Scientific Research Publishing

Int. J. Communications, Network and System Sciences, 2017, 10, 219-226
http://www.scirp.org/journal/ijcns
ISSN Online: 1913-3723
ISSN Print: 1913-3715

Error Searching System with Keyword Extraction and Keyword Fuzzy Matching

Fan Yang, Zhenghong Dong, Lihao Liu

Academy of Equipment, Beijing, China
Email: sarah0824@hotmail.com, dzh.bj@163.com, 10494901@qq.com

How to cite this paper: Yang, F., Dong, Z.H. and Liu, L.H. (2017) Error Searching System with Keyword Extraction and Keyword Fuzzy Matching. *Int. J. Communications, Network and System Sciences*, **10**, 219-226.
https://doi.org/10.4236/ijcns.2017.105B022

Received: April 18, 2017
Accepted: May 23, 2017
Published: May 26, 2017

Abstract

This paper has proposed an error searching method to search the solutions of errors that occurred in the unified commanding platform mix-deployed software (UCPMD). Because those errors belong to different stages or may be happened in different services, applications, IP ports, system software, or different versions of software, and those errors are also can be classified into different types. It is necessary to locate accurate reason that cause an error as well as find out its solution. The proposed error searching system applies Chinese keyword extraction and Chinese fuzzy matching between keywords, which considers the processed keywords as the index to find out the solutions of errors. Besides, the error searching system had made correspondence among errors, reasons, and solutions, and put them to different categories in terms of their characteristics, such that it is easy to manage, search, and use. Among others, we have added specialized thesaurus as the index of keywords, which enriches and completes the searching results. Because of the proposed error searching system evolves keyword extraction and keyword fuzzy matching technologies; it is more accurate to find out user-interested solutions.

Keywords

Database Design, Search Engine, Extraction, Fuzzy Matching

1. Introduction

Unified commanding platform mix-deployed software (UCPMD) integrates 22 sub-system from 4 different institutes, including 92 different software. Because of the differences of underlayer protocol and the differences of standard, there are many errors occurred during the stages of setup, configuration, and operation, which seriously affect the usage. Moreover, because those errors are various, which may be happened in different operation phases, stages, TCP/IP

communication protocol layers, sub-sys- tem software, it is necessary to design a database system which can manage those errors. The proposed method provides a design of error searching database, which can search the errors occurred in the stages of setup, configuration, and operation, and also provides the reason that cause the error as well as the corresponding solution. The proposed method effectively finds out the solutions of errors occurred in the UCPMD platform.

The current error searching systems are various, including On-Board Diagnostics designed for vehicles [1], which can search vehicle errors according to vehicle OBD error code; Computer error searching system [2], which can search the hardware problems, network problems, and software problems, etc. Those error-searching systems focus on errors in specific area, designing specific databases to store the errors as well as the characteristics of those errors. Because UCPMD is used for commanding and ordering between the superiors and the subordinates in specialized field, the design for DB tables, logical structures has to build up according specialized characters of UCPMD platform and the design and definition for keywords requires personalized customization. In that case, the current error searching software cannot be applied to this platform. It is necessary to design a specialized database system to effectively solve the various errors occurred in the UCPMD platform.

This paper proposes an error searching method to search the solutions of errors that occurred in UCPMD, which evolves Chinese keyword extraction [3] [4] and Chinese keyword fuzzy matching [5] technologies, and considers the processed keywords as the index to search errors, the reasons that cause the errors, and the corresponding solutions. Besides, this error searching system had made correspondence among errors, reasons, and solutions, and put them to different categories in terms of their characteristics, such that it is easy to manage, search, and use. Among others, according to the specialization of the system, we have added specialized thesaurus as the index of keywords, which enrich and complete the searching results.

2. Related Work

2.1. Keyword Extraction

We applied IK Analyzer [6] to extract keyword. IK Analyzer is a lightweight Chinese participle and open source develop toolkit based on Java, which combines dictionary participle as well as sematic participle. It adopts "forward iteration finest-grained participle algorithm" [7], to support two ways of participle mode, which are fine-grained and intelligent participles. Intelligent participle supports simple process of ambiguity exclusion [8] and combined output for quantifiers. Besides, IK Analyzer adopts multi-processor analysis mode [7], which can support English letters, digitals, and Chinese characters, etc.

However, this method can only separates words from text, even the unnecessary words, such as "a", "as", "of" etc. It cannot extract meaningful words from the separated words. The good news is that it allows user to configure self-defined "extension stop dictionary" which can make the separation more

intelligent.

