Database Design on Corpus System for Chinese-English Translation of Scientific Papers

Yanming Li[1], Yongchang Ren[2*]

[1]*School of Applied Technology*
University of Science and Technology
Liaoning, Anshan, Liaoning, 114051, China
[2]College of Information Science and Technology Bohai University
Jinzhou, Liaoning, 121012, China

***Abstract* — Scientific paper translation is difficult for Chinese author, and corpus system can provide reference. Database design is the basic work for the system construction, the conceptual structure design, logical structure design and the integrity of the design are designed according to the steps of database in this paper, and inverted index design according to the full text retrieval. Summary structure design described 4 Chinese entities and 4 English entities, and the contact about 3 "one-to-many" and 4 "one-to-one" between the entities with E-R graph; Logical structure design of SQL Server 2005 database management system in accordance with the transformation rules designed the logical structure of data storage table; The integrity of the design which is based on the integrity of theory designed the primary key of each entity in view of the entity integrity, and foreign key in view of referential integrity between entities; Inverted index design took the inverted index data structure as the foundation, explained the process and method of creating the inverted index file. According to the contents of this paper constructed the corpus system for Chinese-English translation of scientific papers with reasonable storage structure and high access efficiency which can satisfy the needs of different staffs.**

Keywords - scientific paper; Chinese-English translation; corpus; database design

I. INTRODUCTION

Scientific papers is also known as the original papers or primary document in information science, it is the scientific analysis, comprehensive research and description about the phenomenon(or problems) in natural science, engineering technology and science and the arts and humanities field that science and technology personnel or other researchers on the basis of scientific experiments (or test) to study, and also some phenomena and questions for further, meanwhile to sum and innovate some results and conclusions, and begin the expression of electronic and written in accordance with the requirements of each science and technology periodicals. Scientific papers belongs to spiritual products, if it were to have more social and economic value, it would get social recognition first, namely scientific papers must have the function of communication. With the development of society and globalization of politics, economy, science and technology, the Chinese-English translation of scientific papers is becoming more and more prominent. Scientific papers relates to various fields of the natural sciences, has the advantages of accurate, objective, formal, rigorous logic etc Therefore, the purpose of translation of scientific papers is not only to let readers know, understand and appreciate the original ideas and style by the translation, more importantly, to report the leading thought of science text to the readers accurately, to achieve communication functions of scientific papers [1,2].

translation or adhere to the original form rigidly, cannot reproduction of the original language appropriately, cause to the conveyed information cannot obtain the anticipated effect, affect the expected function of text. And these mistakes often

The translation is a cross language communication behavior to express the ideological content of one language in another language accurately and completely. Translation is a science, and also a kind of art, which is a language to re-express things that already express in another language correctly and completely, so it is the key of international understanding that the progress of society cannot do without translation. To translation the scientific paper into English accurately, appropriately and fluently and can be understood and accepted by the readers of the translated text is not easy, there will inevitably be translation errors. Common translation errors are three [3]: the one is pragmatic translation errors, which is the highest level of the translation errors, and covers all the translation errors in a sense, because any fault will damage the expected function directly or indirectly. The main reason to produce functional translation errors is neglecting the translational function or readers in the process of translation; the two is cultural translation errors, which is refer to the lack of understanding of the culture background of the target language in communication caused by cultural differences which lead to the error of the choice of language form, namely the conflict that between norms and practices of translation and target culture; the three is linguistic translation errors, which is a specific level of translation errors. The most common language translation errors are reflected in the vocabulary, grammar, idioms and voice, The probability reasons for translation are using inaccurate or wrong expression in the will seriously affect the communication effect of the translation. How to achieve the communicative function of English translation accuracy of scientific papers and what kind of translation theory to analyze and avoid mistakes in

the English translation of scientific papers has become widespread problem.

