Historical Information Retrieval Model based on Period Query Using Time-Variant Extraction of Historical Object

Jun Lee, Yong-Hee Jang, and Yong-Jin Kwon

Abstract—**In providing historical information, current searching service requires users to do repetitive information retrieval and filtering process by supplying search results based on simple keyword query that does not reflect the characteristic of historical information. If the information retrieval service reflecting the feature of time-variant is provided, users would be able to save time and reduce unnecessary effort that is repetitively consumed to acquire desired historical information. Moreover, users can get the systemic retrieving result of historical information reflecting the phrase of the times. Hereupon, in this paper, we suggest theoretical Basis for historical information retrieval and provision by analyzing common character of historical information and defining historical object. Furthermore, we suggest a historical information retrieval model consist of three-classes that are 'Collecting Historical Information, Extracting Time-Variant and structuralization of Historical Information, The Information-Providing Space' to provide users with historical information based on particular period considering one of the common character of historical information. Suggested model, through Crawler and Wrapper, automatically collects historical information from Web and restructure collected historical information to time-series considering the aspects of 'period' and 'relation' that the historical information has. Finally, it carries out structuralization based on correlation of historical objects. Also, by retrieving historical information in particular period using 'Period Query', which can search information correspondent to time-variant, one of the features of historical information, users can obtain search results that dynamically change according to the chosen particular period, although they searched the same historical keyword. Therefore, users can effectively retrieve information by reducing complicated retrieve process when they look up historical information which has the feature of time-variant. In addition, by going through the structuring process, we provide not only required historical information but also additional information such as correlation of historical objects by information graphics that are intuitive user interface. Thus, users could effectively search and acquire historical information.**

Index Terms—**Period query, time-variant, retrieval model, historical object, structuralization.**

I Introduction

With the advent of the Internet and the Web, vast amount of information has been accumulated on the Web. In these flows, historical information is also not an exception. Digitized a variety of ancient documents and book can be viewed easily on the web. Historical information has description form based on time flow and distinctiveness of contents that described fact in the past. Therefore, a differentiated information service provision is required as compared general scholar information. [1], [4]

However, traditional retrieval service or historical information provision service provide simply keyword and directory retrieval. Thus, it does not effectively satisfy requirements that occur in various forms of information retrieval. For example, after retrieving historical information by 'The Second World War', suppose users want to obtain additional information such as people associated with The Second World War, circumstances before and after The Second World War era. In such cases, if using current information retrieval service, users must go through repeat many information retrieval and filtering process, for instance, re-retrieving relevant people, reading many retrieved pages one by one. Also, the historical information, for example the phases of the times of England in 16C, the information retrieval process becomes more complex. Because it can not be specifies retrieval query by keyword. Furthermore, the relationship of people or events changes passage of time. However, current retrieval services do not corresponding these changes and have limitations that provide the same retrieval result for the same keyword. Thus, there are many retrieval requests through collective intelligence or collaboration for historical information that can not get using current retrieval service. To solve this problem, time-variant, which is one of the various features of historical information, reflecting the retrieval service is required.

In this paper, we identify characteristics of historical information and then define historical information in historical information retrieval to provide historical information reflecting time-variant to users. The historical information retrieval model, suggested in this paper, as possible retrieval corresponding time-variant of historical information by users period query, consist of three-classes that are 'The Historical Data Space, The Extracted Time-Variant & Structured Information Space, The Information-Providing Space'. In order to ensure feasibility and practicality of suggested model, in this paper, '500 years of Choson era in Korea' as the target, we implement on web browser applying suggested automatically collecting of historical information. The remainder of this paper is organized as follows. Section 2 explains The Historical Information Retieval Model which we propose, and Section 3 discusses the importance of the proposed model and concludes the paper.

II Proposal of Historical Information Retrieval Model

The historical information retrieval model, by applying the characteristics of historical information, provides

Manuscript received April 12, 2012; revised May 25, 2012.

