Design and Implementation of Keywords Extraction and Management System Based on Java Platform

Teng PAN [1,a] Bing LIU [2] Yanshuang LIU [3] Yunjie LI [4]

[1,2,3,4] PetroChina Pipeline R&D Center, 51th Jinguang Road. Langfang. Hebei. China

[a] zsypt@sina.cn

Keywords: Keyword extraction and management; System design; Software architecture; Java platform

Abstract: This essay discusses the design and implementation of a keywords extraction and management system based on java platform. By using development methodology of domain engineering, the essay focuses on key technology of keywords extraction system and outline some ideas of the software architecture, in which identify basic principles and requirements of the development of keywords extraction and management system. The system functions are implemented with Java language in Eclipse environment. The system is featured with expandability, reliability, suitability and friendly interface demonstrated in application test of engineering practice and plays positive practical value in keywords extraction and management in technical domain.

Introduction

Keyword is a high level summary of a document, presenting the overall content and topic of the document. Keyword extraction refers to how to extract certain words or phrases from a document automatically to present the document's tropic objectively and accurately. Keyword extraction technology enjoys important application value due to its extensive application in various information processing domains such as information retrieval, text categorization/clustering, information filtering, data mining, document abstract, automatic translation and domain concept system establishment [1].

Keyword extraction and centralized management positively enhanced term standardization process in technical technological domain. Statistics shows correct identification of keyword play certain role on regulating and integrate usage of keyword and terminology [2]. Keyword extraction and management is an important work in term database building.

System Requirement Analysis

Design requirements of Keyword extraction and management system is as follows:

1) simple interface, easy operation, stable running;

2) system functions include registration/sign-in, text import, text export, format conversion (auto conversion from word text to txt), text segmentation, keyword extraction, task management and inquire, keyword addition, keyword deletion, keyword modification and keyword inquire;

3) system functions must complete and keyword extraction satisfied to expected requirements;

The system's logic module is Figure 1 as following, which is derived from in-depth requirements analysis and based on system objectives in line with software engineering.

System Structure Design and Software Architecture

Introduction of System Structure and Functions

The primary functions based on the system requirements analysis are registration/sign-in, task management and inquire, data import/export, keyword extraction and keyword management.

1) Registration/sign-in module: the module requires personal information of users to record how users operate the system.

2) Task management and inquire module: every extraction of keyword has to be implemented

under task management. Users can select a task or open task. The module is also featured with historical task query. A task window records created task name, address and creation time of the user.

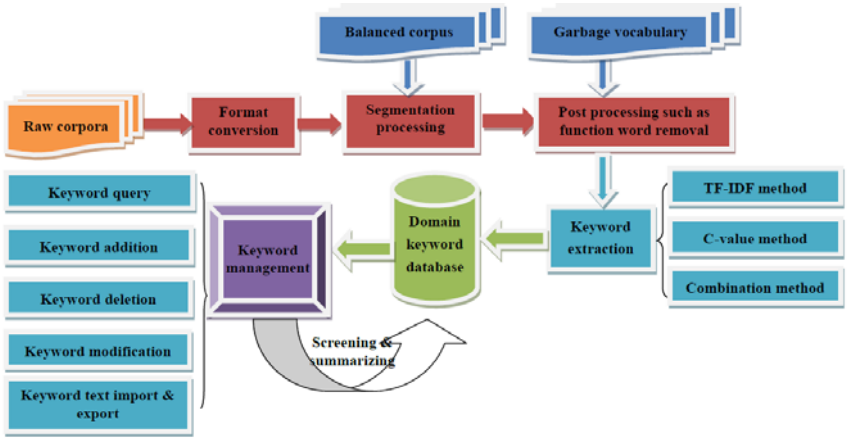

Figure 1 Flowchart of keyword extraction and management system

3) Word segmentation module: it refers to corpus segment, which is the foundation work for keyword extraction. The module is designed to corpus segment in domain text and prepare for keyword extraction.

4) Keyword extraction module: it is keyword extraction which offers TF-IDF method, C-value method and combination method.

