Exploiting RFC2828 as a Domain Vocabulary for Identifying IT Security Literature

WANG Lidong  
Department of R & D  
National Computer network Emergency technical Team/Coordination Center of China  
Beijing, China  
GMwanglidong@gmail.com

QIAN Liping  
Department of Computer Science  
Beijing University of Civil Engineering and Architecture  
Beijing, China  
wangzs@vip.sina.com

Abstract—The volume of published scientific literature available on Internet has been increasing exponentially. Some of them reflect the latest achievement of the specific research domain. In recent years, many projects have been funded aiming to online scientific literature mining, especially in biomedical research. Scientific literature covers most of the hot topics in the research field and has a very large domain-specific vocabulary. The exploitation of domain knowledge and specialized vocabulary can dramatically improve the result of literature text processing. The purpose of this paper is to research on automatic identifying and classifying IT security literature so that IT security related papers can be retrieved from Internet with high accuracy. RFC 2828 provides explanations and recommendations for use of IT security terminology. In this paper, we evaluated the effects of IT security literatures identification with RFC2828 glossary-based feature choice and TF/IDF scheme. Our experimental result shows that its performance is better than the common TF/IDF method.

Keywords—document classification; IT security; vocabulary; RFC2828; literature

I. INTRODUCTION

The volume of published scientific literature available on Internet has been increasing exponentially. Some of them reflect the latest achievement of the specific research domain. These publications are published by their authors online in their homepages, or being organized and indexed by large-scale database system. Some free-access literature databases, such as biomedical domain papers database PubMedCentral and computer sci. & tec. full-text articles database Citeseer, deposit a lot of articles in their repository and serve as rich resources for corpus supporting for massive text processing.

Researches on information retrieval from online literatures can be mainly divided into two areas. The first area is paper structure-based meta-data retrieval. Using customized rules and description criterions (e.g. Dublin Core and ISO 23950 Bib1), meta-data of title, author affiliation and references can be get and then be used for deep thinking works such as author profiling, expert finding, affiliation network analyzing. The second area is paper content-based knowledge mining. By recognizing the meaningful entities and relations in the content, implied knowledge inner- or inter- papers can be discovered and then be used for works such as named-entity recognition, text classification, synonym finding, concept or relation extraction and hypothesis generation. In recent years, many projects have been funded aiming to biomedical literature mining. Ingrid Petric tried to identify potential contributions to a better understanding of autism focusing on articles from database PubMedCentral [1]. D. Shilin presented a method of mining physical protein-protein interactions by exploiting profile feature from full-text articles [2]. L. Yudong identified Protein-Organism-Location relations in the text of biomedical articles [3].

The purpose of this paper is to research on automatically identifying and classifying IT security literature retrieved from Internet. Different feature sets are selected for IT security literature identification under Vector Space Model (VSM). The three vocabularies used to build feature sets include a general terms vocabulary, a controlled key phrases for IT security and RFC2828 glossary. An experimental system was developed to compare their performance.

The rest of the paper is organized as follows. Section 2 is a brief review of text classification. Section 3 describes our methodology. Section 4 presents our experimental results and some discussion. Section 5 is the conclusions.

II. RELATED WORKS IN TEXT CLASSIFICATION

The first step in automatic text processing is to transform a textural document in a representation suitable for machine processing. The most commonly used method, know as VSM, represents a document as a vector in the term space (feature space). In VSM, if a set of features \( \{t_i\}_{i=1}^{n} \) and some kind of weighting method are selected, then a textual document \( d \) can be represented as \( (t_1, w_1; t_2, w_2; \ldots; t_n, w_n) \).

Feature selection plays a crucial role in machine learning methods based on feature vector. Features can be words, phrases, concepts or entity relations. The most common way for feature selection is dictionary-based methods, which use customized terminological resources to locate term occurrences in text. The advantage of this approach is simplicity while the performance is highly related with the words list and their weights. Statistics information or expert knowledge can be considered in feature selection. Statistics information includes term frequency, document frequency, entropy or mutual information. Two common kinds of lexicon are stop-words lists and domain-specific dictionary. All words listed in stop-words will be filtered in text preprocessing and words listed in domain dictionary will be...
used to construct features. In recent years, there are also some researches applying well-defined domain ontology into text classification. However, knowledge engineering approaches are extremely time-consuming and typically very specific - the adjustment to other domains is usually difficult.

