I.J. Wireless and Microwave Technologies, **2012, 5, 34-40**
Published Online October 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2012.05.05

Available online at http://www.mecs-press.net/ijwmt

Keywords Review of IT Security Literature in Recent 20 Years

QIAN Liping[a,*1], LIN Yan[a], WANG Lidong[b,*2]

[a] *Dept. of Computer Science, Beijing University of Civil Engineering and Architecture, Beijing, China*
[b] *National Institute of Network Information Security, Beijing, China*

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 identify the frequently used keywords in IT security literatures. The result then can be utilized to improve the performance of automatic IT security document retrieval, identification and classification. Our method is to query CiteseeX to retrieve source data of paper description information and build an artificially annotated corpus. Over the corpus, we perform words frequency statistics, word co-occurrence analysis, discrimination index computation, retrieval efficiency analysis and thus build a lexicon of IT security based on our experimental result. The lexicon can further be used in improving retrieval performance and assisting new words discovering and document classification.

Index Terms: Lexicon; IT Security; Scientific Literature; Vocabulary; RFC2828

1. 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

* Corresponding author:
E-mail address: [*1] wangzs@vip.sina.com; [*2] gmwanglidong@gmail.com

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 high frequency keywords and key phrases in IT security literature. These keywords and key phrases can then be utilized to improve the performance of automatic document classification and web document recommendation. Our method is to submit some IT security-related keywords and phrases to Internet search engine and further retrieve the necessary items, such as title, publish year and abstract, from the query result. A large annotated corpus was built with the retrieved information. By performing top-k analysis, co-occurrence analysis and search efficiency analysis over the corpus, we construct an IT security lexicon. The validity of the lexicon will be evaluated in the future by experiment of automatic document classification. An experimental system was developed to evaluate the performance of our methods.

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.

2. 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 {ti|i=1...n} and some kind of weighting method are selected, then a textual document d can be represented as (t1, w1; t2, w2; …; tn, wn).

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 curated based on the presence of experimental evidence of Drosophila gene products. 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.

3. Evaluation of IT Security Keywords

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.

3.1. 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]. IETF RFC 2828 (Request For Comment 2828) provides an internally consistent, complementary set of abbreviations, definitions, and explanations for use of terminology related to IT security [9]. But besides somewhat outdated, there are also many non-security terms included to make the glossary self-contained. In this paper, we firstly build a controlled vocabulary, SeedAND in which the items are selected by experts from RFC 2828 and the CFP of some top security-related international conferences in recent years, the taxonomy of Internet attack research and several online network security dictionaries (e.g. http://www.Itsecurity. com/dictionary/). The size of SeedAND is 201.

We perform keywords analysis on all non-stopping words over a large annotated corpus built from some IT security literature in recent 20 years. Based on our experimental result, another IT security lexicon is automatically built. The effectiveness of the lexicon in identifying and classifying new IT security literature will be evaluated in the future.

3.2. Methodology

The experimental corpus is retrieved from CiteseerX by implementing a meta-searching interface. Shallow NLP techniques were exploited to fulfill our purpose.

1) Pre-processing:

Pre-procession includes data clean, stop words removal and stemming. Description page of a paper was deleted from corpus if title or abstract information about the paper was missing. Stop words were removed using Glasgov stop-words vocabulary and stemming was performed using an improved version of Porter's algorithm.

2) Word frequency computing:

We perform keywords analysis on all non-stopping words over the corpus. The whole corpus, C_{All}, is Annotatd by IT security experts and divided into two sets, the positive set C_{Pos}, the negative set C_{Neg}. Word frequency is computed over $C_{All,}$ C_{Pos} and C_{neg} not only in total but also on each year from 1989 to 2010 respectively. By which we get an overview of evolution of hot topics on IT security research. We use TF/IDF weight scheme to measure the importance of a word. There are a lot of variants of TF/IDF weighting methods. Let tf_w be the frequency of phrase w in corpus, n the total number of papers in the corpus, n_w the number of papers containing word w. We choose the common TF/IDF method as in (1) to computer W_w of word w in our corpus C.

$$W_w = \frac{tf_w \times \log(n/n_w + 0.01)}{\sqrt{\Sigma_{w_i \in C}[(tf_{w_i} \times \log(n/n_{w_i} + 0.01)]^2}} \qquad (1).$$

