Automatic Generation of Chinese Phonetic Initial Field in ArcGIS Map Database and its Application

Lianhe Yang , Shanshan Ji

yanglh@tjpu.edu.cn

Abstract—Based on the existed ArcGIS map database, the CPI field is generated automatically for all layers. An amending method is introduced based on phrase, which is used to amend the possible CPI errors caused by Chinese polyphones. As its application, CPI inquiry functionality is added to the original ArcGIS map and further extended into customary abbreviation inquiry functionality, which makes ArcGIS map inquiry efficient and humanized.

Keywords—ABP(Amendment Based on Phrase), ArcGIS, Database, CPI(Chinese Phonetic Initial)

I. Introduction

Geographical Information System (GIS) is a new subject that combines computer science, geography, topography and so on together. GIS rises rapidly in the past 30 years [1], and shows a vast potential market. GIS is studied in almost every vocation, which is fit for their own features [2].

Chinese characters are usually used as labels of ArcGIS maps in China, thus all the place names in the databases are described in Chinese characters. So, when searching a place name, one must enter Chinese characters frequently. Although it is unnecessary to give a whole place name, after all, Chinese character input is trivial, and it needs switching Chinese-English status regularly. In the meantime, there are a lot of repeated codes during Chinese character input, therefore, one must make a choice among them, and this decreases the efficiency of GIS application to some degree.

There is no doubt that using Chinese phonetic initial (CPI) instead of Chinese character itself will improve inquiry efficiency. But the collection of CPI data is a time-consuming work, and polyphone is another problem in the collection process. In this paper, the author develops a general-purpose program, which is used to generate CPIs for place names, and CPI inquiry functionality can be added to an existed ArcGIS map automatically.

II. Data structure analysis in ArcGIS map

ArcGIS is the earliest and the most mature geographical information system, which is developed by ESRI [3]. In the world market, it has the largest market share [2]. SHP map format is one of the standard formats in ArcGIS, and other formats can be converted to this one easily. So, this paper will discuss ArcGIS maps based on SHP format.

Generally speaking, there are lots of layers in an ArcGIS map, such as highway, railroad, river, school, bank, store and so on. Every layer consists of at least three separate files: main file, index file, and data file (alias attribute file). The three files have the same filename, but their extensions are different, and they are SHP, SHX and DBF respectively. DBF file format complies with the standard of dBASE III. It is a free table, and can be manipulated directly by dBASE III, Foxbase, FoxPro and Visual Foxpro.

In an ArcGIS map, although different DBF files in different layers have different structures, the field *Name* can be found in every DBF file structure, and its type is character. For every layer, its DBF file is self-governed, and the structure can be modified and expanded so as to fit for the requirement of CPI inquiry. In order to make the process automatic, all the operations should be completed by programs.

III. Implementation of CPI inquiry

In order to implement CPI inquiry, the attribute file structure in an ArcGIS map database must be extended, i.e., the necessary fields in CPI inquiry must be added automatically by programs. And then the CPI data must be generated, i.e., all the data in the CPI field is appended by the same programs. The following text of this paper will discuss the procedure.

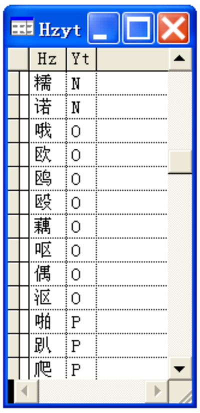
Automatic extension of table structure in a layer

Generally speaking, the relevant files for all layers in an SHP map are put together in a separate folder. A PRG file (command file) can be created and when run, through a loop statement, every DBF file (called original file in the following text) in the folder is opened one by one, and their structures can be extended and their contents can be added.

To extend a table structure, "*modify structure*" command can be used to complete this function through a man-machine interface. But this direct method is not fit for program implementation, so we must try to find another way. In other words, an indirect method should be used to extend a table's structure. First, put the structure of an original table into a "structure" table, add a record to the structure table, and fill in the corresponding fields' contents. Second, create a temporary table from the structure table, and append the records from the original table to it. Finally, delete the original table file, and change the temporary table filename to the original table filename. Thus, the original table structure expansion is completed.

