A Commodity Information Search Model of E-Commerce Search Engine Based on Semantic Similarity and Multi-Attribute Decision Method

Ziming Zeng
Center for Studies of Information Resources, Wuhan University
Wuhan 430072, China
E-mail: zmzeng1977@163.com

This article is supported by "Research on Information Aggregation and Visualization Modeling of New E-commerce Search Engine", MOE Project of Humanities and Social Science in Chinese Universities(No:08JC870011), and also supported by China Postdoctoral Science Foundation(No:20090460988), and by supported by the MOE Project of Key Research Institute of Humanities and Social Science in Chinese Universities (No: 07JJD870220).

Abstract

The paper presented an intelligent commodity information search model, which integrates semantic retrieval and multi-attribute decision method. First, semantic similarity is computed by constructing semantic vector-space, in order to realize the semantic consistency between retrieved result and customer's query. Besides, TOPSIS method is also utilized to construct the comparison mechanism of commodity by calculating the utility value of each retrieved commodity. Finally, the experiment is conducted in terms of accuracy and customer acceptance rate, and the results verify the effectiveness of the model and it can improve the precision of the commodity information search.

Keywords: Semantic vector, Multi-attribute decision, Commodity information search

1. Introduction

Nowadays, the advance of Internet and Web technologies has continuously boosted the prosperity of e-commerce. Through the Internet, it has become daily life for people to online shopping, and the number of people buying, selling and performing transactions on the Web is increasing at a phenomenal pace. With the further development of e-commerce, it will not be easy for customers to single out the best commodity when faced with the massive commodity information in the Internet. Usually, customers utilize various E-commerce search engines to search and compare commodities when they do online shopping in the Internet. Therefore, E-commerce search engines have largely become the main methods for customer to acquire commodity information and relevant services in the course of e-commerce activities. However, common search engines (such as Google, baidu, etc) and keyword-based search are not only low-efficient, but also sometimes the retrieved document contents of web pages are non-relevant with customer's query. The main reasons that result in these problems are: 1) the traditional information search techniques cannot express the semantic information correctly, and the information search based on keyword-matching still causes the semantic inaccuracy of retrieved results. 2) The heterogeneous characteristic of information organization is very obvious because of the diversity of e-commerce platform and the standard deficiency of relevant domain information description. 3) There are still not effective commodity evaluation and comparison mechanism so as to cause the information overload of the retrieval results.

In order to solve the existent problems that traditional information search methods have, many scholars both at home and abroad propose many new web search approaches, which are based on ontology and semantic web. Popov defines a general framework for document search that is supported by ontology, and integrate full-text search with ontology-based methods (Popov, 2004). Pablo presents a semantic search model based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm (Pablo, 2007). TAP utilizes semi-automatic techniques to extract relevant knowledge from free texts and semi-structured data (Guha, 2003). In TAP, the system transfers the document into the semantic web document format after analyzing the semantics of the document, and consequently utilizes the structured knowledge to improve the precision ratio of information. Wang also presents a semantic-similarity based information search method (Wang 2006), which utilizes ontology to describe semantics of the customer queries and Web documents, and computes the semantic similarity between the concept and property of domain knowledge to realize the semantic information search. Zhang presents a semantic annotation method based on ontology, which can be used in intelligent search systems for semantic reasoning (Zhang 2008).

All of these methods can solve the semantic inaccuracy of information to some extent, and can realize the semantic accordance between the customer's query and document information. However, in order to solve the information overload of commodities and provide accurate information search and shopping service for customers, one kind of effective commodity evaluation and comparison mechanism should be constructed, which is based on the realization of semantic retrieval of general information. In order to realize it, the paper presents an intelligent commodity information search model of E-commerce search engine, which integrates semantic retrieval and multi-attribute decision method. First, semantic similarity can be computed by constructing query semantic vector and document semantic vector respectively based on ontology, in order to realize the semantic consistency between retrieved results and customer's query. Besides, TOPSIS method is also utilized to construct the comparison mechanism of commodity by calculating the utility value of each retrieved commodity, and choose the most suitable one for customers. Therefore, the intelligent model not only considers the semantic search problems, but also utilizes the commodity evaluation and comparison mechanism, in order to improve the precision ration of commodity information and provide intelligent shopping services for customers.

