Dead Link Prediction Model Based on Double SVM

Liu Honglan
Beijing Key Laboratory of Knowledge Engineering for Materials Science, China.
School of Computer and Communication Engineering, University of Science and Technology Beijing
Beijing, China
honglanliu@ies.ustb.edu.cn

He Yong
Beijing Key Laboratory of Knowledge Engineering for Materials Science, China.
School of Computer and Communication Engineering, University of Science and Technology Beijing
Beijing, China
wizardhy@gmail.com

Qin Xiaona
Beijing Key Laboratory of Knowledge Engineering for Materials Science, China.
School of Computer and Communication Engineering, University of Science and Technology Beijing
Beijing, China

qinxiaonazi@foxmail.com

Abstract—All the search engines are facing the problem of dead link. Dead link prediction model, which is able to quickly distinguish between normal links and dead links, filter out dead links, ensure the validity of the search results. This paper proposed to construct the SVM dead link prediction model by the effective attributes of the links because the fast update of the links in engine library, this model can quickly identify dead links. Due to the attributes related to the types of web site, independent training prediction model was proposed according to the different types of web site to improve the precision rate, experiments proved that the precision and recall rate of the independent model higher than the uniform model. It is unrealistic that training samples marked completely rely on manual because the independent training samples' amount is large, so using SVM dead link prediction model based on web content to prepare the sample because it's high accuracy, and applicable to all links. It constituted a dead link prediction model based on double SVM by using SVM dead link prediction model twice, which greatly improved the precision rate of dead link prediction, and reduces the prediction time.

Keywords-dead link; Support Vector Machine; the effective attributes;Dead Link Prediction Model;the independent model

I. INTRODUCTION

There are more and more resources on the Internet, and they change faster and faster, so the search engines become more and more important. For a search engine, if the dead links stay in the front of the search results, users will get an invalid, no content or an error pages when click on it. It will waste the users' time and energy, and seriously impact on the user experience. The research on the prediction of dead link, can rapidly distinguish normal links and dead links, filter out the dead links, and ensure the effectiveness of the search results.

The experiments proved that SVM have a better classification effect on the binary classification. This paper used SVM as a machine learning method, studied on the Internet dead link issues deeply. First built the SVM dead link prediction model based on web content to prepare the sample. Then construct the SVM dead link prediction model based on the effective attributes of the links. And proved that the usability of the independent dead link prediction model based on double SVM through comparing the unity model and independent model.

II. VECTOR SPACE MODEL

SVM has a simple geometric significance, when the training sample linear separable, in the training set $D=\{(x_1,y_1),...,(x_n,y_n)\}$,$x \in Rn$,$y \in \{-1, +1\}$,$i=l,2,...,n$, (Formula 1). Solve the hyperplane$<w, x> + b = 0$ (Formula 2)to separate the two types of samples correctly, to meet the requirements of the hyperplane can have more than one. The two types of samples with a maximum interval of Margin hyper planes for optimal hyperplane, it makes SVM have good generalization ability. The sample points nearest from hyperplane determine the position of the optimal hyperplane, nothing to do with the sample points that far away from the optimal hyperplane, sample points closest to the hyperplane is also called support vectors [1].

A. Linearly separable:

Solving $<w, x> + b = 0$, the process is described as follows: In the particular training sample set, solving the normal vector w and b values to make the interface class

interval D (W, b) maximum ,equals to make $\frac{1}{2}\|w\|^2$ has a minimum value. To solve the original problem, as is expressed by the formula: $\min \frac{1}{2}\|w\|^2$ $s.t., y_i\left(w^T x_i + b\right) \geq 1, i = 1,...,n$ (Formula 3). Solving the above problem by introducing Lagrange solution, get the optimal classification function is:

$$f(x) = \left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^T x + b = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \quad \text{(Formula 4)}$$

B. Linearly inseparable:

When the samples of the training set is linearly inseparable, the objective function becomes $\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k} \xi_i$ s.t. $y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, i = 1,...,k$ (Formula 5). Among them, the non-negative parameter ξ_i (Called slack variables), C is a constant greater than zero, and he controlled the punishment degree of error classifying samples, known as the penalty factor [2].

