Application of Refined LSA and MD5 Algorithms in Spam Filtering

Jingtao Sun[1, 2]
1. College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China
Email: sun2651@126.com

Qiuyu Zhang[2] and Zhanting Yuan[2]
2. College of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
Email: sun2651@qq.com

Abstract—**The paper proposes a spam filtering method that uses integrated and refined Latent Semantic Analysis (LSA) and Message-Digest Algorithm 5 (MD5) algorithms to address a series of universal problems in spam filtering, including remarkably lowered filtering precision and notably unbalanced filtering efficiency as a result of lack of latent semantic analysis of mail contents. In introducing LSA, its weighting function is improved by integrating fuzzy membership to improve effectiveness of LSA in processing mail contents. On top of this, MD5 algorithm is used to generate "E-mail fingerprint", thus enabling quick matching and realizing highly efficient and accurate processing of mass-mailing spam. The result of the simulation experiment testifies effectiveness of the method.**

Index Terms—**Latent Semantic Analysis, Message-Digest Algorithm 5, Fuzzy Membership, E-mail Fingerprint, Spam Filtering**

I. FOREWORD

E-mail has gradually become an important mean of information exchange in people's daily life. It is extensively used in business, government organs, universities, secondary school and families. [1] Over the past years, spam carrying undesired information has brought endless annoyance to Internet users, network administrators and Internet service providers (ISP). To guard against spam, people have proposed a variety of spam filtering methods. [2]

Latent Semantic Analysis is a computer technology developed to extract information effectively. [3, 4] It can analyze a huge amount of texts using statistical method to extract and quantize latent semantic information of character word in the file, thereby eliminating the effect of variant character word on text authentication and improving the precision. LSA was originally used in text information searching, and now the study of applying LSA in spam filtering has ignited the interest of some scholars. [5] Yet LSA tends to inherit the Vector Space Model (VSM) and its own characteristics are neglected, which leads to a lack of implantation of apriori or global document information. To address the problem, the paper proposes a method to improve the original weighing function in combination of fuzz membership and refine the application of LSA in spam filtering. Introduction of refined LSA provides a very effective mean to fight spam that contains hidden information. Yet as a matter of fact, most spam spreads in large LANs using mass-mailing mechanism, which are characterized by: the main body or sender address of the spam changes frequently and dynamically, while its text and attachment remain almost unchanged. Therefore, when using the refined LSA to identify Emails containing hidden information, we need to use for reference similar text inspection methods, generating "E-mail fingerprint" of mass-mailing spam. Then we can perform mail authentication by comparing "E-mail fingerprints". In this paper, we will use MD5 algorithm to get "E-mail fingerprint".

II. ANALYSIS OF KEY TECHNOLOGIES

A. LSA

The fundamental theory of LSA is to map the document that high dimensional VSM represents to the low dimensional latent semantic space. The word-document matrix of the original text collection can be approximately represented by the dimension reduction matrix containing only K orthogonal factors, which is generated by performing Singular Value Decomposition (SVD) [6, 7] on the word-document matrix of the text collection.

First, a document library needs to be structured that can be represented by an m×n word-document matrix X=$[x_{ij}]$, in which x_{ij} is a nonnegative value that stands for the frequency the number i word appears in the number j document. As there are a huge number of words and documents and the number of words in a single document is limited, X is usually a sparse matrix. To make the information that the word-document matrix X carries satisfy practical requirement, weighting processing should be conducted on x_{ij} to get a weighted m×n word-document matrix $X' = [x_{ij}']$. In the process, two contributors should be taken into consideration, i.e.

the local weight value L (i,j) and global weight value C(i), representing the significance of number i word in number j document and in the entire document library respectively. Now we get the weighted m×n word-document matrix $X^{'}=[x_{ij}^{'}]$.

$$x_{ij}^{'} = x_{ij} \times L(i,j) \times C(i). \quad (1)$$

When performing SVD on $X^{'}$ (assume m>n, rank(X)=r, K exists and K<<min(m,n), in the F-norm significance, $X^{'}$'s K-rank approximate matrix of $\boldsymbol{X}_K^{'}$ is:

$$\boldsymbol{X}^{'} \approx \boldsymbol{X}_K^{'} = \boldsymbol{U}_K \boldsymbol{\Re}_K \boldsymbol{V}_K^{\mathrm{T}}. \quad (2)$$