Corpus refers to the large-scale electronic text library by scientific sampling and processing, is the basic resource of corpus linguistics, and is the main resource of empiricism language research methods. Since the computer corpus appears, the corpus linguistics has rapidly developed, using corpus to carry on contrastive linguistic studies and linguistic ontology research has achieved plentiful and substantial achievements [4,5]. Becker created the world's first translational English corpus, proposed the use of parallel corpus, multilingual corpus and comparable corpus which can find and determine the semantic features difficult to find by routine method, research text style, language habits, such as language redundancy, lexical co-occurrence, standard degree, coherent form, syntactic pattern, even the using feature of punctuation, and help the translator to select appropriate translational strategies.

In recent years, more and more Chinese teachers and researchers published in the International Journal of English, many units take the three papers (SCI/EI/ISTP) as the most important indicators of performance appraisal and position promoting. Scientific papers translation is always the difficulty for Chinese author and the translated scientific papers resources are precious wealth. The construction of corpus system of scientific papers translation for translation has the following effect [6]: In the stage of understanding, corpus can help the translator to determine the semantic prosody of the original lexical, the semantic prosody of the author and style of the original, so that to improve the correct of understanding and to provide promise guarantee for the faithfulness of the translation; In the stage of expression, can search synonyms, near synonyms sentences, collocations, sentence comparison and so on through corpus retrieval platform, the translator can obtain a lot of reference words, sentences and expressions, to make the output of translation more accurately; In the verification stage, the translator can use translational evaluation system to quantitative analysis of the translation, and find the lack of translation then embellished and perfect through the data comparison. Corpus translation systems need to handle large amounts of data, which are stored in the database, thus the core work of database design and construction of corpus translation system, determines the quality and the success or failure of the system. This paper uses standardized method to design database and provide support for the construction of corpus system for Chinese-English translation of scientific papers.

II CONCEPTUAL STRUCTURE DESIGN

Conceptual structure design also known as conceptual design is based on the requirements specification produced by the demand analysis stage, according to the specific method to abstract as a data model which is not dependent on any specific machine, namely conceptual data model. Conceptual data model gets rid the designer's attention of the complex implementation details, and focus only on organization structure and processing mode of the most important information. The conceptual model, mainly used in the database design stage of system development, is in accordance with the user's view of modeling data and information, to realize the use of entity relationship diagram. The relationship between each entity and the related entities of the description system of conceptual model is the system characteristic and static description. E-R diagram is also called entity relationship diagram, which provides the method to express entity types, attributes and relations, used to describe the concept model of the real world [7,8]. Among them, the entity use rectangle for representation, contact use diamond frame and properties use the oval frame. In this case, use rounded shape frame to represent attribute in order to save space. The E-R model diagram of conceptual structure design of the system database is shown as Fig. (1).

In Fig. (1), there were 8 entities which including 4 Chinese entities and 4 English entities. The four Chinese entities are Paper Type(Chinese), Paper Information (Chinese), Clause Information (Chinese), Word Information (Chinese); The four English entities are Type(English), Paper the Information (English), Clause Information (English), Word Information (English). Each Chinese entity corresponds to an English entity, the connection between entities is "one to one", which is a Chinese entity corresponds to only one Chinese entity, and the connection's name represented by "Exist". Four out of three Chinese entities are "one-to-many", whose name is represented by "Include", for example, contact Paper Type(Chinese) entity with Paper Information(Chinese) is "one-to-many", which means one Paper Type(Chinese) including many Paper Information(Chinese), but one Paper Information(Chinese) corresponds only one Paper Type(Chinese). The connection between the four English entities is exactly the same to Chinese ones, so no more description is needed.