The paper is organized as follows. In section 2, the intelligent commodity search model is described. In section 3, the experiment is described and the results are analyzed. Finally, the conclusions are discussed in section 4.

2. The intelligent Search Model of the System

2.1 System Architecture

The main task of the commodity information search is: (1) semantic matching between retrieved commodity information and customer's query is realized; (2) the comparison and evaluation mechanism of retrieved results is provided to improve the accuracy of information. The system model is illustrated in Figure 1.

The whole information search process is: (1) the main work of query interface is to realize the bidirectional communication between customers and search system. On one hand, a customer can input query conditions via query interface, including query keywords, relevant properties and values; On the other hand, in order to collect and analyze the customer's current needs, query interface ask the customer to express his/her qualitative needs about commodities. Taking a notebook for example, the system ask the customer to express qualitative features about multi-media, graphics display, network communication and interface supporting of a notebook, and give relevant weight values; (2) the query keywords are analyzed and extended semantically based on ontology and the semantic query vector is built; (3) the retrieved web pages are extracted and the semantic document vectors are built by annotating documents semantically based on ontology; (4) the semantic similarity of semantic query vector and document vectors is computed, in order to realize the semantic-matching between documents and query keywords; (5) After extracting the characteristic information of annotated commodity documents, each commodity is evaluated by calculating the utility value based on multi-attribute decision method; (6) the rank of each commodity based on semantic similarity and commodity utility value is calculated, and finally the retrieved results of commodities are ranked.

2.2 Domain Ontology

Ontology was originally a philosophy concept to study the essence of the existence and compositions of objectives (Gruber, 1993). In the artificial intelligence, ontology is a formal, shareable, and explicit specification of a shared conceptualization. Domain ontology is a specialized ontology which is used to describe special domain knowledge. Its goal is to capture the relevant knowledge in the field, provide common understanding of the domain knowledge, identify common recognition of the concepts, and give an accurate definition for these concepts and relationships that exist between them. In the domain of commodity information retrieval of e-commerce search engine, such structure can establish a good level of knowledge. In order to facilitate the search and comparison of commodity information provided by different e-commerce websites, a special ontology of the particular domain must be established for the construction of various commodity classifications and the establishment of commodity information model. In terms of simple commercial websites, domain ontology can be established manually or acquired semi-automatically from the contents of websites. However, in terms of large-scaled commercial websites, the manual construction of ontology is relatively complex, and the domain ontology can be reused and modified based on the existing ontology, in order to adapt to the particular commodity domain. In the paper, a notebook ontology as the basic of the commodity information search is developed, using the Protégé editor. The simplified structure of the notebook ontology is illustrated in Figure 2.

In the Figure 2, the construction process of commodity domain ontology can be described as follows: (1) the domain and scope in the commodity ontology is specified. (2) the concept system of the domain is constructed, including the definition of classes, properties and instances in the domain ontology. (3) the process of the ontological concepts is specialized. (4) the commodity ontology using the Protégé editor, is established, and transferred into the OWL-based document format. (5) the commodity ontology is evaluated and maintained.

2.3 Construction of the Semantic Query Vector

The technical course of the construction can be described as followed: in information search of the particular domain such as notebook commodity, the advantage of the domain ontology in the knowledge representation can be fully utilized to analyze and extend the query keywords of customers semantically, the aim of which is to help search engine to acquire the query intention of customers accurately. Based on it, the semantically extended query conditions are formed, and the semantic query vector can be constructed correspondingly.

