A Survey of Internet Data Mining Technologies Based on Machine Learning

Lin Du[1, a], Yehong Han[2, b]

[1] School of Information Science and Engineering, University of Qilu Normal, Jinan 250014, China;

School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China;

[2] School of Information Science and Engineering, University of Qilu Normal, Jinan 250014, China.

[a]dul1028@163.com, [b]sdzzhyh@163.com

Keywords: Data mining, Machine learning, Rank learning, Topic model learning.

Abstract. As a knowledge discovery in internet, data mining technology is the cross of machine learning and databases technology. Data mining can provide massive amounts of data analysis using machine learning techniques and manage massive amounts of data by the use of database technology. The goal of machine learning is to simulate and realize human learning behavior in order to optimize a performance criterion using past experience and example data. Two kinds of internet data mining technologies which include rank learning and topic model learning are introduced in this paper. As a supervised learning sorting method, the task of rank learning is sorting the retrieved pages for a given query. As a Bayesian model, topic model learning is designed to automatically extract all possible topics and each topic page from the index page, so that the web page and query statement can match on the topic. The latest researches about above areas are reviewed.

Introduction

With the rapid development of computer technology and the great raising of the ability to store data, large amounts of data have accumulated in all of the social areas. To analyze the data and explore useful information become a common requirement. The goal of data mining is to find useful knowledge from massive data. Data mining technology is the cross of machine learning and databases technology. Data mining can provide massive amounts of data analysis using machine learning techniques and manage massive amounts of data by the use of database technology. The goal of machine learning is to simulate and realize human learning behavior in order to acquire new skills, reorganize existing knowledge and continuously improve their performance. Machine learning is the core of artificial intelligence and the fundamental way to make computers be intelligent. The frame of internet data mining is show below.

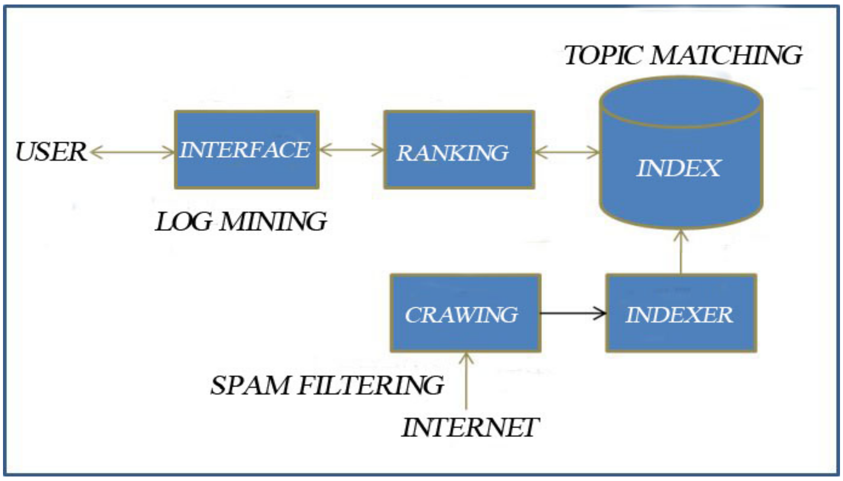


Fig. 1 The frame of web data mining

Internet data mining can be seen as knowledge discovery in internet. The study of machine learning algorithms is programming computers to optimize a performance criterion using past experience and example data .Web data mining technologies based machine learning include learning to rank, topic model learning, matching learning, page importance learning and so on. The rest of this paper reviewed two kinds of mining technologies which include rank learning and topic model learning.

Rank Learning

Rank learning is a supervised learning sorting method. For a given query, the task of rank learning is to sort the retrieved pages. The rank is achieved by four steps: manually labeling training data, document feature extraction, classification learning function and using machine learning models in the actual search system. Machine learning rank principle is shown below.

