Database Research Method for Researches on International Chinese Language Teaching in an Era of Big Data*

SONG Fei, HAN Xiu-juan, XU Ming-hui
Beijing International Studies University, Beijing, 100024, China

The database research method is a method that analyses, generalizes and deduces from the data of subject investigated with database techniques, quantitative statistics and mathematical models. As the big data age comes with the data explosion in modern society, the International Chinese Language Teaching (ICLT) shows signs of sizable data accumulation, remarkable economic property, strong modeling requirements and notable cross-research trends, which thus make this method necessary as a new and independent research method in the researches on this area. Theory bases, applicative areas, available software and data resources, research program designs, as well as their advantages and disadvantages will be figured out in this paper. In the near future, it will bring about a revolution to the international Chinese language teaching.

Keywords: big data, International Chinese Language Teaching, database research method

Introduction

Rapid development of the technology and equipment in data mining, storing and processing in recent years has boosted the fast growth of a group of natural sciences with statistics as a major research method, which ushers in an era of “Big Data” as an important analyzing basis. In contrast, subject to traditional research methods, social sciences and humanities are not sensitive enough to the revolution brought by new methods in an age of “Big Data”. On this background, it will be a significant channel for social sciences and humanities to step up to a new level by seizing opportunities of “Big Data” time and on this basis, by developing from “quality” to “both quality and quantity” and from “coarse graining” to “fine graining”.1

Considering the fast development of ICLT in recent years, its research method should also be applicable to

* Acknowledgements: This work is supported by the project Study of 5D Graded Index System and Graded Lexicon Construction for International Chinese Language Teaching (No. 16YYC028) from Beijing Social Science Fund and the project Researches on the Chinese words and phrases Relationship based on Database and the Teaching Applications (No. 545002/002) from the Social Science project of Beijing Municipal Education Commission.
SONG Fei, Ph.D., Lecture, School of Chinese, Beijing International Studies University.
HAN Xiu-juan, Ph.D., Associate Professor, School of Chinese, Beijing International Studies University.
XU Ming-hui, Postgraduate, School of Chinese, Beijing International Studies University.

1 It is a set of relative concepts between “coarse graining” and “fine graining” that represents the relative difference of the structural division in the theoretical system of different level for the same discipline. Taking physics for example, the evolution from classical theory to quantum theory belongs to a transformation from the macroscopic “coarse graining” research to the quantum “fine graining” research incorporating all details. It is believed in this paper that development of social sciences and humanities will definitely face such conceptual changes; however, compared with natural sciences, a certain theory or rule of it is less universal but its constitution of this phenomenon is usually more complicated, so the “fine graining” process is filled with more difficulties.
the subject characteristics and grasp the fast changing development situation. But a large number of studies in this area are limited to theoretical deductions and short of data, or belong to shallow inductions of external phenomena and are unable to reach a level of mathematical statistics that reveals inner laws, with a lag nature. One the on hand, researchers are required to find the intrinsic correlation through superficial phenomena for the ICLT creative advancement; and on the other hand, they should make forward-looking predictions based on the potentially emerging new situation and problems. Thus, neither pure theoretical deductions nor shallow inductions with a certain lag nature are applicable to the demands of rapidly developing subject.

Meanwhile, ICLT also presents some features in an age of “Big Data”: a certain scale of related data has been accumulated, laying a foundation for the macro-policy making; Chinese language teaching has become an industry in some countries and areas, which can be used to support the interdisciplinary transplantation of econometric theories; modeling demands in the research fields is becoming stronger in which mathematical modeling plays a more and more important role; interdisciplinary development momentum of Chinese ontology research is evident, with the computational thinking being more and more universal in language researches. Therefore, related researchers are required to pay attention to cultivating their data thinking, improving their mathematics and statistics accomplishments, and accommodating themselves the research method revolution of big data time. For the database research method, metrical studies are stressed and reasonable inductions and deductions are advocated based on database building and data analyzing, which corresponds to the demands of ICLT in a new era and plays an influential role in ICLT studies.

