User Page Interest Prediction Based on Improved Collaborative Filtering Algorithm: A Postprint

Authors: Song Bodong, Zhang Lichen

Date: 2018-08-13T00:00:00+00:00

Abstract

Discovering user interest levels from massive datasets constitutes an important approach for implementing targeted information delivery in the e-commerce domain. Based on the sparsity characteristics of big data, this paper introduces the singular value decomposition method into collaborative filtering algorithms to compute and validate webpage interest levels for internet site users, proposing a user webpage interest prediction algorithm based on an improved collaborative filtering algorithm. This algorithm can identify user webpage interest levels and their influencing factors by extracting "false ratings" present in explicit user rating data from web log files. MATLAB simulation results demonstrate that the proposed user webpage interest measurement method based on the improved collaborative filtering algorithm can effectively overcome the sparsity of massive data, showing significant improvements in both prediction accuracy and measurement speed.

Full Text

Preamble

Prediction of User Page Interest Based on Improved Collaborative Filtering Algorithm

Song Bodong, Zhang Lichen
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Discovering user interest in massive datasets is a crucial method for implementing targeted information push in e-commerce. Considering the sparsity characteristics of big data, this paper introduces singular value decomposition into collaborative filtering algorithms to calculate and verify page interest degree for internet site users, proposing a user page interest prediction algorithm based on improved collaborative filtering. The algorithm can extract "false ratings" from explicit user rating data in web log files to discover user page interest and its influencing factors. MATLAB simulation results demonstrate that the proposed measurement method based on improved collaborative filtering effectively overcomes the sparsity of massive data, showing significant improvements in both prediction accuracy and computational speed.

Keywords: big data; singular value decomposition; page interest; collaborative filtering algorithm; data sparsity

0 Introduction

Discovering user interest degree in massive data is an important means for providing personalized recommendations. Currently, user interest degree (ID) acquisition mainly relies on two approaches: explicit feedback and implicit feedback. Both methods collect and compute user page visits, access time, and page operations to reflect user interest. For instance, Xia Yiguo et al. [?] established a neural network model to analyze and calculate user interest by fitting user visit frequency and duration. Li Feng et al. [?] developed an adaptive user interest model based on implicit feedback, calculating user interest through analysis of implicit feedback behaviors. Xing Ling et al. [?] constructed a user interest decay factor measurement model using an improved K-means algorithm. Yin Chunhui et al. [?] proposed a user interest acquisition algorithm based on browsing behavior, interpreting user interest through analysis of page viewing time, speed, and scrollbar operations. Wang Weiwei et al. [?] introduced an interest model based on user behavior patterns and browsing content. Yu Miao et al. [?] proposed a robust constrained information embedding method to construct relationship matrices, reducing matrix sparsity and improving algorithm accuracy. Zhang Yanmei et al. [?] developed a social tagging recommendation algorithm adapted to user interest changes. Chen Chen [?] constructed a personalized interest prediction and discovery model for library small data readers. Li Changbing et al. [?] studied Web access user association rule mining, while Zhang Zhaoheng et al. [?] proposed a collaborative filtering recommendation algorithm based on social network information.

Although these methods investigate user interest collection and measurement from different perspectives, most lack consideration for big data sparsity, resulting in significant discrepancies from actual conditions. Therefore, building upon previous research, this paper considers the impact of massive data sparsity on user interest measurement and proposes a website user interest prediction model based on improved collaborative filtering. This model extracts explicit user rating data from web log files and predicts user page interest degree and influencing factors by calculating differences between any two items in rows and columns.

