Construction Inverted Index for Dynamic Collections Visualization in Thematic Virtual Museums System

Sajarwo Anggai[#1], Ivan Stanislavovich Blekanov[#2], Sergei Lvovich Sergeev[#3]

[#]Applied Mathematics and Control Processes, St. Petersburg State University
7-9 Universitetskaya Naberezhnaya, St. Petersburg, 199034, Russian Federation
[1]sajarwo@gmail.com
[2]i.blekanov@spbu.ru
[3]slsergeev@yandex.ru

***Abstract*— Thematic virtual museum system main goal is providing more relevant information based on user characteristic, behavior and desire. Thematic exhibition as main unique feature very considering on users' whole activities. Barry and Maria define creation of new knowledge, transformative and self-directed experiences, engagement with the full diversity of visitors, and transparency as the source of a viewpoint of the exhibition as museum specific evaluation criteria. Museum exhibition is not book on the wall rather than leading visitors to new attitudes, values and ideas even in very different cultural backgrounds or religious beliefs. The system with interactive information in the thematic exhibition which has been processing and delivering to the user will be leading to the better understanding by traveling in similar fields. In order to accomplish those tasks, this research have been performed and constructed thematic virtual museums inverted index to gain fastest retrieval document collections. The system is using special architecture and multi system indexing in thematic virtual museum inverted index. This index can manage collection in many forms of data structured, semi structured or unstructured which have processed by data access layer engine. In this paper showed a design and modification of the inverted index experimental result for thematic virtual museums system and a sample of exhibition visualization to get more precision by embedding method.**

Keywords— document collections; exhibition; inverted index; thematic virtual museums; visualization.

I. INTRODUCTION

Thematic virtual museum system main goal is providing more relevant information based on user characteristic, behavior and desire. Thematic exhibition as main unique feature is very considering on users' whole activities. Thematic concept can increase user experience, deep knowledge, discovery new related content, explore unexplored, and most important thing is not disturb users by providing irrelevant information.

Barry and Maria define the creation of new knowledge, transformative and self-directed experiences, engagement with the full diversity of visitors, and transparency as the source of a viewpoint of the exhibition as museum specific evaluation criteria. Museum exhibition is not a book on the wall rather than leading visitors to new attitudes, values and ideas even in very different cultural backgrounds or religious beliefs [1].

Interpretation of tangible and intangible cultural and natural heritage can be spread in various recent devices of information and communication technology to the end-users. The system initiative to serve users by providing accurate museum collections information which can be delivered through desktop, web or mobile application in real time. The system can help curator or museum administrator for research, clarify, grouping, visualizing and design thematic event exhibition in some cases.

John H Falk writes that most museum visitors mentioned that visitor go to museums in their free time in order to have fun and enjoy themselves or to see new interesting things in a relaxing and aesthetically pleasing setting [2].

The system with interactive information in the thematic exhibition which has been processing and delivering to the user will be leading to the better understanding by traveling in similar fields i.e. art, archeology, artefact and culture value. Bridging entities relationship by weighting and combining it in vector space also a part of this research.

In order to accomplish those tasks, this research has been performed and constructed thematic virtual museums inverted index to gain fastest retrieval document collections.

II. INDEXING SYSTEM

Inverted index usually known as inverted file is a data structure to store document-term pair which widely used in information retrieval field. Christopher D. Manning define Information retrieval (IR) in academic field study as a technique to find material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [3]. Inverted index often used to process many data form, i.e. structure, semi-structure and unstructured.

Previously front-end interface has been created and it can manage, grant and revoke data access from and to the system using data access layer. There are 105 museums with total

978-1-5386-1874-5/17/$31.00 ©2017 IEEE

24.608 collections from Ministry of Education and Culture, The Republic of Indonesia. Each collection in museum institutions have structure and unstructured data source. This data sources increase day by day depend on museum administrator/curator which managing every museum institution. Those data become hard to maintain just by using the recent technology of database system, therefore it is necessary to develop a system for extracting and processing data sources until become important information and useful for visitors [4].

The actual concern for managing big collections in museum institutions are gaining fast searching relevant information document collections in thematic virtual museums system, therefore this work has constructed an inverted index which aligned with thematic virtual museums information need [5].

A. Forward Index

Forward index task usually using document-term, the structure is very helpful for retrieving a set of information as a payload in order to get fast performance related to virtual museum item collections. The context or information as payload in this index automatically will be processing the documents after data access layer crawler service has finished the job schedule.

B. Inverted Index

An inverted index is a data structure that maps a word, or atomic search item, to the set of documents, or set of indexed units, that contain that word – its postings. Operational IR also concerns on block update speed, access speed, index size, dynamics and scalability as mentioned in [6]. The index data structure contains a pair of term-document or inverted list for each term stores a list of the documents *d* [7].

III. Thematic Virtual Museum Index

The system is using special architecture and multi system indexing in virtual museum inverted index. This index can manage collections in many forms of data structured, semi structured or unstructured which have processed by a data access layer engine.