Automatic generation of CPI data

According to the method above, after CPI field *yt* is extended, the content of this field is null. In order to fill in the field content automatically, a Chinese-CPI contrast table *hzyt.dbf* is needed. There are only two fields, *hz* and *yt*, in this contrast table, denoting Chinese character and its CPI, with width 2 and 1 respectively. All the Chinese characters in GB2312-80 and their phonetic letters are listed in [4]. If you put them into a text file, and use "*append from <text file> SDF*" command, then you can create a contrast table for Chinese characters and its CPIs. By the way, all the letters in field *yt* are lowercase. After running command "*replace all yt with upper(yt)*", all the lowercases will be changed into uppercases, as shown in Fig. 1.



Hz	Yt
糯	N
诺	N
哦	O
欧	O
鸥	O
殴	O
藕	O
呕	O
偶	O
沤	O
啪	P
趴	P
爬	P

Fig. 1 Chinese-CPI Contrast

Theoretically, since the field *yt* of the attribute table for every layer has been extended, and every Chinese character in the field *Name* has the only CPI, it is easy to generate CPI data for every place name. But the issue is much more complicated in practical than in theoretical, because there are a lot of polyphones in Chinese. If a CPI is generated like this, the CPIs of "yin2hang2" will be "YX", and the CPIs of "kuai4ji4" will be "HJ". So, the generated CPI data need to be amended.

Automatic amendment of CPI of polyphones

A. Amendment limits

All the pronunciations (phonetic letters) of polyphones found in GB2312-80 have been listed in [4], and every ployphone's pronunciations are arranged according to its usage frequency. We call the first pronunciation high-frequency pronunciation (HFP), the rest non-high-frequency pronunciation (NHFP). When a polyphone appears in a place name, it pronounces HFP in most cases, and need not to be amended. Furthermore, although some polyphone has different phonetic letters, they have the same CPIs. In this case, it is unnecessary to amend it.

To sum up, the amendment of CPI is only limited to NHFP. For example, the CPIs of "yin2hang2" , "YX", should be amended to "YH", the CPIs of "kuai4ji4" , "HJ", should be amended to "KJ", etc.

B. Amendment approaches

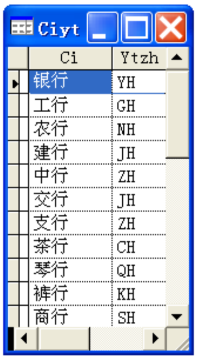
In order to amend CPI, a novel method, known as NHFP polyphones CPI amendment based on phrase (ABP), is introduced in this paper. Now we discuss its principles.

Just as the name suggests, "polyphone" means that a Chinese character has different pronunciations in different context. In fact, through using 2-character-phrase, 3-character- phrase or 4-character- phrase, the pronunciation of a polyphone can be determined. So, the author creates an phrase table with NHFP polyphones, *ciyt.dbf*, for the Chinese characters found in place names. There are only two fields in this table, *ci* and *ytzh*, the former denotes a phrase including NHFP polyphone, and the latter denotes the phrase's CPIs. Both of the two fields have the same type, character, and their widths are 8 and 4 respectively.

The ABP algorithm of CPIs amendment is:

```
For a place name in field Name
  Match every phrase in ciyt.dbf
  If matched
    Replace the corresponding substring in yt with ytzh
```

It can be seen that all the polyphones in *ciyt.dbf* are different initial consonant polyphones (DICP), meaning that a polyphone's HFP and NHFP have different initials. DICP is seldom found in Chinese. Statistics show that there are 93 and 75 DICPs in band 1 and band 2 Chinese-character-base of GB2312-80 respectively, and few of them can be found in place names. It can be concluded that there exist a few records in *ciyt.dbf*, thus it will comsume little time for the program to amend the CPIs in DICPs. Based on the place names of the ArcGIS map of Tianjin City, the author searches all the DICPs automatically, creates a phrase-base with NHFP polyphones in them, as shown in Fig. 2.