1 Network User Page Interest Calculation and Matrix Representation

1.1 Page Interest Calculation

User page interest degree represents the level of user interest in web pages. By analyzing user access page size, visit frequency, browsing time, scrollbar usage, page saving actions, and page byte transmission, we can reveal user page interest. Let $Interest_{i,j}$ denote user i' s interest in page j, n_{ij} represent the number of visits by user i to page j, $t_{i,j}$ denote the average browsing time of user i on page j, and $sb_{,j}$ represent page j' s size. The influence factor of page size $sb_{,j}$ on user browsing time is defined as $T_{i,j} = \frac{t_{i,j}}{sb_{,j}}$, while the influence factor for visit frequency is $N_{i,j} = n_{ij}$. Since it is difficult to determine which factor is more important between $N_{i,j}$ and n_{ij}, this paper employs principal component analysis to obtain weights for these indicators. The page interest calculation formula and standardization formulas are as follows:

$$Interest_{i,j} = \alpha \times T_{i,j} + b \times N_{i,j} = \alpha \times \frac{t_{i,j}}{sb_{,j}} + b \times n_{ij}$$

$$T_{i,j} = \frac{T_{i,j}}{\max(T_{i,j})}$$

$$N_{i,j} = \frac{N_{i,j}}{\max(N_{i,j})}$$

1.2 User Page Interest Matrix Representation

After preprocessing user access page size, visit frequency, browsing time, scrollbar usage, page saving actions, and page byte transmission, the data is converted into a user page interest VSM analysis matrix. This paper treats each returned user access action record as one document and performs identification and segmentation to form different user page interest VSM analysis matrices, as shown in Equation (4):

$$M_0 = \{i \times j\}_{m \times n} \quad (4)$$

where M_0 represents each user' s interest in different pages, i denotes user i, j represents accessed page count, m indicates the number of users, and n indicates the number of page items.

2 Improved Collaborative Filtering Algorithm

2.1 Algorithm Improvement Rationale

The Slope One algorithm is based on the principle that average values can replace rating differences between two unknown individuals. For example, if users A and B' s average rating deviation for items X and Y is $((3-4)+(2-4))/2=-1.5$, this means ratings for item X are generally 1.5 points higher than for item Y. Using this approach, Slope One can predict user C' s rating for item X. Given a user rating matrix, we can define a training set x and calculate the average deviation $Dev_{j,i}$ between items i and j using Equation (5):

$$Dev_{j,i}=\sum_{u\in S_{j,i}(x)}\frac{u_j-u_i}{card(s_{j,i}(x))}$$

where $S_{j,i}(x)$ is the set of users who rated both items i and j, u_i is user u' s rating for item i, u_j is user u' s rating for item j, and $card(s_{j,i}(x))$ is the number of users who rated both items.

By calculating $Dev_{j,i}$ through Equation (5), we obtain a symmetric matrix with real-time rating update capabilities. When predicting user u' s rating for item j, we use the set R_i of items in $S(u)$ (the set of items rated by user u in training set x) that have average deviation $Dev_{j,i}$ with item j to calculate user interest. The specific formulas are shown in Equations (6) and (7):

$$P(u)_j=\bar{U}+card(R_j)\times dev_{j,i}$$

$$R_j=\{i|i\in S(u),i\neq card(S_{j,i}(x))>0\}$$

where $\bar{U}$ is the average rating of items rated by user u in the training set, and $card(R_j)$ represents the average deviation $dev_{j,i}$ between item j and items rated by user u.

Each website contains numerous pages, and users with different interests typically only access pages they find interesting. This results in high sparsity in the page interest matrix, making it difficult to find similar users. This also causes the Slope One algorithm to fail when predicting interest for pages with few user interactions. To address this issue, we must identify a dense matrix with the most similar structure to the user page interest matrix as a foundation, then apply the Slope One collaborative filtering method to filter out matrix sparsity and derive user page interest.

2.2 Algorithm Improvement

Singular Value Decomposition (SVD) is a widely used machine learning algorithm. Li Weijiang et al. [?] proposed a social recommendation algorithm integrating trust propagation and singular value decomposition, which uses feature decomposition to process rating matrices. This effectively avoids data fitting issues, overcomes sparsity and inaccuracy problems in similarity calculation for zero-rating users, and demonstrates good recommendation performance.