Theory Bases of the Database Research Method

Database and the Database Research Method

Database is a specific data set in which various data are aggregated based on a certain data model, and thus, it possesses small redundancy, high independence and good expansibility, and can be shared by all users within a certain domain.

The database research method is a research method to conduct inductive analyses and rational deduction for studying objects by means of database, quantitative statistics and mathematical modeling.

Academic Ideas Behind the Method

From this method point of view, language rules, social laws and economic laws in ICLT are hidden behind the language realities (phenomena), social realities (phenomena), and economic realities (phenomena), and the more extensive realities are, the more universal the reflected laws become. In this method, complete and well-organized aggregations and analyses for realities, instead of preconceived views and theories, are advocated to reveal the inner laws as much as possible; simple quantitative descriptions are opposed and logical induction

---

2 For instance, the going-global of the Confucius Institute plays a decisive influence on ICLT in recent years. It has been just ten years since the birth of the first Confucius Institute in 2004 till now. From the perspective of time span, accumulation of various data is just able to reach a critical point where a certain type of mathematical modeling and regression analysis are effective and feasible according to the observed values.

3 E.g., with the expanding volunteer teachers of the Confucius Institute, it is unrealistic to dispatch field supervisors or experts to assess and grasp the classroom teaching situations of thousands of volunteer teachers each year. Instead, it becomes particularly necessary to build an “interaction model for classroom discourses” that can truly and effectively reflect the interactive situation of classroom discourses, realize automatic quantitative analyses for enormous teaching lessons, and measure their rough teaching situation quickly and effectively by data modeling.
and deduction based on quantitative analyses are emphasized. It neither belongs to an absolute induction method, nor a pure deduction method in essence. But it is an organic combination of the both, of which the former is applied into the database building and data analyzing and the latter is applied into the popularization or predictions after successful modeling or theoretical improvement.

Undeniably, metrical studies play a decisive role in this method, and it has become an inevitable trend to emphasize on quantitative analyses and discard the simple mental analyses for the academic research in the new stage. Besides, rigorous mental analyses should be based on a large quantity of realities that have achieved digital storage and adequate quantitative analyses.

**Required Professional Knowledge to Grasp This Method**

Knowledge in linguistics, Chinese teaching as a foreign language, statistics and economics, as well as corresponding computer knowledge (including database software, corpora processing software, programming software, mathematical modeling software and data analyzing software) are required for ICLT researches.

**Applicable Research Areas of This Method**

This method is applicable to extensive research areas, and it requires following some thought for topic choices. Abstractly speaking, it is particularly suitable for the following three areas:

First, areas that develop in time series, for example, there is a hypothesis that trade volumes between China and a foreign country is positively correlated to the number of Chinese language learners in this foreign country (WU, 2012), which is verified by extracting related data of those years. At this moment, by building the database and establishing vector auto-regression (VAR) model with the data in it, the positive correlation is attainable and it is also possible to figure out whether there is a causality and what kind of causality it is (for instance, whether it is linear; which department’s trade volumes has the biggest impact on increasing Chinese language learners; whether the number of learners can rise with the increase of trade volumes; and how many learners should increase when the trade volumes of this year is known; and what disturbance vectors are). Equally, according to the data related to the annual expenditures of the Confucius Institute Headquarters as well as local popularization of Chinese, reasonable resource allocation can be determinable to help international communication of Chinese and the prospect is predictable.

Second, area for contrastive analyses at the same cross section, e.g., analyzing the data from “the Confucius Institute Database” established by College of International Education, Minzu University of China (MUC), and taking the year 2010 as a cross section, contrastive analyses are conducted for the operational situations among inter-continents, countries and the Confucius Institutes. As a result, some related hypotheses are verified; some data usually without explicit concepts are obtained; and some data unexpected at the beginning are concluded, such as: “in 2010, there were averagely 16.55 Chinese language classes in the global Confucius Institutes; calculating by the students in a standard academic year⁴, Suan Dusit Rajabhat University Suphanburi Campus completed the maximum Chinese language teaching workloads; only 33.87% courses were open to social