C. Non-linear problem:

For the nonlinear classification, by introducing the kernel function $K\left(x_i, x_j\right)$, the sample data of original space is mapped to high-dimensional feature space by nonlinear transformation, in the high dimensional space to find optimal or generalized optimal classification plane. Classification function becomes:

$$f(x) = \text{sgn}(\sum_{i=1}^{k} \alpha_i y_i K(x_i, x_j) + b) \quad \text{(Formula 6)}$$

Currently popular kernel functions are polynomial kernel function, the radial basis function kernel (RBF core), and linear kernel function:

Polynomial Kernel: $K(x_i, x_j) = (\langle x_i, x_j \rangle + R)^d$. Where d is an integer, representing the order of the polynomial.

RBF Kernel: $K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2})$ The kernel will be the original space for the infinite dimensional space mapping.

Linear Kernel: $K(x_i,x_j)=x_i \cdot x_j$.This is actually the inner product in the original space [3].

III. SAMPLES ACQUISITION STAGE (SVM DEAD LINK PREDICTION BASED ON THE WEBPAGE CONTENT)

The whole process of the sample acquisition (the dead link prediction based on the webpage content) can be divided into the following sections: Data acquisition, preprocessing of the webpage, feature extraction, text representation, selection and training of the classifier.

A. Data acquisition

For the normal Webpage data, randomly selected 5000 normal Webpage through the analysis of hao123，360 site .Combined with constructing and tagging dead links manually, dead links number reached 4848.The total sample number nearly ten thousand, almost the same number of positive and negative class, the whole belongs to a good training samples.

B. preprocessing of the Webpage

All of the HTML tags removed, leaving only the text, use the ICTCLAS software package for the segmentation of the text.After pretreatment, got words of the original text feature.

C. feature extraction

The untreated original feature space consume the computing resources and affects the classification accuracy. Webpage feature extraction is the process to extract features from the original feature space. The main methods of feature extraction are TF, DF, IG, MI, X2 statistics. Set the VSM dimension is 250, the kernel function is RBF, through experiment, the precision rate, the recall rate, the F1 values of X2 statistics are higher than other feature extraction methods.

D. VSM representation

After feature extraction, still not formatted data, classification algorithm is still unable to directly handle. Using the classic VSM (Vector Space Model) for web data mapping. Set feature extraction method is DF, the kernel function is RBF, the dimension of the feature space from 100 to 400, step 50, comparing the experimental results, see that dimension is 250 has the best classification results.

E. Selection and training of the classifier based on SVM

Libsvm is a simple, easy to use and fast and efficient SVM pattern recognition and regression software package that design and develop by Taiwan University Lin Chih-Jen associate professor etc. Now widely used in academic research and industry. This paper will use libsvm to carry out experiments. Classifier training process is as follows:

1) *formatted data sets.* Libsvm data format requirements in units, each line format is <label 1: v_1 2: v_2 3: v_3 ... n: v_n>.

2) *the data scale.* Data scale is the meaning of the scope of the every dimension with normalization.

3) *selection of kernel function.* Set feature extraction algorithm is x^2 statistics, VSM dimension is 250, RBF kernel has the best classification results through the experiment, precision reached 0.94, and the recall rate and the F1 value are all above 0.91, higher than other kernel functions, specific data reference to the following test.

4) *Grid search parameters.* For the SVM algorithm, in the linear inseparable and nonlinear cases there are two parameters C and σ need to be set in advance, this experiment with use the grid search method for finding the right parameters of C and σ. ⑤ *cross-validation.* The sample set is divided into training and test sets, used to verify the model ability of unknown data classification [4]-[10].

IV. SVM DEAD LINK PREDICTION BASED ON THE ATTRIBUTES OF THE LINKS

A. Data acquisition

The previous work done by SVM dead link prediction based on web content to prepare the samples' data for the SVM dead link prediction based on the attributes of the links.

B. *Attributes feature selection*

TABLE I. CHARACTERISTICS OF EFFECTIVE

numeric field	Type	Remark
link_depth	int	Links depth
page_type	int	Page type (a value represents a type)
click	int	Clicks Frequency
show	int	Show times
url_type	int	Link type (0 for static,1 for dynamic)
dir_depth	int	Directory depth
sex	int	Porn sensitivity
political	int	Political sensitivity

Dead link prediction based on attributes, is mainly artificial judgment which attributes are effective, no longer needed complex feature selection algorithms, mainly based on the artificial experience + model validation. Training for useful feature model, based on the characteristics of reflecting the link attribute, can distinguish web pages. Based on the artificial experience of effective characteristics are shown in Table 1.