In the formula, $\boldsymbol{U}_K = (u_1, u_2, \ldots, u_K)$ is an orthogonal matrix, and $u_1, u_2, \ldots, u_K$ are left singular vectors of $\boldsymbol{X}_K^{'}$ as well as the eigenvectors of $\boldsymbol{X}_K^{'}\boldsymbol{X}_K^{'T}$. $\boldsymbol{\Re}_K = diag(r_1, r_2, \ldots, r_K)$ is a diagonal matrix. $r_1, r_2, \ldots, r_K$ is the singular values of $\boldsymbol{X}_K^{'}$, and $r_1 \geq r_2 \geq \ldots \geq r_K > 0$. $\boldsymbol{V}_K = (v_1, v_2, \ldots, v_K)$ is an orthogonal matrix, in which $v_1, v_2, \ldots, v_K$ are singular vectors of $\boldsymbol{X}_K^{'}$ as well as the eigenvectors of $\boldsymbol{X}_K^{'}\boldsymbol{X}_K^{'T}$.

B. Calculation method for fuzzy membership-based latent semantic weight

(1) Weight calculation method

At present, LSA generally adopts traditional weight calculation method, which divides weight into two parts: one is called local weight (marked as $L(i,j)$). It emphasizes the significance of a certain word in a certain document. The simplest form of definition is to use word frequency as its quantified expression. The other part is called global word weight (marked as $C(i)$). It emphasizes the significance of a certain word in the entire text collection and represents the significance of the role of a certain word in differentiating documents. Global word weight is usually calculated using statistical method. An important task of LSA is to extract semantic information, i.e. latent semantic relationship between words. Latent relationship between two words with greater weights is more likely to be considered as important semantic relationship by LSA and thus retained. So the weighting function $M(i,j)$ is expressed in the following formula:

$$M(i,j) = L(i,j) \times C(i). \quad (3)$$

Nevertheless, the method only considers local weight and global word weight and overlooks the contribution of documents in differentiating words, which results in lowered accuracy in text identification. Therefore, it is an important direction of study on improving accuracy of LSA in text identification to define a more effective weight calculation method.

(2)Expansion of LSA weight calculation method

Given the above-mentioned problems and in combination of documents that provide more information to words, influence on basis vector of latent semantic space should be amplified; while the influence of documents that provide less information to words on basis vector of latent semantic space should be diminished. This information induction thought will be introduced in the definition of global weight of documents to expand calculation of LSA weight.

Semantics of a document is differentiated by the semantic of words it contains, while semantics of words is closely related to the theme of document, in which they appear. Fuzzy border exists between different themes. Different preference and interests lead to varied focuses of different themes, i.e. certain priori knowledge exists for different themes. Therefore, the author introduces the priori knowledge in the definition of global weight of document, thus further amplifying or diminishing the influence on basis vector of latent semantic space.

Global word weight $C(i)$ is an induction of horizontal information of matrix X and global document weight we defined $S(j)$ is an induction of vertical information of matrix X. Therefore, the weigh calculation formula can be expanded to:

$$M^{*}(i,j) = L(i,j) \times C(i) \times S(j). \quad (4)$$

If the effect of global document weight in the entire weight definition is not considered, take $S(j) = 1$.

We define global document weight and obtain new weight expression in the following way:

Define P themes $H_1, H_2, \ldots H_p$ in text collection U, each with certain priori weight Q_{H_i} $(i = 1, 2, \ldots, p)$. Each document in the collection can be expressed as a m -dimensional vector, which is marked as $u = (u_1, u_2, \ldots, u_m)$ and called characteristics index vector. It is widely believed that more information a certain document provides to the collection, the greater its role in text identification and the higher its global weight is. So we can define the global weight of a document using the product of priori weight Q_{H_i} and the membership of u to H_i $H_i(u)$. It is expressed as:

$$S(j) = H_i(u) \times Q_{H_i}. \quad (5)$$

Here, $H_i(u)$ is constructed as below:

a) Select k_i samples from characteristic index vectors of class theme H_i, define $h_{ij} = (h_{ij_1}, h_{ij_2}, \ldots, h_{ij_m})$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, k_i$. In the formula, h_{ij} indicates the characteristic index vector of the number j sample in H_i. h_{ij_k} Indicates the measured data of the number k characteristic index of the number

j sample in H_i, $k = 1,2,\ldots,m$. According to practical problems involved in the article, we use word frequency as the measured data of characteristic index to perform calculation.