We calculate weight over C_{All}, C_{Pos} and C_{Neg} respectively and use discrimination index DI to describe the difference between the words over two different corpus. Suppose for corpus C1, the first k_1 maximum weight words consists W_{C1}, and for corpus C2, the first k_2 maximum weight words consists W_{C2}. $W_{C1}=\{w_{11}, w_{12}, \ldots, w_{1k1}\}$, $W_{C2}=\{w_{21}, w_{22}, \ldots, w_{2k2}\}$, Let wt_{ij} be the weight value of w_{ij}, then $DI(W_{C1}, W_{C2})$ is computed as in (2):

$$DI(W_{C1}, W_{C2}) = \frac{\sum_{w_{1x} \in W_{C1} - W_{C2}} wt_{1x} + \sum_{w_{2y} \in W_{C2} - W_{C1}} wt_{2y}}{\sum_{w_{ij} \in W_{C1} \cup W_{C2}} wt_{ij}} \qquad (2)$$

3) Word co-occurrence statistics:

In statistical NLP methods, word co-occurrence shows probabilistic word association. It has been extensively used in information retrieval and document classification systems and is an effective way for improving retrieval performance and assisting new words discovering. The top-k words co-occurrence on years usually implies the evolution of hot topics on IT security research. We use a two word collocational window to capture bigram word co-occurrence. The weight of each co-occurrence is also computed with TF/IDF-like scheme and the discrimination index of word co-occurrence over different corpus is computed as in (2).

4) Retrieval effectiveness computing:

A major problem in today's prevailing search engine is the user has to identify the item that describes the subject of interest to achieve an effective search. In this paper, we do "AND" query instead of "PHRASE" query when collecting source data. The "AND"-mode query requires all terms in the submitted phrase appear in the result, while "PHRASE"-mode searches for the exact phrase as the submitted one. The retrieval effectiveness of each submitted phrase is computed over the annotated corpus as in (3), where n(pi) stands for the number of papers retrieved with phrase and, of this, $df_{pos}(p_i)$ for the number of papers annotated as positive.

$$Eff\,(p_i) = df_{pos}(p_i)/n(p_i) \tag{3}$$

4. Experimental Results and Discussion

4.1. Prototype System

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

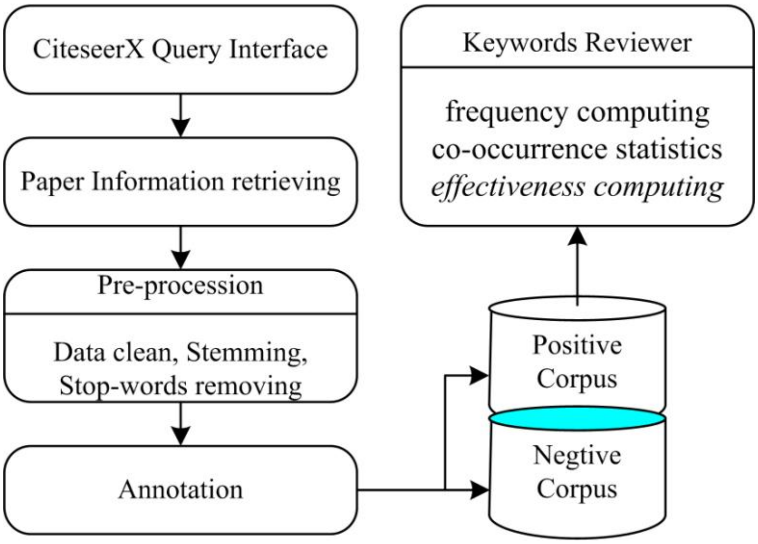

Fig. 1. System architecture

4.2. Experimental Corpus

The experimental corpus is collected by querying CiteseerX with keywords in Seed_{AND} and accessing the description page corresponding to paper URL parsed out from result page. CiteseerX 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. CiteseerX supports phrase search and allows browsing the database using citation links.

Our system implemented an automatic querying interface to CiteseerX. It submits 201 phrases in Seed_{AND} one by one to CiteseerX and processes the result. The description page links of the listed publications are extracted out. Each description page link is further accessed to retrieve the necessary items including paper title, abstract, published year. In this way, we collect information of 18,604 papers after duplication data clean, of which there are 4963 (60.03% positive) with unknown published year and 626 (24.60%) published in 1990 and before. Annual distribution of the retrieved paper from 1991 to 2010 is (total/positive%): 174 (28.16%), 234 (27.35%), 319 (32.29%), 440 (30.91%), 508 (34.25%), 651 (42.24%), 773 (40.49%), 868 (47.24%), 914 (50.11%), 1085(54.84%), 1049(54.53%), 1164(60.40%), 1137(62.09%), 1072(69.68%), 1053(72.93%), 754(74.67%), 436(72.02%), 217(71.89%), 138(57.97%), 33(57.58%).

4.3. Experimental Results and Discussion

The total 18,608 papers are annotated by IT security experts and split into two sets: the positive set is IT security-related and has a total of 10336 papers, the negative set contains the other 8272 papers. After pre-processing, 45159 words, also the Allwords vocabulary, were generated from all papers. There are only 33727 and 26719 words generated from positive corpus and negative corpus respectively.