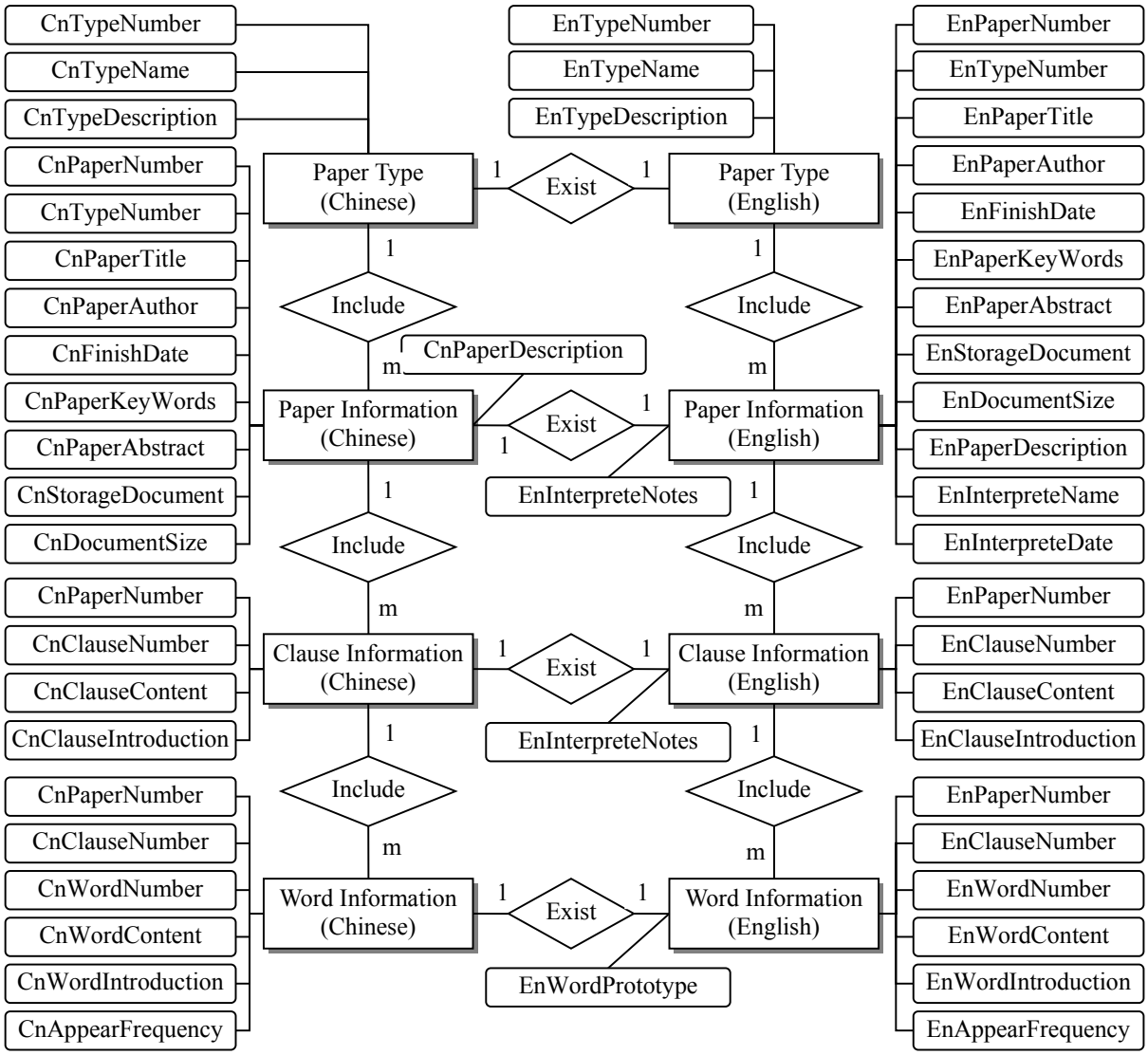

Figure 1 E-R diagram of system

III LOGICAL STRUCTURE DESIGN

The process from the conceptual structure design to the logic structure design is an important part of the database design. The logical structure design is appropriate or not, directly affects the function and efficiency of the entire database system. The design of this stage is initiative optimize in database system, when the database meets the needs of the conditions, the overhead of time and space is the best performance, and ensure good performance of system operation. The E-R model obtained in the stage of conceptual design is used to reflect the data demand mode of the user, which has nothing to do with the specific data model and DBMS. In order to establish the database required by the user, the conceptual model needed to be converted into some specific data model which support by DBMS. Its task is the process which converts the conceptual model into a specific data model supported by DBMS, meanwhile need to consider the performance characteristics of the data model and concrete of DBMS data model [9-11]. Logical structure design needs to deal with two aspects of the problem:

