3rd Korea-Japan Joint Workshop on Pattern Recognition (KJPR2008) , Seoul, Korea, pp.101-102 (2008. 11).

A Prototype of Information Retrieval System to Read Mokkans

Jun Takakura†, Somayeh Sherini†, Akihito Kitadai†, Masaki Nakagawa†,
Hajime Baba†† and Akihiro Watanabe††

†Tokyo University of Agriculture and Technology
Naka-cho, Koganei, Tokyo, Japan
E-mail: j.takaqula@gmail.com
††Nara National Research Institute for Cultural Properties
Nijo-cho, Nara, Japan
E-mail: hajime@nabunken.go.jp

Extended Abstract

We present a prototype of information retrieval system to support reading historical *mokkans*.

"*Mokkan*" is a generic name given to a kind of Japanese historical documents that have handwritten characters on wooden tablets. Over 320,000 historical *mokkans* made and used in the 8th century have been excavated in Japan. Most of the *mokkans* were used as shipping tags. For that reason, we can draw valuable information of the material flow, local product names, place names and personal names at the time from the translations of the *mokkans*. However, decoding the *mokkans* with damaged or broken parts is difficult even for expert readers (archeologists and historians). Especially in the process to read unreadable character patterns to complement the translations, they have to obtain large amounts of information from another ancient documents or books of archaeology/history.

We can find some recent researches of information processing for decoding historical documents [1] [2]. Especially, quick and easy-to-use information retrieval to support reading fragmented text on historical documents has shown its effectiveness.

Unfortunately, stained and damaged *mokkans* have a number of missing or misused characters, and transpositions of words. To support reading such *mokkans*, we have proposed Extended Aho-Corasick method (EAC) [3]. The Aho-Corasick method that is the base of the EAC is a fast text retrieval method [4]. The EAC provides robust text retrieval even if the keywords contain missing, misused and wrong ordered characters.

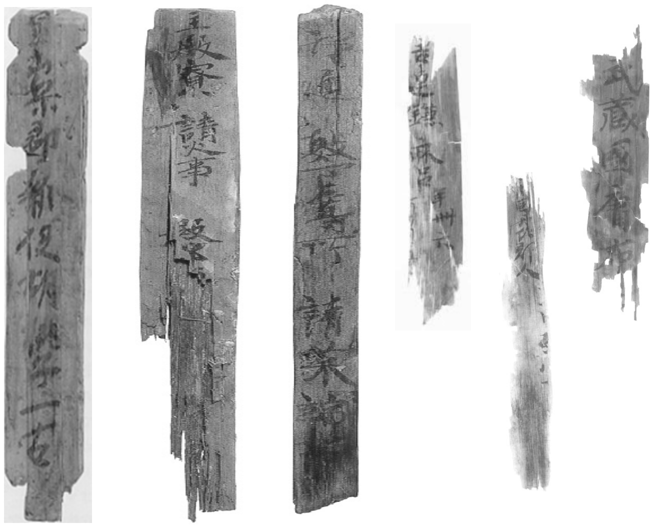
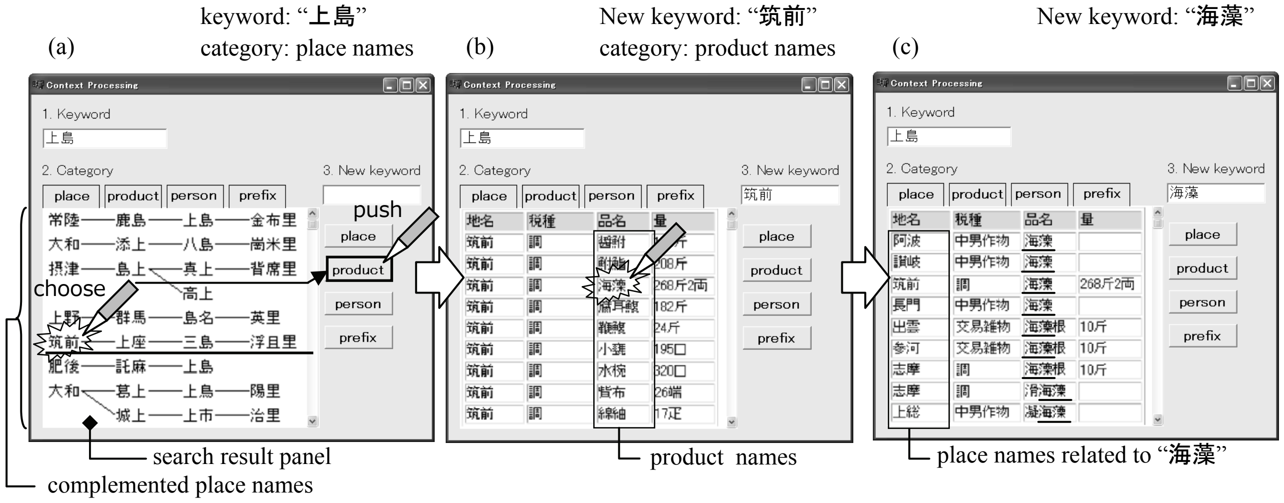


