Research on Computer Crime Evidence Retrieval Method based on Ontology

Xuezhi Chi [1, 2, a]

[1]Shandong Police College, Jinan 250100, China

[2]Shandong Normal University, Jinan 250014, China

[a]chixuezhi@126.com

Keywords: Ontology; Semantic expansion; Evidence retrieval; Computer Forensics.

Abstract. The traditional evidence retrieval method is not able to identify the word semantic in forensics user question, only make the mechanical matching, in order to improve the efficiency of evidence retrieval, get satisfactory retrieval results, this paper proposes an ontology-based evidence retrieval method, through the concept extraction from forensics data and query, concept interconnection graph construction, to realize the semantic expansion. The experimental results demonstrate that this method is superior to the traditional keyword-based retrieval method and able to perform evidence retrieval on a conceptual level.

Introduction

With the development of computer crime is on the increase, as an important means to combat computer crime, computer forensics is becoming more and more important, computer forensics includes data collection, protection, extraction, analysis and submitted to the court, in the forensics process, The evidence retrieval is one of the most important step, because of computer crime forensics data is often very large, data sources and data forms are not the same, how to carry out the retrieval work to assist computer forensics is the key work of computer crime forensics.

Traditional information computer crime evidence retrieval, most are based on keyword matching and inverted index table, only matching search keywords based on forensics user input. now, with the development of technology, the computer crime forensics add some natural language processing system to deal with more complex questions, but because of the defect model, The traditional method is not able to identify the word semantic in forensics user question, only make the mechanical matching from the larger amount of forensics data, due to the large amount of computation and returned set , so for forensics users not only have no actual meaning, but also because of the lack of certain relevance, so that the result of traditional evidence retrieval is not satisfactory.

The traditional evidence retrieval has the following deficiencies:

(1)Precision and recall is not guaranteed. Because the retrieval question is too free and simple, causing the search returns information too much, precision and recall are not guaranteed, the reason: No constraints on the meaning of the retrieval word, the computer does not recognize the semantic. The form of query does not regulate, the computer cannot understand the real retrieval intention. Simple characters matching cannot provide retrieval based on concept.

(2) The retrieval results lack of unified output form, cannot be shared and reused. The current evidence retrieval display content is some information list, rather than the evidence with semantic to answer user expectations, need manual further selection, analysis, induction and consolidation in order to get the final conclusion, so bad availability.

(3)Cannot search the hidden information in the forensics data. The so-called hidden information refers to those who have not been directly stated, but implied some important information in the forensics data.

To solve those problems of above, this paper regards domain ontology as the foundation of semantic understanding, propose evidence retrieval model based on ontology, perform evidence retrieval on massive forensics data, so that obtain the ideal evidence retrieval results with semantic features.

The evidence retrieval method based on Ontology

Traditional evidence retrieval model. Traditional evidence retrieval purposes, mainly based on the user's query that is keyword, find to meet forensics user requirements associated information from massive forensics data, search results that exclude irrelevant information, return evidence information that users are concerned about. Retrieval model is to determine whether the returned information is relevant to the query, and the related information is sorted. According to the different relevance determination method, the formation of different evidence retrieval model, the traditional evidence retrieval model are the following three types: Boolean model, vector space model and probabilistic model.

Concepts and definitions related to Ontology. Ontology has four layers of meaning:

(1) The conceptual model. By abstracting some phenomena in the objective world related to the concept and obtain the model, the representation meaning of independent of the specific environment condition.

(2) Explicit. The concepts used and constraints on the use of these concepts are clearly defined.

(3) Formal. Ontology is computer-readable.

(4) Share. Ontology is the embodiment of the recognition of common knowledge, reflect the concept set of recognized in related areas, it is addressed to groups rather than individuals, the goal of ontology is to capture the domain knowledge, provide for the common understanding of the domain knowledge.

Definition 1: A concept can be expressed as a six tuple: C= (W, U, S, P, H, WA)

W: Vocabulary Used to express the concept

U: Class (field) which this concept belongs to

S: Synonym for W

P: Attribute sets of the concept

H: The lower word set of the concept

WA: Element weights of the concept

Definition 2: Precision and recall

Precision and recall are the two most frequently used measures of retrieval performance evaluation.

Ontology-based evidence retrieval. Ontology has a good concept hierarchies and support for logical reasoning. When, for example, a user enters the term “virus” into a search, the search supplies results such as “virus system” because the query term contains “virus,” even though the user intends to retrieve results about virus “attack”. Many of the current evidence retrieval systems do not specify the semantic relationships of information; therefore, users are less able to find related evidence. When a user searches a system that possesses a large number of resources, this problem gets worse. The user must then review lengthy lists of irrelevant resources and make a decision regarding relative relevance. However, ontology builds semantic relationships among these terms; thus, users can review related evidence information and search for evidence based on ontology.

