A RECOMMENDER SYSTEM BASED ON ONE-CLASS CLASSIFICATION

HỆ THỐNG TƯ VẤN DỰA TRÊN MÔ HÌNH PHÂN LỚP

Nhan Cach Dang[1], Duong The Bui[1]

[1] Ho Chi Minh City University of Transport (UT-HCMC), Vietnam

Abstract: *In the early days, outliers were considered anomalous, possibly erroneous observations, which should be identified and removed from the analysis, when building the model. However, analyzing such outliers might be useful in some case.*

In this paper, we develop a recommender system to identify and explain ab-normal accesses to our system, the Online Learning System – based on Moodle (OLS) and the online Educational Management System (EMS). Our approach is proposed on the basic of one-class classification and Support Vector Machine to process and classify raw text data. Our experiments demonstrate the utility and accuracy of the system in dataset from the usage of our system.

Keywords: *Recommender System; text mining; TF-IDF; vector space model; support vector machine; One-Class SVM.*

Classification number: *1.4*

Tóm tắt: *Trước kia, khi xây dựng mô hình, các giá trị bất thường, khác biệt với những dữ liệu quan sát thì cần được xác định và loại bỏ khi phân tích. Tuy nhiên, trong một số trường hợp, phân tích những khác biệt này mạng lại giá trị hữu ích trong những năm gần đây. Trong bài báo này, chúng tôi phát triển hệ thống tư vấn cảnh báo để xác định, giải thích các truy cập bất thường vào hệ thống của chúng tôi, Hệ Thống Đào Tạo Trực Tuyến – dựa trên nền tảng Moodle (OLS) và Hệ Thống Quản Lý Đào tạo (EMS). Cách chúng tôi tiếp cận dựa trên mô hình phân lớp (One-Class Classification) và Máy học hỗ trợ vectơ (Support Vector Machine, SVM) để xử lý và phân loại dữ liệu. Các thí nghiệm thể hiện tính hữu ích và tính chính xác của hệ thống trong tập dữ liệu từ việc sử dụng hệ thống của chúng tôi.*

Từ Khóa: *Hệ thông tư vấn, khai phá văn bản, TF-IDF, máy hỗ trợ vector, mô hình phân lớp, One-Class SVM*

Chỉ số phân loại: *1.4*

1. Introduction

Searching for anomalous observations (called outliers) is as old as the data analysis itself. There is no need to argue that data models deduced from data contaminated with outliers may yield very poor image of the structure of such data. More and more applications new a day are built on the basic of finding outliers.

This paper presents a recommender system for data analysis and decision support making (DSS) that we need frequently to judge whether the observed data items are normal or abnormal. Our approach is proposed on the basis of the Vector Space Model and Support Vector Machine [1, 2] with One-Class classification (OSVM) to process and classify raw text data. Although OSVM has been studied for problems of text classification recently [3], applying it to analyze log file of data usage from the education management system on university is also useful. Our experiments on our two systems, Education Management System and the Online Learning System, demonstrate the utility and accuracy of the Recommender in these systems.

The rest of the paper is organized as follows. Section 2 introduces a background and related work of this trend. Section 3 presents the Recommender system based on one-class classification. Section 4 introduces datasets for evaluating this Recommender. Section 5 summarizes experiments and results, followed by the conclusion in Section 6.

2. Background and Related Work

2.1. Vector Space Model

Vector space model or term vector model [4] is an algebraic model for representing text documents as vectors of identifiers. The elements of this vector expresses the relevance ranking of words or some word frequency function as the appearance or absence of each word in the document.