(1) Conversion rules. Conversion rules including the method of the entities and connections. The conversion rule of entity is "an entity converted into a relationship model. Attribute of the entity is attribute of relationship, the code of the entity is the code of relationship"; The conversion rules of connection are including binary connection such as "one-to-one", "one-to-many", "many-to-many" and "multiple" , in this system there is only "one-to-many", and the conversion rule is " A connection on one-to-many can be changed into an independent relationship model, which can also be merging with the relational schema of multi-terminal

corresponding. If the conversion is an independent relationship model, the code of each entity connected with the connection and attributes of themselves are converted to the attributes of relationship, and the code of relationship is the code of multi-terminal entity."

(2) Data type. This system uses SQL Server 2005 database management system, and the DBMS provides more than 20 kinds of data types, this system has selected five kind of data type [12]: One is the integer types, which choose "Int", such as the attribute of "CnAppearFrequency", and each value occupies 4 bytes of storage space; Two is the exact decimal type, which choose "Decimal", such as the attribute of "CnDocumentSize", and the storage space occupied by data is determined according to the integer and decimal digit; Three is the date and time types, which choose "Datetime", such as the attribute of "CnFinishDate", the data consists of valid date and time, occupies 8 storage bytes; The four is a binary type, choose "Varbinary", such as the attribute of "CnStorageDocument", the occupation of the number of bytes is "Max", which is the new requirements of the version of SQL Server 2005, before this the previous version used Image type to store binary data of any type; The five is the character type, choose the attribute of "Varchar", and most of the database used the type. Character data composed of any combination of letters, symbols and numbers. When the character data is less than 8000 with specific data indicates the occupied byte number, or represented by "Max", which is also the new requirements of the version of SQL Server 2005. In the previous version, when the data is greater than 8000 characters, stored as a "Text" type.

The logical structure design of system database is shown in Fig. (2).

IV INTEGRITY DESIGN

Database integrity refers to consistency, correctness, effectiveness and compatibility of the data of the database in logic [13]. Mainly from two aspects to understand the integrity of data: the accuracy of data, namely the value of each field must satisfy certain data type, range and restriction; the consistency of the data, namely the value of each attribute of the related entities must match each other. In database application system, ensuring the integrity of data is the basic requirements for the design of database. Data integrity is usually including the integrity of entity, referential integrity and user-defined integrity. The user-defined integrity is designed for each field, and this paper only designs the entity integrity and referential integrity.

(1) Entity integrity. The requirement of entity integrity is the primary key field of every table cannot be empty or repeated values. Entity integrity refers to the integrity of row which requires the entire row in the table has a unique identifier, called the primary key. The primary key whether can be modified, or the entire the column whether can be deleted, dependent upon the integrity required between the primary key with the other tables. Rule of Entity integrity is "the main attributes corresponding to all the primary key of basic relations are not null value". In the 8 tables of logical design in chart 2, the design of primary key is respectively as follows: The primary key of "Paper Type (Chinese)" table is "CnTypeNumber";The primary key of "Paper Type (English)" table is "EnTypeNumber"; The primary key of "Paper Information (Chinese)" table is "CnPaperNumber"; The primary key of "Paper Information (English)" table is "EnPaperNumber"; The primary key of "Clause Information (Chinese)" table is "CnPaperNumber and CnClauseNumber"; The primary key of "Clause Information (English)" table is "EnPaperNumber and EnClauseNumber";The primary key of "Word Information (Chinese)" table is "CnPaperNumber, CnClauseNumber and CnWordNumber"; The primary key of "Word Information (English)" table is "EnPaperNumber, EnClauseNumber and EnWordNumber";