The basic idea of the ontology-based evidence retrieval is:

(1) With the help of domain experts, establish related domain ontology;

(2) collect the forensics data in the sources, and the data collected by the prescribed format are stored in the metadata database (RDB, KDB, etc.);

(3) For the query requests from user retrieval interfaces, do the semantic extension based on ontology for user forensics query, query converter convert the query request to the prescribed format?

(4) Retrieval results returned to the user through custom processing.

User queries can be extended to a semantic vector, every concept described by their own attributes, as shown in Fig. 1.

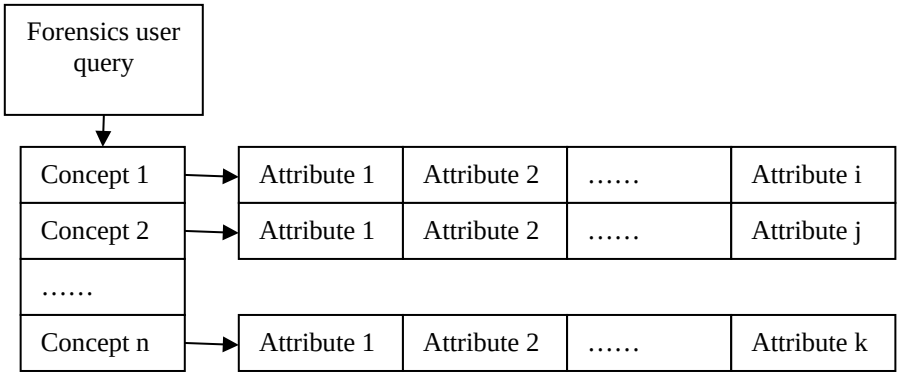

Fig. 1 Semantic vector of forensics user query

keywords that input in the user query, sometimes directly is the concept in the ontology, and sometimes is a concept description attribute or limit, so we need to separate these areas, and unity into semantic vector, as the expansion of user query. The user evidence query needs semantic expansion based on ontology.

Firstly, analysis the structure of the user query, the user query structure can be divided into three types through statistical : T mode, O mode and T+O mode, in mode T, each term is not in the ontology, we use the method of context oriented statistical to semantic extension; in mode O, query is composed of concept, relationship and instances and so on, so user's query requirements can be directly obtained according to the definition in the ontology ; T+O mode can be layered, first the user query requirements can be obtained according to the presence definition in the ontology, and then in the query results, using terms that does not exist in the ontology to query. To expansion the related content of concept in the ontology, this paper will build a connected graph of concepts, as the semantic reference of the concept mutual extension, the process is shown in Fig. 2.

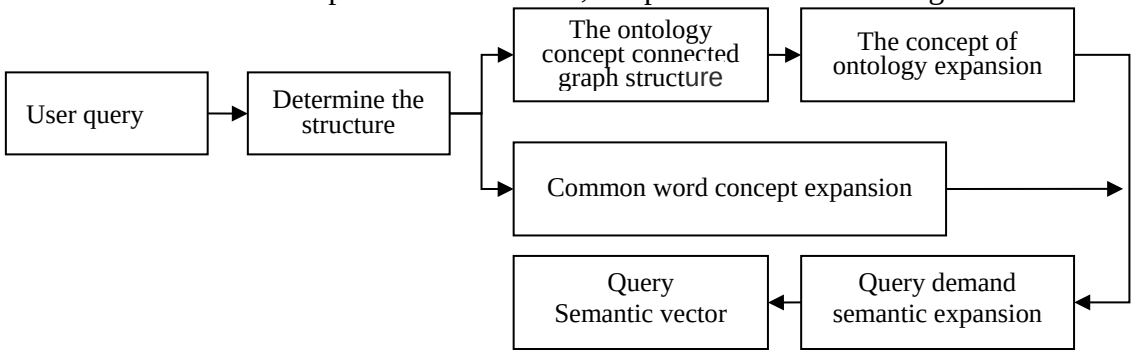

Fig. 2 Concept expansion process

The algorithm of ontology concept connected graph construction is as follows:

Input: training evidence data set

Output: edge weights of the ontology concept connected graph

Begin

(1) Initialization of ontology concept connected graph;

(2)Take the training forensic data set of corresponding ontology domain, statistical the concept number D_i in evidence data;

(3) For each D_i

A) for any two concepts of C_j and C_k appeared in D_i, take the low frequency number f of occurrences as the two concepts appear simultaneously in D_i;

B) If C_j and C_k is connected in the ontology concept map, accumulate the number f;

C) If C_j and C_k is not connected, then is connected with Cj and Ck, and assign f to it;

(4) take the maximum of the number on all edges as the denominator in the graph to get the edge weight wj, k that connected C_j and Ck;

End.

Thus, obtain the ontology concept connected graph with weights, in the time to expand the concept of the ontology, the method can be used for expansion.