The authors are with Department of Telecomm and Info. Engineering, Korea Aerospace University, Goyang-city, Gyeonggi-do, Korea.
(Tel.: +82-2-300-0458; fax: +82-2-3159-0458. e-mail: jun@grrc.kau.ac.kr)

historical information and relevance historical information efficiently for users who want to retrieve. Thus, the model supports the user who can be easier to acquire historical information. This retrieval model is modeling that is series of process about historical information acquisition and processing on the web, extracting and structuring, information display based on graphic interface. As ‘Fig. 1’, Suggesting model consist of three-classes.

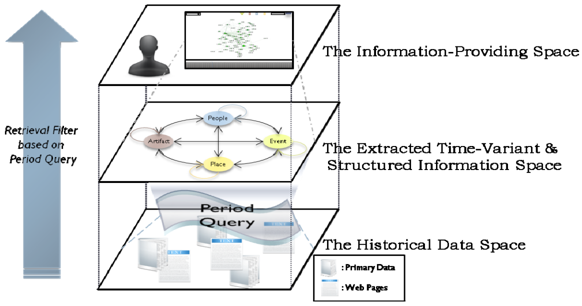

Retrieval Filter based on Period Query
The Information-Providing Space
People
Artifact
Event
Place
The Extracted Time-Variant & Structured Information Space
Period Query
The Historical Data Space
: Primary Data
: Web Pages

Fig. 1. Suggested historical information retrieval model.

A. The Historical Data Space

Historical information, including factor of period, happened in the past is a record containing the information. Ancient documents, a kind of historical information, are digitized, there are on the web, and the other historical information such as web pages that reinterpreted and recorded historical facts by the user exists on the web. In this space, in order to extract time-variant of historical information, define historical object by applying the characteristic of historical information. Also through defined the historical object, historical information is collected, pre-processed, and built a database.

B. Definition of Historical Object

Describing the form of historical information, as we’ve seen in Section 1, has characteristic that describe with relevance of historical information, as an example of name, event. For example,

- *King Taejo* retransferred its capital to *Gyegyeong*, the capital of the Goryeo dynasty that Choson had replaced. In the 5th year of *Taejong* (*1405*), the capital was moved back to *Hanyang* (Seoul), and a new palace was built east of the principal palace, '*Changdeokgung*.' – Encyclopedia of Korean, Changdeokgung
- His Successor *Won-Gyun* was killed and defeated by the Japanese military at *the Battle of Chilchonryang* on *July, 1597*. As a result, … Admiral *sun-sin LEE* fought 133 ships of the enemy, and he has achieved victory that defeat 31 ships *(Battle of Myeongnyang)*. – Encyclopedia of Korean, sun-sin LEE, as those described above.

In case of first example, it is described with the name *King Teajong* who carried out crucial role in composition of *Changdeokgung*. The second example, by describing the achievements of *sun-sin LEE*, it is described with the name *Won-Gyun*, the place *Myeongnyang* and the event *Battle of Chilchonryang*. We know, through these examples, historical information is described with particular types of relevance historical information.

Also required retrieval objects, as historical information have characteristic that is limited particular types. For example, ‘www.history.com’ [2], popular website as providing US historical information, provides historical information using featured topic classification. It is categorized by people, places, events, eras, et al, and ‘Korea History Information Integration System’ [3], popularly used in Korea historical information retrieval system, is categorized by people, place, artifact, book, research, et al. To summarize it as occupancy rate, it is made up of name(60%), artifact(25%), place(10%),etc(5%). Through the above facts, we know that historical information is described and retrieved mainly focus on people, artifact, place and event. Hereupon in this paper, we regard historical object as historical information. Suggested and defined historical object is as follows:

- *People: Described in the history of the name refers to a real person.*
- *Artifact: Left over from the past that type of tangible / intangible cultural asset*
- *Place : existence of a place in history*
- *Event : root or main event as flow of history*

