Journal of Global Business and Social Entrepreneurship (GBSE)
Vol. 1: no. 3 (2017) page 12–23| gbse.com.my | eISSN 24621714|

MODELLING DATA MANAGEMENT CONTENT FOR PAPERWORK AND REPORT ORGANIZER USING DATA RETRIEVAL AND EXTRACTION WITH e-NOTIFICATION

NurSuhailayani Suhaimi
suhailayani@melaka.uitm.edu.my

FadilahEzlinaShahbudin
fadilahezlina@melaka.uitm.edu.my

FaiqahHafidzahHalim
faiqahhalim@melaka.uitm.edu.my

AnisAfiqahSharip
anis588@melaka.uitm.edu.my

Abstract

The issue of organizing the information storage in handling and managing data is crucial. Over the years, filing has been used as one of the ways to manage data. Since the organization has to manage and monitor the data collection and data changes manually, it led to the issues of data redundancy, including the inconsistency and inaccuracy of data. Moreover, data in physical files has been exposed to security issues as they are easily access even without the authorization. As the technology is evolving, there are many ways to protect the data from unauthorized access, including the introduction of Database Management System. These technologies are able to manage the data efficiently as multiple authorized users can access and modify the data. As a consideration regarding the data security issues as highlighted above, any changes made to the data can be tracked and information retrieval could be improved. With the aim to model data management content for paperwork and report organizer, this system allows quick access for users to construct paperwork based on the template. Additionally, some features such as search and document retrieval for printing are provided thus it makes it easier for user to submit the document to respective department. Other than to reduce data redundancy and discrepancy issues, this system is able to assist users to reduce time for managing paperwork. It is also provides the benefits including easy management and access of historical data, indexed data storage, and efficient data retrieval for reporting purposes.

Keywords: data retrieval, data extraction, data management

Introduction

Management generally being acknowledge as organizing an organization systematically. Every organized system entails a proper management process. Organizing data could be tedious when it is involving complex and numerous variables. The tremendous amount of data with multiple data-types could lead to the issue of data integration problems in the data warehouse. Meanwhile, the on-going advancement of information overflow has made it challenging to acquire priceless information even on the web. At the moment, database does not just serve the purpose of storing data. The data being stored will eventually require to be retrieved and produce knowledge, not just typical information or raw data. This paper discusses issues and processes regarding modelling a conceptual template of online paperwork creator with e-notification alert. The role of data retrieval and extraction will be demonstrated during construction of the report where the knowledge formed based on the inserted paperwork information. In order to avoid information load in the report construction, data retrieval and extraction methods is applied. For alerting purposes, notification feature has been embedded in the system as to acknowledge user on any new or updated data. Numerous issues have been discussed in literature reviews and the process of the research is depicted in the methodology section. The conceptual model serves the purpose of minimizing the problem of data losing, data discrepancy and data redundancy. The targeted purpose of this research is producing efficient information systems to aid in managing and handling the data storage with notification feature embedded.

Literature Review

Information in digital format is mostly unstructured except for the information in the form of databases (Nagarkar and Kumbhar, 2015).Data modelsare connected to each other and represent how they are processed and stored inside the system. Several issues in modeling this research have been considered to be minimized by data retrieval and data extraction.

1.1. Problems in Data Management

The implication of time in information production and consumption has been known in information retrieval study. Processing and accessing database resources accessible on Internet are occasionally complicated, especially when textual content is concerned. An inexperienced user may require a general explanation of the contents obtainable in the database in order to decide if the information is beneficial for his or her search requirements. Issues in data management always involved with dirty data which are noisy, redundant and discrepancies problems. Data being stored not always managed in proper manner, especially when being integrated with other databases. Another problem occurs during the retrieval phase where it usually display back whatever stored in database, not processed into information. Most required information to be displayed is to serve the purpose of decision

making. Thus the data being retrieved and extracted supposed to be a complete knowledge, no longer raw data.