b) Calculate mean sample of k_i characteristic index vector h_{ij} $i = 1,2,\ldots,p$; $j = 1,2,\ldots,k_i$ selected from theme H_i using formula $h_i = (h_{i1}, h_{i2}, \ldots, h_{im})$, in which $h_{is} = \frac{1}{k_i}\sum_{j=1}^{k_i} h_{ij_k}$, $k = 1,2,\ldots,m$.

c) Construct membership function of class theme H_i . Assume that $u = (u_1, u_2, \ldots, u_m) \in U$, calculate distance between u and h_i $d(u, h_i)$, and make $D = \max\{d_1(u,h_1), d_1(u,h_2), \ldots, d_1(u,h_p)\}$, membership function of H_i is calculated in the formula:

$$H_i(u) = 1 - \frac{d_i(u,h_i)}{D},\ i = 1,2,\ldots,p \qquad (6)$$

A new LSA weight calculation method is obtained through the above steps. Using the method to perform weighted conversion of word-document matrix, the matrix obtained can still been shrunk through Singular Value Decomposition and K-rank approximate matrix. Therefore, the expanded LSA weight calculation method is consistent with the traditional one in terms of format.

Compared with traditional LSA weight calculation method, the expanded method not only overcomes data sensitivity in LSA, but also implants priori information in basis vector of the latent semantic space to avoid lack of flexibility of SVD.

C. MD5 algorithm

MD5 algorithm (Message-Digest Algorithm 5) was developed in early 1990s by R. Rivest of MIT Laboratory for Computer Science and RSA Data Security Inc. It is simply a compression function and does not have any parameter. Message m, no matter how long, will turn into a 128-bit sequence after MD5 algorithm. Typical application of MD5 algorithm is to generate the Message-Digest of a message to protect it from being falsified. [8, 9]

The process is MD5 algorithm is:

Step 1: filling position

In SHA algorithm, positions of message m will be filled so that the reminder when dividing final number of bits of message m by 512 is 448. That is to say, the number of bits filled makes the total number of bits 64 bits less than a multiple of 512. To fill the positions, add a 1 first, and then add 0 until the above requirement is satisfied.

Step 2: expanding length

After the position is filled, affix a 64-bit number to the end, which represents the length of the original message m. The length of the outcome message would be a multiple of 512.

Step 3: initializing variables

There are four variables - A, B, C and D, all are 32 bits in length. After initialization, A = 0X01234567, B = 0X89abcdef, C = 0Xfedcba98 and D = 0X76543210.

Step 4: processing information

First, define four auxiliary functions: F(X,Y,Z)=(X&Y)|((~X)&Z), G(X,Y,Z)=(X&Z)|(Y&(~Z)), H(X,Y,Z)=X^Y^Z, I(X,Y,Z)=Y^(X|(~Z)). X, Y and Z in the functions are all 32 bits in length. If the corresponding bits of X, Y and Z are independent and even, all bits in the result will also be independent and even.

Input of each round of data processing is a 512-bit variable and the output is a 128-bit variable (ABCD). Change the variable and construct table T [1…64]e in each round using one quarter of sine function, as shown below:

$$\mathrm{T[i]} = \lfloor 2^{32} |\sin(i)| \rfloor\ ,\quad i = 1,2,3...,64$$

In the formula, i is the radian, radian is a 32-bit data used as random number.

Step 5: output

ABCD obtained after the above steps are the output results. They are stored in an uninterrupted sequence and occupy a total of 16 bytes and 128 bits. A is the lowest bit and E is the highest bit. The 16 bytes output is in a hexadecimal sequence.

III. SPAM FILTERING METHOD INTEGRATING REFINED LSA AND MD5 ALGORITHMS

In the paper, refined LSA and MD5 algorithms are integrated in the spam filtering method to efficiently and accurately filter spam. Introducing semantic analysis and "E-mail fingerprint" will make spam filtering more flexible and adaptable. The main idea includes the following aspects:

(1) Analysis of variant character word

In spam filtering, refined LSA technology is adopted to reduce dimension and get approximate matrix $\boldsymbol{X}_K^{'} = (doc_1^{'}, doc_2^{'}, \ldots, doc_n^{'}) = (word_1^{'}, word_2^{'}, \ldots, word_m^{'})^T$. We can compare the similarity between any two words using such common methods as: included angle cosine value, dot product of vectors and correlation coefficient, etc. According to formula (2), in K -dimensional vector space, a word can be represented as $word_j^{'} = (u_{j1}r_1, u_{j2}r_2, \ldots, u_{jK}r_K)^T$, in which u_{jh} stands for number h component of vector u_j . To get similarity of vector $word_j^{'}$ and $word_l^{'}$, we calculate correlation coefficient ρ using the following formula:

$$\rho = \frac{\sum(word_j^{'} - \overline{word_j^{'}})(word_l^{'} - \overline{word_l^{'}})}{\sqrt{\sum(word_j^{'} - \overline{word_j^{'}})^2 \sum(word_l^{'} - \overline{word_l^{'}})^2}} \qquad (7)$$

If a queried word w is not contained in X , we need to compare the similarity of the word w with any word in X . For this purpose, the word w should be projected into space X . As there is not a line in $\boldsymbol{U}_K$ that represents the queried word, we need to use SVD Fold-in [10] to add such a line:

$$U_w = w * V_K * \Re_K^{-1}. \quad (8)$$

In this wise, once a queried word is folded in, it can be added to the existing word vector collection, thus making it possible to conduct similarity comparison between it and other words in the collection.

(2) Generation of "E-mail fingerprint"

Analysis of spam using the refined LSA technology will help eliminate large amount of "noise" words that are useless to spam filtering, and get the hidden character words that affect precision of spam filtering. To ensure that "E-mail fingerprint" of the inspected E-mails can be acquired quickly, highly efficiently and accurately using MD5 algorithm, and that the text information can be reflected more effectively without damaging the relationship between words, we use the character words acquired using refined LSA as the sampling points in the E-mail text, and introduce sliding window character extraction algorithm [11] to reconstruct words near the sampling points and further expand character selection scope, which enables extracted character words to reflect characteristics of the text more accurately. With the MD5 algorithm, a character code of certain length is generated from the character words acquired using the sliding window character extraction algorithm. These character codes are the "E-mail fingerprints".

(3) E-mail filter

We will store the "E-mail fingerprints" acquired using the above method in a MySQL database. The database mainly stores tab-files table and includes "files" and "characteristic" fields. The "files" field is used store file information and the "characteristic" field is used to store digital fingerprint information generated to avoid file repetition from occurring.

Data is imported into tab-files table in the following steps:

(a) Conduct processing to get mail document M, and calculate mail characteristic of M ID-CTM. (b) Check whether there is identical mail characteristic ID as CTM in the database. (c) Skip the document if there is and go to (a) to process the next mail until all mails are imported in the database. (d) If there is not, save the mail document and corresponding mail characteristic ID in tab-files table and go to (a) to process the next mail until all mails are imported in the database.

Now the focus of spam filtering has shifted to "E-mail fingerprint" comparison. In the experiment, we use the universal quick matching algorithm to compare "E-mail fingerprints".

IV. Experiment and Conclusion

In analysis of the performance of spam filtering, selection of language material library is extremely important. There are now some authoritative standard language material libraries overseas, such as PU1 language material library. Yet in China, an authoritative standard language material library is nowhere to be found. In this case, we choose from extensive sources 1561 spam mails and 721 legitimate mails and conduct the experiment on a computer with PM2.1G CPU and 2G memory.

MIME standard has now become the mainstream standard for E-mails on the Internet, and mail theme and contents are usually in Base64 and QP (Quote-Printable) code. Before mail identification, decoding should be performed in accordance with encoding mode of the theme and contents to remove HTML tags and attachments in the mail and reserve pure text contents in the mail text. This is followed by Chinese word division, filtering of prohibited words, removal of words of extremely high or low frequency as well as other preprocessing of the mails to generate a 7312×2282 word-document matrix X . Then the matrix is weighted by weighting function $M(i, j)$ to get X' . After this, SVD is used to perform base conversion on the VSM space to generate latent semantic space X_K' . During the process, selection of dimension reducing factor K has a direct influence on the efficiency of the latent semantic space model and similarity between X_K' and X' following dimension reduction. If the value of K is too small, useful information will be lost; if the value of K is too large, the calculation volume will increase. Therefore, an optimal K value should be selected in accordance with actual text collection and processing requirement. The article uses contribution rate δ as the criterion to assess the K value selected, i.e. $A = \mathrm{diag}(a_1, a_2, \ , a_n)$, and $a_1 \geqslant a_2 \geqslant \ \geqslant a_t = \ = a_n = 0$, contribution rate δ :