2.2. Keyword Fuzzy Search

Lucene is a developing toolkit for full text search engine [7], which supports for Java development. It provides Fuzzy searching (FuzzyQuery) function [9]. The reason why this paper applied FuzzyQuery for fuzzy searching is because FuzzyQuery makes use of similarity matching, which can recognize two similar words. FuzzyQuery makes use of the best string matching technical based on Damerau-Levenshtein Distance algorithm [10] to compute the transfer steps from one word to another, which is considered as the basis of marking similarity. If the similarity is less than a set value (normally, the value is 0.5), then the two words are considered as similar.

3. Proposed Method

3.1. System Function Design

Figure 1 shows the system function design. The error database system contains two modules, which are search engine and database, where database has three subfunction modules explained as follows.

1) Data import/enter

This function supports two ways of importing data. One is importing by Excel file directly, and the other is entering data by administrator.

2) Keyword fuzzy search

This function supports for the fuzzy matching between the extracted keywords and the keywords in keyword table. The proposed method involves keyword extraction and keyword fuzzy matching technologies, which can obtain more accurate related results.

For fuzzy matching, this paper has involved two ways of fuzzy matching. The first is literally similar, which means if two keywords have the same characters, then they are similar. The second is that word meanings are similar, which means even if there is no same character between the two keywords, but they

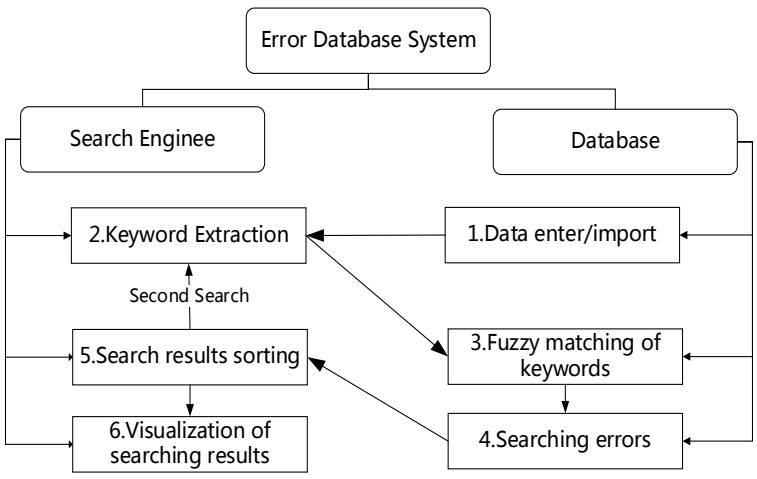

Figure 1. System function diagram.

have similar meaning, then they are still similar. By using two ways of fuzzy matching, the searching results are more accurate; otherwise, parts of searching results are missing, which might be the solution of the searched errors.

For example, when carrying out keyword extraction, if the operator types are "人民群众的基础" (the basis of people), then the extracted keywords are "人民"、"群众"、"基础" ("people", "masses", "basis").

Second, we carry out fuzzy matching for the extracted keywords "人民" ("people"). And the results are "人们"、"人群"、"公民"、"民众" ("people", "crowd", "citizen", "populace") if we apply the first way of fuzzy matching.

And the results are "百姓"、"基层" ("common people", "grass roots") if we apply the second way of fuzzy matching.

Then we apply the keywords obtained by both ways of fuzzy matching, then we can find out the corresponding errors, reasons, and solutions.

3) Find out the corresponding error, reason, and solution according to the extracted keywords.

Search engine includes 4 sub-function modules.

1) Keyword extraction

To extract useful keywords from the input contents. Keyword extraction technology can automatically make participle for the searching contents, and then extract keywords as the searching index. Besides, considering that the way of descriptions are different between the input content and the keywords in database, it is necessary to do fuzzy matching to extract keywords, and find out the related errors.

The searching order is first to carry out fuzzy matching with Chinese thesaurus of Lucene, and find out similar keywords, and then use the found keywords to do fuzzy matching with the keywords in database to find out the corresponding keywords. If there is no appropriate keyword in Chines thesaurus, then directly carry out fuzzy matching with the keywords in database.

2) Sorting for the searching results

This method sorts the searching results according to matching degree.

3) Second search

The system also supports second search from the already searched results. Similarly, this paper first applies keywords extraction, and then carries out fuzzy matching. Next is to find the corresponding solutions in database, and then to sort the orders of the searching results to make the results finer.