---

⁴ From: On the Index System of Confucius Institute Assessment, Wu Yinghui, Educational Research, 2011(8), and it is mainly used to measure the Chinese training workloads completed by a Confucius Institute within a year or several years, with the calculation formula: students in a standard academic year = (the number of learners x the learning days)/160 days, based on 5 days per week, 20 days per month and 4 months per semester.
learners in Asia, being the minimum, but in South America, the proportion reached high up to 71.91%; examination-orientated courses of the Confucius Institutes in the South Korea accounted for 11.92% among the entire courses, having the highest proportion; in addition, all these courses are targeted to HSK; among these courses, primary and intermediate courses accounted for 80.68%, and so on. Thus, it would be very hard to discover those specific differences among hundreds of campuses without contrastive analyses after the building of database.

Third, areas of which the quality traits facilitate the datamation, interaction study of ICLT classroom discourses is an example. A system analyzing classroom discourses interaction just based on data linkage was once established by way of function combinations with Excel, and interactions of volunteers’ teaching conditions in classrooms could be analyzed automatically by coding different class discourses interactions. Though many aspects such as the teaching activity, the fore-and-aft teaching activity combination, teachers’ control styles in the classrooms, applications of teachers’ specific teaching methods, reflections of students and the actual effects of classroom teaching seemingly only belong to quality traits, they can be digitalized into different dimensions and reflected by data to a certain degree, and the results come from the encoding of classroom teaching activities (See Figure 1).

By enhancing the building of database in those areas, data sources and studying methods can be provided for both quantitative and qualitative researches in ICLT, and significant supports can be offered to produce related research results based on the database research method, and the positive role can be played in expanding the new areas.

Current Available Software and Database Resource

Software Resource

Those computer software including database software (such as MS Excel, MS Access and SQL Server), text processing software (such as MS Word, MS Text and ICTCLAS), database management software (like Antconc), and statistical software (such as SPSS and Eviews), web crawler tools, programming software (like Python) and

---

5 The data comes from: Minzu University of China “International Chinese Language Teaching Database” (http://tpi.muc.edu.cn).
big data processing software (like Hadoop) can be available for this method. Though they belong to several different categories and possess very different characteristics, all the above-mentioned software are needed for collaboration at different stages, in different areas and to different layers during the application of this method.

For example, Access can be used for simple calculation but possesses powerful data management function; and Excel can be used for simple data management but possesses powerful calculation function (SU, 2010). Namely, even extraordinarily complicated work could be finished if their respective strength is given a reasonable play. During the construction of a large-scale corpus, Access is a better choice because its recording ceiling can reach high up to 2 billion records (with a single mdb no more than 2G) and it is convenient to establish relations among databases. But as for a Flanders Interaction Analysis System (FIAS) with more complicated calculation (to be mentioned in the following example)⁶, Excel is a priority because its function has a similar effects to the programming.

For another example, major function of ICTCLAS include Chinese word segmentation, part-of-speech tagging, as well as named entity recognition, new word recognition and user dictionary support; AntConc is used for corpora analyses based on text files and mainly possesses three function: word searching, vocabulary generating, and topic terms; and Eviews can be applied to find statistical relations among data quickly and predict future value (FV) based on the relations obtained.

As for web crawler tools, programming software and big data processing software, of which the primary function for this method is corpus (data) excavation, data management and cloud computing aimed at data, respectively, will not be discussed further here.

Database Resource

At present, there are many database resources for Chinese language researches, and especially in recent years, it presents a trend of fast expansion, of which at least dozens of databases have been established relying on R&D institutions such as colleges and universities and most of them are corpora. The main reason for the rapid growth is that database technology becomes widespread and computer processing capacity is improved, and the root cause is that data volumes are expanded rapidly triggered by big data and researchers are paying more and more attention to this method. Some of well-known Chinese language databases are: contemporary Chinese language (CCL) corpus of the State Language Work Committee, CD-ROM database of People’s Daily, CCL corpus of Center for Chinese Linguistics PKU, HSK dynamic composition corpus of Beijing Language and Culture University, Chinese language corpus of Harbin Institute of Technology, global Chinese language teaching (CLT) materials base of Sun Yat-sen University and CCL real text processing corpus of Xiamen University and so on. And College of International Education MUC has also strengthened the building of database since 2010 and has established an ICLT database, which is the sole all-fields ICLT database in China now.