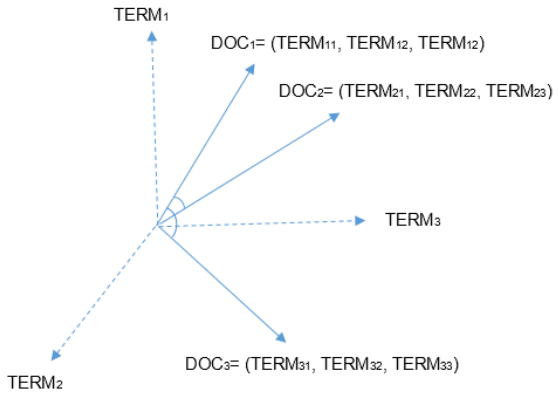

Fig.1. *Presentation vector space model.*

This model presents text documents as points in n-dimensional Euclidean space. Each unique term in the document collection corresponds to a dimension in the space. Documents are viewed as points in a hyperspace whose axes are the terms used in the document vectors. The location of a document in the space is determined by the degree to which the terms are presented in a document. The similarity between two documents is defined as the distance between the points corresponding to them or the angle of the vectors, as shown in Fig. 1.

The measure TF-IDF (Term Frequency - Inverse Document Frequency) is often used because of its effectiveness in the field of text mining. This is a common method to evaluate and rank the importance of a word in a document. Details of this measure are shown in the following sections.

2.2 TF-IDF measure

TF-IDF, the short term of frequency–inverse document frequency, is a statistical measure reflecting how important a word is to a document in a collection or corpus [5]. Just like the name, TF–IDF is the product of two statistics, Term Frequency (TF) and Inverse Document Frequency (IDF).

TF (Term Frequency) – the number of times the words appears in the Document. It is measured which raw frequency divided by the maximum raw frequency of any term in the document:

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

Where:

- f(t,d) is the frequency, that is ,the number of times the word t appears in the Document d,
- max {f(w,d):w∈d} is maximum raw frequency of any term in the document.

IDF (Inverse Document Frequency) is a reciprocal of the number of Documents in which the word occurs. The inverse document frequency is a measure of whether the term is common or rare across all documents. For example, in a corpus of fashion documents, the term “fashion” or “model” will appear all over the corpus. When it is already a popular term, it will not provide much information. IDF is cal-culated as:

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Where:

- |D| : total number of documents in the corpus D,
- $|\{d \in D : t \in d\}|$:number of documents where the term t appear(it means $tf(t,d) \neq 0$).

If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to

$1+|\{d \in D : t \in d\}|$

Mathematically, the base of the log function does not matter and constitutes a constant multiplicative factor towards the overall result. In other words, the change in the base of the log function will not change the ratio between IDF results.

$$tfidf(t,d,D) = tf(t,d) \times idf(f,D)$$

A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the

weights hence tend to filter out common terms, and keep the importance term (or keyword).

For the purpose of retrieving useful information related to fashion from online social network data, we use TF-IDF measure for transforming data to space vector model. However, we propose a technique that reduces the dimensional specifications of this Vector Space Model to train an SVM (Support Vector Machine) classifier.

2.3 Validation measuring for retrieved documents

Precision, recall, and the F measure [6, 7] are set-based measures. They are computed by using unordered sets of documents. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each such set, precision and recall values can be plotted to give a precision-recall curve, such as the one shown in Fig. 2.

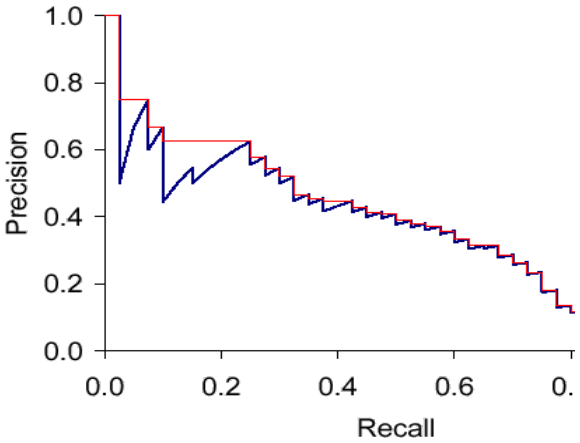

Fig. 2. *Precision/Recall graph1.*

For classification tasks, the terms true positives, true negatives, false positives, and false negatives (see Fig. 3) compare the results of the classifier under test with trusted external judgments. The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to whether that prediction corresponds to the external judgment. They are defined as an experiment from P positive instances and N negative instances for some conditions. The four outcomes can be formulated as showed in