The effective attributes of the links are not so much, so use these attributes to classify dead link is still relatively weak. It can be increased through collecting attributes to add features. The principle is clustering model through the shape of a URL, then statistics each URL patterns for some data , it is concluded some statistical data, The properties of the data spell together as a single URL attribute again , This can increase more effective attribute data. The set of attribute data are shown in table 2.

TABLE II. THE SET OF ATTRIBUTE DATA

Field	Type	Remark
success_crawl_ratio	double	The proportion of successful crawl
invalid_redir_ratio	double	The proportion of jump invalid
weight10_ratio	double	The proportion of empty short pages
update_fail_ratio	double	The proportion of update Failed
get_fail_ratio	double	The proportion of crawl failed
low_value_ratio	double	The proportion of low-value
content_dead_ratio	double	The proportion of content dead links
change_bad_ratio	double	The proportion of links deterioration
change_good_ratio	double	The proportion of link becomes good
change_content_ratio	double	The proportion of links deterioration

C. *Selection and training of the classifier based on SVM*

It is the same with last model, classifier choose libsvm toolkit, too.

V. DEAD LINK PREDICTION MODEL BASED ON DOUBLE SVM

The whole process of the Dead Link Prediction Based on Double SVM can be divided into the following sections：The samples acquisition phase, model learning phase and model forecast phase, as shown below Fig .1:

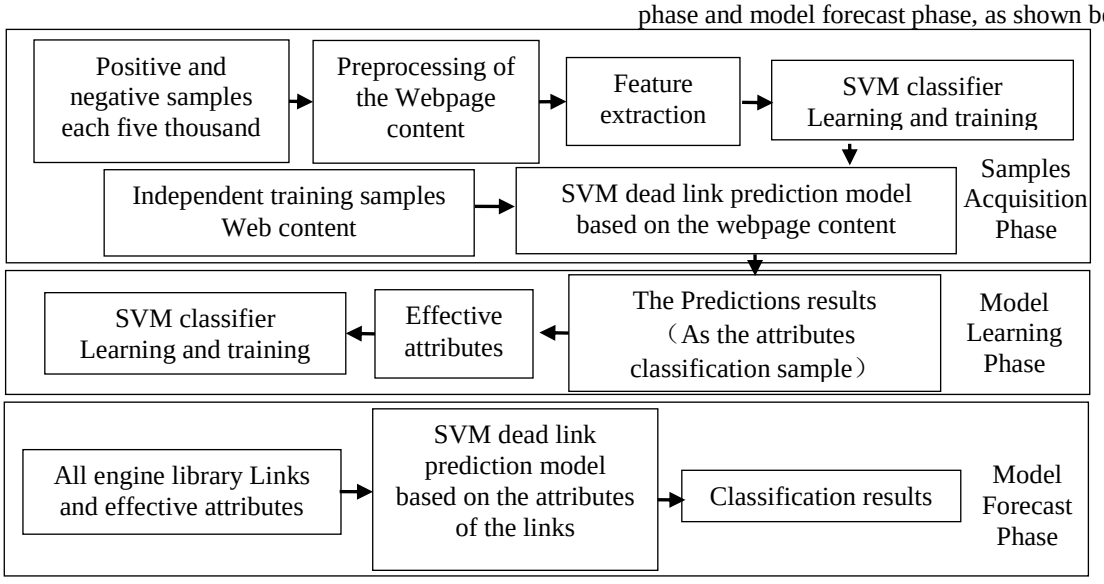

Figure 1. dead link prediction based on double SVM

VI. EXPERIMENTAL COMPARISON

A. Experiment evaluation method

For the classify problem of dead link, the target class is dead links, so dead link is defined as a positive class, and a normal link is defined as a negative class .see table 3.

TABLE III. TP, FP, FN, TN DEFINITION

	Positive class	Negative class
Are classified out	TP	FP
Not classified out	FN	TN

The formula of the precision is P= $\frac{TP}{TP+FP}$, represents the ratio of positive class among all the classified samples.

The formula of the recall is R= $\frac{TP}{TP+FN}$, represents the ratio of positive class among the class which was judged to positive class.

The formula of the F1-measure is F_1= $\frac{2*PR}{P+R}$, It is derived from the definition $\frac{1}{F_1}=\frac{1}{P}+\frac{1}{R}$, F1-measure is known as the comprehensive evaluation index, and it considers the precision rate and recall rate. The value of F1 can be integrated to judge the performance of a classifier is good or bad. This article will use F1-measure approach to evaluate the performance of the classifier.