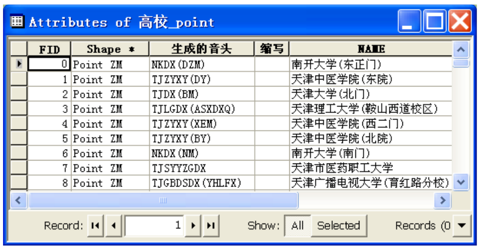
Ci	Ytzh
银行	YH
工行	GH
农行	NH
建行	JH
中行	ZH
交行	JH
支行	ZH
茶行	CH
琴行	QH
裤行	KH
商行	SH

Fig.2 Polyphone-CPI Contrast

Up to now, through extending attribute table structure, generating CPI data, amending polyphones initials, all the CPI data for all the attribute tables of all the layers have been generated, and CPI inquiry functionality has been implemented in an ArcGIS map.

IV. CPI INQUIRY EXAMPLE

Take school layer for example, the generated CPIs are shown in Fig. 3. When you search the map, if you enter CPIs "tjgydx", then the result will be shown as Fig. 4. After click "TJGYDX(DM)", the place name will be put on the center of the map area, as shown in Fig. 5, meaning that the CPI inquiry functionality has been added to the ArcGIS map. Of course, there is no need to enter the whole name of CPIs, ArcGIS can automatically match the string entered.

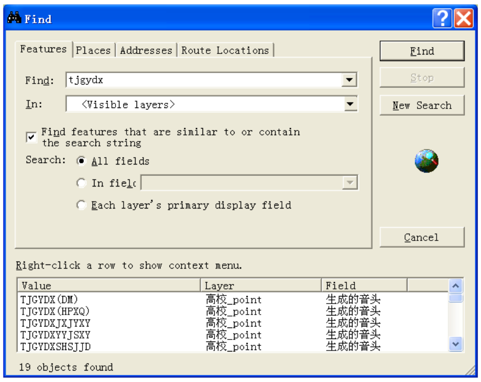


Attributes of 高校_point

FID	Shape *	生成的音头	缩写	NAME
0	Point ZM	NKDX(DZM)		南开大学(东正门)
1	Point ZM	TJZYXY(DY)		天津中医学院(东院)
2	Point ZM	TJDX(BM)		天津大学(北门)
3	Point ZM	TJLGDX(ASXDXQ)		天津理工大学(鞍山西道校区)
4	Point ZM	TJZYXY(XEM)		天津中医学院(西二门)
5	Point ZM	TJZYXY(BY)		天津中医学院(北院)
6	Point ZM	NKDX(NM)		南开大学(南门)
7	Point ZM	TJSYYZGDX		天津市医药职工大学
8	Point ZM	TJGBDSDX(YHLFX)		天津广播电视大学(育红路分校)

Record: 1 Show: All Selected Records (0

Fig.3 Generated CPIs



Find

Find: tjgydx

In: <Visible layers>

☑ Find features that are similar to or contain the search string

Search: ● All fields
○ In field
○ Each layer's primary display field

Right-click a row to show context menu.

Value	Layer	Field
TJGYDX(DM)	高校_point	生成的音头
TJGYDX(HPXQ)	高校_point	生成的音头
TJGYDXJXJYXY	高校_point	生成的音头
TJGYDXYYJSXY	高校_point	生成的音头
TJGYDXSHSJJD	高校_point	生成的音头