1.2. Data Retrieval

Information retrieval (IR) is the task of acquiring information resources pertinent to an information need from a stockpile of information resources. By using metadata or full-text indexing, searches in the collection of information resources can be done. Iswarya and Radha (2014) stated that IR is the process of salvaging related or desired information from a large pool of data source. According to Otegi, Arregi, Ansa and Agirre (2014), in order to catalogue and retrieve documents, the traditional IR system use keywords or tag is used. Keyword retrieval proved to be unsuccessful when dissimilar but closely associated words are utilized in the query and the pertinent document. Information retrieval concerns to getting back or retrieving information stored in various storage media, exactly in the same way it is stored. Standard text mining systems usually work on "categorized documents" rather that unprepared documents (Feldman, Fresko, Kinar, Lindell, Liphstat, Rajman, Schler and Zamir, 1998). These documents were tagged with terms that are able to identify their respective content for retrieval. For data retrieval, a database system provides a query or called data manipulation by Aho and Ullman (1979) to retrieve information from database. The data retrieved by a single query can range from a small simple subset of database to a large complex subset. In order to retrieve a text data, a query language should provide physical data independence where the results of a query should not depend on the representation of the data.

Many previous data retrieval studies have used data-inversion routines based on the principles of the TDMAfit program developed by Stolzenburg and McMurry. The measured growth factors are assumed to fall into a few, well defined separated modes, which are log-normally distributed and assumed to be linear. These types of data retrieval are described by an arithmetic mean given by the growth factor value and the standard deviation of the mode, multiplied with the value of a second calculated parameter, the growth dispersion factor. These parameters were determined by fitting 1–3 normal distributions of diameter growth factors to the measurement data would suggest using Gaussian model to retrieve text.However, the gaussian fitting method does not give a detailed error analysis on the data set, nor does it give a measure of the resolution of the measurements. It is thus possible to apply a bi-modal gaussian fit to a measurement distribution where the splitting of the modes is finer than the resolution of the measurements. However, the gaussian fitting method does not return a growth factor distribution, but merely the positions and widths of the growth factor modes.

1.3. Data Extraction

Information Extraction (IE) is one of the tasks in text mining. According to Grishman (2015), IE is the procedure of examining text and recognizing statements of linguistically defined entities and its interrelationships. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them.Knowledge extraction is the creation of knowledge from structured and unstructured sources.The resulting knowledge needs to be in a machine-readable and machine-interpretable format and must represent knowledge in a manner that facilitates inferencing. Although it is methodically similar to information extraction, the main criteria are that the extraction result goes beyond the creation of structured information or the transformation into a relational schema. It requires either the reuse of existing formal knowledge or the generation of a schema based on the source data. According to Balakrishna, Werner, Tatu, Erekhinskaya & Moldovan (2016), knowledge extraction converts document content into semantic triples

Methodology

Management is a process that used various methods to ensemble the activities of planning and monitoring the performance. Step by step of each activities conducted involve an agile process of data storage and access. The main process is focus on data retrieval and data extraction. The activities conducted in this research are illustrated in the flowchart below:

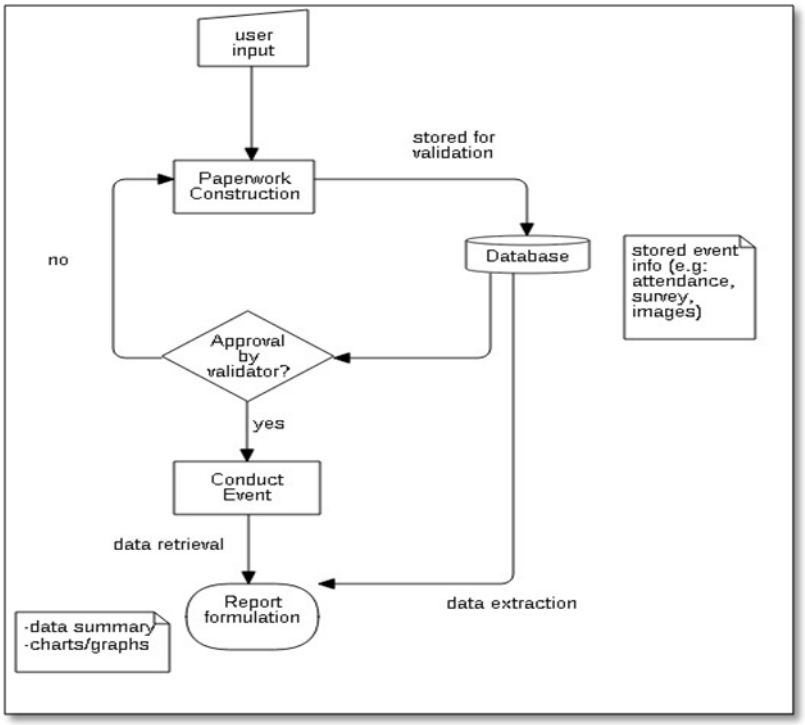

Figure 1: Basic Flowchart for Data Management Model

1.4. Paperwork Construction

In order to construct the paperwork from the template provided by this research model, sample of paperwork stored as data in the database. The information required by the paperwork may have some of it suggested by the prototype and some of it from the user input. Certain input may require open-ended data from the user to make it more flexible. The gathered data from the user input to construct the paperwork then stored for report retrieval.

