ACTA MONGOLICA
CENTRE FOR MONGOL STUDIES
National University of Mongolia

Tom. 7 (291) 2007 023–028

# Memory-based dictionary browser for Mongolian toponym

*Purev Jaimai*
*Odbayar Chimeddorj*
*Ravdan Enhbayar*

## Introduction

Various searching systems are the parts of commonly used computer programs. The searching systems provide facilities to find and retrieve needed materials from a large scale of information in a short time. So, the development of such system is still used in electronic dictionaries and phonebooks as well as NLP [1, 6] and Internet environment. However, the fundamental methods for searching systems were found earlier, the new systems are developed in the way that these methods are appropriately used for data structures and features of new systems. We have developed a search program for the electronic form of 8 volumes dictionary of Mongolian toponym.

Primary data used by our browser, comprises 232,495 place names in total, which is about 19,353 Mbytes in size [7]. The number of word tokens in the data is 1,852,090 and 138,591 of them are headwords with extent of 12,905 Mbytes.

The primary data has 5 fields including Mongolian place names and their transcription in Roman, the type name, aimak and somon[1] names where the place is settled (see Table 1).

*Table 1. The primary data structure of the browser*

| Place name | Transcription in Roman | Type | Aimak name | Somon name |
|---|---|---|---|---|
| Аав толгой | Aav tolgoy | Толгой | Төв | Сэргэлэн |
| Их авдрант уул | Ih avdrant uul | Уул | Хөвсгөл | Галт |
| Шагшуурга цав | Shagshuurga tsav | Цав | Өвөрхангай | Сант |

[1] Aimak is the largest administrative division of Mongolia, while somon is a territorial administrative unit subordinate to an aimak.

The browser was written in C# language, thus it supports Unicode environment. Regarding the browser structure and design, it consists of following four parts: (1) for loading the primary data and head words into memory, (2) inputting the words to search, (3) searching, and (4) displaying retrieved data (see Figure 1). The structure and function of each part were discussed in the next sections in detail. Furthermore, the retrieved data could be stored into .doc or .txt file, and printed.

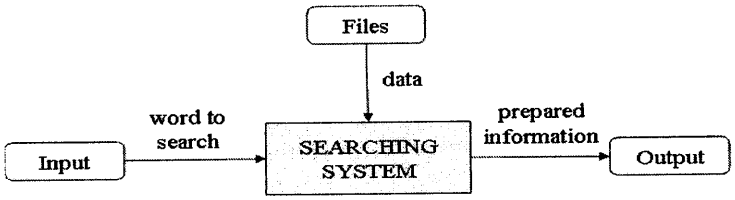



*Figure 1. The general structure of MonPlace browser*

## The browser structure and design

The hash-table and concordance methods [1,2,3,4,5] are implemented in the browser as a core part of data structure. For quick loading the data from files into browser's memory section without any pre-processing them, the data is prepared previously in files (see Figure 2) in a way that its structure is appropriate for concordance.

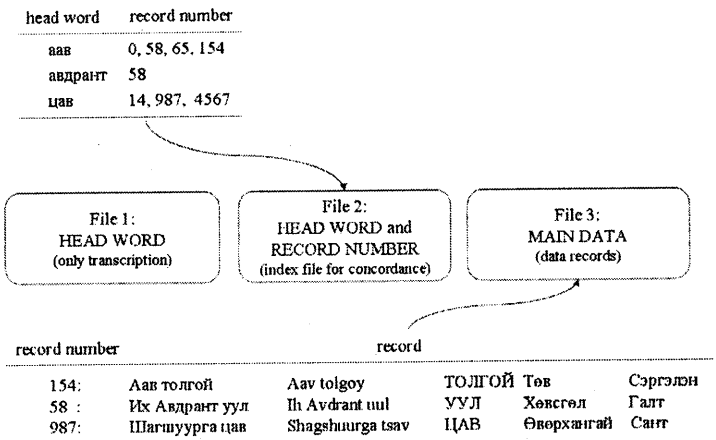
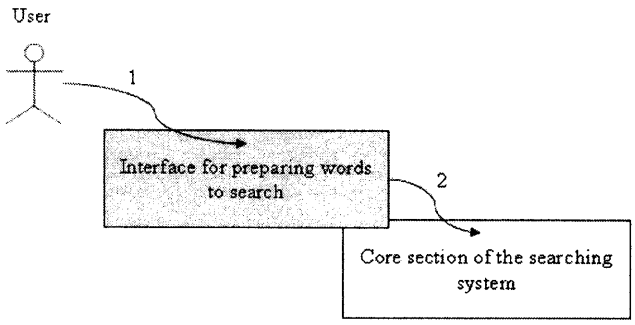