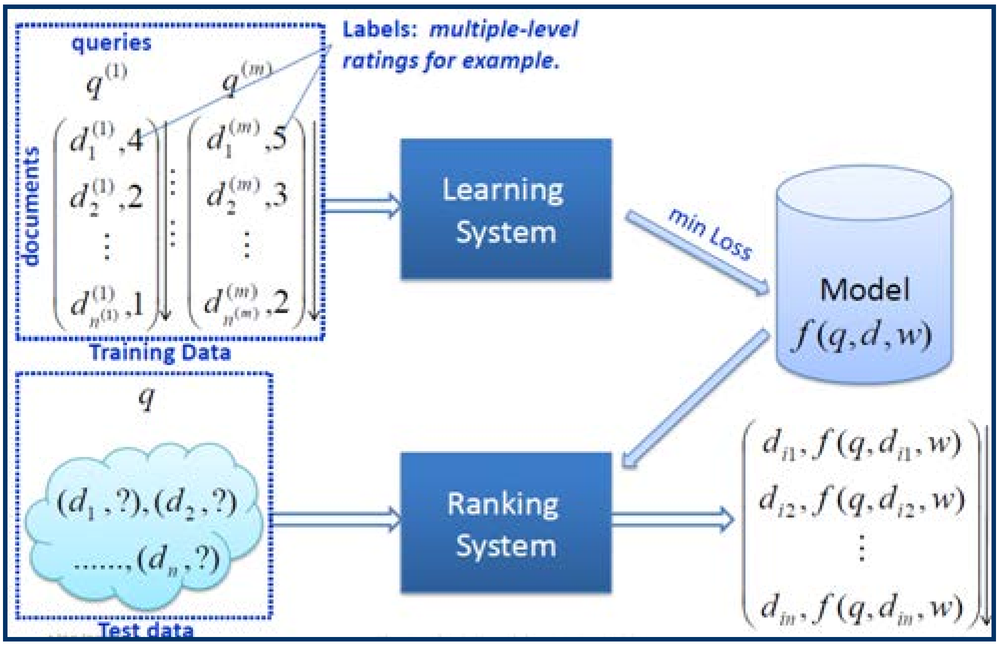
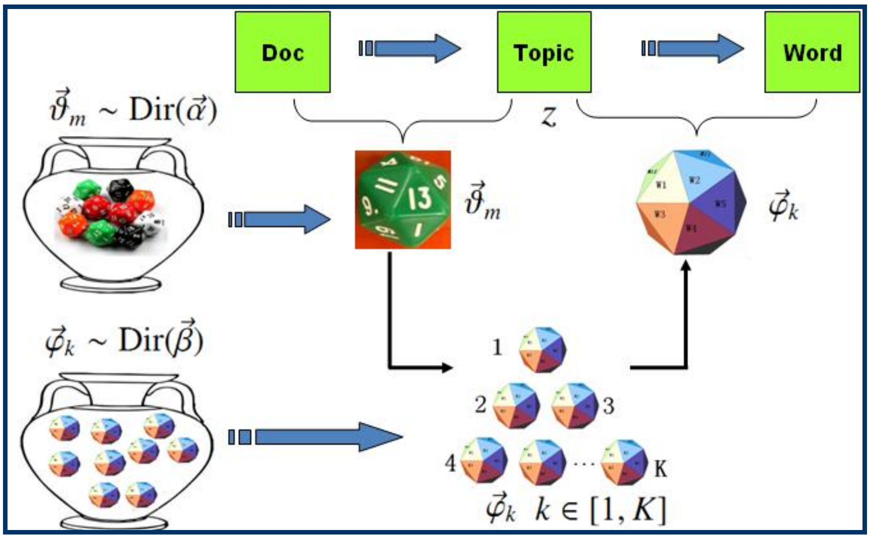


Fig. 2 Rank principle

The learning process for the labeling training set is to select rank method, determine the loss function and obtain the sort of model by minimize the loss of function of the target parameters. Prediction result is inputted in rank model in prediction process. The relevant scores are acquired. The final order of predicted results is achieved by the rank of above scores. Rank creation is the core problem in rank learning. It is usually seen as a supervised learning task. Hang Li [1] gives specific explanations on rank learning creation and rank learning aggregation. Many methods have been proposed for ranking creation. The methods can be categorized according to the loss functions they employ and the techniques which employ, such as the SVM based, Boosting based, and Neural Network based approaches. What's more, the author introduces several rank learning methods which include Prank, OCSVM, McRank, Ranking SVM, IRSVM, GBRank, RankNet, AdaRank, SVM MAP, SoftRank, LambdaRank, LambdaMART and so on. Weiwei Zong et al.[2] propose pointwise extreme learning machine and pairwise extreme learning machine to learn relevance rank learning problems. What's more, extreme learning machine type of linear random node is proposed together with kernel version of extreme learning machine to be linear. In the paper [3], a novel rank learning model which simultaneously utilizes visual features and click features is proposed. Based on large margin structured output learning, visual consistency is integrated with click features through a hyper graph regularized term. Cross-modal ranking is a corer topic that is imperative to many applications involving multimodal data. So as to boost cross-media retrieval, to discover a joint representation for multimodal data and learning a ranking function is important. In the paper [4] a novel approach is proposed to find latent joint representation of pairs of multimodal data via a conditional random field and structural learning in a leastwise ranking manner. All of the correlations among multimodal data are captured in terms of their sharing hidden variables. Entity ranking refers to retrieving and ranking related objects and entities from different structured sources in various scenarios. Based on machine learned ranking models using an ensemble of pair-wise preference models, Chang sung Kang et al.[5] present an extensive analysis of web scale entity ranking. Structured knowledge bases, entity relationship graphs and user data are used to derive useful features to facilitate semantic search with entities directly within the learning to rank framework. The detailed feature space analysis and a suite of novel features in the context of entity ranking are presented.