Apart from open database resources, other accesses to data such as network data excavation and recording transcription are also included.

---

⁶ It belongs to a statistical modeling aimed at the interaction activities of classroom discourses. Through a digital code system of Flanders Interaction Analysis System (FIAS), a series of events in classrooms are recorded in chronological order and then connected to be a time sequence to present the teaching structures, behavior patterns and styles in the classroom.
Research Program Design

Research program should be well designed in the beginning for the successful application of this method in ICLT research, and several main steps are shown as below:

Explicit Research Object

For instance, taking “Linguistic Elements in U.S. CLT Materials” and “Management Structure of Confucius Institute” as research objects, their program designs in many aspects such as the research thought, data types, labeling fields and even the database building tools will be quite different from each other, so an explicit research object at the very start is the foundation of the rest program design.

Clear Research Thought

ICLT research cover various aspects like Chinese ontology, Chinese teaching as a foreign language and international communication of Chinese, so specific thought need to be fit for the corresponding field after the object is confirmed. Taking the aforementioned two objects as an example, for the first one, it could roughly be as below: “exhaustive collection of U.S. CLT Materials—materials datamation—Access as the database building tool—the building of five sub-database, namely, chapter, sentence, phrase, word and field according to different linguistic elements—labeling of linguistic elements in every sub-database—data analyses of linguistic elements in each level—conclusions”; however, for the second object, the thought could be easier, including “exhaustive collection of the Confucius Institute’s curriculums—Excel as the building tool—the building of Management Structure database of Confucius Institute—data analysis of the database—conclusions”. Therefore, research thought should be clear according to characteristics of various objects; or otherwise, it may end in vain due to any incomplete procedure or improper setting.

Comprehensive Data Collection

Data collection is the key to build a database, but without data, the structure design will be useless no matter how perfect it is. In fact, the issue of sparse data is very common, and there are usually two solutions: by algorithm like singular value decomposition and clustering to mitigate it; or by data retrieval and excavation which are important research fields in computational linguistics. Data retrieval is a process to retrieve and extract the required accurate data directly from the expected data source according to research demands; Data excavation is from statistics, on-line analytic processing, and information retrieval, machine learning, expert system (dependent on the past empirical laws) and pattern recognition, etc., and thus, it can be applied into both data collection and analyses.

Building of Database in Compliance with Principles

Clear elated concepts before the building of database. Record and field. The integrated framework of a structurized database is made up of records and fields.

One “record” means one “row”. As shown in Tab. 1, “1, U.S., Free Talks, Anything Goes, Senior, 1, Chinese New Concept and Social Phenomenon and Comment” are considered as one record and thus located in one row.

One field means one “line”, and as another significant structure in a database, it is used to save every point of recorded information by item. As the above example, every part segmented by a comma should mean “Serial Number, Confucius Institute Name, Year and The Number of Chinese Language Curriculum” field, respectively.
Those important, independent, deep-analysis-required and polymerizable information points should all be recorded as respective field in a database. While building, some basic information points can be set up independently and then added according to the demands (SU, 2010).

**Redundancy.** Data redundancy means unnecessary repetition of data, and the precondition for database construction is to strengthen the independence of data and reduce redundant data, for example, there are 32 lessons in a book in the process of building a teaching material database. Then, it contains both the publishing information and the concrete contents of every lesson. If these two kinds of data are put into a same database, the publishing information will appear 32 times together with the contents of every lesson, but such repetition that is known as redundancy is meaningless and even affects database performance. Hence, it is workable to respectively build a database for them and then make an association between them with teaching material names as associated fields.

**Rigidity and flexibility of data.** Rigid data is referred to those with unified formats and explicit values in each record of the same field, and they usually do not need further labeling, for example, data in a format of “yyyy-mm-dd” in the field of “the establishing data of every Confucius Institute”. Flexible data is referred to those without sufficiently explicit formats, purposes or usages during data entry, like data in the field of “the mission of every Confucius Institute”, but it can be transformed to be rigid data by further labeling and processing from different angles.