B. Kernel function selection

Libsvm needed to choice the kernel function, the following experiments were conducted on the kernel function selection. Setting x2 statistical feature extraction algorithm, VSM dimension of 250, select the kernel function that precision rate, recall rate and F1 value are the highest as the next test kernel function.

TABLE IV. KERNEL FUNCTION SELECTION

Kernel Function	Precision	Recall	F1
RBF	0.94	0.91	0.92
Polynomial Kernel	0.89	0.90	0.89
Linear Kernel	0.84	0.83	0.83

The experiment results show that the RBF kernel has the best classification effect.Then do the next experiments by using the SVM classifier of RBF kernel.

C. experimental model approach

The goal of the SVM dead link predict model based on attribute is to forecast all the links in link library. Unified model is that all the links treat as the same kind of data, only one model need to be trained in theory. Extract part of the data from all of the simple, and the rest to do the test set. , this is no problem under the breadth of consideration. But from the perspective of more fine, Link characteristics of different sites are not the same, Such as video website links is faster than information release site changes, using separate model will be better.

Experiment with five big site data, compare the performance of the uniform model and independent model. Take the sample of each positive class, negative class reach 5000, use libsvm training and prediction. The results are shown in table 5

TABLE V. COMPARISON UNIFIED MODEL AND INDEPENDENT MODEL

Site	Unified Model			Independent model		
	Precision	Recall	F1	Precision	Recall	F1
www.tudou.com	0.78	0.83	0.80	0.86	0.91	0.88
www.sina.com.cn	0.77	0.75	0.76	0.87	0.88	0.87
tieba.baidu.com	0.93	0.94	0.93	0.93	0.93	0.93
blog.163.com	0.81	0.79	0.80	0.92	0.93	0.92
www.csdn.com	0.68	0.66	0.67	0.89	0.88	0.88

From the results can be clearly seen, independent model had better performance than the unified model. Therefore, it can improve the overall precision rate and recall rate through independent training strategies. To the big sites, use the corresponding site link data training model independently, for the small site, unified model is enough.

VII. CONCLUSION

In the rapidly changing information era，The Dead Link Prediction Model Based on Double SVM has more research value on improving time and accuracy. The method of applying SVM learning machine to predict dead link provides guidance on the link scheduling crawl, let the random blindly grasping change to targeted grasping, more useful links can be scheduled in the limited resources. But the detection cost of SVM need to be improved. The prediction of dead link still has many problems to be solved. The next step of work can be done after SVM dead link prediction based on the attributes of the links using SVM dead link prediction based on the webpage content to improve the accuracy.

REFERENCES

[1] Gu Yaxiang, Ding Shifei. Support vector machine [J]. Research progress in computer science, 2011, 38 (2):14-17.

[2] Scholkopf B, Smoda A J. Learning with kernels [M].Cambridge, MA: M IT Press, 2002

[3] Chew H G, Bogner R E, Lim C. Dual-i support vector machine with wirer rate and training size biasing [A]. In: Proceedings of 26th IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing) 2001 [C]. Salt Lake City, USA: IEEE, 2001: 1 269- 1 272.

[4] Lin ChunFu, Wang ShengDe. Fuzzy support vector machines [J]. IEEE Transaction on Neural Network, 2002, 3 (2):464- 471.

[5] Li Xiaoyu, Zhang Xinfeng, Shen Lansun. The support vector machine (SVM) and the research progress of [J]. Measurement technology, 2006, 25 (5):7-12.

[6] R.Pearson, Goney, and J.Shwaber. Imbalanced Clustering for Micro-array Time-Series[C].Proc. Int'l conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II,2003

[7] PHUA C, ALAHAKOON D. Minority report in fraud detection: Classification of skewed data [J]. S IGKDD Exploration s, 2004, 6 (1) :50259

[8] ZHENG Zhaohui, WU X, SR IHAR I R K. Feature selection for text categorization on imbalanced data [J]. S IGKDD Explorations, 2004, 6 (1) : 80289.

[9] COHEN G, H ILAR IO M, SAX H, eta. Data imbalance in surveillance of nonsocial infections[C] / /Proc of the 4th International Sym2posium on Medical Data Analysis (ISMDA '03) . Berlin: [s. n.] ,2003: 1092117

[10] CHEN Jianxun, CHENG T H, CHAN A L F, eta. An application of classification analysis for skewed class distribution in therapeutic drug monitoring the case of vancomycin [C] / /Proc of Workshop on Medical Information Systems (IDEAS2DH 04) . Beijing: [s.n.] , 2004: 35239