Different features have different importance in a document and thus a weighting method is needed to show their contribution. The TF/IDF (Term Frequency/Inverse Document Frequency) weight is often used in text processing. Other simple weighting methods include 1/0 to indicate whether a term is appearing in the document and some TF/IDF weight variants (e.g. probabilistic TF/IDF). TF is related to the term count in a given document and IDF is a measure of the general importance of the term in whole corpus. Although very popular in text processing, the TF/IDF model has its drawbacks. Being inability to exploit implicit information of different classes, TF/IDF method shows poor accuracy for classifying documents that are not so much different to each other. Dimension reduction technique must be considered in TF/IDF to speed up text processing.

A variety of techniques have been used for solving the text classification problem. Among them, SVM delivers state-of-the-art performance in text classification and other real-world applications. It works efficiently with instances that implicitly belong to a high dimensional feature space and obtain comparatively high accuracy. Over-fitting can also be avoided in SVM formulation by requiring positive and negative training instances be maximally separated by the decision hyper-plane. Other methods such as Maximum Entropy Model, Hidden Markov Model and Bayesian Theory have also been studied and achieved noteworthy performance in text processing. In recent years, Kernel-based methods have become popular. Some kernel functions suitable for text processing include Vector Space Kernel, Bags-Of-Words Kernel, Latent Semantic Kernel. Although most recent text classification systems use machine learning, but when training examples are not available, handcrafted rules remain the preferred technique. Yeh once ran a text mining competition as part of the KDD Challenge Cup 2002. The task was a curation problem to evaluate papers from the FlyBase data set and determine whether the paper should be included. The best performing entry used a set of manually constructed rules based on POS (Part-Of-Speech) tagging, a lexicon, and semantic constraints determined by examining the training documents. Another well performing approach looked for manually chosen “keywords” and computed the distance between keywords and gene names [4].

Natural Language Processing (NLP) plays an essential role in text processing. NLP can process information on syntactic, semantic or pragmatic level. Syntactic deals with the structure of symbols, the related concept includes term frequency and co-occurrence. Semantic level deals with the meanings of symbols, a common architecture of semantic classification (e.g. ontology) is built to describe the properties of and relations between entities and events in document. Pragmatics has to do with context-dependent features of language. Currently, the combined exploitation of syntactic structures and semantic knowledge has effectively improved the performance of text processing tasks, while pragmatics is still a field under research and not widely applied to text classification due to its complicated knowledge representation and reasoning mechanism.

III. DOMAIN-SPECIFIC VOCABULARY-BASED IT SECURITY LITERATURE IDENTIFICATION

Compared with Web pages, scientific papers are always well organized, presented the ideas in a clear, logical way and the scientific and technical terminology is standardized. But it doesn’t lower the level of difficulty of text processing tasks since scientific literature is also lack of formal structure and presented with natural language. In addition, scientific literature covers most of the hot topics in the research field and has a large domain-specific vocabulary. The exploitation of domain knowledge and specialized vocabulary can dramatically improve the result of literature text processing [5]. Ramakrishnan leveraged the availability of a controlled vocabulary called MeSH and domain knowledge in the form of the UMLS and combined them with NLP techniques for relationship extraction. Their experiment showed that domain knowledge can be effectively combined with NLP techniques to achieve good effect. In this paper, we will evaluate whether IETF RFC 2828 (Request For Comment 2828) glossary can be used as a domain-specific lexicon for identifying IT security literature [6].

A. Selection of an IT Security Domain Vocabulary

Keywords listed in scientific articles can serve as a source of domain vocabulary. But English publications seldom particularize them. As an emergent branch of IT, security has a dynamically changing terminology without available semantic taxonomy and domain-specific vocabulary. Even there is a fundamental ambiguity in the use of word “IT security. The research and construction of IT security vocabulary usually focuses on taxonomy of vulnerabilities and Internet attacks [7, 8]. RFC 2828 provides an internally consistent, complementary set of abbreviations, definitions, and explanations for use of terminology related to IT security [9]. Although some non-security terms are also included to make the glossary self-contained, it is relatively complete.

