Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 56 (2016) 476 – 480

www.elsevier.com/locate/procedia

9th International Conference on Digital Enterprise Technology - DET 2016 – "Intelligent Manufacturing in the Knowledge Economy Era

A Chinese Topic Crawler Focused on Customer Development

Tong Wu[a], Yanchun Liang[a,b], Chunguo Wu[a], Shifeng Piao[a], Deyin Ma[c], Guozhong Zhao[d], Xiaosong Han[a,d,*]

[a] College of Computer Science and Technology, Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, Jilin University, Changchun 130012, China

[b] Zhuhai Laboratory of Key Laboratory for Symbol Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai 519041, China

[c] School of Business, Jilin University, Changchun 130012, China

[d] Daqing Oilfield Personnel Development Institute, CNPC, Daqing 163000, China

* Corresponding author. Tel.: +86 137 5669 8685. *E-mail address:* hanxiaosong@jlu.edu.cn

Abstract

This paper presents a Chinese topic crawler focused on customer development, in order to meet the needs of users for more accurate and particular Internet information. The concept of meta-search engine is introduced, and the keywords are expanded by the ontology of HowNet. Through the web crawler, preprocessing and classification, the information on customer relations can be divided into three categories: company, platform and meaningless. Numerical experiments show that satisfactory results can be obtained in some particular information-seeking areas. The average accuracy for classification is more than 80%, which can meet customer needs in most cases.

Peer-review under responsibility of the scientific committee of the 5th CIRP Global Web Conference Research and Innovation for Future Production

Keywords: Topic crawler; Customer development; Meta-search engine; Semantic similarity

1. Introduction

Currently, the general search engine based on keywords search has been widely applied to daily life, and people's life and work are inseparable from its support. However, with the development of science and technology and the improvement of quality of life, more and more people pursue higher accurate and professional personalized search engine. Hence the topic web crawler emerged, which was adapted to the specified topic and the individual needs, and became an important research direction of the search engine and web information mining [1]. In this paper, we proposed a Chinese topic crawler in view of the customer relationship management, which was used to search and develop customer relations of the specific fields.

The current customer relationship software contains limited information, because their resources are obtained from either a long-term accumulation or online yellow pages, such as Alibaba.com and HC360.com. In this paper, the information collected from yellow pages is called platform information. When we search information through the search engines, there are usually lots of websites that has nothing to do with the company's customer information. This kind of information refers to meaningless information. Though more company information can be obtained through search engines, the difficulty is how to filter out the needed information, which is also the main content of this paper.

There are another four sections in this paper, in section "Literature review", we review some methods related to web crawler; in section "Topic crawler model", we describe the proposed topic crawler and show each step; in section "Experimental results and analysis", we demonstrate the performance of our method. At last, section "Conclusion" gives the summary.

Peer-review under responsibility of the scientific committee of the 5th CIRP Global Web Conference Research and Innovation for Future Production
doi:10.1016/j.procir.2016.10.093

2. Literature review

The application and development of the topic models, aresearch field of natural language processing, is the basis to build the web crawler system [2].Many researchers put forward improved strategies based on keywords matching. Ouyang[3] proposed ontology-based theme portrayed crawling strategy, through the establishment of mapping mechanism to map the ontology semantic to the keywords list, and the strategy combined with the ontology reasoning programs to complete the sorting and fetching of web links. Peng [4] proposed a self-adaptive topic crawler method which was based on link-contexts. Li [5] put forward an improved method which combined the topic relevance predict algorithm and SVM classification algorithm with HITS algorithm, to solve the problem of information retrieval in the special areas. Huai [6] put forward a named entity linking frame based on the probabilistic topic model, to solve the text ambiguity problem, and the proposed method provided a new way of keywords filtering. Deng [7] changed the problem from topic-specific information discovery to website topic classification, and proposed a strategy which crawled webpages as less as possible though using external links to discover websites. In view of the multi-topic crawling, Zhong [8] put forward the idea of dividing topic rules into atomic structure, and designed different allocation strategies for built-in and general search engines. Xu [9] obtained the semantic boundary and the concept collection of the topic areas based on the Wikipedia category system, where the authors exploited the search engine for collecting the topic-related web information source, and finally built a specific-topic concept network.

Most of the existing web crawler are aimed at English web pages, the Chinese customers also need a kind of web crawler to search information, so in this paper, we propose a Chinese topic crawler to meet the need of customers.

3. Topic crawler model

The proposed topic crawler regarded HowNet [10] as a primary file of the Chinese ontology library. We expanded the keywords by string screening and synonyms searching on HowNet, so as to form a local ontology library of the specific field. Then on the meta-search engine, the field library was used to crawl customer information with related keywords.

The overall processing was summarized in Figure 1. The four main steps are: 1) keyword expansion; 2) information crawling;3) text preprocessing: duplicate data were filtered out after crawling and IKAnalyzer [11] was employed for word segmentation processing;4) classification: the Naive Bayesian algorithm was used as classifier. The operating procedures and the output of each step will be introduced in detail in the following sections.

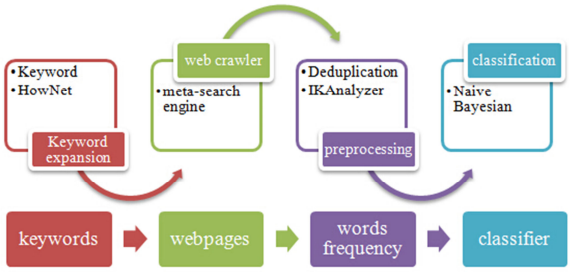

Fig. 1. Algorithm flowchart.

3.1. Keywords expansion

This paper is aimed at company information in various industries, which needs the specific area knowledge ontology. HowNet [10] is an online commonsense knowledge base which is used to reveal the relationship of intrinsic concept and the intrinsic properties relationship of concept. It contains 66182 most used Chinese words in daily life, which can support the specific area knowledge base to expand keywords. Figure 2 shows the flow chart of the keyword expansion process.

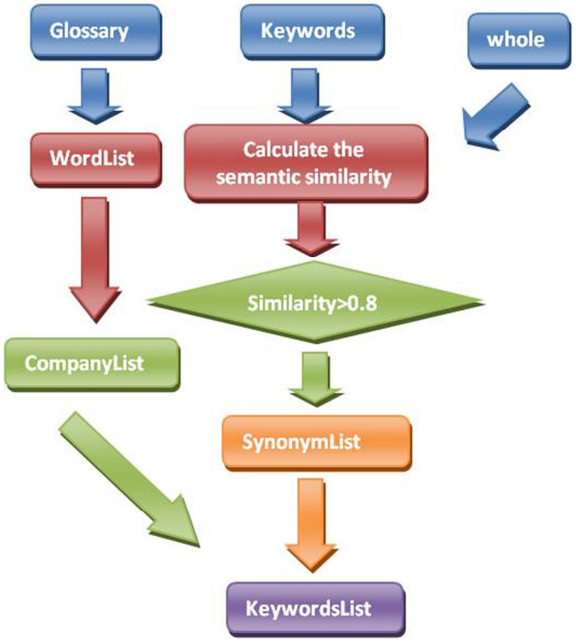

Fig. 2. The establishment of ontology knowledge base (Keywords expansion).

The establishment of ontology knowledge base is divided into two steps: firstly, in the HowNet glossary file, all the Chinese words contains all of the term “company” will be collected, and save the results to the “companyList” file. Secondly, calculate the semantic similarity between the keywords entered by users and each Chinese word in the “wordList” file. Then, select the Chinese words which similarity value exceeds 0.8 and save the Chinese words to