This paper employs the SVD algorithm to first construct an $m \times n$ matrix R from collected data and decompose it into three matrices: $R = T_0 S_0 D_0$, where $S_0 = diag(\delta_1, \delta_2, \cdots, \delta_r)$. Assuming $\delta_1 \geq \delta_2, \cdots, \delta_r \geq 0$, then T_0 and D_0 form $m \times r$ and $r \times r$ orthogonal matrices respectively ($T_0^T T_0 = I, D_0 D_0^T = I$). The variable r is the rank of matrix R ($r \leq \min(m, n)$), and S_0 forms an $r \times r$ diagonal matrix. By setting all r values greater than 0 and arranging them in descending order, we can generate a simplified approximate matrix.

Typically, matrix $R = T_0 S_0 D_0$ requires T_0, D_0, and S_0 to be full rank, which has computational complexity and operational overhead. By applying SVD to S_0 and retaining only the k largest singular values while replacing others with 0, we obtain a matrix with k singular values ($k \times r$). Further removing rows and columns with value 0 from S_0 yields diagonal matrix S. Similarly, reducing matrices T_0 and D_0 to T and D through dimensionality reduction gives us matrix $R_k = TSD$, where $R_k \approx R$ [**?**, **?**, **?**].

We then apply collaborative filtering to normalize user interest matrix R and obtain matrix R'. Using SVD on R' yields output matrices T, D, and a user matrix $TS^{1/2}$ ($m \times k$) describing user relationships in k-dimensional space, along with an interest degree matrix $S^{1/2}D$ ($n \times k$). By introducing collaborative filtering to compute $TS^{1/2}$ and $S^{1/2}D$, we can calculate user page interest degree.

2.3 Experimental Validation

2.3.1 Interest Degree and Sparsity Calculation This paper analyzes Taobao (https://www.taobao.com) web log data from June 2017. After data cleaning and processing using the aforementioned method, the page interest formula is shown in Equation (9):

$$Interest_{i,j} = 0.7071 \times \frac{t_{i,j}}{sb_{,j}} + 0.8561 \times n_{i,j}$$

This yields user page interest matrix M_0 with 105 users and 168 pages. A partial user page interest matrix is shown in Table 2 .

The sparsity of matrix M_0 is calculated as $1 - \frac{105\times168}{105\times168} = 0.9468$, indicating extreme sparsity but within acceptable network data characteristics. After standardization, matrix M_0 becomes M_{10}. As shown in Figure 1 [Figure 1: see original paper], as matrix sparsity increases, the Mean Absolute Error (MAE) value also increases. MAE is defined as follows: let $r(u,i)$ be the predicted rating for an item and $\hat{r}(u,i)$ be the user' s actual rating. MAE is the average of absolute differences across N such rating pairs:

$$MAE = \sum \frac{|r(u,i) - \hat{r}(u,i)|}{N}$$

A smaller MAE value indicates more accurate user rating prediction and higher prediction quality.

1) **Recommendation Generation**: The standardized data matrix serves as the basis for SVD implementation ($T_0^T T_0 = I, D_0 D_0^T = I$). Vector space methods then calculate similarity in $TS^{1/2}$ to obtain nearest neighbors and corresponding recommendation sets for user page interest.
2) **Interest Degree Calculation**: Multiplying matrices $TS^{1/2}$ and $S^{1/2}D$ yields item inner products, enabling calculation of user u' s interest in item i:

$$I(u,i) = \bar{u} + TS^{1/2}(u) \times S^{1/2}D'(i)$$

where $\bar{u}$ is user u' s average rating across rated items.

Through Equation (8), we obtain a user rating matrix without missing values, even for new users and items, making the user page interest matrix 100% dense and resolving data sparsity issues. User interest in other site pages can then be calculated from this matrix.

This paper performs SVD on the singular value matrix M_1 such that $M_1 = T \times S \times D$. Matrix M_1 is 105×168, yielding T as a 105×105 matrix, S as 105×168, and D as 168×168.