**Main principles of building database.** *Reserving research space for those with clear demands.* Construction of database should serve concrete research demands, but it does not mean the structure cannot be changed once the design is finished. Sometimes, it may serve more than one purpose and whereby make a series of achievements, so it is necessary to spare enough expansion space in the design. For instance, researchers hope to study the Confucius Institute from different aspects, including some basic information, pecuniary condition, administrative situation, curriculums, activities, and media reports and so on. It is foreseeable that there will definitely be redundancy if all these data are put into a database, so separation is demanded. But only a part of related data can be used in the separated “Curriculum Database of the Confucius Institute” owing to different view of research, such as the field of “Curriculum Name” for curriculum category study and the key field of “Teaching Hours” for teaching hour study. Therefore, at the beginning of design, it is necessary to collect related available information sources as much as possible to set fields for as many information points as possible and reserve research space.

**Choosing tools for building database according to data characteristics.** As previously mentioned, those tools have their own strengths and functions, which require researchers to select the most suitable tools according to data characteristics (including whether the data records exceed tools’ upper limits, data formats, as well as calculating amount and methods required for data analyses).

**Careful selection of data doormats and field modification as required.** Data format can be set for the field as required, such as number, date, text, remark, character string, and the setting of each format like field size, default value, validity rule, the required field and null string are different, so it requires selecting the appropriate format according to data characteristics.

In addition, researchers should improve the design of database structure as much as possible before the building, but it is hard to avoid omissions or exceptions, so perfection is unnecessary. In this case, it is advisable
to add or delete fields as required on a basis that the original database structure can work normally so as to reduce
the loss of data entry time to the largest extent.

**Redundancy reduction to the greatest extent and reasonable segmentation.** The solution to the
aforementioned redundancy is to segment one database into two or more relational databases reasonably, taking
“Country Specific Chinese Materials” as an example, with structure and data of the redundant material database
shown in Table 1:

<table>
<thead>
<tr>
<th>ID</th>
<th>Targeted Country or Area</th>
<th>Chinese Name</th>
<th>Foreign Name</th>
<th>Difficulty Degree</th>
<th>Class Order</th>
<th>Text Title</th>
<th>Text Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U.S.</td>
<td>Free Talks</td>
<td>Anything Goes</td>
<td>Senior</td>
<td>1</td>
<td>Chinese New Concept</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>2</td>
<td>U.S.</td>
<td>Free Talks</td>
<td>Anything Goes</td>
<td>Senior</td>
<td>2</td>
<td>Non-excessive Beauty Economy</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>3</td>
<td>U.S.</td>
<td>Free Talks</td>
<td>Anything Goes</td>
<td>Senior</td>
<td>3</td>
<td>Cellphone Hazard</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>31</td>
<td>U.S.</td>
<td>Chinese Society</td>
<td>A Kaleidoscope of China</td>
<td>Senior</td>
<td>1</td>
<td>A Complete Story - A Sino-U.S. Consumption Comparison</td>
<td>Culture and Comparison</td>
</tr>
<tr>
<td>32</td>
<td>U.S.</td>
<td>Chinese Society</td>
<td>A Kaleidoscope of China</td>
<td>Senior</td>
<td>2</td>
<td>Buying Houses or Having Kids Rights of Pursuit Unequal to Punishment Exemption for Citizen</td>
<td>Daily Life</td>
</tr>
<tr>
<td>33</td>
<td>U.S.</td>
<td>Chinese Society</td>
<td>A Kaleidoscope of China</td>
<td>Senior</td>
<td>3</td>
<td>Rights of Pursuit Unequal to Punishment Exemption for Citizen</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>67</td>
<td>U.S.</td>
<td>New Heights</td>
<td>No</td>
<td>Senior</td>
<td>1</td>
<td>the so called Chinese threat</td>
<td>Politics and Economy</td>
</tr>
<tr>
<td>69</td>
<td>U.S.</td>
<td>New Heights</td>
<td>No</td>
<td>Senior</td>
<td>3</td>
<td>National Security and Network Freedom</td>
<td>Science and Technology</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