Figure 1. Damaged Mokkans

In this research, we made a prototype of information retrieval system using EAC (EAC-system). The EAC-system accepts incomplete translations of *mokkans* as the keywords. Also, we constructed databases of four categories: local product name, place name, personal name and prefix (dignity) of the personal name. From these databases, the users of the EAC-system can obtain text strings similar with the keyword.

Figure 2 shows an example of crossover information retrieval between the categories of place name and product name. The keyword at the start point of the processing is “上島”. By choosing the “place (place name)” tab on the search result panel, the EAC-system displays the complemented place names consists of state-county-city-town parts (Figure 2-(a)). In this example, the user of the EAC-system who is interested in one of the place name “筑前-上座-三島-浮且里” (that contains the characters of the keyword) chooses “筑前 (the state name)” and pushes the “product” button on the right side of the search result panel. Therefore, product names related to the state name and the additional information of the product names are shown in the search result panel (Figure 2-(b)). By choosing one of the product manes (the “海藻” means seaweed), the user can obtain other place names related to the product names (Figure 2-(c)). Such information retrieval is useful since a shipping tag has to show both “From” and “To” place names.



(a) 1. Keyword: 上島

常陸——鹿島——上島——金布里
大和——添上——八島——崗米里
摂津——島上——真上——背席里
　　　　　　　　高上
上野——群馬——島名——英里
筑前——上座——三島——浮且里
肥後——託麻——上島
大和——葛上——上島——陽里
　　　　城上——上市——治里

(b) 1. Keyword: 上島　3. New keyword: 筑前

地名	税種	品名	量
筑前	調	醤鮒	[illegible]
筑前	調	鮒鮨	208斤
筑前	調	海藻	268斤2両
筑前	調	鱶耳鰒	182斤
筑前	調	鞭鰒	24斤
筑前	調	小甕	195□
筑前	調	水椀	320□
筑前	調	貲布	26端
筑前	調	綿紬	17疋

(c) 1. Keyword: 上島　3. New keyword: 海藻

地名	税種	品名	量
阿波	中男作物	海藻	
讃岐	中男作物	海藻	
筑前	調	海藻	268斤2両
長門	中男作物	海藻	
出雲	交易雑物	海藻根	10斤
参河	交易雑物	海藻根	10斤
志摩	調	海藻根	10斤
志摩	調	滑海藻	
上総	中男作物	凝海藻	

Figure 2. An example of the context processing

Key words Mokkan, Historical document, Support system, Information retrieval.

Acknowledgement

This research is being supported by Grant-in-Aid for Scientific Research (S)-20222002 and Grant-in-Aid for Young Scientists (B)-19720202.

References

[1] M.S. Kim, K.T. Cho, H.K. Kwag and J.H. Kim, “Segmentation of Handwritten Characters for Digitalizing Korean Historical documents”, *Proc. 6th DAS*, Florence, Italy, pp. 114-124, 2004.

[2] B. Gatos, I. Pratikakis and S.J. Perantonis, “An Adaptive Binarization Technique for Low Quality Historical Documents”, *Proc. 6th DAS*, Florence, Italy, pp. 102-113, 2004.

[3] A. Kitadai, K. Nishijima, K. Saito, M. Nakagawa, H. Baba and A. Watanabe, “Context Processing to Read Text on Damaged Wooden Tablets”, *Proc. 10th International Workshop on Frontiers in Handwriting Recognition*, La Baule, France, vol. I, pp. 581-586, 2005.

[4] A.V. Aho and M.J. Corasick, “Efficient string matching: An aid to bibliographic search”, *Commu. of ACM*, vol. 18, no. 6, pp. 333-340, 1975.