(2) Referential integrity. Referential integrity is the data of the primary key in a table corresponding to the foreign key in another table, which ensure the effectiveness of the connection between the tables in the database and prevent the loss of data or spread of meaningless data in the database. In the process of the development of software, if referential integrity is not guaranteed, it will be disastrous for data [14,15]. Between four Chinese entities and four English entities are four connections on "one-to-one", which is realized through the establishment of 4 foreign key constraints. Such as "Paper Information (Chinese) and" Paper "form Information (English)" table, through the attribute of "CnPaperNumber" and the attribute of "EnPaperNumber", establish the foreign key constraint named "FK_EC_Paper".There are 3 connections on "one-to-many" links 4 Chinese entities, through the establishment of 3 foreign key constraints to achieve. Such as "Paper Information (Chinese)" table and "ChineseClauseInformationTable" table, through the "CnPaperNumber" attribute, establish the foreign key constraint named "FK_Cn_PaperClause". The 4 English entities also built 3 foreign key constraints. To sum up, totally built 10 foreign key constraints, has been marked by the arrow in Fig. (2).

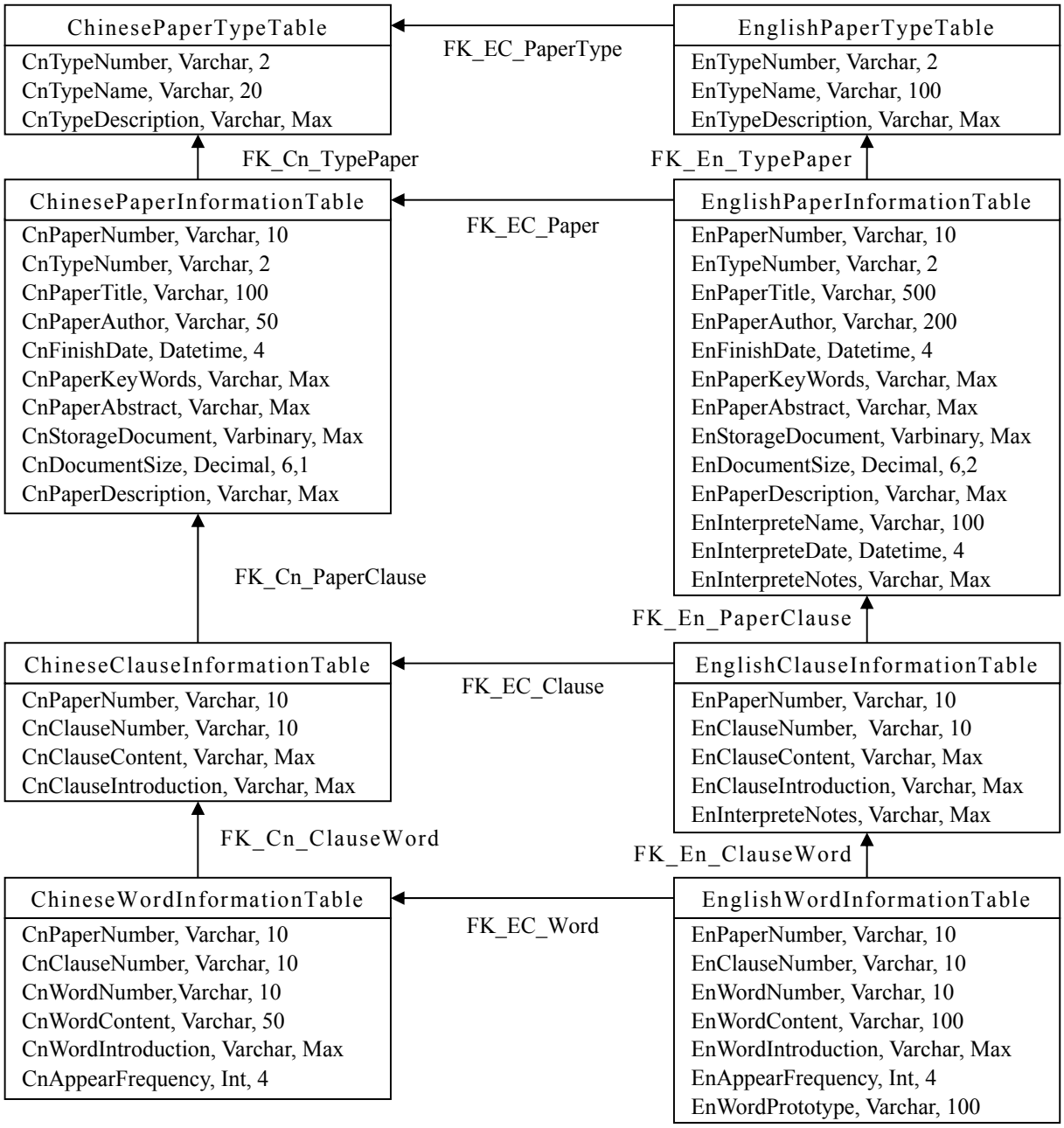

Figure 2. Logical structure model

v INVERTER INDEX DESIGN

The core function of corpus system is retrieval; inverted index is an important technique to improve the efficiency of full text retrieval. Inverter index stems from practical applications need to find records according to the attribute values, each item in the table include an attribute value and the address of each record with the attribute values. Because it does not determine the value of the attribute by records, but to determine the position of record by the attribute value, therefore is called the inverted index. The file with the inverted index called the inverted index file, referred to as the inverted file [16].