There are redundant data in several fields such as “Targeted Country or Area, Chinese Name, Foreign Name
and Difficulty Degree” within bold areas, which recurs with every lesson. However, such unnecessary repetition
that may affect the efficiency should be wiped out. Thus, the solution is to split one material database into two
relational but independent databases, with the separation way shown in Table 2:
Table 2

_A Solution to the Sample of Redundant Materials_

<table>
<thead>
<tr>
<th>ID</th>
<th>Targeted Country or Area</th>
<th>Chinese Name</th>
<th>Foreign Name</th>
<th>Difficulty Degree</th>
<th>Class Order</th>
<th>Text Title I</th>
<th>Text Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U.S.</td>
<td>Free Talks</td>
<td>Anything Goes</td>
<td>Senior</td>
<td>1</td>
<td>Chinese New Concept</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>2</td>
<td>U.S.</td>
<td>Chinese Society</td>
<td>A Kaleidoscope of China</td>
<td>Senior</td>
<td>1</td>
<td>A Complete Story - A Sino-U.S. Consumption Comparison</td>
<td>Culture and Comparison</td>
</tr>
<tr>
<td>3</td>
<td>U.S.</td>
<td>New Heights</td>
<td>No</td>
<td>Senior</td>
<td>1</td>
<td>the so called Chinese threat</td>
<td>Politics and Economy</td>
</tr>
</tbody>
</table>

... ... ... ... ... ... ...

<table>
<thead>
<tr>
<th>ID</th>
<th>Chinese Name</th>
<th>Class Order</th>
<th>Text Title I</th>
<th>Text Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Free Talks</td>
<td>1</td>
<td>Chinese New Concept</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>2</td>
<td>Free Talks</td>
<td>2</td>
<td>Non-excessive Beauty Economy</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>3</td>
<td>Free Talks</td>
<td>3</td>
<td>Cellphone Hazard</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>31</td>
<td>Chinese Society</td>
<td>1</td>
<td>A Complete Story—A Sino-U.S. Consumption Comparison</td>
<td>Culture and Comparison</td>
</tr>
<tr>
<td>32</td>
<td>Chinese Society</td>
<td>2</td>
<td>Buying Houses or Having Kids</td>
<td>Daily Life</td>
</tr>
<tr>
<td>33</td>
<td>Chinese Society</td>
<td>3</td>
<td>Rights of Pursuit Unequal to Punishment Exemption for Citizen</td>
<td>Social Phenomenon and Comment</td>
</tr>
<tr>
<td>67</td>
<td>New Heights</td>
<td>1</td>
<td>the so called Chinese threat</td>
<td>Politics and Economy</td>
</tr>
<tr>
<td>68</td>
<td>New Heights</td>
<td>2</td>
<td>Economic Development and Democracy</td>
<td>Politics and Economy</td>
</tr>
<tr>
<td>69</td>
<td>New Heights</td>
<td>3</td>
<td>National Security and Network Freedom</td>
<td>Science and Technology</td>
</tr>
</tbody>
</table>

In this way, redundant data in the original database is effectively eliminated and the data association efficiency is increased by establishing “One-to-Many” relationship based on the common field of “Chinese Name” in two databases.

**Coexistence of rigid and flexible data for storage and labeling.** Considering that the exploitation depth could vary from time to time due to different researchers’ point of view for the use of data, this principle of accepting the fact that rigid data and flexible data coexist might be the most economical way to get the best results; for example, some researchers may want to figure out which Confucius Institutes emphasize on the mission of “Chinese culture communication”, and they also hope to make statistics about which Institutes’ missions have been found their location in the society of the host country. And if soft data is forced to be separated into rigid data only for a neat database, there could be three negative effects: 1. Complexity and redundancy of the database are increased; 2. Data becomes sparser than before; 3. Semantic information can be lost easily. That’s, if it is hard

---

7 Referring to a relation between two tables in the relational database, of which a single line in the first one can be related to one or more lines in the second one, but a line in the second can only be related to a line in the first.
to process soft data for the moment, then, keep the flexibility for further labeling when it is required later.