Topic Model Learning

Topic model learning is designed to automatically extract all possible topics and each topic page from the index page, so that the web page and query statement can match on the topic. As a Bayesian model, topic model strictly adhere to the Bayesian probabilistic framework. Topic model has features of unsupervised learning as a production model. The topic is easy to be understood by human and the characteristics of semantics implied documentation set can be discovered by using a lot of existing internet data. The frame of topic model is shown below.



Doc
Topic
Word
$\vec{\vartheta}_m \sim \mathrm{Dir}(\vec{\alpha})$
z
$\vec{\vartheta}_m$
$\vec{\varphi}_k$
$\vec{\varphi}_k \sim \mathrm{Dir}(\vec{\beta})$
1
2
3
4
K
$\vec{\varphi}_k \;\; k \in [1, K]$

Fig. 3 The frame of topic model

Topic model is one of the generative models with broad application in natural language processing and machine learning. A discrete data set is generated by topic set hiding in the data set. The data set includes a collection of documents and pictures sets. Each topic in above set is word probability distributions. For each document, a topic proportion is firstly extracted. Then above model select a topic for the position of each word. A word from topic word distribution corresponding is selected to fill the position. Above steps are repeated until entire collection of documents is produced. Topic model learning can boost the performance of information retrieval, but real application is limited because of scalability issues. Although Scaling to larger document collections via parallelization is an active area of research, almost all of the methods require drastic steps such as vastly reducing input vocabulary. In the paper [6], regularized latent semantic indexing is proposed for parallelization. It is as effective as existing topic models, and scales to larger datasets without reducing input vocabulary. Above model formalizes topic modeling as a problem of minimizing a quadratic loss function. It lets the learning process to be decomposed into multiple sub-optimization problems which can be optimized in parallel. Labeled latent dirichlet allocation which belongs to supervised topic model lacks considerations of the label frequency of the word which is crucial for multi-label classification. In order to solve above problem, Ximing Li et al.[7] propose a novel model named censored prior topic model. Class-feature-censored suggests a discriminative label word vector that takes the label frequency of the word into account. The censored prior topic model uses this Class-feature-censored vector as prior for label-word distributions. With the development of interactive communication through internet, micro blog becomes a dominant medium in social media. Traditional topic detection methods are unable to achieve a desirable level of performance because of high flood of meaningless tweets and other characteristics of micro blogs, traditional topic detection methods are unable to achieve a desirable level of performance. In the paper [8], a model named multi-attribute latent dirichlet allocation in which the time and tag attributes of micro blogs are incorporated into LDA model is proposed. Above model can decide whether a word should appear in hot topics or not through using a time variable about the time attribute. Applying tag attribute allows the model to rank the core words high in results so that the expressiveness of outcomes can be improved over the traditional LDA model.

Summary

Internet data mining technology is the cross of machine learning and internet databases technology. With the rapid development of software industry and web interaction, there has been an amount of increasing demand of valuable data mining. Two kinds of data mining technologies which include rank learning and topic model learning are reviewed in this paper. We strongly believe that, in the near future, this research field will be paid more and more attentions and will promote the fundamental theories research in the related fields.

Acknowledgement

In this paper, the research was sponsored by the Natural Science Foundation of Shandong Province, China (Grant No. ZR2013FL010).

References

[1] Hang Li, Learning to Rank for Information Retrieval and Natural Language Processing, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2011.

[2] Weiwei Zong, Guang-Bin Huang, Learning to Rank with Extreme Learning Machine, Neural processing letters, 39(2) 155-166, 2014.

[3] Yu, J, Learning to Rank Using User Clicks and Visual Features for Image Retrieval, IEEE Transactions on cybernetics, 45(4) 767-779, 2014.

[4] Fei Wu, Cross-Modal Learning to Rank via Latent Joint Representation, IEEE Transactions on Image Processing, 24(5) 1497-1509, 2015.

[5] Changsung Kang, Dawei Yin, Learning to rank related entities in Web search, Yahoo Labs Neurocomputing, 2015.

[6] Quan Wang, Jun Xu, Regularized latent semantic indexing, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011.

[7] Ximing Li, Jihong Ouyang, Xiaotang Zhou, Centroid prior topic model for multi-label classification, Pattern Recognition Letters, 62(1) 8-13, 2015.

[8] Guolong Liu, An Improved Latent Dirichlet Allocation Model for Hot Topic Extraction, 2014 IEEE Fourth International Conference on Big Data and Cloud Computing , 470 – 476, 2014.