*Figure 2. The data file structure of MonPlace browser*

The sections for inputting search words and displaying information are the main interfaces of our browser. The first interface is used to transfer words typed by user with some attributes like color(s) and conjunction conditions into the searching section (see Figure 3).

A user could type several words with their colors and and/or conditions between them, or select a word from the headwords list. After the search is completed, the browser displays the highlighted word(s) with their colors.

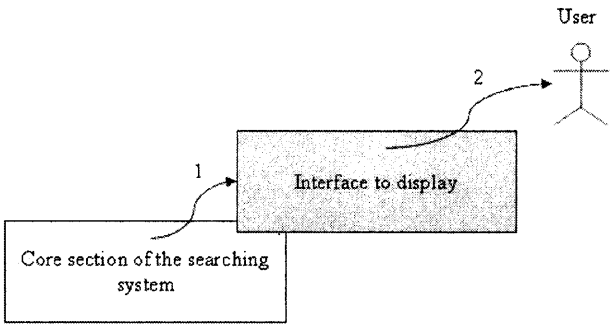



(1) To type or select the words with their colors and / or conditions

(2) To transfer the words retrieved in previous level

*Figure 3. The structure of the headwords inputting section*

The second interface that is shown in Figure 4 displays the records that are retrieved and then prepared either in text view or table view format.




(1) To transfer the retrieved and prepared information

(2) To display the transferred information

*Figure 4. The structure of the retrieved information displaying section*

The headwords and the primary data are implemented in memory as hash-table and array variable, respectively. The hash-table section of the browser is a core section and it also performs several functions like calculating and preparing conditions for multiword searching (see Figure 5).

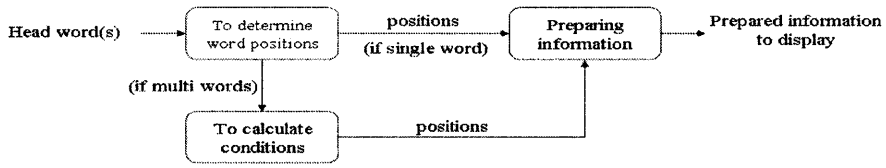



*Figure 5. The core section of the browser*

As mentioned above, the sections for preparing and displaying of retrieved word are the main interfaces of the browser (Figure 6).