**Labeling as Required**

Labeling previous soft data (sometimes, rigid data also included) according to the demands is an unbridgeable step between the database construction and data analysis. For instance, when they are finding which text topics are related to school life, researchers need manual labeling based on the field of “Text Title”, because they might screen inaccurate key words just on a basis of “Text Title” field of which the naming characteristics are unknown for them in advance. There are mainly two labeling methods: building some new fields, such as “School Life” and “Tourism”, code data of which the records belong to school life will be input in the corresponding field (like “Yes” or “Y”), or otherwise, distinctive code data will be input (like “No” or “N”); or building a new field, such as “Category” in which the category name of text titles will be directly input (like “School Life” and “Tourism”, the same category for the same name). After that, it will be rather convenient to extract those records about school life according to the topic field.

**Reasonable Data Analysis**

Data analysis is a significant step in this method, and whether the anticipated results can be achieved depends on the analytical ability to a large extent. Statistical accomplishment, computer skill and modeling ability of researchers mainly decide their data analytical capacity, but a grasp of fundamental types of data analysis also can help deepen the understanding. It can be classified into descriptive analysis and mathematical-modeling-based analysis, of which the former belongs to shallow quantitative relation analysis; and the latter is to seek correlations among data changes in different fields and then build mathematical models with fields as variables, for example, input-output effectiveness model of the Confucius Institute with various investment of Hanban as variables. However, descriptive analysis is achieved by means of calculating some eigenvalues (such as mean, mode and median). For instance, through calculation, it can be found that “there were averagely 16.55 Chinese language classes in the Confucius Institutes around the world in 2010”, thus to know the conditions.

**Research Conclusion Extraction**

Conclusion extraction is also a key part, and also the fundamental difference between the database research method and quantitative statistics-based data descriptive method, of which the former requires sufficient application of induction and deduction and the latter is only describing in this part. Researchers need sufficient and credible data to draw limited conclusions, and should be good at discovering the hidden qualitative conclusions behind the quantitative relations.

**The Advantages and Disadvantages of This Method**

**The Advantages**

The greatest advantage is to apply the methodology of natural science into social science, in which quantitative analysis is emphasized and simple mental analysis with strong subjectivity and randomness for the traditional social science researches is also discarded, so it is boosted closer to arguable, computable and deducible research.

It has its own advantages compared with other similar methods:
**Comparison with the corpus-based analytical method.** Contained in the database research method, the corpus-based analytical method is usually applicable to language studies, and namely, its related applications are also based on linguistic elements, with very limited range of application. However, if all the methods are classified into different levels, then the database research method belongs to a fundamental level, just as mathematics is the basis of all the rest science and engineering subjects, so its range of application is extensive without any strict dead corner.

**Comparison with the quantitative statistics-based method.** Firstly, there is an intersection between the quantitative statistics-based method and the database research method, which means they are independent of database to conduct descriptive analyses just replying on data or draw conclusions through mathematical modeling. Nevertheless, the database research method is based on the building of database with multiple dimensional approaches, and besides, those databases can be associated with each other, so the application modes are more flexible.

Secondly, their fundamental difference is that the database research method requires sufficient induction and deduction but the other one pays more attention to descriptions. While making conclusions, researchers have to get limited deductions depending on sufficient and credible data, and they also have to be good at revealing the hidden qualitative conclusions behind those quantitative relations.

**The Disadvantages**

The disadvantages are mainly shown as below:

**An issue of data authenticity.** Considering data sources are not always authentic, it is hard for researchers to guarantee the authenticity of data each time. Once those misleading data is not discovered in time, all the studies later can be overturned entirely and easily.

**An issue of data sparsity.** Despite data is true and credible, researchers are still faced with a dilemma of too much sparse data. If so, they must consider how to ensure the generalization of study results. Good news is that more and more related researches have been conducted in those years and it can be mitigated by algorithm to some extent.