The inverted index in advance for information build index table, each item of index table consists of a word and relevant information of the word, because it is not to find words through a document, but to determine document through the words, so called it the inverted index. The main advantages are: when treated the complex queries of multi-keyword, complete logical operation as intersection, union of

query first in the inverted list, and access the record after get the result. It does not have to access each record randomly, but convert the inquiry of record into the operation of set address, so as to improve the speed of search [17].

A. Data Structure on Inverter Index

Inverted file is data structure to describe a term set element (Teams) and a collection of documents element (Docs) of corresponding relationship [18]. Term sets are used to save dictionary tables of vocabulary all the query keywords, these keywords are extracted from the indexed document, data structure can be described as:

$$\begin{gathered} < k_1, < df_1, < d\_list1 >> \\ k_2, < df_2, < d\_list2 >> \\ \cdots\cdots \\ k_i, < df_i, < d\_listi >> \\ \cdots\cdots \\ k_n, < df_n, < d\_listn >>> \end{gathered} \tag{1}$$

In the formula, k_i is representing the keyword of the order of *i* in dictionary tables of vocabulary, df_i represents the number of *i* included in the list of documents, $< d\_listi >$ represents the information of the list of documents about the *i* keyword.

The document sets are used to save the information in the document list, data structure described as:

$$\begin{gathered} < d_1, f_1, < p_{11}, \cdots, p_{1f} > \\ d_2, f_2, < p_{11}, \cdots, p_{2f} > \\ \cdots\cdots \\ d_i, f_i, < p_{i1}, \cdots, p_{if} > \\ \cdots\cdots \\ d_n, f_n, < p_{n1}, \cdots, p_{nf} >> \end{gathered} \tag{2}$$

In this formula, d_i represents the ID of document in order *i* where the keyword appeared, f_i represents times of the keyword appeared in order *i*, $< p_{n1}, \cdots, p_{nf} >$ the appeared set of position.

B. Establishment on Inverted Index File

The index construction is a process from positive table generates inverted table. When the index set up, get the inverted list. The creating process is shown in Fig. (3)[19].

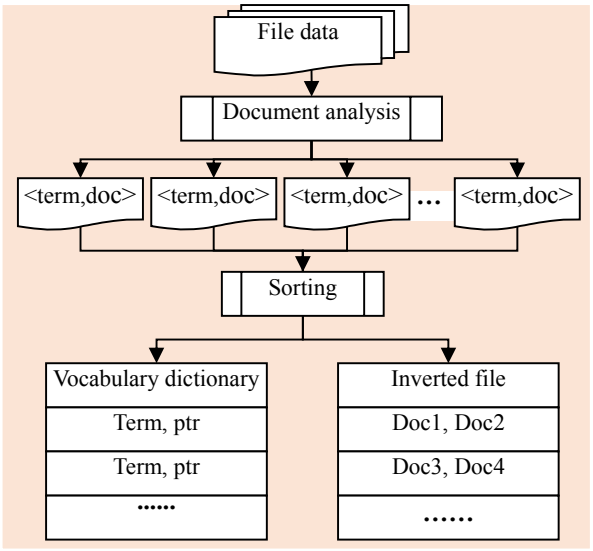

Figure 3. Establishment on inverted index file

The creating process in three steps is described as below:

The first step: Analysis the document into words term marker by lexical analysis. English segmentation is relatively simple, can be separated by spaces. But what need special attention is the process of word segmentation can reduce prototypes of English word. Through three steps: One is according to space / symbol segmentation, using regular expressions is easy to implement; Two is eliminating the stop word, which is similar to high frequency as the a/an/and/are/then, high frequency words will produce great interference for calculation formula based on the word frequency, so need to filter; Three is Stemming, such as deformation of singular and complex, -ing and -ed, but in the calculation of correlation, as the same word. For example, apple and apples, doing and done are the same word, the purpose of stemming extract is to merge these abnormal. Stemming has 3 major mainstream algorithm, Porter Stemming, Lovins Stemmer and Lancaster Stemming.

The second step: Using the method of random hyperplane hash to get rid of repeated word term. For a vector *v* of dimension *n*, to get a sign of digit number $f(f << n)$, the algorithm is as below: randomly generate vector *v* of dimension *n* as $r_1, r_2, \cdots, r_f$ in number f, for each vector r_i ,if the dot product of v and r_i more than 0, the digit number *i* of final sign is 1, otherwise is 0. Randomly generated *f* hyperplane in dimensional *n*, each hyperplane divides space of vector *v* into two parts, if v above the hyperplane is given a 1, or get a 0, then together obtained 0 or 1 in number *f* into a signature about dimensional f. If the angle between the two vectors *u* and *v* is θ , then the probability of a random hyperplane separate the angle is θ/π , so the probability of corresponding digit of signature different about *u* and *v* is θ/π . So you can use number of different corresponding digit to the signature of two vectors, namely the Hamming distance to measure the different degree of two vectors. The specific approach is to construct

function hash, establish a unique ID for each word, then according to the ID to count the repetition of words.

The third step: Inverted generate a list of words. Gets the word number, number of the words from the beginning of 0, and a continuous distribution, if the word exists in the dictionary, returns the number corresponding the dictionary; If it is not in the dictionary, insert in the dictionary after assigning a new number; Updating the inverted index, if the word exists in the dictionary, update only the corresponding inverted list, if not present in the dictionary, the dictionary of inverted table increases at the same time, it can guarantee the size of inverted index same to the size of dictionary. Converts the document into the vector composed of words' number, because in the process of constructing the index has represented the document by the words' number, and dictionary information is no longer needed, which can be release the occupied space in order to save space.

VI CONCLUSIONS

Corpus translation studies means based the specific target of research, to establish a corpus based true translational corpus, including monolingual comparable corpus, bilingual / multilingual parallel corpus, translational corpus and some other types. Translational corpus needs to be detailed annotation of translation text, the translator's information and other factors, the corresponding corpus needs to do aligned process of sentence or some level for two kinds of corpus; Comparable corpus needs to marked the style, theme, the author, the translator and other factors. Corpus translation studies take electronic text as the foundation, take the computer statistics as means, do a large range or specific scope of description for various translational phenomena, on the basis of the full description, inquiry two languages and converted process, characteristics and rules of them, then analysis and explain the translational phenomena or verify hypotheses about translation. In essence, corpus translation studies are the product of interdisciplinary combine Descriptive Translation Studies with Corpus Linguistics.

The content of this research, fill up the blank in the translational corpus. The use of corpus for Chinese-English translation of scientific papers can promote the reform of scientific papers in translational studies, guide the translation of scientific papers to deepen, get rid of limitation of scientific papers translation, expanding thoughts of scientific papers translation. Corpus system on Chinese-English translation of scientific papers in the contents of this paper has reasonable storage structure and access efficiency which can satisfy the needs of different staffs. The actual application according to need improve and optimize performance of database., by building an inverted index to improve the speed of full-text retrieval and creating a unique index to ensure the consistency of the data or through the trigger to realize the associated operation and other means.