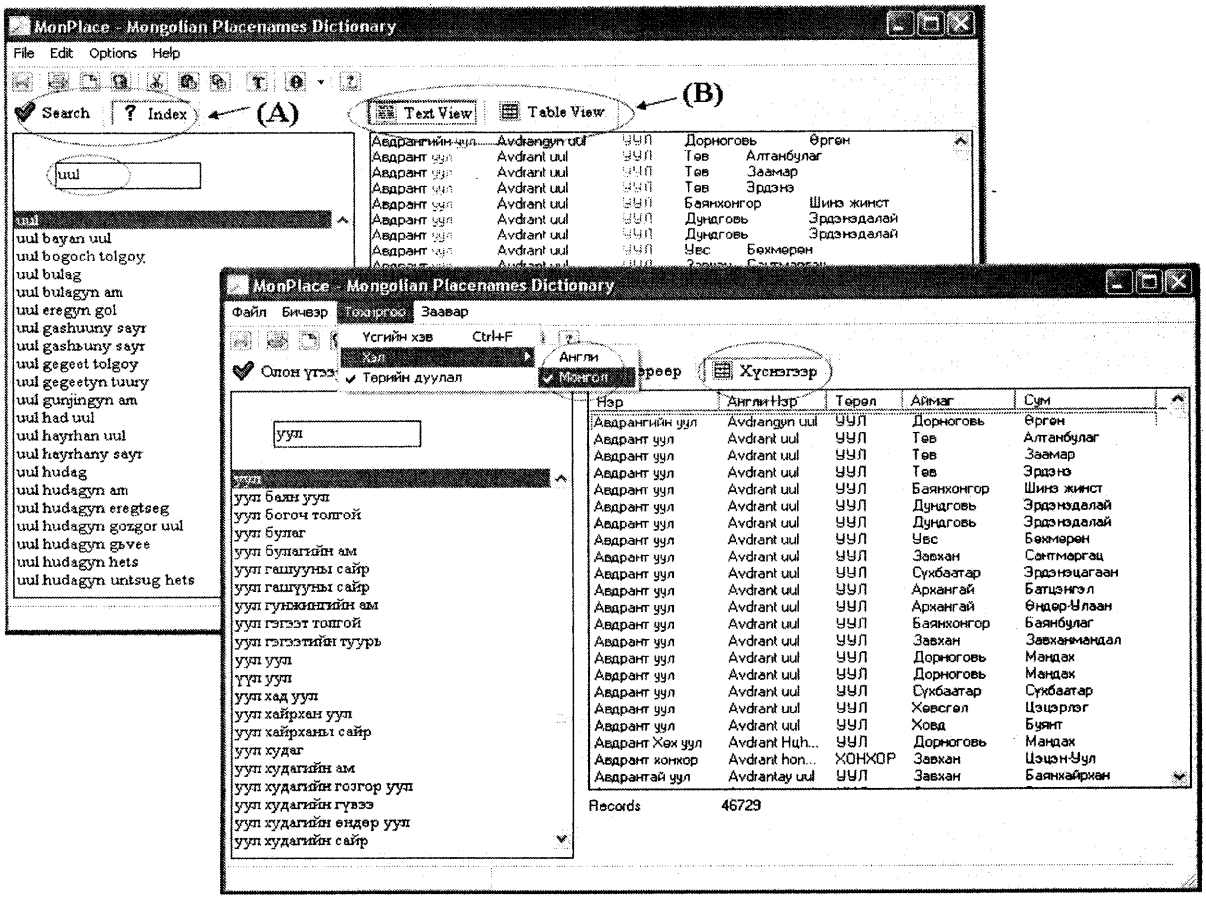



*Figure 6. MonPlace browser interface*

The search query and displaying information interfaces are on the left (A) and right (B) sections of Figure 6, respectively. If the browser language is selected as English, the headword should be typed in Roman. For inserting Mongolian special characters like ө, ү, е, ё, ц, ч, ш, ь(ъ), ю, я, ий, ы, there is a toolbar that contains their transcriptions like ö, ü, ye, yo, ts, ch, sh, y, yu, ya, y, y respectively[7].

## Result

Searching system is one of the main cores in the modern information age. Nowadays, the basic methods for searching activity are studied adequately, but in practice, the appropriate methods are needed for searching systems with diffrent features. The searching system for the electronic version of Mongolian toponym dictionary that consists of 8 volumes is based on the hashtable and concordance. The usage of these methods provides a quite fast result and retrieval the words in a large scaled text. But it is taking a little more time to display the prepared information on the screen. The browser is tested on the two computers, Pentium IV (CPU 2.66GHz, RAM 512MB) and Pentium III (CPU 1.1GHz, RAM 256MB) (see Figure 7). For searching "уул " (mountain) word that is repeated 46 thousand times in the

primary data, the browser spends around 6.5 seconds on Pentium IV and 18.5 seconds on Pentium III respectively.