1.5. Retrieval and Extraction

The fundamental of Information retrieval can be classified into Boolean, vector, probabilistic and inference network model. The most frequent method used to retrieve text is by locating the documents that contain a certain search term or keyword to search all documents for the specified string. This is called full text scanning. We compare the documents against the corresponding characters of the document and calculate the similarity of desired and stored. Although the simple implementation, the extraction then take place to filter only required knowledge to be presented in the report.

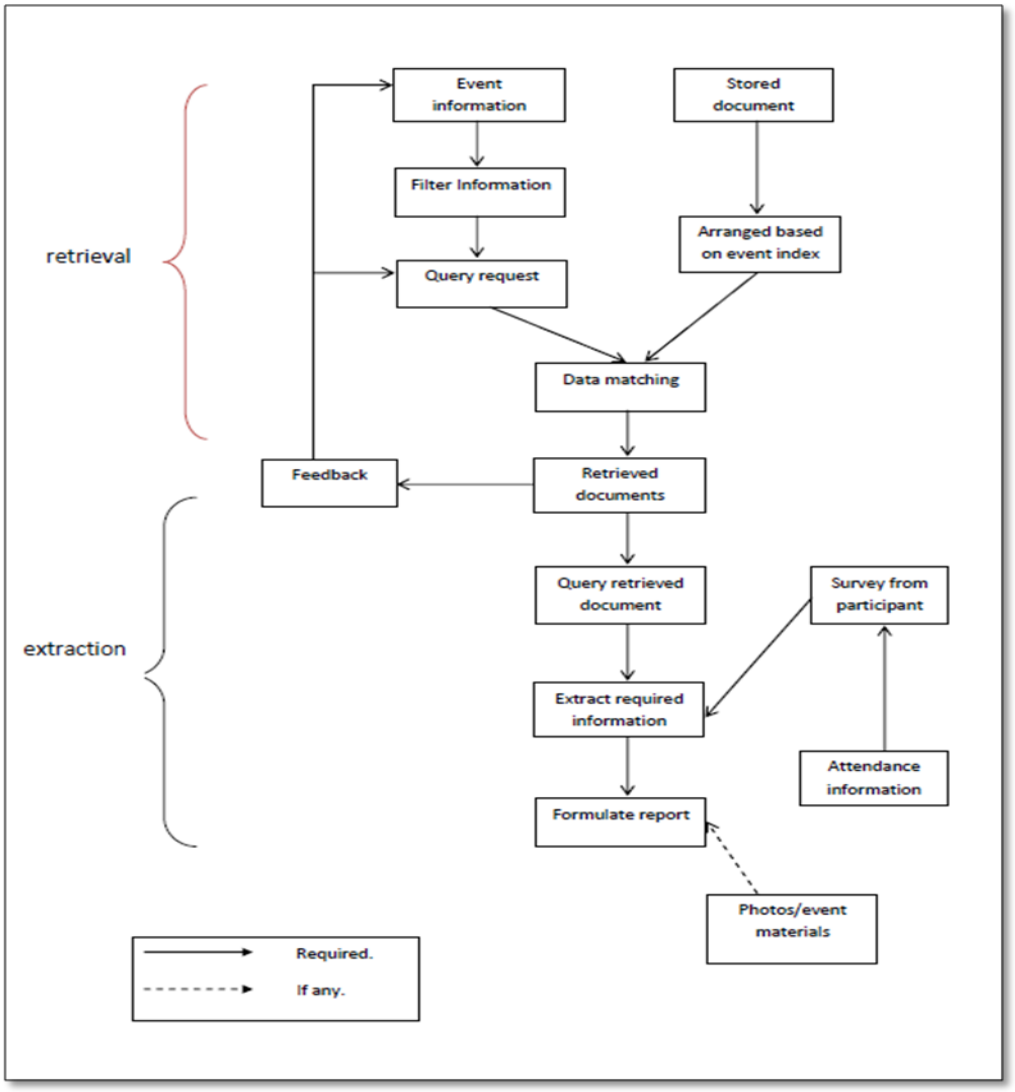

Figure 2: Data Retrieval and Extraction Flow

The statements of the data retrieval are as follows:

(1) $w \leftarrow k\ (\ w_1, \ldots\ , w_n)$
Where w and w_1 are word-valued variables and k is a keyword that produces from query $w_1, \ldots\ , w_n$ a particular word $w_1(x_1)\ w_2(x_2), \ldots\ w_n(x_n)$ where $w(x)$ is the x^{th} component of word w.