Table 1 shows the first 10 maximum weight word co-occurrence in decreasing order over C_{All}, C_{Pos} and C_{Neg}.

Table 1. Maximum weight word co-occurrence

Corpus	**Co-occurrence**
C_{All}	Internet drafts, key agreement, access control, sensor networks, buffer overflow, web services, identity based, intrusion detection, data mining, ad hoc
C_{Pos}	Internet draft, buffer overflow, key agreement, web services, access control, ad hoc, sensor network, intrusion detection, identity based, key exchange
C_{Neg}	Internet draft, ad hoc, real time, risk management, rule based, risk assessment, peer peer, end end, world wide, large scale

The discrimination indexs on the first 200 words are as follows: DI(C_{All}, C_{Pos})=0.126, DI(C_{All}, C_{neg})=0. 208, DI(C_{Pos}, C_{neg})=0.454. The discrimination indexs on co-occurrence are as follows: The discrimination indexs on word are as follows: DI(C_{All}, C_{Pos})=0.730, DI(C_{All}, C_{neg})=0.197, DI(C_{Pos}, C_{neg})=1.718. The above DI value suggests that our lexicon is more representative.

In our retrieval effectiveness computing, the first 10 phrases are: block cipher (99%), rbac(98%), biometric authenticatin(96%), masquerade attack(96%), ciphertext(95%), digital watermarking(95%), certificate revocation(94%), diffie hellman(94%), hijack attack(94%); and the last 10 are: green book(2%), british standard 7799(5%), manipulation detection code(6%), system resource(6%), system availability(7%), data integrity(7%), carnivore(8%), social engineering(8%), one time password(8%), clark Wilson model(9%).

5. Conclusions

The exploitation of domain knowledge and specialized vocabulary can dramatically improve the result of literature text processing. The primary focus of this paper was to construct a domain-specific lexicon for identifying IT security literatures. The result then can be utilized to improve the performance of automatic IT security document retrieval, identification and classification. Since scientific literature covers most of the hot topics in the research field and has a large domain-specific vocabulary, our method is to select some generally accepted IT security keywords, submit each of them to CiteseeX, parse the search results and the relevant paper description page. Items as titles, abstract, authors and publishing year are extracted to build our corpus. The corpus is annotated by IT security experts. Over the corpus, we perform words frequency statistics, word co-occurrence analysis, retrieval efficiency analysis. A lexicon is recommended based on the above result. The DI value suggests that our lexicon is more representative. The effectiveness of the lexicon in identifying and classifying new IT security literature will be evaluated in the future.

Acknowledgements

This paper is partly supported by project 2010Z14.

References

[1] Ingrid Petric, Tanja Urbancic, Bojan Cestnik. Discovering Hidden Knowledge from Biomedical Literature. Informatica, vol. 31, pp. 15-20, Mar. 2007.

[2] D. Shilin, H.Minlie, W.Hongning, Xiaoyan Zhu. Profile-feature Based Protein Interaction Extraction from Full-Text Articles. Proc. of the 7th BIOKDD Workshop in the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2007, pp. 42-49, published online: http://bio.informatics.iupui.edu/ biokdd07/BIOKDD07_Proceedings.pdf.

[3] L. Yudong, S.Zhongmin, S.Anoop. Exploiting Rich Syntactic Information for Relation Extraction from Biomedical Articles. Proc. of the Annual Conference of the NAACL-HLT 2007, Companion Volume, Apr. 2007, pp. 97-100.

[4] A. M. Cohen, W.R.Hersh. A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformaties, vol. 6, Mar. 2005, pp. 57-71, doi:10.1093/bib/6.1.57.

[5] L.Venkata Subramaniam, Sougata Mukherjea, Pankaj Kankar. Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application. Proc. of the 12th International Conference on Information and Knowledge Management, ACM, Nov. 2003, pp. 410-417, doi: 10.1145/956863.956941.

[6] C. Ramakrishnan, K. J. Knchut, A. P.Sheth. A Framework for Schema-Driven Relationship Discovery from Unstructured Text. Lecture Notes in Computer Science, Springer Berlin/Heidelberg, vol. 4273/2006. International Semantic Web Conference Nov. 2006, pp. 583-596, doi: 10.1007/11926078.

[7] Simon Hansman. A Taxonomy of Network and Computer Attacks. Computers & Security. Vol. 24, Feb. 2005, pp.31-43, doi:10.1016/j.cose.2004.06.011.

[8] Ali Abbas. A Comprehensive Approach to Designing Internet Security Taxonomy. Proc. of the Canadian Conference on Electrical and Computer Engineering, 2006 (CCECE’06), IEEE Press, May 2006, pp. 1316-1319, doi: 10.1109/CCECE.2006.277393.

[9] R. Shirey. Internet Security Glossary, RFC 2828, The Internet Engineering Task Force, May 2000.