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive (tp)	False positive (fp)
	Predicted condition negative	False negative (fn)	True negative (tn)

Fig. 3. *P positive instances and N negative instances for some condition.*

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percent-age.

$$precision = \frac{tp}{tp + fp}$$

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

$$recall = \frac{tp}{tp + fn}$$

Accuracy is the proximity of measurement results to the true value.

$$\text{Accuracy} = \frac{tp + tn}{(tp + tn + fp + fn)}$$

F measure is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score is the harmonic mean of precision and recall:

$$F = \frac{recall \times precision}{(recall + precision)/2}$$

Commonly used evaluation measures including Recall, Precision, F-Measure and Rand Accuracy are applied due to their origin in Information Retrieval. In fact, sometimes we can not directly use these measures to compare two lists of ordered documents returned because of independence of the internal order of the documents [8]. To measure the quality of an ordered list of documents, the average precision of all the relevant documents in the ordered list can be calculated.

2.4 One-Class Classification

One-Class classification tries to identify objects of a specific class amongst all objects, by learning from a training set

1 http://nlp.stanford.edu/IR-book

containing only the objects of that class. The traditional classification tries to distinguish between two or more classes with the training set containing objects from all the classes. The term One-Class classification (OCC) was introduced by Moya and Hush [9], some researchers have applied OCC but some of them use other terms to define one class classification such as Outlier Detection [10], Novelty Detection [11] or Concept Learning [12]. Many works can be found in scientific literature, such as in [13] where news recommendation for users based on One-class classification is proposed. They used OCC to calculate the similarity between domain and user interesting. In the research [3, 14, 15], authors are over view the technique use One-Class Classification show at the Fig.4.

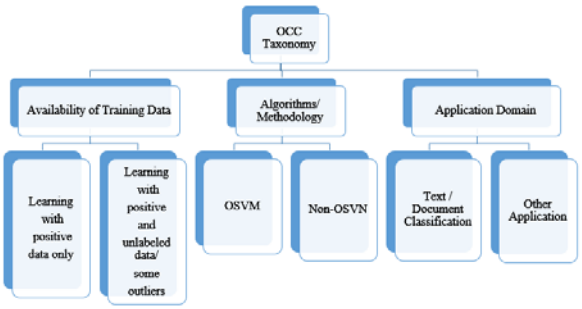

Fig. 4. *A Taxonomy for the study of OCC techniques [15].*

Simply, one-class classification process includes some steps: (1) data preprocessing; (2) model machine learning; (3) classification processing in training model and (4) result interpretation and reporting. Data pre-processing is an important step in the data mining process. Data pre-processing includes cleaning, normalization, trans-formation, feature extraction, selection and transforming the text data into space vector model. Machine learning focuses on prediction, based on known properties learned from the training data.

3. Recommender System based on One-Class

This recommender system will observe the access to our systems by checking the details of information of the users. such the ID of the user to login the EMS, there are some relevant information such as: computer name; Operating system (OS) or version OS, IP address…Process of our Recommender can be divided into three phases: the general phase in which we measure the probability each of every user's ID that usually appears on the computer or when it appear and add them to the reference table data. This task will done every day to update the information. The main phase involves one-class classification to train model of the system. The training data we used was collected from the data of our system. In the final phase, the system will detect and recognize the strange or irregular access to our systems. It will alert to the administrator or email to the user.