C. Collecting and Pre-Processing of Historical Information

In order to retrieve fast and accurate historical information corresponding to user’s query, historical information should be collected in advance, and pre-processing of the collected historical information is required. Historical information, historical objects are used as retrieval query, through crawler, automatically is collected from traditional search engines. Also through wrapper, digitized ancient documents are collected. Collected historical information, through the above process, is performed classification and text pre-processing in order to correspond to user query such as period and keyword. Classified historical information, through sort by period and historical object, is stored database. In the text pre-processing, through removal of common nouns that are not used for structuralization and determination of frequency that reduces computation time of correlation algorithm, pre-processed historical information is stored database.

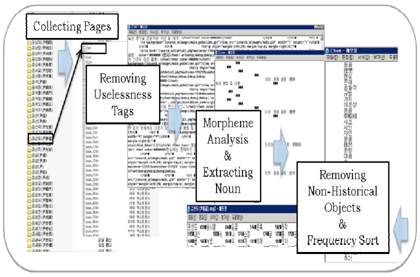
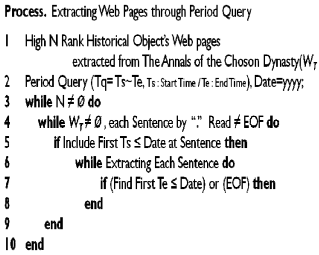

Collecting Pages
Removing Uselessness Tags
Morpheme Analysis & Extracting Noun
Removing Non-Historical Objects & Frequency Sort

Fig. 2. Example of pre-processing of collected web pages



```
Process. Extracting Web Pages through Period Query

High N Rank Historical Object's Web pages
        extracted from The Annals of the Choson Dynasty($W_T = \{W_1, W_2, \ldots, W_k\}$);
Period Query (Tq= Ts~Te, Ts : Start Time / Te : End Time), Date=yyyy;
while N ≠ Ø do
    while $W_T$ ≠ Ø , each Sentence by "." Read ≠ EOF do
        if Include First Ts ≤ Date at Sentence then
            while Extracting Each Sentence do
                if (Find First Te ≤ Date) or (EOF) then
            end
    end
end
```

Fig.3. Process of extracting historical information corresponding period query at web pages

D. The Extracted Time-Variant and Structured Information Space

The extracted time-variant & Structured information space is the space, which extracts historical information and structure extracted historical information corresponding to user's query. In this paper, in order to extract time-variant which is one of the various features of historical information, defines period query that is retrieval query based on period. Also extracted historical information, using period query or keyword query, through applying correlation algorithm, is structured based on extent of correlation between historical objects. Also extracted historical information, using period query or keyword query, through applying correlation algorithm, is structured based on extent of correlation between historical objects. Then in this space, using context factors, define detail relationship between historical objects which are structured by degree of correlation.

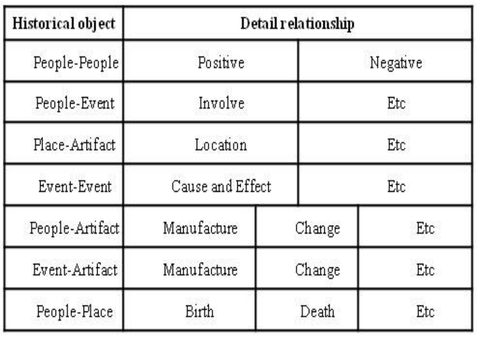
E. Characteristic and Definition of Historical Information

Historical information provision service is required the provision of additional services which are differentiated from the general academic information because of characteristic such as uniqueness of contents and morphological of data, so that it considered based on information requirement from the perspective of users. Accordingly, in this paper, we will define clear historical information and examine characteristics of historical information in order to provide efficient historical information. First, in order to define the historical information, looking at the dictionary definition of the word 'history',

- All the things that happened in the past, especially the political, social, or economic development of a nation.
- The events that took place from the beginning and during the development of a particular place, activity, institution etc.
- A record of something that has affected someone *or* been done by them in the past.