4) Visualization of searching results

This is to display the searching results visually.

3.2. Flowchart

Figure 2 gives the whole flow chart of the error searching system which contains 5 stages, which are keyword extraction, fuzzy matching, find out error ID, soring searching results, and visualization of searching results, respectively. As searching process is the focus of this paper, which only evolves stage1, 2 and 3, so we explain the 3 stages in more detail as **Table 1**.

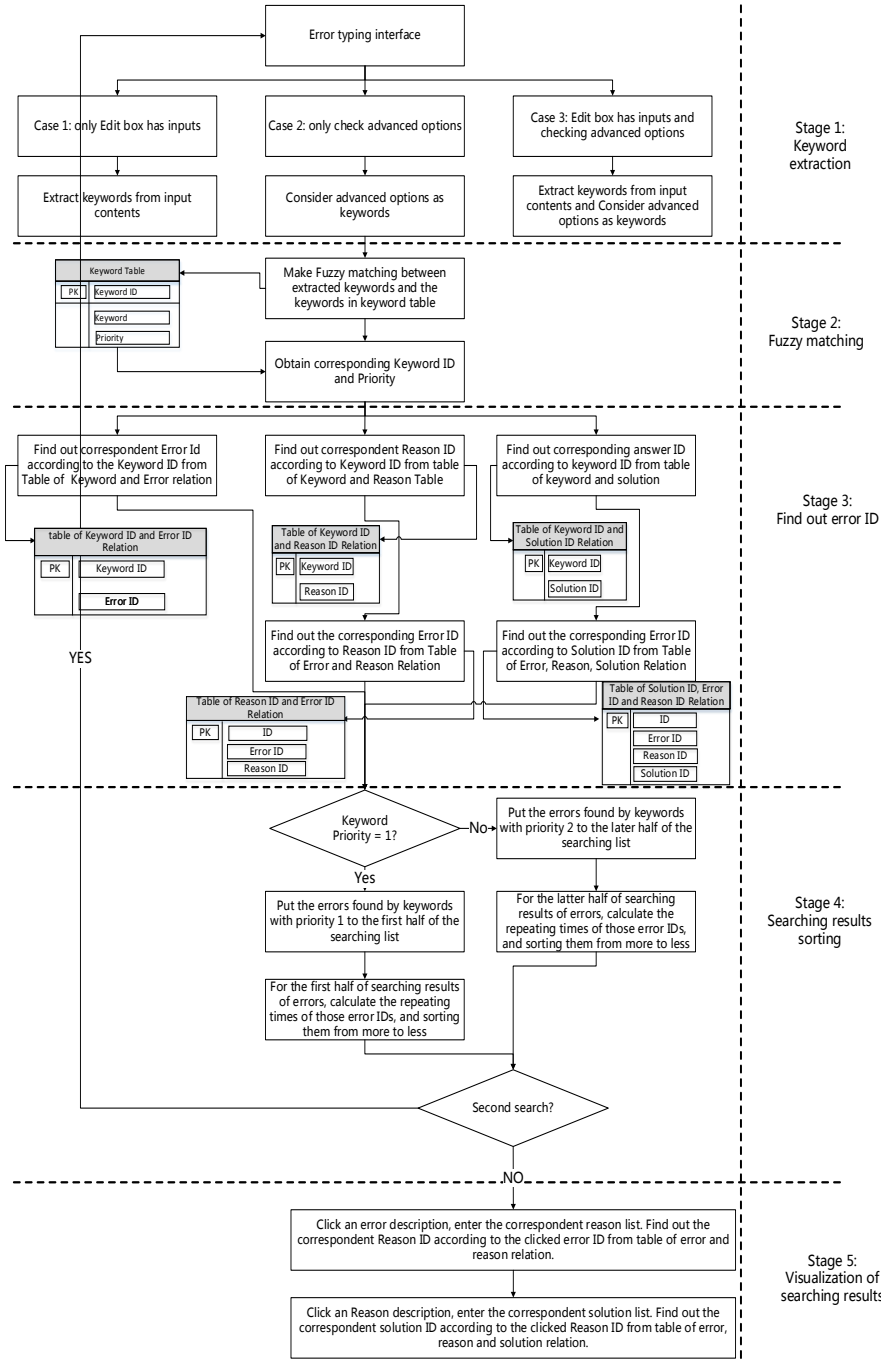

Figure 2. System flowchart.

4. Experiments and Prototype

This paper has implemented a prototype of this searching system for verifying if the searching is valid and useful.