**An issue of model building defects.** The biggest problem for the quantitative method is that there are defects in building models, which is thus also called as "pseudoscience". It is a dilemma in the process of global modeling for this method, because any false premise can lead to a false conclusion, and then all the research results will become ridiculous once the reality modeling fails. However, human beings never stop finding those theories to explain the grand unification of the world; instead, they have achieved considerable progresses in global modeling on this road. Every significant revolution and breakthrough in science makes people move one step forward toward to “the ultimate modeling” of the world. Those who contribute to the progress of speech recognition technology, text retrieval technology, as well as automatic word segmentation technology and accuracy rata of input methods never show a disdainful attitude toward the original machine translation algorithm due to its inaccuracy. Today, social science researchers have to face enormous changes in social science research brought by big data instead of just relying on their own mind assumptions. Therefore, inaccurate descriptions of realities and incorrect predictions of the future based on data might be caused by the limited human’s capacity of applying data instead of data defects.
An issue of impossible absolute accuracy. Because of data authenticity and sampling relativity, it is hard to conduct absolutely accurate descriptions for facts in this method, which is an issue of the afore-mentioned “fine graining” realization. In fact, social sciences and humanities have been transformed from simply mental analysis into empirical study in which “fine graining” research is achieved for once and various subject branches are derived and developed. However, social sciences and humanities theories are still too far away from the final “fine graining” theories. Even for natural science, there is still a long road. Naturally, there will be too many difficulties to deal with if researchers are thinking to realize the final “fine graining” in social sciences and humanities or even if they are trying to explain the relative “coarse graining” theory based on the current “fine graining” degree. Thus some of them regard it as “an impossible mission”. But as a matter of fact, its role in many respects like document researches that is “hypercritical” of details is not blocked at all. On the contrary, more and more document researchers are keeping an eye on the building of document database.

Thereinto, data authenticity and model building are two critical issues that need to be solved during the research of this method.

A Revolution to ICLT Research Methods by This Method

It is an essential constituent for the database research method in ICLT research methods. First, database technology, regardless of hardware or software, is qualified to meet ICLT research demands; for example, it is no longer as hard as before to build a corpus with hundred millions of words in a short time. On the one hand, it only takes some time to extract the required data on the internet purposefully with web crawler tools, and once parameters are carefully set, the test work will be done by web crawler tools; on the other hand, resources integration of big data corporates are also available for corpora, such as search engines like “Google and Baidu”, input methods like “Sogou and iFLY”, encyclopedia like “Wikipedia” and social networks like “Renren and Kaixin 001”, all of which acquire numerous corpora beyond the reach of those academic research institutions while they are conducting their main business. But it has some economic cost and its realization depends on commercial cooperation. Nevertheless, as long as researchers themselves possess enough capacities, they can break shackles of database technology no matter which method they adopt.

Next, this method has shown some superior characteristics that might overturn conventional methods. There is a bold hypothesis that the development situation of the “Chinese Popularity” around the world can be assessed according to data acquired by retrieving relevant entries of Chinese language learning in the countries and districts of the rest of the world from Google’s searching database. Similarly, a corpus with large-scale first-hand spoken language transcriptions is available in iFLY speech database; an enormous semantic categorization lexicon database has been established by Sogou, which can detect new words for the first time; and knowledge element system database of Wikipedia can be improved by global users, etc. All of these are just the tips of iceberg for the database research method to serve ICLT research.

Then, ICLT research has a strong demand for data research. As previously mentioned, it is a promising new subject, so it lacks sufficient research data and methods, and it is hard to keep up with the rapid development of Chinese language teaching around the world. And the database research method is just targeted to solve those problems: Corresponding information can be obtained by means of data excavation for an issue of limited research data; Current economics models and data econometrics can be utilized because ICLT possesses obvious
economic characteristics and has a close relationship with economy development; As an emerging subject with limited research methods, it is quite necessary to build its unique research methodology system based on those fundamental subjects like mathematics; And with the rapid development and increasing new issues, high-fitting models for various problems is in urgent need to make predictions for the reference of decisions.

Therefore, the issue is not about whether this method is applicable to ICLT research but ICLT researchers. And it is unwise to turn a blind eye to the upcoming revolution brought by data. With its unique concepts and means, the database research method will contribute to both ICLT research method to a new level and also making ICLT research scientific.

References