In the paper, the semantically extended method centers around the initial query keywords, and makes the initial query keywords to have synonyms extension, upper extension and lower extension respectively based on the domain ontology. In the method, the synonyms extension is to acquire all the synonyms relevant to initial keywords as the new extended keywords. The upper extension is to acquire upper keywords adjacent to initial keywords as the new extended keywords. Similarly, the lower extension is to acquire lower keywords adjacent to initial keywords (including concept properties and instance values) as the new extended keywords. Therefore, when processing a customer's query, the system can describe and extend the initial keywords semantically based on domain ontology, and acquire the customer's query intention more accurately. For example, when a customer purchases a notebook, he can input "HP commercial use display 17 inches IEEE802.11" as query keywords. When extracting the query keywords, the system can utilize domain ontology to decide that the customer's query intention is to retrieve commodity information relevant to the ontological concept "notebook". By semantically upper extending the keywords "HP" and "commercial use" respectively, the system can deduce that notebook brand that the customer needs is HP, and the product orientation is for commercial use. Similarly, by semantically lower extending the keyword "display" and upper extending the keyword "17 inches", the system can also deduce that display size that the customer needs is 17 inches. Moreover, by semantically upper extending the keyword "IEEE802.11", the system can deduce that wireless network card that the customer needs complies with the IEEE802.11 standard, and the performance feature should support wireless network communication.

After the semantic extension process of query keywords is accomplished, the extended customer's query condition can be represented as a semantic vector $Q = \{C_1, C_2, ..., C_i, ..., C_t\}$ (C_i denotes the ith query keyword). The corresponding weight $W_i (1 \le i \le t)$ can construct the semantic query vector, which is defined as $WQ = \{W_1, W_2, ..., W_i, ..., W_t\}$. Each weight of the vector represents the semantic importance of relevant keyword in the query. In the construction process of semantic query vector, the system can provide human-machine interfaces, which will help customers to further determine the accurate query intention in order to update the semantic query vector in real time.

2.4 Semantic Annotations and Vector Construction of Documents

The search engine utilizes web crawler to choose some typical B2C websites (dangdang.com, amazon.cn, etc.). At present, a Wrapper is used to extract commodity attribute information from the commodity pages, and transfer into OWL formatted documents by tanking advantage of the commodity ontology.

In the research work, the Vector Space Model (VSM) is used, which treats a document as a collection of keywords and considers about the frequency information. In the process of semantic document annotation, the system treats the customer's query keywords as index terms in documents, and assigns weights to index terms in documents, which is based on the frequency of occurrence and location of index terms. Therefore, the semantic vector of a document can be defined as $D_k = \{C_1, C_2, ..., C_i, ..., C_t\}$, where C_i $(1 \le i \le t)$ is an index term of the document D_k. In the VSM of the documents, the index terms is often assigned a responding weight W_{ik}, which can be calculated by most frequently used $tf - idf$ scheme (Dumais, 1991):

$$W_{ik} = tf_{ik} \times idf_i = \frac{freq_{ik}}{\max_i freq_{ik}} \times \log(\frac{m}{n_i}) \tag{1}$$

Where $freq_{ik}$ is the number of occurrences of term C_i in document D_k; $\max_i freq_{ik}$ is the maximum number of occurrences of all the terms in document D_k; m is the number of documents in the system, and n_i is the document frequency for term C_i in the document set D. According to the formula (1), the weights of each document in the document set D can be calculated. Therefore, in the vector space of documents, the semantic vector of document D_k can be further denoted as $D_k = \{W_{1k}, W_{2k,}, ..., W_{ik}, ..., W_{tk}\}$.

2.5 The Intelligent Information Search Algorithm

(1) Computation of semantic similarity

After the semantic query vector and document vector are constructed respectively, the search engine utilizes cosine correlation method to measure the semantic similarity between the customer's query and commodity documents, so as to evaluate the semantic matching between the customer's needs and retrieved results. The semantic similarity can be qualified by the cosine of the angle between these two vectors, that is,

$$sim(D_k, WQ) = \frac{D_k \bullet WQ}{|D_k| \times |WQ|} = \frac{\sum_{i=1}^{t} W_{ik} \times W_i}{\sqrt{\sum_{i=1}^{t} W_{ik}^2} \times \sqrt{\sum_{i=1}^{t} W_i}} \tag{2}$$

In the formula (2), D_k is the semantic vector of document, and WQ is the semantic vector of extended keywords. Obviously, the similarity value is limited in $[0,1]$.