5) Keyword management module: the module offers such functions as keyword upload, query, addition, deletion and modification.

Figure 2 is the system functional structural.

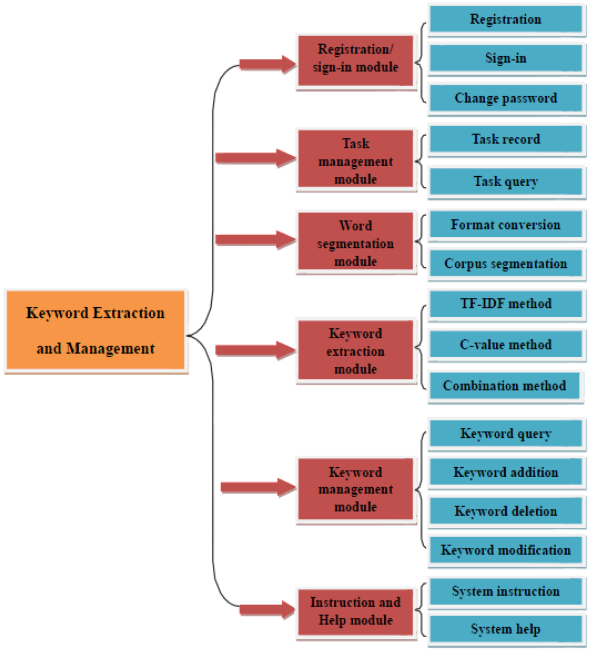

Figure 2 System functional structural

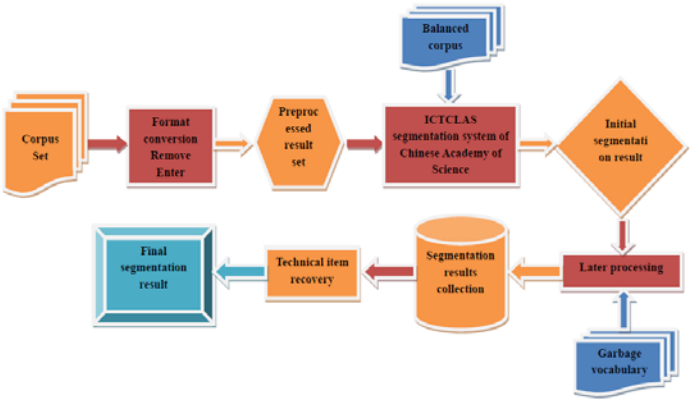

Figure 3 Corpus segmentation flow

Key Technology of Software Architecture

The key of the system is how to accurately describe word segmentation module and keyword extraction module.

Corpus Segmentation Module

Corpus segmentation flow is shown in Figure 3. First is to convert given Word text to Txt file through preprocessing handle. Next is to segment preprocessed text clustering with ICTCLAS, word segmentation tool of Chinese Academy of Science, followed by initial filtering of segmented

words for technical words and symbol, such as obsolete word and punctuation. Last is rational combination and splitting of segmented word to determine proper boundary of domain words in order to achieve best segmentation processing size.

Keyword Extraction Module

It is observed that technical domain keyword is featured with close combination of each component that constitutes technical domain keyword and certain word regular pattern of class combination, language regularity on one hand. And on the other hand is higher frequency of presence in their domain text in comparison to common text due to the strong territoriality of keyword, and even frequency distribution in each text of their domain, namely frequency property [3]. The current extraction method of technical domain primarily adopts rule based method, statistic based method or combination of the two methods due to such features.

Rule based method in this system uses a popular algorithm: C-value parametric method [4]; Statistic based method uses TF-IDF (Term Frequency-Inverse Document Frequency) parametric method [5].

