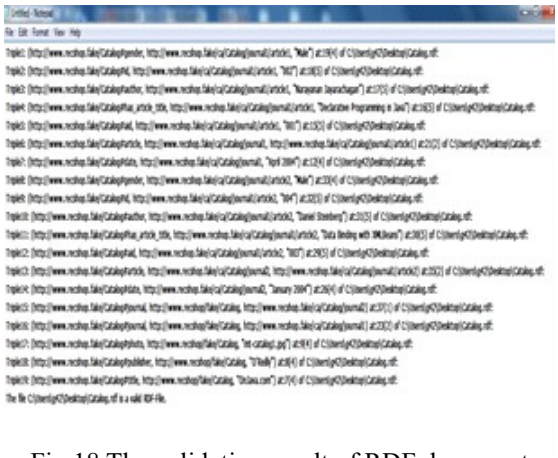

Fig 18 The validation result of RDF document

6.3 Comparative Analysis between Existing Methods and Proposed Method

In order to overcome the heterogeneity among the information systems, data exchange systems have been proposed. The data exchange systems (also known as data transformation/translation systems) restructure the data from the source according to a global schema. In recent research works, semantics are exploited to bridge the heterogeneity gap among the information systems and provide semantic integration and interoperability. Table 3 provides an overview of the comparative analysis of the methods. The system described in each row is specified in the first column (System), the Environment Characteristics are shown in columns 2-3 and the supported Operations are shown in columns 4-6. The environment characteristics include the Data Models of the underlying data sources (2nd column) and the involved Schema Definition Languages (3rd column). The operations include the Schema Transformation operation (4th column), the indication for the Use of an Existing Ontology (5th column) and the Data Transformation mechanism (6th column). If a schema transformation operation is supported, the value of the third column is the operation description; if the method does not support schema transformation, the value is "no". If the value of fifth column is "yes", the method supports mappings between XML Schemas and existing ontologies and, as a consequence the XML data are transformed according to the mapped ontologies. Finally, if a data transformation mechanism is provided, the sixth column has its description as value and if the value of the fifth column is "no", the system does not provide a data transformation mechanism.

Table 3 Comparative Analysis between Existing Methods and Proposed Method

System	Environment Characteristic		Operation		
	Data Model	Schema Definition Language	Schema Transformation	Use Existing Ontology	Data Transformation
Klevin (2002)	XML/RDF	XML Schema RDF Schema	no	no	XML - RDF
WEESA (2004)	XML/RDF	XML Schema OWL	no	yes	XML - RDF
Ferdinand et al. (2004)	XML/RDF	XML Schema OWL-DL	XML Schema - OWL-DL	no	XML - RDF
Garcia & Celma (2005)	XML/RDF	XML Schema OWL-Full	XML Schema - OWL-Full	no	XML - RDF
Gloze (2006)	XML/RDF	XML Schema OWL	no	no	XML - RDF
XS2OWL (2007)	XML/RDF	XML Schema OWL-DL	XML Schema - OWL-DL	no	XML - RDF
Thuy et al. (2007)	XML/RDF	DTD OWL-DL	DTD - OWL-DL	no	XML - RDF
Janus (2008)	XML/RDF	XML Schema OWL-DL	XML Schema - OWL-DL	no	no
DTD2OWL (2009)	XML/RDF	DTD OWL-DL	DTD - OWL-DL	no	XML - RDF
XS2OWL 2.0 (2011)	XML/RDF	XML Schema OWL	XML Schema - OWL	no	XML - RDF
The proposed approach	XML/RDF	DTD RDF Schema	DTD – RDFS OWL	no	yes

7. CONCLUSIONS

To answer the increasing demands on RDF repository, we carefully studied the existing RDF data management systems, identified the preferred properties of an RDF repository and proposed to take advantage of the latest XML data storage and efficient query processing techniques. There is a wide variety of graph patterns that can be matched through SPARQL queries, which reflects the variety of the data that SPARQL was designed to query. As a result, SPARQL can efficiently extract information hidden in non-uniform data and stored in various formats and sources. This paper describes 11 transformation rules that the rule of mapping for DTD to RDFS/ OWL and transforming algorithm from XML document to RDF data. In addition, our approach is efficient for time consuming in translation from XML to RDF documents for supporting Semantic Web applications in various domains.

ACKNOWLEDMENT

We would like to thank the University of Computer Studies, Meiktila for their support.

REFERENCES

[1] S.Decker, S.Melnik, F.V.Harmelen, D.Fensel, M.Klein, J.Broekstra, M.Erdmann, and I.Horrocks, "The Semantic Web: The Roles of XML and RDF", 2000, IEEE Internet Computing.

[2] Sergey Melnik, "Bridging the gap between RDF and XML",1999, available at http://www.db.stanford.edu/melnik/rdf/syntax.html .

[3] Michel Klein, "Interpreting XML via an RDF Schema", 2002, Database and Expert Systems Applications.

[4] I.F. Cruz, H. Xiao and F. Hsu, "An Ontology-Based Framework for XML Semantic Integration", In IDEAS'04: Proceedings of the International Database Engineering and Application Symposium, 217-226, (2004).

[5] P.T.T. Thuy, Young-Koo Lee, Sungyoung Lee and Byeong-Soo Jeong, "Transforming Valid XML Documents into RDF via RDF Schema", International Conference on Next Generation Web Services Practices, IEEE, October 2007.

[6] Peter Patel-Schneider and Jerome Simeon, "The Yin/Yang Web: XML syntax and RDF Semantics", 11th International WWW conference, Hawaii, 2002.

[7] Refsnes Data, "Introduction to DTD", 1999-2007, available at:

[8] http://www.w3schools.com/dtd/dtd-into.asp.

[9] Dan Brickley, R.V Guha and Brian McBride, "RDF Vocabulary description language 1.0: RDF Schema", W3C, Feb 2004, available at: http://www.w3.org/TR/2004/REC-rdf-schema.

[10] Frank Manola, Eric Miller, "RDF Primer", W3C Recommendation, February 2004, available at: http://www/w3.org/TR/REC-rdf-syntax.

[11] T.Bray, J.Paoli and C.M. Sperberg-McQueen, "eXtensible Markup Language(XML) 1.0", W3C Recommendation, Feb 1998, available at: http://www.w3.org/TR/REC-xml.

[12] P.T.T. Thuy, Young –Koo Lee, and Sungyong Lee, "Exploiting XML Schema for Interpreting XML Documents as RDF", 2008 International Conference onServices Computing