A controlled vocabulary which we call “KeySecWords” is constructed in this paper besides RFC2828 glossary. These terms are selected from the CFP of some top security-related international meetings in recent years, the taxonomy of Internet attack research and several online network security dictionaries. (e.g. http://www.itsecurity.com/dictionary/). The total number of our KeySecWords is 140.

A subset of another English dictionary from SIL International Linguistics Department is used in our experiment. Compared with the concepts of RFC2828 glossary and KeySecWords, we call it GeneralTerms.

B. Methodology

Three feature sets are constructed with terms separately from GeneralTerms, RFC2828 Glossary and KeySecWords
and are applied to identifying IT security literature. The experimental corpus is retrieved from CiteSeer by implementing a meta-searching interface. Shallow NLP techniques were exploited with TF/IDF scheme to compare the classification result for the three feature sets.

1) Pre-processing: Pre-processing includes data clean, stop words removal and stemming. The retrieved PDF files are first converted to text. Some PDF files were created from images and generated code instead of text were during conversion. These files were deleted from corpus. Space characters of a few of successfully converted text files are missing during text conversion and word cutting were performed with SIL vocabulary. Stop words were removed using Glasgow stop-words vocabulary and stemming was performed using an improved version of Porter’s algorithm.

2) Documents representing: We use VSM to represent documents. Each document was separately represented by features with phrases from GeneralTerms, RFC2828 Glossary and KeySecWords. The weight of each feature was computed with TF/IDF scheme. GeneralTerms is constructed with the words generated after pre-processing and feature reduction.

3) Feature weighting: We use TF/IDF scheme for feature weighting. There are a lot of variants of TF/IDF weighting methods. Let $t_f$, $d$, $n_t$ be the frequency of term $t$ in document $d$, $n$ the total number of documents in the collection, $n_t$ the number of documents including term $t$, we choose the common TF/IDF method as in (1) to compute $W_{t,d}$ of feature $t$ in document $d$.

$$W_{t,d} = \frac{t_f \times \log(n/n_t + 0.01)}{\sum_{i \in d}[(t_f \times \log(n/n_i + 0.01))]^2}$$  \hspace{1cm} (1)

4) Document similarity computing: We use Cosine to measure the similarity of two documents. Suppose document $d_i=<w_1, w_2, \ldots, w_m>$, $d_j=<w_1, w_j2, \ldots, w_m>$, then Cosine similarity $S(d_i, d_j)$ of $d_i$ and $d_j$ is given in (2)

$$S(d_i, d_j) = \frac{\sum_{k=1}^{m}(w_k \times w_k)}{\left(\sum_{k=1}^{m}w_k^2\right)\left(\sum_{k=1}^{m}w_k^2\right)}$$ \hspace{1cm} (2)

5) Document classifying: We use k-Nearest Neighbor (k-NN) classifier for classification. k-NN exhibits encouraging performance in text classification and does not rely on prior probabilities. We choose 2-NN algorithm for our purpose. The two class is marked with $C_1$ and $C_2$, with $C_1$ the 350 positive examples of IT security-related publication and $C_2$ the 200 negative examples of other domain. All examples were marked with IT security professionals. Let $C_1\{d_1,1, d_1,2, \ldots, d_1,s\}$, $C_2\{d_2,1, d_2,2, \ldots, d_2,t\}$, the distance between $d_i$ and $C_j$ $(j=1 \ or \ 2)$ is given in (3)

$$D(d_i, C_j) = \frac{1}{||C_j||} \sum_{j=1}^{||C_j||} \sqrt{\sum_{s \in F}(w_{i,s} - w_{j,s})^2}$$ \hspace{1cm} (3)

$d_i$ is classified into class $C_i$ if $D(d_i, C_1) < D(d_i, C_2)$, $C_2$ otherwise.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Prototype System

For the purpose of evaluation, we built a prototype system and created test datasets. The architecture of the prototype system is shown in Figure 1.

B. Experimental Corpus

The experimental corpus is collected by querying CiteSeer and downloading from paper URL parsed out from result page. CiteSeer is a scientific literature digital library and search engine that focuses primarily on the literature in computer and information science. It indexes PostScript and PDF research articles on the Web and uses ACI to automatically create a citation index that can be used for literature search and evaluation. CiteSeer supports phrase search and allows browsing the database using citation links.