(2) Computation of commodity utility value

Based on the semantic search, the commodity evaluation and comparison mechanism should be established, in order to single out suitable commodity. The system utilizes a multi-attribute decision making approach, which is derived from TOPSIS (technique for order preference by similarity to ideal solution), to calculate the utility value of each commodity for the customers (Balabonovic,1997). Its mathematical model can be described: $P = \{p_1, p_2, ..., p_m\}$ as the vector of the commodity information that has been searched on Internet, $Y = \{y_1, y_2, ..., y_n\}$ as the qualitative feature vector of the commodities, and the utility value of the commodity $p_i (1 \le i \le m)$ about the attribute $y_j (1 \le j \le n)$ can be denoted as $f_{ij} = f_j(p_{ij})$, which represents the relative performance of the commodity p_i in the qualitative feature i. Therefore, the decision matrix that consists of $m \times n$ f_{ij} can be denoted as:

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix} = (f_{ij})_{m \times n} \tag{3}$$

In order to facilitate mutual reference between the multi-attributes easily, the decision matrix should be normalized, which can be followed by the formula (4):

$$f_{ij}^{'} = \frac{f_{ij}}{\sqrt{\sum_{i=1}^{m} (f_{ij})^2}} \tag{4}$$

After normalizing the decision matrix, the value of $f_{ij}^{'}$ is limited [0,1]. In this way, the utility of the commodity $p_i (1 \le i \le m)$ can be measured by:

$$U_i = \frac{S_i^-}{S_i^* + S_i^-} \tag{5}$$

Where

$$S_i^* = \sqrt{\sum_{j=1}^{n} [\omega_j (f_{ij}^{'} - f_{j\_best}^{'})]^2} \tag{6}$$

$$S_i^- = \sqrt{\sum_{j=1}^{n} [\omega_j (f_{ij}^{'} - f_{j\_worst}^{'})]^2} \tag{7}$$

In the above equations, n is the number of commodity qualitative features; $f_j^{'}$ is the normalized performance value of a commodity in the feature dimension j; $f_{j\_best}^{'}$ and $f_{j\_worst}^{'}$ are the best and worst performance value(normalized(in the same dimension, respectively; and ω_j means the customer's relative need in this feature. Based on the above measurement, the search engine will calculate the utility of all the searched commodities and therefore establish the comparison and evaluation mechanism of commodity information.

(3) Integrated commodity search algorithm

Integrating semantic similarity and commodity utility value, the composite score $\Pr oRank_i$ of commodity $p_i (1 \le i \le m)$ can be computed by the formula (8):

$$\Pr oRank_i = \lambda \cdot sim(D_i, WQ) + (1-\lambda)U_i \qquad (0 \le \lambda \le 1) \tag{8}$$

According to the formula (8), when $\lambda = 1$, the system will only provide the ability to have semantic information search, but can't provide the comparison mechanism of searched commodity information. On the contrary, when $\lambda = 0$, the system will only provide the comparison mechanism of commodity information, but not considering the semantic matching between searched commodity information and customer's query. So the algorithm of integrated intelligent information search can be described as followers:

Algorithm: Product_Rank()

Input: Customer's query keywords; qualitative feature needs and relative weight of commodities

Output: URL of $top-k$ commodities

```
  Init(Q); Load(Q);    // semantic extension of query keywords, initiate and load customer's semantic query vector
  Begin
  for each $D_i$ do
      if log($D_i$) = 0 then  //  document vector $D_i$ of the commodity $P_i$ is not processed
      {
Load($D_i$);     //  have semantic processing of $D_i$, and load it
              GetSimarity(Q, $D_0$);    // compute the semantic similarity between $D_i$ and
                                         customer's query keywords $Q$
              GetProduct_Utility($P_i$);  //compute the utility value of the commodity $P_i$
                                           based on TOPSIS method
              GetProduct_Rank($P_i$);     // compute composite score of the commodity $P_i$
              InsertTo(Result);   //  insert searched results into set Result in descending order
          }
Output(Result, $top-k$);    // output $top-k$ commodities URL as the searched results to
                               target customer
End
```

3. The experiment

We have developed a prototype system of e-commerce search engine based on Java platform and Jakarta Lucene library. To construct the annotated data repository, the system utilizes web crawler to search focused commodity information related to notebook from predefined B2C E-commerce websites (dangdang.com, amazon.cn, etc.) and downloads searched information to the local database. About 3000 HTML Web pages of notebook commodity are downloaded from relevant websites as experimental corpus, and the Wrapper is utilized to extract commodity information from retrieved Web pages, then the semantic annotation of documents can be realized based on domain ontology.