$$\delta = \sum_{i=1}^{k} a_i \Big/ \sum_{i=1}^{t} a_i . \quad (9)$$

The contribution rate δ , proposed in reference of related factor analysis concept, indicates the degree, to which the K-dimensional space represents the entire space. Fig 3 shows that the closer the K value is to the rank of matrix A, the smaller $\| A - A_K \|_F$ is and the closer A_k is to A. Yet as the value of K continues to increase, its influence on δ will decrease or even disappear. Analysis indicates that when the value of K increases to a certain level, nearly all important characteristics of word-document matrix are represented. In this case, further increasing K value will only introduce noise. When K=400, the degree of representation is almost the same as when K=500. Yet when K=400, less time is consumed. So, we choose K=400.

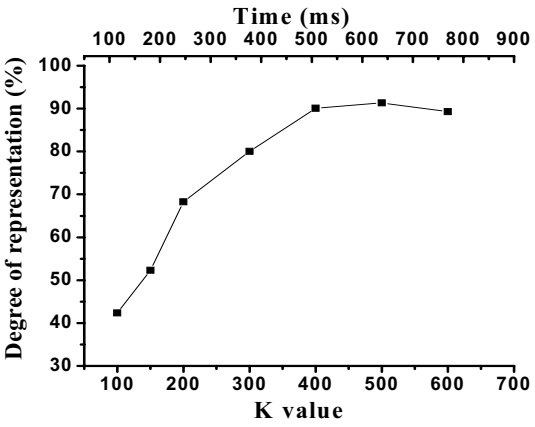

Figure 1. Analysis of K value

When generating "E-mail fingerprint", configuration of size of the sliding window will also affect performance and efficiency of the LSA and MD5 algorithms in spam filtering. As shown in Figure 2, the larger the sliding window is, the better the filtering system performs. This is because when the window becomes larger, more characteristics will be selected and more document characteristics will be represented. In the mean time, both recall rate and precision will improve. Yet the large the window is, the slow the operation speed and the longer the operation time will be. When the window size is 2, an optimal balance will be stricken between performance and operation speed.

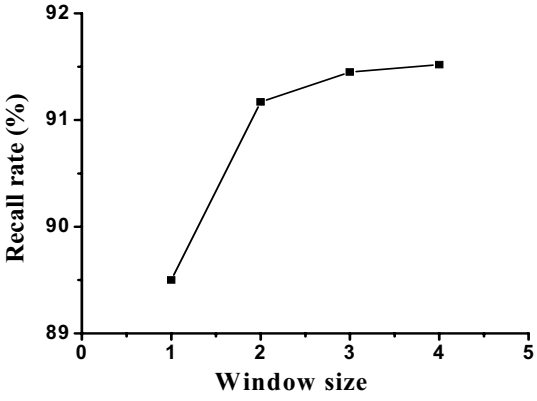

Figure 2. Analysis of window size

The paper analyzes the precision and recall rate of the spam filtering method that combines refined LSA and MD5 algorithms. [12] Precision is the correctness rate in judging spam mails. It reflects the ability of a spam filtering method in identifying spam correctly. The better the precision is, the less legitimate mails will be identified as spam. Recall rate refers to the proportion of spam mail detected. It reflects the ability of a spam filtering method in detecting spam mail. The better the recall rate is, the less undetected spam will be. The precision and recall rate represent two different aspects of spam filtering quality, and the same priority should be given to both of them. Therefore, there is a new evaluation index, namely F1 test value. Formulas of the above evaluation indexes are as below:

Recall rate: $R = \frac{N_A}{N_S} \times 100\%$

Precision: $P = \frac{N_A}{N_A + N_B} \times 100\%$

F1 value: $F = \frac{2RP}{R+P} \times 100\%$

In the formula, N_A is the number of spam correctly identified; N_S is the actual number of spam mails; N_B is the number of legitimate mails identified as spam by mistake. To further verify effectiveness of the method, we conduct another experiment to compare the spam filtering method that combines refined LSA and MD5 algorithms and SVM and Naïve Bayes mail filtering method. Table 1 shows the comparison result in the experiment.

