13th Global Congress on Manufacturing and Management, GCMM 2016

Research on search engine of knowledge adaptation system based on large scale data set

Zhou Zhiwei[a], Guo Xiaomin[a], Wei Lin[a], Zhang Mengmeng[a], Lian Zhikang[a],

Zhao Yamin[a], Zhang Jianhua[a]*

[a]Zhengzhou University, No.100 Science Avenue, Zhengzhou 450001, China

Abstract

The world has entered the era of knowledge economy, knowledge is becoming the core of social production. Any organization or individual can only continue to strengthen the application of knowledge, in today's changeable, fast changing, change in the competition to stand out. Provides a rich source of data explosion of information and knowledge for knowledge application, knowledge management research direction of past and existing knowledge, needs the shortcomings, which introduced the large-scale data adapter based on knowledge. Large data sets in the memory not only need to consume a large amount of storage space, but also put forward higher requirements for data retrieval, and the existing search engine is not very good to meet the needs of knowledge adaptation system. Based on large data sets, this paper mainly studies the retrieval system of knowledge adaptation system. The selection of the algorithm for the search engine, the analysis of the deficiencies, and put forward a targeted improvement program-kNN algorithm to increase the pre process, expand the scope of application and effect of the engine. Finally, the efficiency and effectiveness of the model are verified by the example of forest cover.

1. Introduction

In recent years the global acceleration into the era of Knowledge Economy[1,2], the most valuable production

---

* Corresponding author. Tel.: +1-583-811-6237.
  E-mail address: 957131915@qq.com

resources in social production activities has been turned into Knowledge. The rapid development of information technology has brought the explosion of data volumes, but due to the amount of data on a huge scale, large data sets not only need to consume large amounts of storage space in the storage, at the same time, the data matching and retrieval are there will be a bottleneck.

In view of this, we proposed the knowledge search engine adaptation system based on large-scale data sets. Among them, the author of searching engine to do some modification, such as using the rough sets and the variation coefficient method of combining the distance formula of kNN algorithm used by the condition attribute weights allocation, make it more relevant applications. Use k - means to do division algorithm to accelerate the efficiency of search engine, etc. Therefore, this article research results have certain practical value, improvements to the enterprises in knowledge management, knowledge applications link has a certain reference value.

2.Search engine

Search engine is the foundation of the whole knowledge adaptation system stage, efficient and accurate retrieval out with unanswered questions matching degrees higher case set has important influence on the quality of the fit for follow-up, now commonly used retrieval algorithm mainly are: nearest neighbor, inductive reasoning and knowledge to guide strategy three categories.

Table 1. Comparation of common case retrieval algorithm.

	kNN	Inductive reasoning strategy	Knowledge guidance strategy
Prior training	No	Yes	Yes
Training time-consuming	No	Longer	Medium
Retrieval efficiency	Medium	Faster	General
Algorithm complexity	Easy	More complex	Complex
Interpretability	Poor	Strong	General
Implementation difficulty	Easy	General	Difficult

Although some system design of case retrieval and storage system when they do different optimization, but the overall strategy still belongs to the category of the above three methods. According to the comparison table 1 case retrieval algorithm and the design features of massive knowledge adaptation framework, this paper, the author selected kNN algorithm for case retrieval.

3.Adapter case sets determined based on large data sets

3.1.Case similar level characterization and method

In order to characterize the target case C_q and case similarity between case, usually the condition attribute values ofthe two after the similarity calculation. Commonly used the similarity calculation of Euclidean distance, min had distance, markov distance and vector inner product[3-5], but generally adopts Euclidean distance to do calculation. For under test case C_q (no solution to the target case temporarily), its conditions set $a_{q\_cj}(j=1,2...,m^{\varphi})$ Which the number m^{φ} for cases under the condition of limited conditions. By comparing the retrieval algorithm source in case similarity $\text{sim}(C_i, C_q)$ between the case C_i and case under test C_q. If $\exists \text{sim}(g) \in \text{sim}(C_i, C_q)$ and $\text{sim}(g) = 1$ set up in the case-based reasoning, to retrieve cases in the case bank indicates that the source C_i and target case C_q are identical in all conditions, at this time may consider direct reuse in solving unanswered question C_q, namely "empty adaptation".