The primary factor which is used to evaluate the performance of the search engine is the precision of commodity information search. Similar to the precision computation of general information retrieval, the precision of commodity information search can be computed as the formula (9):

$$\text{Precision} = \frac{\sum_{Accurate} Document}{\sum_{Searched} Document} \tag{9}$$

In the formula (9), $\sum_{Searched} Document$ is the number of al the searched commodity documents, and

$\sum_{Accurate} Document$ is the number of searched commodity documents that satisfy the customer's query needs (including customer's query semantic and the qualitative feature preferences of commodity). According the formula (8), a threshold can be set to compute the precision of the system, which figures out the total number of commodity documents that satisfy the inequality $\Pr oRank_i \geq Threshold$. The higher the precision is, the better the search performance of the system is. Obviously, the parameter λ will influence the precision of the system and the sorting order of searched commodity information. In order to select the suitable value of λ , the relevant simulating experiment has been done. Based on the initial analysis of the experiment, the relation between the parameter λ and the precision of the system is illustrated in Figure 3. Seen from the Figure, the precision of the system is the highest when $\lambda = 0.4$. Therefore, it can be inferred that the utility value computation based on TOPSIS method can give more influence on the commodity search performance than semantic similarity computation.

Besides, different from general information search, the system of commodity information search first integrates semantic similarity computing and utility value computing, and recommends $top - k$ commodities information to customers. Customers can choose several suitable commodities and add them to shopping cards for purchasing. In order to verify the effectiveness of search algorithm presented in the paper, a new assessment criterion of commodity information search can be defined in the paper, which is called Customer Acceptance Ration (CAR). It can be described as the following formula:

$$CAR = \frac{\sum_{ShoppingCa\ rt} \{P_1, P_2, ... P_M\} \bigcap \{Top\ -\ K \text{ Product}\}}{\{Top\ -\ K \text{ Pr } oduct\}} \tag{10}$$

In the formula (10), $\{Top - K \Pr oduct\}$ is the set of $top - k$ commodities that the search algorithm outputs, and $\sum_{ShoppingCart} \{P_1, P_2, ... P_M\}$ is the set of commodities that the customers add to shopping card. Therefore, the elements in the intersection of two sets would be the commodities that customers choose and add to shopping card from $top - k$ recommended commodities. CAR reflects the degree to which customers accept recommended commodities generated from the search algorithm. The higher the value of CAR is, the more the search algorithm satisfies customers' practical shopping need in the environment of e-commerce.

In the experiment, in terms of the CAR, we can compare the search algorithm (parameter $\lambda = 0.4$) with the traditional information retrieval algorithm based on semantic vector model (parameter $\lambda = 1$). The parameter k in $top - k$ can be set $10 \sim 100$ different values respectively, and the systems samples 10 customer's query at random, then calculates the mean value of CAR. The mean value of CAR is illustrated in Figure 4.

Seen from the Figure 4, in different values of k , CAR of the commodity search algorithm is all higher than CAR of the traditional semantic information retrieval. Especially when the value of k is relative low, the difference of two algorithms is much more obvious. This is because that the commodity information search algorithm also presents comparison and evaluation mechanism of the commodity contents and quality, not only considering semantic matching between searched document and customers' query keywords. The retrieval results of the commodity information search algorithm are more accurate, and more satisfy customers' shopping needs. Therefore, the algorithm presented in the paper is more effective than the traditional information retrieval algorithm, and can be used as the search algorithm of e-commerce search engine, in the online shopping environment of e-commerce.