TABLE I. RESULT OF COMPARISON BETWEEN REFINED LSA AND MD5 ALGORITHMS, SVM ALGORITHMS AND NAÏVE BAYES ALGORITHMS

	Recall rate	Precision	F1 value
SVM	91.45%	89.71%	90.57%
Naïve Bayes	76.37%	94.21%	84.36%
Refined LSA and MD5	95.17%	93.83%	94.5%

In the experiment, the method uses refined LSA and MD5 algorithms before spam filtering, which results in a little longer filtering time. Yet the final experiment result shows that although the precision of the method is a little lower than that of the Naïve Bayes method, its recall rate is significantly better than the other two methods and its F1 value is also higher. This shows that performance and efficiency of the method is better than the SVM and Naïve Bayes methods.

VI. CONCLUSION

The paper proposes a spam filtering method that combines refined LSA and MD5 algorithms. The method, which is intended to address certain defects of the traditional LSA, uses for reference the definition of fuzz membership and refines definition method of LSA weight. It also integrates MD5 algorithm, sliding window algorithm and database technology and solves the problem of traditional spam filtering method – low efficiency and inaccuracy in filtering mass-mailing spam. The simulation experiment shows that the method has better performance than that of SVM and Naïve Bayes spam filtering methods, thus inventing a new approach for spam filtering.

REFERENCES

[1] B. Hoanca, "How good are our weapons in the spam wars?" Technology and Society Magazine, vol. 25(1), 2006, pp. 22-30.

[2] Xiao Jie, Huang Zhanglong, "Analysis of E-mail Security Technology," Journal of Xiaogan University, vol. S1, 2007, pp. 47-50.

[3] LIN Hongfei, YAO Tianshun, “Text Browsing Based on Latent Semantic Indexing,” Journal of Chinese Information Processing, 2000, pp. 241-245.
[4] J. R. Bellegarda, “Exploiting latent semantic information in statistical language odeling,” Proceedings of the IEEE, vol. 88(8), 2000, pp. 1279-1296.
[5] Zhang Qiuyu, Sun Jingtao, “Technology of Spam Filtering Based on Latent Semantic Analysis,” The Systemics and Informatics World Network, vol. 04, 2007, pp. 1265-1270.
[6] D. I. Martin, J. C. Martin, M. W. Berry, “Out-of-core SVD performance for document indexing,” Applied Numerical Mathematics, vol. 57, 2007, pp. 1230–1239.
[7] GAI Jie, WANG Yi, WU Gangshan, “The Theory and Application of Latent Semantic Analysis,” Application Research of Computers, vol. 03, 2004, pp. 9-12.
[8] K. Jarvinen, M. Tommiska, J. Skytta, Hardware Implementation Analysis of the MD5 Hash Algorithm, System Sciences, 2005, pp. 298a- 298a.
[9] B. Preneel, P. C. Van Oorschot, “On the security of iterated message authentication codes,” IEEE Transactions on Information Theory, vol. 01, 1999, pp. 213-216.
[10] Guo Junfang, http://140.122.185.120/PastCourses/2003F-InformationRetrievalandExtraction/Present_2003F/2003F_LSI_郭榮芳.pdf.
[11] Yi Fasheng, Wang Yan, Xia Mengqin, Zeng Jiazhi, “An Improved Algorithm of Slide Window Protocol Based on GACK,” Computer Engineering, vol. 14, 2006, pp. 92-94.
[12] Zhu Qiaoming, Li Peifeng, et al., Chinese Information Processing Technique, Tsinghua University Press, 2005.

Jingtao Sun Doctor student. He was born in Daqing Heilongjiang province in 1981. He has published many academic papers in domestic core magazine and international conference. His research interests include: information security, Chinese text classification, Anti-Spam etc.

Qiuyu Zhang Associate professor and master tutor. Vice dean of School of computer and communication in Lanzhou University of Technology, director of software engineering center, vice dean of Gansu manufacturing information engineering research center, director of “software engineering” characteristic research direction and academic group of Lanzhou University of Technology. His research interests include: image processing and pattern recognition, multimedia information processing, information security, software engineering etc.

Zhanting Yuan Professor and doctor tutor in the School of Computer and Communication Engineering, Mr Yuan received his MASc in June 1989 from Artificial Intelligent and Robot Research Center of Xi’an Jiaotong University. Now Professor Yuan is the Science Leader of computer science and technology, director of Gansu manufacturing information engineering research center, administrative director of Chinese electrical higher education, and the first batch of Chinese new century “Bai qian wan” person with ability. The research interests of professor Yuan include: image processing and pattern recognition, computer vision, Software engineering, information security etc.