Our system implemented an automatic querying interface to CiteSeer. It submits each phrase in KeySecWords to CiteSeer and processes the result. Titles and description page links of the listed publications are extracted out. Each description page link is submitted to CiteSeer again to retrieve the deposit link and citations. PDF files were downloaded from the deposit link. The citations are parsed for extensive searching to collect more documents. In this way, we collect more than 16000 PDF files. After duplication and data clean, we get 5620 documents. Another 32 phrases belonging to fields of computer network, machine learning and parallel computing were selected to collect another 1240 examples. The total 6860 documents are split into two sets: one for training and one for testing. The training set S1 has a total of 500 documents classified by IT security experts, out of which 300 are positive (IT security-related and 200 are negative. The testing set S2 has 5860 documents and is subdivided into two sets S21 and S22. S21 has 500 documents classified by IT security experts, out of which 276 are positive and 224 are negative.
In pre-processing phase, 223534 words were generated from all documents. Our feature reduction steps are as follows: firstly, words less than 3 characters and appearing in less than 3 documents were removed. Secondly, we computed the product of term frequency and document frequency of each word of the 121937 words generated from last step and sorted them in descending order. The first 15000 words are selected to construct our GeneralTerm vocabulary. Table 1 shows the dimension of the three feature spaces constructed from vocabulary GeneralTerm (G), KeySecWords (K) and RFC2828 Glossary (R) respectively.

**TABLE I. COMPARISON OF DIMENSION OF THREE FEATURES SPACES**

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>G</th>
<th>K</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>15000</td>
<td>140</td>
<td>1050</td>
</tr>
</tbody>
</table>

C. Experimental Results and Discussion

For each feature space, feature weights were assigned using TF/IDF scheme as in (1). The similarity of two documents is computed using cosine similarity as in (2). For each document in $S_1$, we computed its distance with $C_1$ (IT security-related) and $C_2$ (Others). Let $tp$, $fp$, $tn$ and $fn$ be the true positive, false positive, true negative and false negative respectively, by reviewing the classification result of every document in $S_2$ manually, we use macro-precision (MP), macro-recall (MR) and macro-f1 (MF) with (4) - (6) to compute the classification performance metrics, as shown in Table 2.

\[
MP = \frac{tp}{tp + fp + tn + fn}/2. 
\]

\[
MR = \frac{tp + tn + fn}{tp + fp}/2. 
\]

\[
MF = \frac{MP \times MR \times 2}{MP + MR}. 
\]

**TABLE II. CLASSIFICATION RESULT OF THREE FEATURE SETS**

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>151</td>
<td>157</td>
</tr>
<tr>
<td>FP</td>
<td>142</td>
<td>41</td>
</tr>
<tr>
<td>TN</td>
<td>82</td>
<td>183</td>
</tr>
<tr>
<td>FN</td>
<td>125</td>
<td>119</td>
</tr>
<tr>
<td>MP%</td>
<td>45.57</td>
<td>64.91</td>
</tr>
<tr>
<td>MR%</td>
<td>45.66</td>
<td>64.86</td>
</tr>
<tr>
<td>MF%</td>
<td>45.62</td>
<td>64.86</td>
</tr>
</tbody>
</table>

According to the result in Table 2, the performance varies across the three feature set. RFC2828 glossary-based method performs the best and reaches an F-measure of 88.46, a significant improvement over the other two. Although KeySecWords-based method exhibits better result than the general term-based method, its false negative is relatively a little higher. Its performance benefits much from the true negative score.

V. CONCLUSIONS

The primary focus of this experiment was to evaluate a domain-specific lexicon for identifying IT security literatures. Scientific literature covers most of the hot topics in the research field and has a large domain-specific vocabulary. The exploitation of domain knowledge and specialized vocabulary can dramatically improve the result of literature text processing. Experimental results demonstrated that RFC2828 glossary can effectively improve the performance of IT security literature identification compared to the common statistical approach-based feature choice scheme. It showed that RFC2828 glossary was suitable for the intended purpose of IT security literature identification.

ACKNOWLEDGMENT

This research was partially supported by the general projects of Sci. & Tec. Development of Beijing Municipal Education Commission under grant KM200610016004.

REFERENCES