The algorithm of ontology concept expansion is as follows:

Input: ontology concept connected graph, the concepts needed to expand
Output: The result of the expansion of the concept
Begin
(1) The loading of ontology concept connected graph;
(2)If there is the concept C that need to extend, then executing step 3, otherwise to 4;
(3) Arrange the adjacent nodes of concept C according to weight, take first K nodes as the expansion of the C, K is the preset number, turn to step 2;
(4) Pairs of K concepts and extract the attribute description of the first k nodes as content of the extension vector;
End.

Experimental results and analysis

This paper use USA Lincoln laboratory experimental data (a total of 9000 connection record) as background data of criminal behavior ontology model, producing model contains 7 kinds of crime classification. In order to evaluate the retrieval performance of queries, we will divide the experiment data into 10 groups, each group have 900 records. Each data set is used as test data sets, other data set used to generate the concept of aggregation. Below is the detailed experimental results. Recall and Precision of the two methods are shown in Table 1:

Table 1 Recall and Precision of the two retrieval methods

Similarity	Recall			Precision		
	keyword -based retrieval	ontology -based retrieval	Increase index	keyword -based retrieval	ontology -based retrieval	Increase index
0.9	2.3	5.7	3.4	82	94.2	12.2
0.8	4.6	9.1	4.5	85.1	89.3	4.2
0.7	10.2	17.8	7.6	89	88.3	-0.7
0.6	14.7	25.2	10.5	89.3	94.6	5.3
0.5	25.8	38.3	12.5	89.1	95.2	6.1
0.4	35.6	53.9	18.3	81.8	94.7	12.9
0.3	56.1	77.4	21.3	77.2	85	7.8
0.2	73.9	94	20.1	67.6	74.4	6.8
0.1	91.3	100	8.7	62	62.1	0.1
0	92.3	100	7.7	60	59.4	-0.6
average	40.7	52.1	11.5	78.3	83.7	5.4

Recall Curves of the two different retrieval methods are shown in Fig. 3:

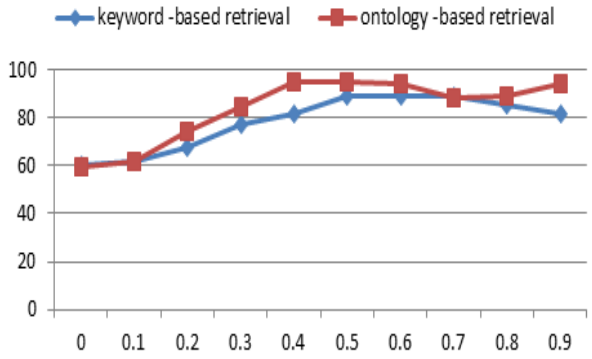

Fig. 3 Recall Curves of the two different retrieval methods

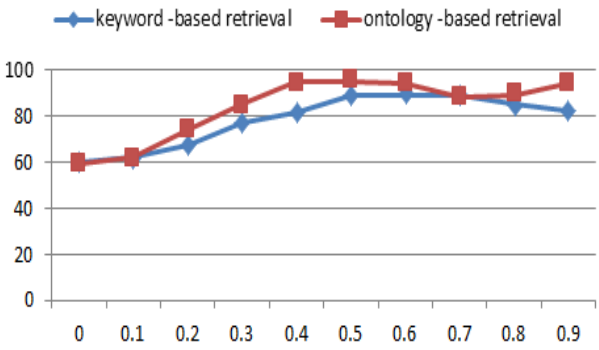

Fig. 4 Precision Curves of the two different retrieval methods

Precision Curves of the two different retrieval methods are shown in Fig. 4:

As we can see from the above experimental data, compared with the traditional keyword-based retrieval method, the recall ratio of the ontology-based evidence retrieval method increased by 11.5%,

in terms of precision, increased by 5.4%. Therefore, we can think that the ontology-based evidence retrieval method is superior to the traditional keyword-based retrieval method.

Conclusions

In this paper, we presented a new ontology-based evidence retrieval method for computer forensics. The ontology-based evidence retrieval method is able to incorporate semantic information in the information retrieval process, then perform evidence retrieval on massive forensics data. The experimental results demonstrate that this method is superior to the traditional keyword-based retrieval method and able to perform evidence retrieval on a conceptual level.

References

[1] Park, Heum ,Cho, SunHo, Kwon, Hyuk-Chul .“Cyber forensics ontology for cyber-criminal investigation, Lecture Notes of the Institute for Computer Sciences. p. 160-165, 2009.

[2] Saad, Sherif. “Method ontology for intelligent network forensics analysis”, PST 2010: 2010 8th International Conference on Privacy, Security and Trust, p. 7-14, 2010.

[3] Brinson Ashley, Robinson Rogers. “A cyber forensics ontology: Creating a new approach to studying cyber forensics”, Digital Investigation, p. 37-43, 2006.