ACKNOWLEDGMENT

This work is supported by the social science fund project of Liaoning province, China (No. L14BTY002).

REFERENCES

[1] X. L. Tian, "Strategies in the Chinese-English Translation of Scientific Paper Under the Guidance of Skopos Theory," Journal of Northwest University (Philosophy and Social Sciences Edition), vol. 39, No. 5, pp. 165-167, 2009.

[2] X. M. Wang, "Application of Functional Translation Theory in Scientific Paper C/E Translation," Journal of Shenyang Jianzhu University(Social Science), vol. 14, No. 1, pp. 92-95, 2012.

[3] X. M. Wang, "Study of Translation Errors in Scientific Paper C/E Translation based on Skopostheorie," Northwest Medical Education, vol. 21, No. 4, pp. 773-777, 2013.

[4] Velislava Stoykova. Teaching Corpus Linguistics. Procedia - Social and Behavioral Sciences. vol. 143, No. 08, pp. 437-441, 2014.

[5] Blanka Frydrychova Klimova. Using Corpus Linguistics in the Development of Writing. Procedia - Social and Behavioral Sciences, vol. 141, No. 01, pp. 124-128, 2014.

[6] G. L. Yu, "Applications of Corpus in Translation Studies," Foreign Language and Literature, vol. 26, No. 02, pp. 117-121, 2010.

[7] Sudha Ram, Vijay Khatri. A comprehensive framework for modeling set-based business rules during conceptual database design. Information Systems, vol. 30, No. 02, pp. 89-118, 2005.

[8] Y. Zhang, "Database design on enrolling management system of college sports class majors," Information Technology, vol. 38, No. 6, pp. 42-45, 2014.

[9] Y. Zhang, "Eliminating Process of Formalization in Logical Database Design," Computer Systems & Applications, vol. 22, No. 6, pp. 179-181, 2013.

[10] H. M. Chen, "Logical structure design of database," Fujian Computer, vol. 28, No. 10, pp. 214-215, 2012.

[11] Ken England, Gavin Powell, "Logical Database Design for Performance," Microsoft SQL Server 2005 Performance Optimization and Tuning Handbook, 2007.

[12] Sergio Flesca, Filippo Furfaro, Francesco Parisi, "Consistency checking and querying in probabilistic databases under integrity constraints," Journal of Computer and System Sciences. vol. 80, No. 7, pp. 1448-1489, 2014.

[13] Michael V. Mannino, "Database Design Application Development and Administration (Second Edition)," McGraw-Hill Education, 2004.

[14] L. J. Zhao, "The Data Referential Integrity of Analysis in SQL Server 2000 Databases," Office Informatization, vol. 16, No. 24, pp. 40-41, 2010.

[15] Wonmook Jung, Hongchan Roh, Mincheol Shin, Sanghyun Park. Inverted index maintenance strategy for flashSSDs: Revitalization of in-place index update strategy. Information Systems. vol. 49, No. 04, pp. 25-39, 2015.

[16] W. N. Dai, "Study and Implementation of Inverter index on Hadoop," Master's degree of University of Electronic Science and Technology, 2013.

[17] Daniele Broccolo, Lorenzo Marcon, Franco Maria Nardini, et al. Generating suggestions for queries in the long tail with an inverted index. Information Processing & Management. vol. 48, No. 02, pp. 326-339, 2012.

[18] K. F. Wang, L. B. Huang, "Tearms in Corpus-Based Translation Studies," Journal of Sichuan International Studies University, vol. 23, No. 6, pp. 101-105, 2007.

[19] Wonmook Jung, Hongchan Roh, Mincheol Shin, Sanghyun Park. Inverted index maintenance strategy for flashSSDs: Revitalization of in-place index update strategy. Information Systems. vol. 49, No. 04, pp. 25-39, 2015.