There are several tasks to perform inverted index in this virtual museum system such as tokenizing, removing stop word, stemming and process to index structure.

Museum index assigns unique Id for each document in the collections as document identifier (docId). Construction of dictionary term makes this system fast to search every cooccurrence document in the specific term as a key. In order to gain the best performance this index used the Hash table to store terms to perform O(1) lookup time according to [6].

A. Embedding, Loading, and Storing

Embedding some information as payload in inverted index data structure is one of the main concern related to the specific area in virtual museum, this information will be stored together with inverted-list. There are several techniques for compressing inverted lists [3] [7]. The thematic virtual museum inverted index store data in file format which contain key-term and sorted *docId* by using compressing technology Varint encoding [8] for reducing file size on the disk. This system using Solid-State Drive (SSD) for fast read/write file compared to Hard Disk.

B. TF-IDF Weighting

Thematic virtual museum inverted index using *Term Frequency (TF)* and *Inverse Document Frequency (IDF)* [3] as base scoring method to determine important information in museum collections. The output of this statistical calculation will be sorted as ranking from highest to lowest.

C. Vector Space Model (VSM)

Similarity measure is better than traditional TF-IDF, the cosine measure which each document is assigned a numeric score indicating similarity with the query, and then the documents that score the highest are displayed as answers [7]. Salton and McGill define cosine vector similarity formula between query and document as shown [9].

$$cosine(q,d) = \frac{\sum_{k=1}^{t} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{t} w_{qk} \cdot \sum_{k=1}^{t} w_{dk}}}$$

The thematic virtual museum inverted index library implements this cosine similarity method as a part of modules as a complement for the fast searching document by comparing the similarity of users' query.

D. Index Adaptive Module

The system engine design which has been developed contain several independent program modules to perform the tasks, this index can be integrated or serve another complex adaptive ranking retrieval system without modifying virtual museum inverted index core.

Modification of standard inverted index structure and adding some functions allowing this system to process information need more complex by combining fastest term-document pair for building another retrieval structure.

IV. Experiment and Result

In order for testing this virtual museum inverted index, the engine has crawled museum information from "Collection Museum Registration System", Directorate Cultural Heritage Preservation and Museum, Ministry of Education and Culture, The Republic of Indonesia, which contain 24.608 items.

TABLE 1. DOCUMENT COLLECTIONS

No	Type	Item
1	Numismatic / Heraldic	2746
2	Ethnography	10107
3	Ceramics	4108
4	Fine Art / Handicraft	1516
5	Philology	2902
6	History	1394
7	Archeology	1153
8	Technology / Modern	682
Total		24608

Those document collections which have been obtained then stored in MongoDB database. The next step, the information is processed and assign unique *docId* documents into inverted index, the result as shown in Fig. 1.

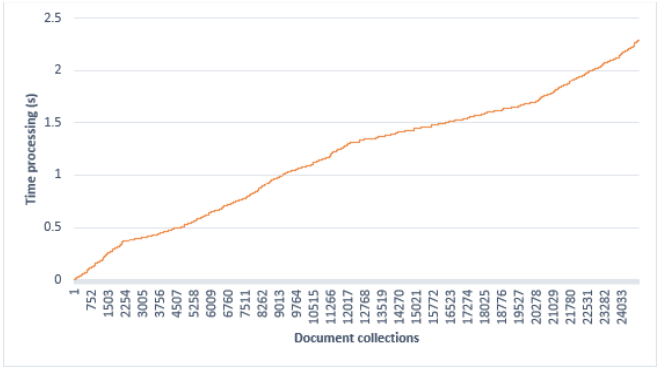

Fig. 1. Time processing document collections into inverted index

Processing documents and input into virtual museum inverted index for 24.608 collections require processing time approximately 2.29 seconds by using Intel Xeon Processor 2620v4, memory DDR4 32Gb, and Kingston Savage SSD 256Gb. This index intends to be used for serving all user's activities in front end server. Further, this inverted index has tested by using query document to document as this research main goal to approximate thematic collections related to user's activity and document collections.

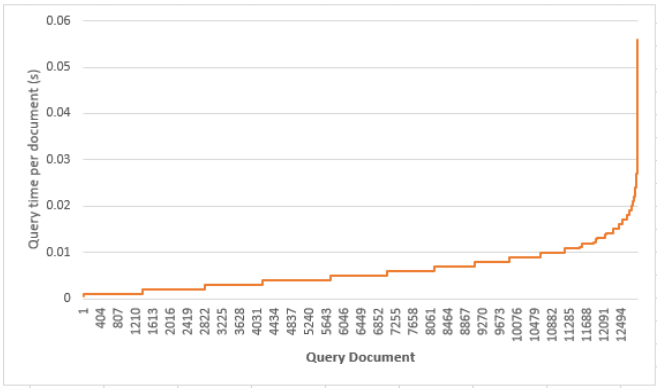

Fig. 2. Elapsed time processing document to document query

The result in Fig. 2, show that given a query "a document to a document" require time less than one microsecond per query, this is aligned with the main goal to design virtual museum inverted index data structure to serve more fast and relevant.