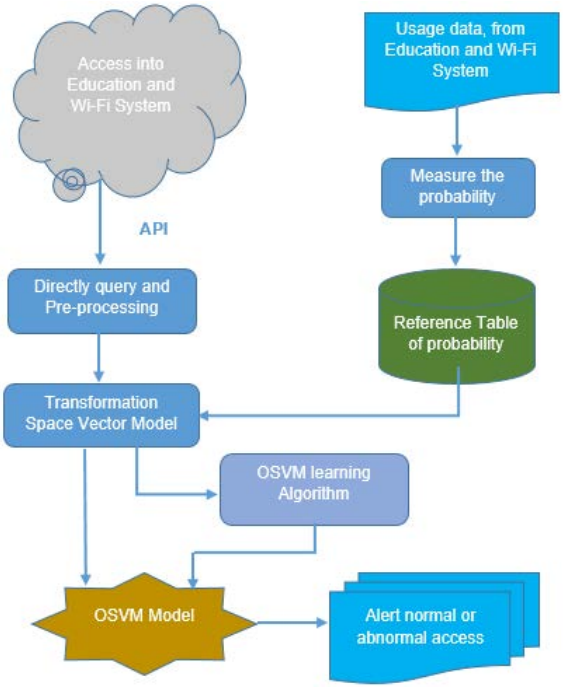

Fig. 5. *Process of the Recommender System.*

Fig. 5 presents the process of our Recommender system that makes the table reference and trains a model for detecting access from our EMS and OLS. In the stage of model training, user's data including computer name, MAC Address; Operating System are collected through the use of API protocol when he/she accesses to the system. Text data are transformed into Space Vector Model by using IF-IDF measure. In the next stage, we also collect data with some new access to the system. And we try to detect abnormal access. In the next session, we present in a more detailed way the data used

in the experiments carried out to test and evaluate the effective-ness of this Recommender.

4. Data model scenario

To perform the experiment on the proposed Recommender, we use the data from two systems of EMS2 and OLS3 University. First, there are over 12.0000 accounts in the o and there are over 30.000 access to my EMS from 2014 to 2016 to check our Recommender.

Fig. 6. *Report logs of users access to the cource.*

Training dataset are built based on data of two system providers. That is, we use well-known access to these systems. In addition, we use some strange access such as the information of the login to the education with new laptop. Fig. 6 presents information of the user access to these systems. Based on well-known information we analyze the data to find abnormal access to systems.

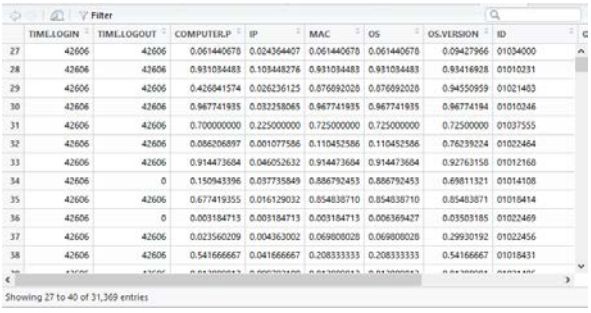

Fig. 7. *The probability of appearance of ID in the computer in the reference table.*

5. Experimental Results

With the data sources described in the previous section, we perform several experiments to test the effectiveness and accuracy of the proposed Recommender. From each source (EMS and OLS). We accessed the system with strange computers, we had trained the model with usage data before that and using this strange information to test the model. We use both Accuracy and F measure to evaluate the accuracy of filtering. Because F-measure is derived from Recall and Precision, we also show the two measures for reference purpose.

For each run, we use one ID to train and test with over 32.000 rows data. Fig. 7 and Fig. 8 show that the high accuracy was achieved for 10 random ID.