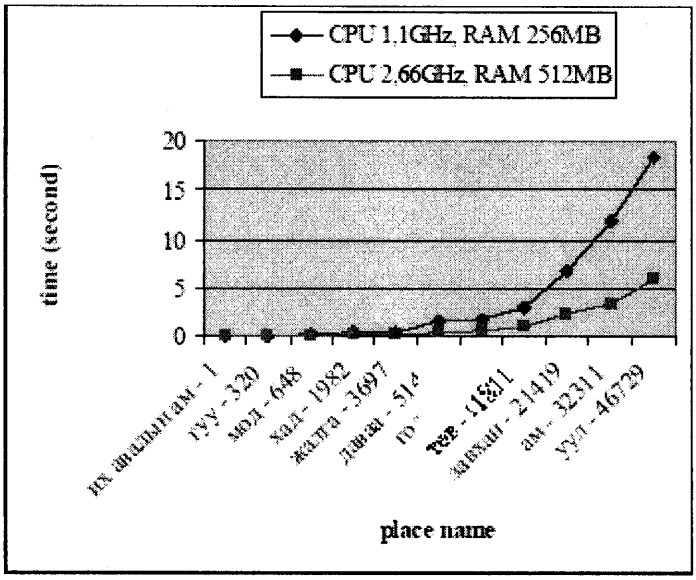



*Figure 7. The test result of MonPlace browser*

Approximately, the Pentium IV spends 80% of the time for preparing retrieved information and then displaying it on the screen. So, for reducing this time, the only part of information that could be fitted into the screen is displayed. As a result, the searching time on Pentium IV is reduced from 6.5 seconds to 1.1 second (see Figure 8).

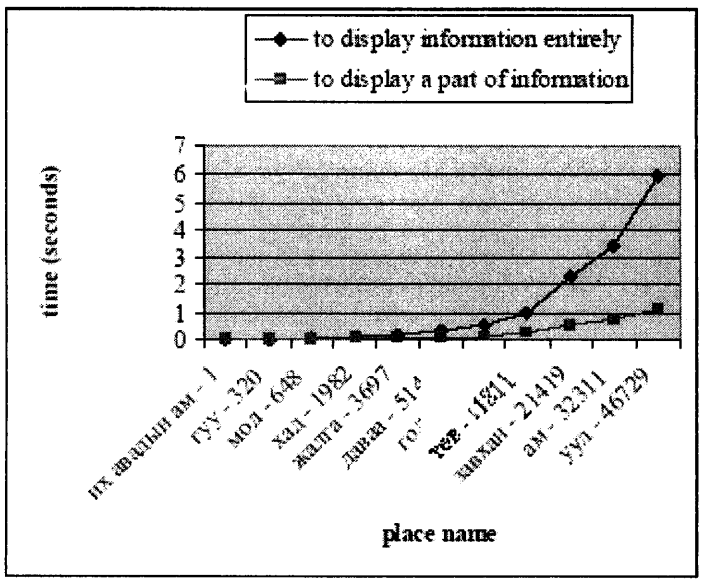



*Figure 8. A Comparing of times needed to display retrieved data*

The methods and principles used to MonPlace can be used for other kind of searching systems.

## ABSTRACT

This paper presents methods and principles used in a computer-based dictionary of Mongolian toponym. The search engine reads the headwords that are previously built from the database of Mongolian toponym in the concordance form and then writes it into the memory as hash-table structure, and uses it for further searching. Search word could be selected from the headwords table or typed in Mongolian or its Roman transcription form. Therefore, the system could perform multiword searching with *and/or* condition. The system itself was written in C# language, thus it supports the Unicode environment. Key words: Mongolian toponym dictionary, geographic information system


## References


[1] Donald E.Knuth. The Art of Computer Programming, Sorting and Searching. Second Edition. 1998.

[2] Christopher D.Manning and Hinrich Schütze. Foundations of statistical natural language processing. The MIT Press Cambridge, Massachusetts London, England.1999.

[3] Information retrieval: Data structures and algorithms /edited by William B.Frakes, Ricardo Baeza-Yates. Prentice Hall, Englewood Cliffs, New Jersey 07632. 1992.

[4] Mark Allen Weiss. Data Structures and Algorithm Analysis in C, Second Edition. Edinburgh University Press. 1997.

[5] Tony McEnery and Andrew Wilson. Corpus Linguisitics. Edinburgh University Press.1997.

[6] Purev J. Research work: Corpus for Mongolian language. Ulaanbaatar, 2006 (in Mongolian).

[7] Thematic dictionary of Mongolian geographical names /editorial director Oshgog Enhbayaryn Ravdan. 8 volumes, 5 books with CD. Ulaanbaatar, 2004. (in Mongolian).