V. VISUALIZATION

A. *Collections in Map*

Information visualization is one of concern in thematic exhibition in order to deliver and serve high quality experience to the visitors.

Design map plan and leveling will help visitors to classify information which has previously received from information traveling visualization. Let visitor move to the area inside the map which could be formed to the internal or external exhibition with there are no predefine items have assigned. Providing more space as a proportion to give object transformation is considerable as part of esthetics. The room design maybe contains object on the floor, wall or in some neared-location or can be reached by users as shown in Fig.3.

Fig. 3. Design Exploration Room for Collection Exhibition

Give the floor planning with good navigation will make users easier to use this system. In virtual it can be providing area by grouping the object, even this particular area didn't exist on the real museum institution.

B. *Dynamic Collection Visualization*

When the users visiting a virtual museum, the system should determine what is the information trying to convey? what the motivated topics which make them enjoy and satisfied? In this situation, thematic virtual museum trying to create the system to maintain interactive information. The collections will appear after automatic user query have been processed. The exhibition which has been displayed should be drawn users' attention and make more attractive by triggering with relevant collections visualization.

The degree of similarity will be leading user from room to room or an area to area of exhibition as shown in Fig. 4. It can be pre-labeling data as a temporary or a permanent exhibition for each object in spread area. The system will be responsible for redisplaying contents by marking user traversed area.

Fig. 4. Dynamic Collections Area

Creating specific functional area design for scheduled thematic event exhibition is recommended to perform flexible objects which have determined include newly arrived collection

in some museums. Along the process, user respond and time spent in the room or area where collections appeared must be recorded as a feedback parameter. User also can be projecting some group of collections base on a manual query to the system with or without their base information.

The dynamic collection will collect visitor visit experiences in order to be processed near real time or next time when visitors will come back again on this system. Trying to understand the visitors to far and more complex by defining some variables is very useful in background inverted index process. Quantitative measures such as demographics provide too little information about visitors in relation to museums to be a useful variable for describing and understanding the museum visitor experience [2].

C. User Interaction

User interaction always recorded by the system on the fly and this information in near real-time will be automatically processed in the background. The result of those process will generate more accurate information related to visitor characteristics, behavior and desire. Defining embedded information into inverted index is helpful for delivering more accurate exhibitions in the specific fields.

When users are walking around the space, the system provides a feature to control some object which has an ability to be controlled, this approach is mean to increase user sensitivity and reaction to gain new knowledge or discovery. Design direct or indirect appearance of the objects can be projected over a period of time. Understanding user needs and wants is critical for exhibition space to be not only attractive and effective for display but also accessible [1].

Working to design ideal interface to the end-user is very crucial things, affected visitor will return another time if visitors satisfied using the application or will be leaving because some of poor user interface design, those all approximation for design user interface visualization.

VI. Conclusion And Future Work

In this paper showed a design and modification of the inverted index experimental result for thematic virtual museums system and a sample of exhibition visualization to get more precision by embedding method. This research will continue to implement this index from contextual to concept by using latent semantic indexing (LSI) projection in future work and improving user interface for more attractive.

Acknowledgment

The authors would like to thank Dr. Gatot Hari Priowirjanto for his help and support on this research.

This work is supported by Department Technology of Programming, Applied Mathematics Control and Processes Faculty, Saint Petersburg State University, Southeast Asian Ministers of Education Organization (SEAMEO), Southeast Asian Ministers of Education Organization Regional Open Learning Centre (SEAMOLEC), Beasiswa Unggulan, Ministry of Education and Culture Republic of Indonesia.

References

[1] B. Lord and M. Piacente, Manual of Museum Exhibitions 2nd Edition, Maryland: Rowman & Littlefield Publishers, 2014.

[2] J. H. Falk, Identity and the Museum Visitor Experience, Walnut Creek, Calif: Left Coast Press, 2009.

[3] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, New York, NY, USA: Cambridge University Press, 2008.

[4] S. Anggai, I. S. Blekanov and S. L.Sergeev, "Design Muntoi web-based framework and search engine analytics for thematic virtual museums," *2015 4th International Conference on Interactive Digital Media (ICIDM),* pp. 1-5, 2015.

[5] S. Anggai, I. S. Blekanov and S. L.Sergeev, "Management and information processing at virtual museum," *2014 IEEE 4th International Conference on System Engineering and Technology (ICSET),* pp. 1-5, 2014.

[6] D. Cutting and J. Pedersen, "Optimization for dynamic inverted index maintenance," *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 405--411, 1990.

[7] A. Moffat and J. Zobel, "Self-indexing Inverted Files for Fast Text Retrieval," *ACM Trans. Inf. Syst.,* pp. 349--379, 1996.

[8] D. Lemire and L. Boytsov, "Decoding billions of integers per second through vectorization," *Software—Practice & Experience,* vol. 45, no. January 2015, pp. 1--29, 2015.

[9] G. Salton and C. Buckley, "Term Wweighting Approaches in Automatic Text Retrieval," *Department of Computer Science, Cornell University,* 1987.