4. Conclusions

The paper presents an original commodity information search model, which integrates the semantic similarity computation based on semantic vector-space model and commodity utility value computation based on TOPSIS method. The search model not only realizes semantic retrieval of commodity information, but also has the commodity selection and comparison mechanism, so as to provide the intelligent information search services for customers' online shopping. Furthermore, relevant experiments have been done to verify the effectiveness of the commodity search algorithm in terms of assessment criteria, such as the precision and Customer Acceptance Rate (CAR). The experimental results show that the algorithm presented in the paper is superior to the traditional algorithm of semantic information retrieval and is more suitable for customers' shopping needs. At present, we develop a notebook oriented e-commerce search engine based on the integrated search algorithm, and the system has shown some potential for e-commerce applications. Based on present research work, we will improve the search model and make it to provide more accurate information search services for customers' shopping in the Internet.

References

B.Popov, A.Kiryakov, D.Ognyanoff, D.Manov and A.Kirilov. (2004). KIM – a semantic platform for information extraction and retrieval. *Natural Language Engineering*, 10(3-4): 375-392

Balabonovic M & Shoham Y. (1997). Fabl: content-based collaborative recommendation. *Communications of the ACM*, 40(3); 66-72

Gruber, TR. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2): 199-220

Pablo Castells & Miriam Fernandez, etc. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transaction on Knowledge and Data Engineering*, 19(2): 261-272

R.V.Guha, R.McCool and E.Miller. (2003). Semantic Search. Proceedings of the 12th International World Wide Web Conference (WWW03), 700-709

S, Dumais. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers*, 23(2): 229-236

Wang jin & Chen en-hong. (2006). Semantic Similarity-based Information Retrieval Method. *Pattern Recognition and Artificial Intelligence*, 19(6): 696-701

Zhang gong-jie & Huang sui. (2008). Research and Implementation of Semantic Indexing based on Ontology. *Computer and Engineering and Design*, 29(8): 2078-2080

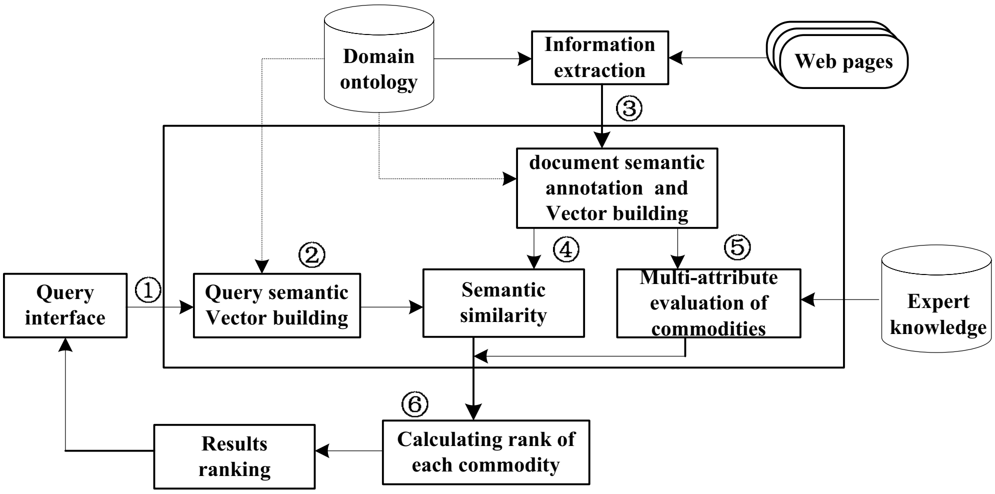
Domain ontology
Information extraction
Web pages
③
document semantic annotation and Vector building
②
④
⑤
Query interface
①
Query semantic Vector building
Semantic similarity
Multi-attribute evaluation of commodities
Expert knowledge
⑥
Results ranking
Calculating rank of each commodity