C-value looks at relations between term contexts, which affected by three factors: a) high density occurrence of current string in corpus; b) number of term candidate including current string; c) type of term candidate including current string, the algorithm method as formula (1):

$$C-value(S)=\begin{cases} F(S), & S \text{ is maximum string} \\ F(S)-T(S)/C(S), & S \text{ is non-maximum string} \end{cases} \quad (1)$$

In the formula, *S* is candidate string, *F(S)* is the occurrence frequency in the text, *T(S)* is frequency of all the parent string of *S* in the text, and *C(S)* is the quantity of all parent string of *S*. From the definition above, if *S* is maximum string then it does not exist in the parent string, as *C-value(S) = F(S)*; If *S* is substring, C-value parameter takes full account of net relations between substring *S* and all parent strings. For example, when maximum string S_1="China petroleum" and substring S_2= "China", if $F(S_1) = F(S_2)$, then $C\text{-}value(S_1) = F(S_1)$ and $C\text{-}value(S_2) = 0$. Therefore, C-value takes account of word embedding and distinguishes parent string and substring, which plays important role in extraction of long terms.

TF-IDF method is a common algorithm of term extraction and takes full advantage of word distribution information in the overall text. TF (Term Frequency) refers to occurrence frequency of certain entry in the text. DF(Document Frequency)refers to the total quantity of certain entry in the corpus document. The algorithm is as Formula 2:

$$TFIDF(t_{i,j})=\frac{n_{t,j}}{\sum_k n_{k,j}} * log\frac{|D|}{|d|} \quad (2)$$

In the formula, $t_{i,j}$ is the number *i* entry occurs in document *j*, $n_{i,j}$ is frequency of $t_{i,j}$ occurrence in the document. $\sum_k n_{k,j}$ is the quantity of all vocabulary in the file. $|D|$ is the total of file in corpus library, and $|d|$ is the quantity of files that include $t_{i,j}$.

System Implementation

The Java platform based keyword extraction and management system uses the above development principles and algorithm structure in Eclipse environment. System hardware and software platform is:

(1) Hardware environment: Client end, CPU >1.7GHz; memory capacity: >1GB; hard drive: >250GB.

(2) Software environment: Client end OS, Windows XP、Windows 7.

(3) Development tool: The system built based Java language in Eclipse environment.

The system interface is shown as Figure 4.

Published by Atlantis Press, Paris, France.

Figure 4 System interface

System Evaluation and Conclusion

System Evaluation

The advantages of Java based keyword extraction and management system are:

(1) Good expandability

The system can process keyword extraction and management for various technical domain corpuses to meet needs and requirements of multiple technical domain users as the system logic processing is for technical technologic domain corpus.

(2) High run efficiency

The system client end connects to database directly which does not requires internet connectivity and broadband width requirement. The respond speed is related to algorithm computing speed.

(3) High accuracy

The tests conducted by developer and authoritative software testers shows that the system achieved over 92% correctness in corpus segmentation and over 90% correctness on keyword extraction.

Conclusion

The essay analyzes overall requirements of keyword extraction and management system, outlines software development principles and key technologies and presents the delivery of the entire system architecture according to software engineering. The system is easy to expand by using different database, with very stable algorithm module and concise interface.

The system has put into operation in PetroChina Pipeline Research & Development Center. The system successfully extracts oil industry vocabulary and becomes helpful assistance to oil industry terminology standardization.

References:

[1] Jiaheng Zheng, Jiaoli Lu, Keyword Extraction Method Study, J. Computer engineering, 31 (2005) 194-196.

[2] Zhike Wang, Haixia Chen, Technical Term Standardization and Role. J. Term Standardization and Information Technology, 3(2007) 10-12.

[3] Wenjing Zhang, Yinhong Liang, Term Extraction Technology Study, J. Information Technology, 3 (2008) 6-9.

[4] Yinhong Liang, Wenjing Zhang, Youcheng Zhang, C Value and Term Extraction Combined with Mutual Information, J. Computer Application and Software, 27 (2010) 108-110.

[5] Lang Zhou, Liang Zhang, Chong Feng, et al, Term Extraction Method Based on Word Frequency Distribution Change Statistic, J. Computer Science, 36 (2009) 177-180.

Published by Atlantis Press, Paris, France.