2.3.2 5-Fold Cross-Validation 5-fold cross-validation is a machine learning technique used to avoid overfitting by training and validating simultaneously. It randomly divides training data into k folds, training k times to find the optimal solution. To compare the effectiveness of the improved algorithm, this study employs 5-fold cross-validation to compare the traditional Slope One algorithm with the improved Slope One collaborative filtering algorithm (Slope One-after-SVD). The results are shown in Figure 2 [Figure 2: see original paper].

Calculating the average yields:

$$MAE_{SlopeOne} = 0.551, \quad MAE_{SlopeOneAfterSVD} = 0.193 \quad (11)$$

Both the average calculation and experimental results demonstrate that under unavoidable data sparsity, the proposed improved collaborative filtering algorithm (Slope One-After-SVD) significantly outperforms the traditional Slope One algorithm in accuracy. Analysis of training set 3 and validation set 3 reveals distinctly opposite characteristics in the simulation curves, indicating smoother handling of data sparsity and effectively avoiding validation failure due to unknown or underestimated sparsity.

3 Conclusion

This paper introduces singular value decomposition into collaborative filtering algorithms to calculate and verify internet site users' page interest degree based on big data sparsity characteristics. Experimental results show that under unavoidable data sparsity, the proposed improved collaborative filtering algorithm substantially improves prediction accuracy across different training and validation sets. However, since the experiments used only a specific small dataset, the improved algorithm' s applicability requires further validation across different datasets and data volumes.

References

[1] Xia Yiguo, Liu Youhua. An improved method of user interest calculation and user interest correction [J]. Computer Applications and Software, 2014, 34(1): 46-55.

[2] Li Feng, Pei Jun, You Zhiyang. Adaptive user interest model based on implicit feedback [J]. Computer Engineering and Application, 2008, 44(9): 76-79.

[3] Xing Ling, Song Zhanghao, Ma Qiang. User interest model based on mixed behavior interest [J]. Application Research of Computers, 2016, 33(3): 661-664, 668.

[4] Yin Chunhui, Deng Wei. User interest acquisition based on user browsing behavior analysis [J]. Computer Technology and Development, 2008, 18(5): 37-39.

[5] Wang Weiwei, Xia Xiufeng, Li Xiaoming. An interest model based on user behavior [J]. Computer Engineering and Application, 2012, 48(8): 148-151.

[6] Yu Miao, Yang Wu, Wang Wei, et al. For micro-blog, multi entity sparse relation data joint clustering [J]. Communication Journal, 2016, 37(1): 151-159.

[7] Zhang Yanmei, Wang Lu. A social tagging recommendation algorithm adapted to user interest changes [J]. Computer Engineering, 2014, 40(11): 318-321.

[8] Chen Chen. Library small data reader personalized interest prediction and discovery model construction [J]. Library Forum, 2017, 37(5): 98-105.

[9] Li Changbing, Ling Yongliang, Wang Erjing. Web access user association rule mining based on interest degree [J]. Computer Engineering and Design, 2017, 38(4): 852-856.

[10] Zhang Zhaoheng, He Xiaowei, Chen Yongbing. Collaborative filtering recommendation algorithm based on social network information [J]. Computer Technology and Development, 2017, 27(12): 28-34.

[11] Lemire D, Maclachlan A. Slope One predictors for online rating-based collaborative filtering [J/OL]. (2008-09-15) https://arxiv.org/pdf/cs/0702144.pdf.

[12] Huang Yichun. Slope One recommends algorithm to improve [J]. Modern Computer, 2017, 34(35): 24-27.

[13] Li Weijiang, Qi Jing, Yu Zhengtao, et al. A social recommendation algorithm integrating trust propagation and singular value decomposition [J]. Computer Engineering, 2017, 43(8): 236-242.

Note: Figure translations are in progress. See original paper for figures.