Figure 1. The intelligent search model of the system

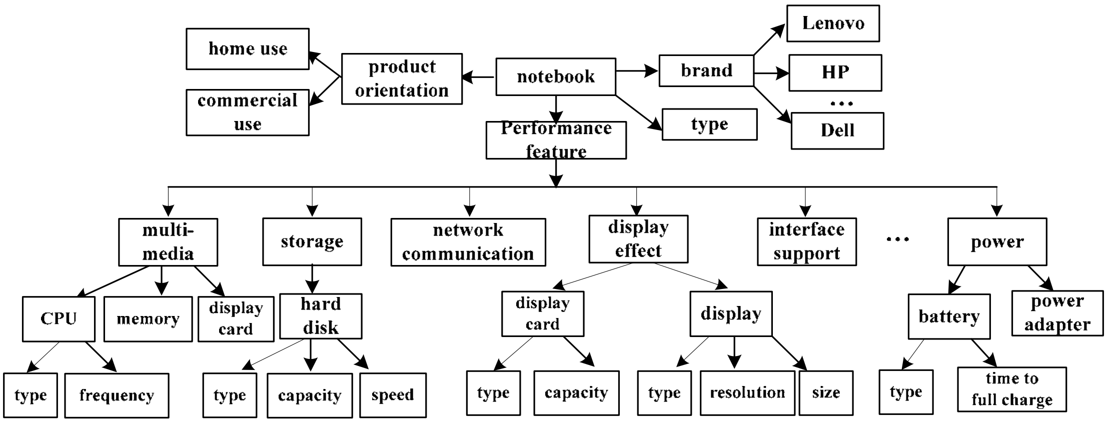

Figure 2. Part of domain ontology of notebook commodity

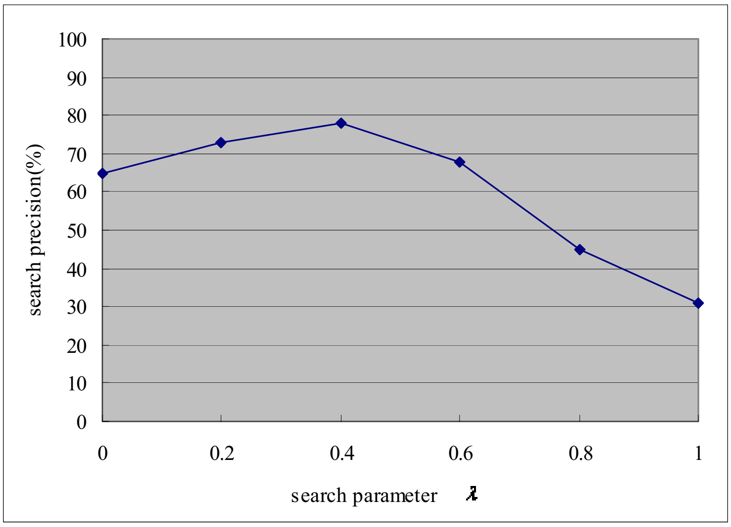

Figure 3. The relation between parameter λ and precision of the system

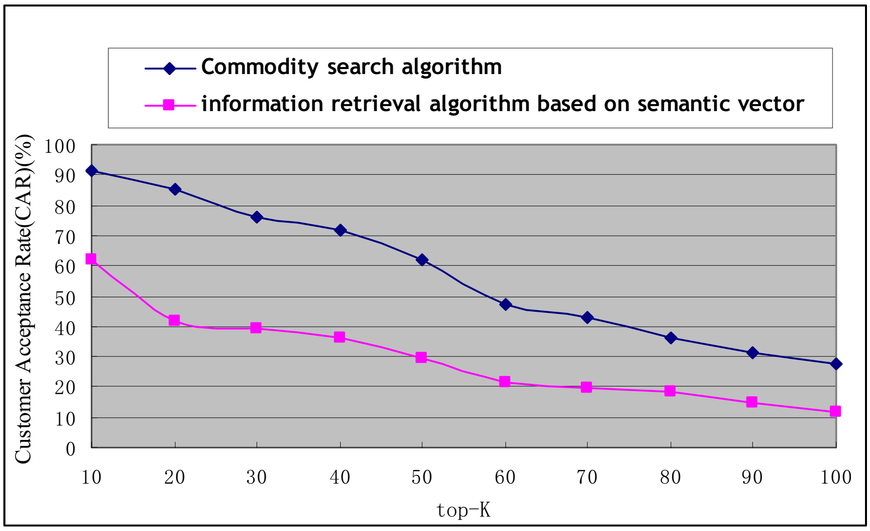

Figure 4. The comparison of two algorithms in terms of Customer Acceptance Rate (CAR)