We have shown the prototype by **Figure 3-6**. **Figure 3** shows the searching input interface, where user can type an error description in the edit box in the middle of this page, and click "故障诊断" (error diagnose). Or the user can use advanced search to refine the search contents by selecting advanced options, including error stages, type of error, error layers, occurred system, occurred soft

Table 1. Explanation on Database searching flowchart.

Searching stages	Contents	
Stage 1 **Keyword extraction**	Case 1	Only typing Edit box: extract keywords from the typed contents (several keywords, including the characters of errors as well as normal keywords)
	Case 2	Only check the advanced searching options. Considering each option as a keyword, and those advanced options can help to accurately locate the stages and locations that errors occurred.
	Case 3	Edit box has inputs and advanced options have been checked as well. To extract keywords from the contents in edit box, and consider the checked options as keywords.
Stage 2 **Fuzzy matching**	Carry out fuzzy matching between extracted keywords and the keywords in the keyword table, find out the correspondent keyword ID and priority in the keyword table. First, to fuzzy match with the words in the word database provided by Lucene, find out similar words, and then use those similar words to fuzzy match with the keywords in keyword table, and find out the correspondent keywords. If there is no such similar words in Lucene database, then directly fuzzy match the extracted keywords with the keywords in keyword table.	
Stage 3 **Searching for error ID**	1) Find out Error ID according to relation of keyword ID and error table; 2) Find out Reason ID according to relation of keyword ID and reason table; 3) Find out the correspondent Error ID according to relation of Reason ID and Error ID table; 4) Find out Solution ID according to relation of keyword ID and solution table; 5) Find out the correspondent Error ID according to relation of Solution ID and Error ID table;	

Figure 3. Search input interface.

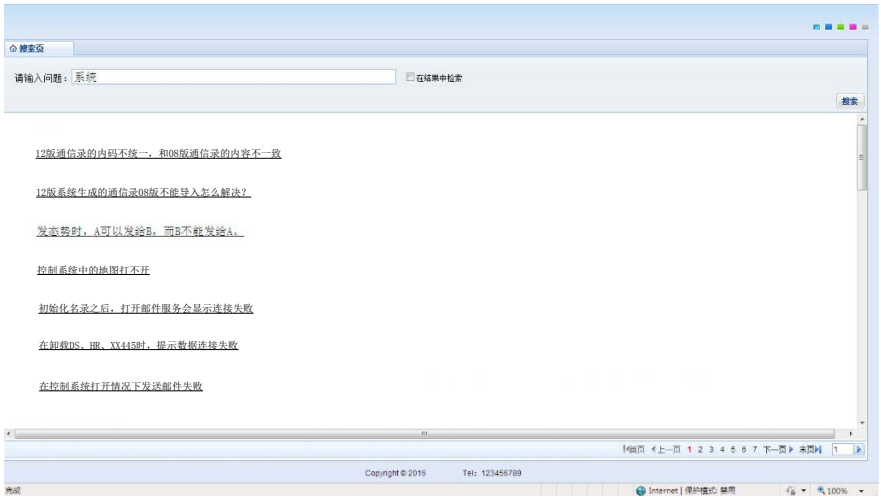

Figure 4. Search results sorting page.

ware, etc. And **Figure 4** shows the searching results sorting page. After searching for keyword "系统" (system), the searching results are list as **Figure 4**. By clicking any searched item in the list, such as the 3rd one. The reason that causes this error is shown as **Figure 5**. By clicking the reason shown in **Figure 5**, the corresponding solution is shown as **Figure 6**. User can search out interested results in this way.

5. Conclusion

This paper has proposed an error searching method to search the solutions of errors that occurred in the UCPMD. This method applies Chinese keyword extraction and Chinese keyword fuzzy matching technologies to find out user interested searching results. The searching results come from errors, reasons, and solutions, which means as long as an indexed keyword appears in any of the descriptions of errors, reasons, solutions, the corresponding set of error, reason,