Dictionary of Contemporary English, Pearson[6]

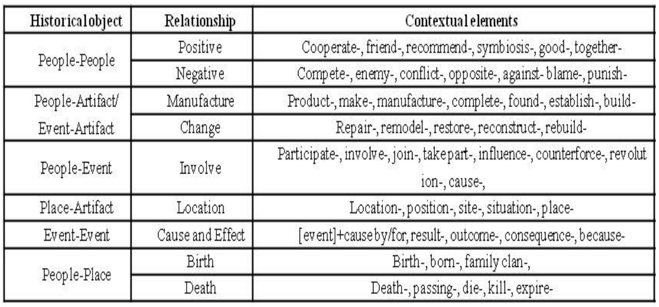
TABLE I: DEFINITION OF DETAIL RELATIONSHIP BETWEEN HISTORICAL OBJECTS



Historical object	Detail relationship		
People-People	Positive	Negative	
People-Event	Involve	Etc	
Place-Artifact	Location	Etc	
Event-Event	Cause and Effect	Etc	
People-Artifact	Manufacture	Change	Etc
Event-Artifact	Manufacture	Change	Etc
People-Place	Birth	Death	Etc

Defined as above, we can confirm that. In this paper, we define historical information based on the word 'history' to use precisely meaning of historical information for future work. From definition of the word 'history', two characteristic can be detected in common. One is meaning recorded facts based on some time in the past, another is describing particular objects such as people, event or artifact. In other words, it means a facts or records about objects existing at any particular time in the past. Hereupon, in this paper, the historical information is defined as 'information about objects in the correspondence at past a certain period' and we use this definition in the development process.

TABLE II: DEFINITION OF CONTEXTUAL ELEMENTS (IMPLYING DETAIL REALTIOINSHIP)



Historicalobject	Relationship	Contextual elements
People-People	Positive	Cooperate-, friend-, recommend-, symbiosis-, good-, together-
	Negative	Compete-, enemy-, conflict-, opposite-, against- blame-, punish-
People-Artifact/ Event-Artifact	Manufacture	Product-, make-, manufacture-, complete-, found-, establish-, build-
	Change	Repair-, remodel-, restore-, reconstruct-, rebuild-
People-Event	Involve	Participate-, involve-, join-, take part-, influence-, counterforce-, revolution-, cause-,
Place-Artifact	Location	Location-, position-, site-, situation-, place-
Event-Event	Cause and Effect	[event]+cause by/for, result-, outcome-, consequence-, because-
People-Place	Birth	Birth-, born-, family clan-,
	Death	Death-, passing-, die-, kill-, expire-

F. Proposal of Period Query

The period query, using in this paper, be simply created only the user to choose a certain period. This query is a retrieval query that is not form of keyword query and based on period which historical information has. Period query can search to target a certain period. Therefore, period query is easy to retrieve time-variant of historical information which has characteristics that meaning of information and relationship of historical objects vary depending on time-variant, for instance, relationship people in 1630, the phases of the Japanese invasions of Korea. Also, Period query can retrieve the historical information that users don't know to retrieve keyword and can't be set keyword. Because of period query is generated by choosing a certain period. Hereupon, in the paper, we can retrieve not only keyword retrieval of historical information, through period query, but also time-variant that historical information has.