#	ID Account	Accuracy
1	Admin	0.9242
2	1034000	0.8922
3	1044001	0.9570
4	1022456	0.8249
5	1021483	0.9767
6	1022468	0.9848
7	1037555	0.9631
8	1022464	0.6917
9	1020452	0.9399
10	1022464	0.6799

Fig. 8. *Experimental results that show our filters have high accuracy in the experimental Education cases.*

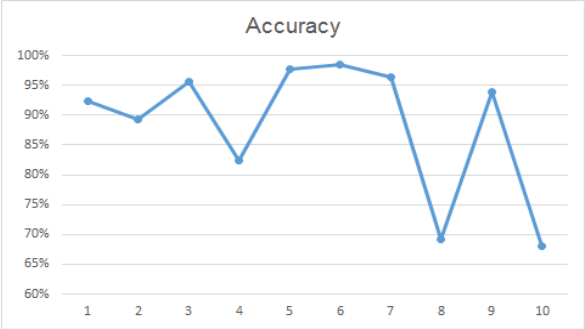

Fig. 9. *Experiments with random ID accounts. The high accuracy was achieved for many accounts.*

The Information in the Fig. 9 shows experimental result with usage data access to the EMS. We use one class to train the model and test it with some random ID ac-counts. The Accuracy measure in our experiment is high, as shown in this chart.

6. Conclusion

In order to provide intelligent recommendation and personalize service for users on the system, in this research, we built a Recommender System based on One-Class classification that can alert the irregular access to the system. The Recommender System tested with datasets from our EMS and OLS shows a good performance.

2 http://gv.ut.edu.vn
3 http://courses.ut.edu.vn

The outlier information is very useful for various application domain, e.g., Outlier Detection; Novelty Detection or Concept Learning. In our future work, we will apply this model to detect outlier in online Social network data. We also employ Hadoop Framework to deal with Big Data when considering environmental data or bio information data❑

Reference

[1] Ikonomakis, M., S. Kotsiantis, and V. Tampakas, Text classification using machine learning techniques. WSEAS Transactions on Computers, 2005. 4(8): p. 966-974.

[2] Sebastiani, F., Machine learning in automated text categorization. ACM computing surveys (CSUR), 2002. 34(1): p. 1-47.

[3] Bhatt, J. and N.S. Patel, A SURVEY ONE CLASS CLASSIFICATION USING ENSEMBLES METHOD. International Journal for Innovative Research in Science and Technology, 2015. 1(7): p. 19-23.

[4] Turney, P.D. and P. Pantel, From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 2010. 37(1): p. 141-188.

[5] Rajaraman, A., J.D. Ullman, and J.D. Ullman, Mining of massive datasets. Vol. 77. 2012: Cambridge University Press Cambridge.

[6] Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

[7] Fawcett, T., An introduction to ROC analysis. Pattern recognition letters, 2006. 27(8): p. 861-874.

[8] Han, J., M. Kamber, and J. Pei, Data mining: concepts and techniques: concepts and techniques2011: Elsevier.

[9] Moya, M.M. and D.R. Hush, Network constraints and multi-objective optimization for one-class classification. Neural Networks, 1996. 9(3): p. 463-474.

[10] Ritter, G. and M.T. Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern Recognition Letters, 1997. 18(6): p. 525-539.

[11] Bishop, C.M., Novelty detection and neural network validation. IEE Proceedings-Vision, Image and Signal processing, 1994. 141(4): p. 217-222.

[12] Japkowicz, N., Concept-learning in the absence of counter-examples: an autoassociation-based approach to classification, 1999, Rutgers, The State University of New Jersey.

[13] Cui, L. and Y. Shi, A Method based on One-class SVM for News Recommendation. Procedia Computer Science, 2014. 31: p. 281-290.

[14] Khan, S.S. and M.G. Madden, One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review, 2014. 29(03): p. 345-374.

[15] Khan, S.S. and M.G. Madden, A survey of recent trends in one class classification, in Artificial Intelligence and Cognitive Science2009, Springer. p. 188-197.

Ngày nhận bài: 29/03/2018
Ngày chuyển phản biện: 01/04/2018
Ngày hoàn thành sửa bài: 21/04/2018
Ngày chấp nhận đăng: 28/04/2018