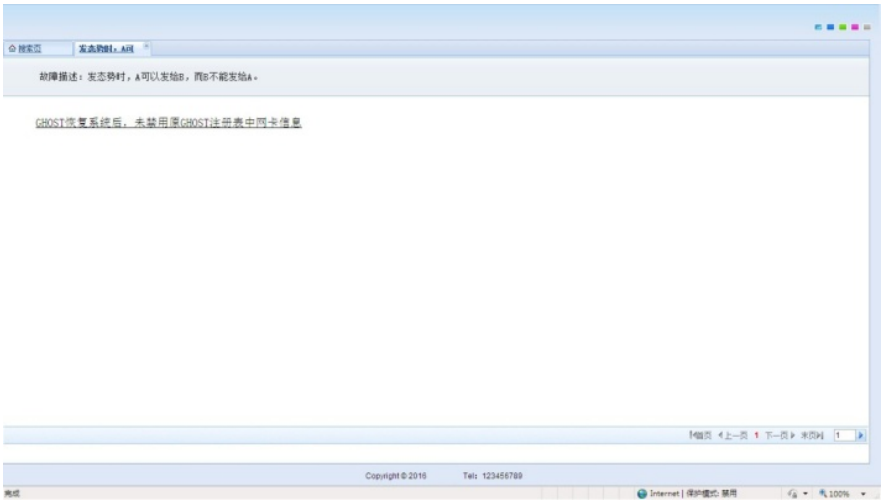

Figure 5. A selected error and the reason that causes the error.

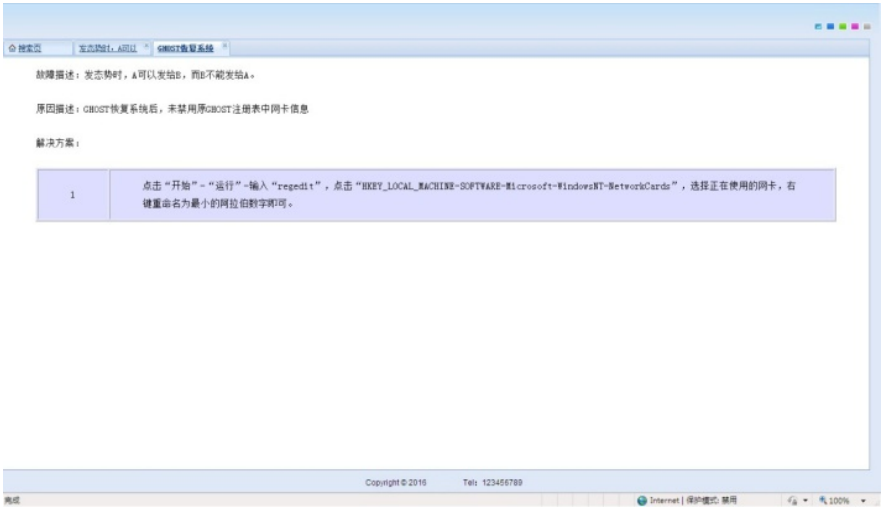

Figure 6. A selected error, the reason that causes the error, and the corresponding solution.

and solution would be listed in the searching results. We also provide a prototype of the method to show the effectiveness and correction of this method.

References

[1] Wang, J.H., Fang, M.D., Gao, J.D., Lu, H.Y. and Dai, C.B. (2006) Basic Principle and Application of On-Board Diagnostics for Gasoline Fuelled Vehicles. *Automotive Engineering*, **28**, 491-494.

[2] Lu, C., Yang, Y.-H. and Xu, G.-M. (2008) Exploitation of Computer Problem Repair and Require on Web System.

[3] Wang, L.-X. and Huai, X.Y. (2012) Semantic-Based Keyword Extraction Algorithm for Chinese Text. *Computer Engineering*, **38**.

[4] Fang, J., Guo, L. and Wang, X.D. (2008) Semantically Improved Automatic Keyphrase Extraction. *Computer Science*, **35**.

[5] Wang, J.-F., Wu, X.-J., Xia, Y.Q. and Zheng, F. (2007) An Approximate String Matching Algorithm for Chinese Information Retrieval Systems. *Journal of Chinese information Processing*, **21**.

[6] Bai, Y.-C., Fu, W. and Xin, Y. (2014) Research and Simulation of Distributed Search Engine Based on Hadoop and Nutch. *The* 19*th National Young People Communication Academic Annual Symposium*.

[7] Gao, C.J. (2013) Research on Lucene Search Engine Based on PSP-BP Neural Network. China University of Petroleum (East China), Master Degree Thesis.

[8] Liu, Y.Z. (2005) Research on Chinese Auto Participle Exclude Ambiguity Algorithm. Chongqing University, Master Degree Thesis.

[9] Hu, H.B. (2015) The Implementation of a Variety of Sorting Methods Based on Lucene. *Computer Knowledge and Technology*, **11**, 57-59.
http://en.wikipedia.org/wiki/Damerau-levenshtein_distance