G. Structuralization of Historical Information

As we've seen Section 2.1.1, the historical object describes relevance of another historical object. Thus, in the paper, by efficient providing historical information, retrieval results of historical information is provided not retrieval results based on ranking of webpage but structured retrieval results that determine historical objects corresponding retrieval query. To do this, it is necessary that extracted historical information based on retrieval query apply appropriate determination of correlation algorithm, for example, 'Euclidean Distance' algorithm (1).

$$\sqrt{}\,((\alpha_1 - \beta_1)^2 + ... + (\alpha_N - \beta_N)^2\,) \;=\; \sqrt{}\,\Sigma\,(\alpha_\iota - \beta_\iota)^2 \quad (1 \le \iota \le N) \qquad (1)$$

Process of determination of correlation is fulfilled based on exists common historical objects in the same document. In addition, in this process not only determines relevance of historical objects but also, for structured results, define detail relationship of historical objects that connected to the relationship which means. Detail relationship of historical object is determined based on frequency of word that holds the meaning of relationship at context, and we define ten kinds of detail relationships for four kinds of historical objects.

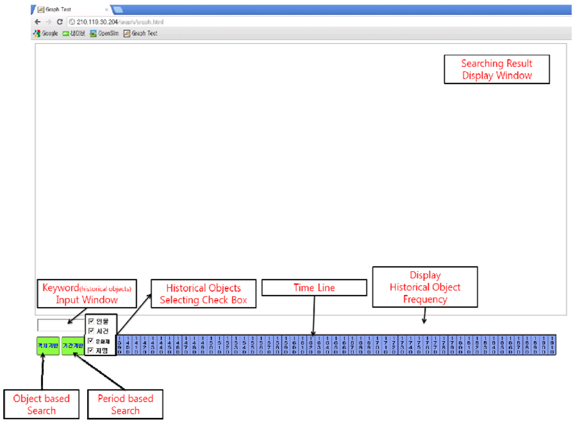

Searching Result Display Window
Keyword(historical objects) Input Window
Historical Objects Selecting Check Box
Time Line
Display Historical Object Frequency
Object based Search
Period based Search

Fig. 5. Composition of user interface.

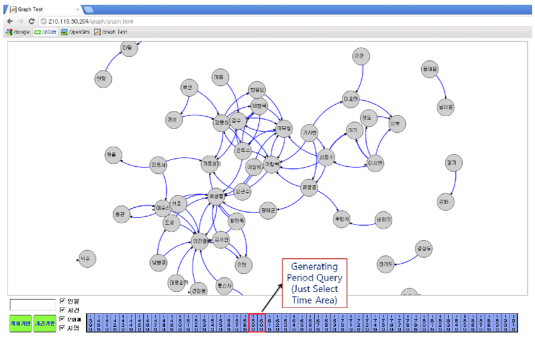


Fig. 6. Example of searching result by period query(1590-1610)

H. The Information-Providing Space

The information-providing space is the interface space to provide historical information for users intuitively. At this historical information is extracted and structured historical information corresponding users query in the Extracted time-variant & Structured information space. In this space, historical information is expressed form of graphic interface between relationships of structured historical objects instead of list form of webpage in traditional web search engines. Hereupon, user can directly or indirectly grasps, though between relationships of structured historical objects, the phases of the times or relationship people that user wants, instead of information filtering that user opens and reads webpage one by one. Also, by expressing frequency of selected or retrieved historical object on timeline, user immediately retrieves and catches historical information by selecting major involving event or period of retrieval object. Moreover, user can directly retrieves historical objects using keyword query as well as using period query, and to retrieve and structure only the desired type of historical objects for user, user can select the type of historical objects.

I. Composition of Graphic User Interface

Vertex of graph is corresponding to each historical objects, connected edge depending on existence of correlation. At this edge represent information depending on detail relationship between historical. Furthermore, for efficient retrieving historical information, two retrieval methods are provided. The one is 'historical objects retrieval based on period' using period query, another is 'period filtering retrieval based on historical objects' using keyword that is historical objects.

Moreover, users can select the type of historical objects which users want to structure. Also, there is a timeline that can intuitively generate period query. The timeline is not merely generating period query but also representing space for frequency of historical objects which currently being searched (Fig. 5, 6, 7, 8).