(2) insert (w, R) and delete (w, R)
Where w is a word variable and R a relation variable.

(3) $R \leftarrow S$
Where R is a relation variable and S is a relational variable or constant. The constant can be the specific date or title of event data going to be retrieved for report purposed.

(4) **for***w* in *R***do**<statement>
Where w is a word variable whose scope is local to the **for** statement and R is a relation variable. Any word variable assigned within the **for**statement is assumed local to that **for** statement. Statement is refers to the whole array records for the inserted event data.

(5) **if***s(w)* then <statement> [else <statement>]
Where w is a word variable and s is a sentence built from the arrays of words to retrieve particular record in most similar statements.

The pseudo code statements of the data retrieval are as follows:

```
input:
        M = set of data extracts from the requirements parameters
output:
        N = set of queries describing extracts

begin
        for each w in M
                (Nw)1 = leftseparator (w)
                (Nw)2 = rightseparator (w)
        end for

        S = sentence (N)
        P = null

        for each s in S
                addrecords(records(s), P)
        end for

        for each w in M
                for i = firstrecord(P) to lastrecord(P)
                        if (matches(w, records(i,P))
                                (Nw) i+2 = 1
                        else
                                (Nw) i+2 = 0
                        end if
                end for
         end for

end

display requested output
```

1.6. Indexing Information

Several indexing techniques such as signature file and inversion indices have been used widely nowadays. This research implement inversion indices where each document

represented by a list of keywords which described the contents. Fast retrieval can be obtained due to inverted keywords. Even the keywords stored alphabetically and may represent the same meaning; a list of pointer has been set to maintain the qualifying documents in the postings file.Let *M* be a composite keys of a record (represented as possible infinite vector) with positive integers M_i, where $\sum Mi \geq N$. To code integer word, $w \geq 1$ relative to *M* we arranged the record according to:

$$\sum_{i=1}^{N-1} Mi < w \leq \sum_{i=1}^{N} Mi$$

Some consideration has been made to extend indexes to provide greater functionality, and the effect this has on space requirements. The consideration to increase the retrieval performances being applied such as improving ranking methods and indexing on word sequences. Some of the inactive data (more than ten years event's date) will be stored in the archive in order to ensure only current data remains at top. However, the inactive data is not deleted due to historical data may be useful for future references when it comes to auditing purposes. The frequency of the similar event data being stored may rearrange the data index to be an active record and pull out back from the archive.

1.7. Searching and Formulate Report

There are several searching algorithms available such as linear search, binary search, blind search and heuristic search. This research used binary search to finds specified position of the element by using keyword. The algorithm compares the search key value with the element of array it stored and returns the information to formulate document. This formulated document then sorted into report. Regular expressions are widely used as a precise, succinct notation for specifying a text search, with a straightforward efficient implementation. The search is performed as if all possible suffixes of the string were tested for a prefix matching the records; the longest suffix containing a matching prefix is chosen, and the longest possible matching prefix of the chosen suffix is identified as the matching records.

Result and Discussion

Lots of information retrieval environments require indexing of the text by groups to increase consistency. This research has served the aim to achieve effective data management in constructing paperwork and formulate report by using the data retrieval and extraction. This model is succeed in reducing the information overload and eventually removes the data discrepancy. The problem of data losing and inconsistencies also minimized. The difference of data management with and without data retrieval is obvious. One of the strength using data retrieval in this research is the ability to represent the dependency between terms; this is the main issue to extract the report from the right paperwork. The linking between terms can represent a number of modified reports. As a result the constructed paperwork and report can

be produced simultaneously. At the same time the representation is flexible depending on user requirement in avoiding information loads. This is evident by the fact that data can be managed efficiently by this approach. The problems of data management are eventually minimized and improve productivity.

```
input:
    X : set of example strings

output:
    minimum FSA (finite state automation)

begin
    A = prefix thee acceptor from X
    for j = successor (firstrecord(A) to lastrecord(A))
            for I = firstrecord(A) to j
                    if matches(i,j)
                            retrieve(A,i,j)
                            extract(A)
                            exit(i-loop)

                    end if
            end for
    end for
    return A
end main
```

The data validated the approach by applying it to extract data from existing database containing a wide variety of data types. The extraction algorithm is applied to each set of event records and manually checked by their indexed if there is redundant similar event stored in the database.