19 objects found

Fig.4 Entered CPIs

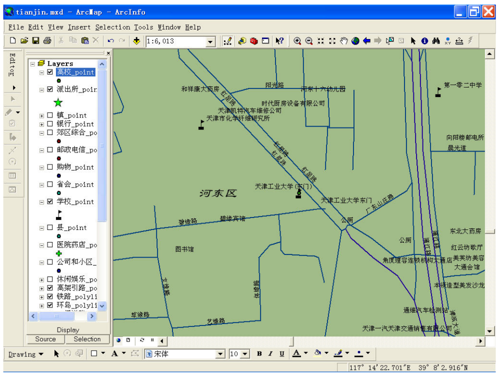


Fig.5 Inquiry result

V. CPI INQUIRY EFFICIENCY ANALYSIS AND OPTIMIZATION

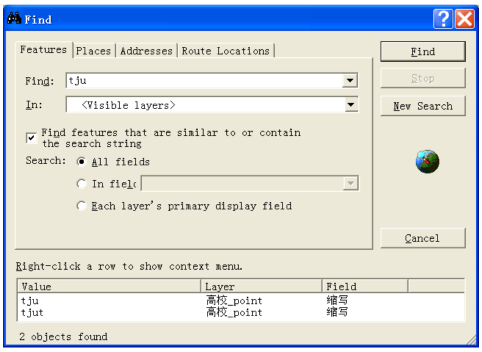
When searching a map, it can extremely accelerate the rate to use CPI instead of Chinese character. The most popular Chinese character coding input method is spelling input method, and its average code-length is 3.516 (including phrase input) [5]. But for CPI method, its code-length is only 1, the efficiency has been increased more than 2.5 times, and in the same time there is no need to switch Chinese-English status. Take "Tianjingongyedaxue" for example, it needs 18 keystrokes to use spelling input method, but it only needs 6 keystrokes to use CPI approach. It can be seen that entering CPIs instead of Chinese characters will increase the efficiency greatly.

After all, the aim of CPI inquiry is to decrease keystrokes in place name inquiry. However, a lot of keystrokes are still needed. Can we have a shortcut?

When the database of an ArcGIS map is opened, it can be found that *Name* field consists of place names and unit ones, and the latter has a very high proportion. Today, the Internet is so popular that every unit has its domain name, especially for universities. For example, the domain name of Tianjin Polytechnic University is *tjpu.edu.cn*, and *tjpu* can be viewed as a customary abbreviation. As a substitute for CPI in map inquiry, it is intuitionistic, easy to remember, furthermore, it is shorter, and when it is used in inquiry, the efficiency will be higher than any before.

In order to be compatible with CPI approach, another field *sx* can be added on the database table based on the field *yt*.

To some extent, customary abbreviation has some side effect. Since it is too short, if a string is the substring of CPI or another customary abbreviation, the repeated code will appear. When the case occurs, all the repeated codes will be listed, and users need to select the necessary item among them. For example, the string "tju" (Tianjin University) is included in the string "tjut" (Tianjin University of Technology), and when a user enters "tju" to inquire about "Tianjin University", "Tianjin University of Technology" will be listed, just as shown in Fig. 6. For users, this side effect is not worth mentioning.



Find

Find: tju

In: <Visible layers>

☑ Find features that are similar to or contain the search string

Search: ● All fields
○ In field
○ Each layer's primary display field

Right-click a row to show context menu.

Value	Layer	Field
tju	高校_point	缩写
tjut	高校_point	缩写

2 objects found

Fig.6 Customary abbreviation input

Although customary abbreviation may bring some side effect, it makes the code-length further shorter, accords with people's habit, and is more humanized. So, it is better than pure CPI approach. In other words, customary abbreviation is an optimization of CPI method.

It should be pointed out that pure CPI data can be generated automatically, but customary abbreviation can not. So, handwork is needed to create the corresponding data.

VI. Conclusion

It is yearn for GIS users to increase inquiry efficiency. CPI inquiry is based on decreasing keystrokes and optimizing the original inquiry functionality. We draw some conclusions from the research above:

1) CPI inquiry increases efficiency more than 2.5 times, and avoids switching between Chinese and English input.

2) As an improved approach, customary abbreviation is more humanized and more efficient than CPI.

3) The automatically generated CPIs may have some possible errors, which can be amended by ABP method.

4) Appending CPI functionality for an existed ArcGIS map can be implemented by programs automatically, and no manual work is needed.

(1) References

[1] Shi Wei. ArcGIS Geography Information System Details. Beijing: Mechanical Industry Press, 2009.

[2] Zheng Guizhou, Chao Yi. Geography Information System Analysis and Application. Beijing: Electrical Industry Press, 2010.

[3] Pan Yongdi, "Anatomy of ArcGIS files and writing code," Journal of Guizhou Meteorology, Vol.30, Dec. 2006, pp. 36-39.

[4] GB2312-80 Chinese-phonetic letters Contrast. http://blog.csdn.net/heiyeshuwu/archive/2007/05/23/1622397.aspx

[5] Simplified Chinese character phonetic input method with three keystrokes for two characters. http://blog.sina.com.cn/s/blog_4c5e244301000bwp.html

Lianhe Yang was born in Tianjin, China, in 1965. Phd, professor, School of Computer Science and Software, Tianjin Polytechnic University, Tianjin, China.