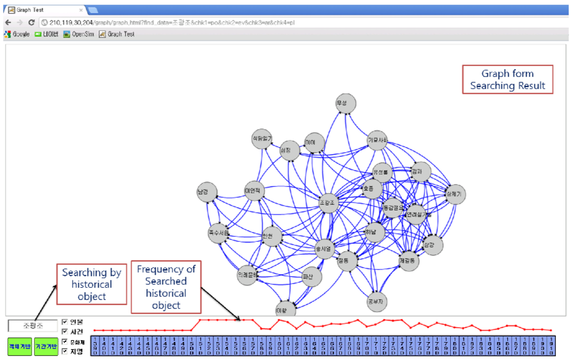

Graph form Searching Result
Searching by historical object
Frequency of Searched historical object

Fig. 7. Example of searching result by historical object(jo gwang-jo)

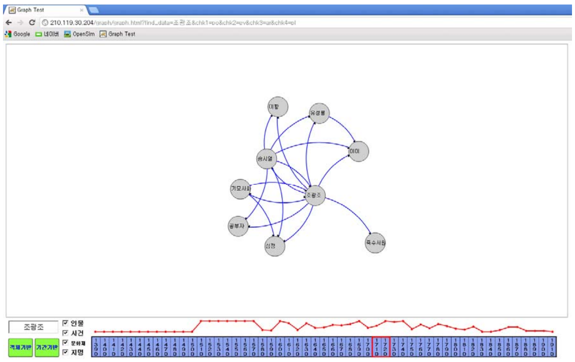


Fig. 8. Example of searching result by period query with historical object (1700-1710&Jo gwang-jo)

III Discussion and Conclusion

In this paper, we propose historical information retrieval model that extracts historical information using period query which is based on period, provides historical information in the form of relationship which is automatically determined. This retrieval model has three important characteristics in terms. First, the process for acquiring historical information of users is that minimizing. In other words, users simply select the period without specific or enter keyword by generating a query, the historical and relation information can be obtained. Also, because of characteristic of historical information, it is often to retrieve relation historical information after retrieve that user wants historical information. For this reason, in this paper, by automatically determining and providing historical and relation information, the user's historical information acquisition and filtering process is greatly reduced. In addition, user that did not know the relevant historical information can be provided has the advantage. Second, intuitive perception of historical information is that it is possible. Users to obtain the desired information, all of the pages user look at the process of retrieving information has to be done, because traditional search engines provide just list of web page thumbnails. However, user interface, in this paper, by having the interface that intuitive grasp of the human form of the graph, relevant information at a glance has the advantage. Third, the retrieval performed by traditional search engines on the proposed model combines the development of retrieval services is possible. The achieve this, first, each search engines built into the index for the keywords in the database additional period to build a database index, and if it have additional database that calculate the relevance of each keywords, the new search service that can retrieve time-variant of information will be possible.

ACKNOWLEDGEMENTS

This work was supported by the GRRC program of Gyeonggi province. [Augmented Broadcasting Service Platform Development for Ambient Service]

REFERENCES

[1] N. Garera and D. Yarowsky, "Structural, transitive and latent models for biographic fact extraction," in *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 300–308, 2009.

[2] History. Com, A E Television Networks, LLC, URL: [Online]. Available: http://www. history.com/

[3] Korea History Information Integration System, URL: [Online]. Available: http://www. koreanhistory.or.kr/

[4] M. Yamamoto, Y. Takahashi, H. Iwasaki, S. Oyama, H. Ohshima, K. anaka, "Extraction and Geographical Navigation of Important Historical Events in the Web," in *Proc. of W2GIS 2011*, LNCS 6574, pp. 21–35, 2011.

[5] J. M. QUEEN, "Some Methods for Classification and Analysis of Multivariate Observations," In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297

[6] The Longman Dictionary of Contemporary English Online - LDOCE, URL: [Online]. Available: http://www.ldoceonline.com