Notification features added to acknowledge the user regarding updated on new data inserted into the record and require immediate response from authorize user. Authorize users consist of validator of the paperwork, treasurer executive (to revise budget needed) and also the applicant of the event's paperwork. This feature is an automated method for notifying users upon changes in some real-time data is disclosed. Treasurer executive is notified once the new application is requested, validator will be prompted once the treasurer executive has revised the budget and the user will be alerted once the paperwork is approved.

This research provides a managed device for a network that performs event notification such as follows:

(1) The markup language comprises an extensible markup language

(2) The event notification message including executable code that when executed causes at least one action to be performed

(3) The event related information including a location pointer to locate further information on the network related to the particular management event that was approved

(4) Communication logic that operates according to Transmission Control Protocol/Internet Protocol (TCP/IP)

(5) The event notification logic formulating the event notification message into a HyperTextTransfer Protocol (HTTP) post transaction for transmitting on the network via the communication logic.

Conclusion

Data management involves retrieving and extracting data from its warehouse whereby data is not simply being retrieved as performed by the search engine. This research project has successfully implemented the conceptual model to manage event's paperwork and report. The extracted data can be used to support decision making as high-level executives of the organization can easily review and provide feedback to the users. The inclusion of e-notification provides the ability to reach the reviewers easily as they will be notified through email when new submission is available. Since most mobile phone comes with email application, reviewers can receive the notification on their mobile phone as long as their devices are powered on and connected to the Internet. Hence, the process of managing paperwork and report is less hassle as everything is stored in the database and can be access easily at any time. However, there are lots of rooms for improvement. Hence, future works may include implementing hybrid Data Mining techniques in knowledge discovery.

Acknowledgements

We would like to take this opportunity to express my profound gratitude to Universiti Teknologi MARA and Ministry of Higher Education of Malaysia (KPM) for funding this research. This research is under the Research Acculturation Grant Scheme (RAGS) with file number 600-RMI-RAGS 5/3 (1/2014).

References

Aho, A. V., & Ullman, J. D. (1979).Universality of data retrieval languages.In *Proceedings of the 6th ACM SIGACT-SIGPLAN symposium on Principles of programming languages* (pp. 110-119).ACM.

Al-Nazer& Ahmed Ali (2006). Collaborative autonomous interface agent for personalized Web search.*Dissertation.King Fahd University of Petroleum and Minerals.*

Balakrishna, M., Werner, S., Tatu, M., Erekhinskaya, T., & Moldovan, D. (2016). K-Extractor: Automatic Knowledge Extraction for Hybrid Question Answering. *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (pp. 390-391). IEEE.

Clarke, C. L., & Cormack, G. V. (1997).On the use of regular expressions for searching text. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, *19*(3), 413-426.

Cubison, M. J., Coe, H., &Gysel, M. (2005). A modified hygroscopic tandem DMA and a data retrieval method based on optimal estimation. *Journal of aerosol science*, *36*(7), 846-865.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., &Zamir, O. (1998).Text mining at the term level. *Principles of Data Mining and Knowledge Discovery* (pp. 65-73). Springer Berlin Heidelberg.

Grishman, R. (2015). Information extraction. *Intelligent Systems, IEEE*, *30*(5), 8-15.

Iswarya, P., &Radha, V. (2014). Speech and text query based Tamil-English Cross Language Information Retrieval system. *Computer Communication and Informatics (ICCCI), 2014 International Conference on* (pp. 1-4). IEEE.

Lerman, K., Knoblock, C., & Minton, S. (2001, August). Automatic data extraction from lists and tables in web sources.In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* (Vol. 98).

Nagarkar, S. P., & Kumbhar, R. (2015). Text mining: An analysis of research published under the subject category ‘Information Science Library Science’in Web of Science Database during 1999-2013. *Library Review*, *64*(3), 248-262.

Nelson, Donald R. (2002). Method and apparatus for providing automated notification to a customer of a real-time notification system." *U.S. Patent.* No. 6,496,568.

Otegi, Arantxa, et al. "Using knowledge-based relatedness for information retrieval." *Knowledge and Information Systems* 44.3 (2015): 689-718.