Computing and Informatics, Vol. 29, 2010, 617–627

SEMI-SUPERVISED LEARNING FOR PERSONALIZED WEB RECOMMENDER SYSTEM

Tingshao Zhu

Graduate University of Chinese Academy of Sciences
Beijing, China 100080
e-mail: tszhu@gucas.ac.cn

Bin Hu, Jingzhi Yan, Xiaowei Li

School of Information Science and Engineering
Lanzhou University, Lanzhou, China

Revised manuscript received 24 March 2010

Abstract. To learn a Web browsing behavior model, a large amount of labelled data must be available beforehand. However, very often the labelled data is limited and expensive to generate, since labelling typically requires human expertise. It could be even worse when we want to train personalized model. This paper proposes to train a personalized Web browsing behavior model by semi-supervised learning. The preliminary result based on the data from our user study shows that semi-supervised learning performs fairly well even though there are very few labelled data we can obtain from the specific user.

Keywords: Web behavioral modeling, data mining, computational cyberpsychology

1 INTRODUCTION

While the World Wide Web contains a vast quantity of information, it is often time consuming and sometimes difficult for a web user to locate the information found relevant. This motivates us to build an effective Web recommendation system to

assist the user in finding relevant pages with respect to his/her own interests. We call such pages *information content pages* or ICpages for short.

The earlier *WebIC* publications [15] have done the research on the training/testing of general "browsing behavior model" from the entire population, to predict each specific user's information need based on the current browsing session.

Since different users have diverse interests, it is critical for recommender system to generate personalized useful recommendations with respect to each user's interest and behavior. It is more appropriate to construct a personal recommendation learning system to generate recommendations for each individual. It is expected that a personalized model can provide more realistic personalized recommendations, therefore the quality of service will decrease as the system generates recommendations from the more specific self-trained model to the more general population model.

Moreover, to learn a general behavior model, a large amount of labelled data must be provided beforehand. However, very often the labelled data is limited and expensive to generate, since labeling typically requires human expertise (e.g. conducting user study). It could be even worse when we want to train personalized model, because it becomes more difficult to collect the same amount of labelled data from one individual as the data from a large population.

To overcome the limit of labelled data during the training and still to obtain high quality of recommendation prediction, Semi-Supervised Learning [4, 13] has been introduced, and it has recently attracted a considerable amount of research. The supervised learning, mostly applied in general population model, basically involves two steps: training and prediction. The training process uses all the labelled data to infer a general prediction function, then the prediction process uses the general function to infer labels for unlabelled data. Semi-supervised learning differs by learning from labelled and unlabelled data simultaneously, and makes predictions in one step. Given a data set

$$\mathcal{X} = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\}$$

with the first l points labelled, and all the rest unlabelled. The goal is to predict the labels of the unlabelled points. Basically it takes advantage of the relationship among labelled and unlabelled data to make more accurate predictions than purely using labelled data for training in supervised learning method.

Therefore, we propose to make recommendations in a personalized Web browsing behavior model by semi-supervised learning. The personalized prediction problem in the Web recommender system is naturally fit into the semi-supervised learning setting by using our previous user study. The preliminary result based on the data from our user study shows that semi-supervised learning performs fairly well to produce personalized recommendation for each user, even though there are very few labelled data we can obtain from the user.

The paper is arranged as follows. Section 2 discusses the related work. Our approach uses the data from a user study (LILAC), which will be described briefly in Section 3. We also demonstrate the key steps to build a personalized Web recom-

mender system by semi-supervised algorithm in Section 4. Section 5 introduces our experiment design to evaluate the performance of semi-supervised learning, using the data collected from LILAC. Finally, we will conclude this paper with the future work.

2 RELATED WORK

There is a great deal of research on generating recommendations for Web users, describing systems that make predictions on specific Web sites (see association rules [1]), some based on specific hard-selected words [3, 8, 2], and others that seek useful complete-web recommendation [15]. This section summarizes several related approaches, and discusses how they relate to our work.

Pirolli and Fu [11] try to identify information need based on the *SNIF-ACT* model: production rules and a spreading activation network. These SNIF-ACT production rules resemble the patterns that we are attempting to learn. However, our system differs by *learning* personalized model for each user; hence our systems do not rely on any prior knowledge of the whole population.

Amanda et al. [12, 7] analyzed Web queries passed to the Excite search engine, and found several interesting characteristics of Web search behaviors – e.g. most of the search involves very few search terms, few modified queries, and rarely use advanced search features, etc. This differs from our research as our focus is on finding useful information based on information gathered innocuously for a specific user, rather than characterizing how users interact with search engines.

The Letizia [9] agent helps a user browse the Web by using heuristics to infer a user interest from browsing behavior. Watson [5] observed users interacting with everyday applications and then anticipated their information needs using heuristics to automatically form queries. The heuristics used by Letizia and Watson are hand-coded. While they may be personalized for each individual, we expect models learned from actual user will be more accurate.

The earlier *WebIC* publications [16, 15] focused on ways to learn general browsing patterns, corresponding to a large population. This paper significantly extends those earlier results by learning each individual's interests and behavior separately. We propose to apply the semi-supervised learning method [14] to predict each individual's information need based on the personalized model. The semi-supervised learning method uses the global information from the labelled and unlabelled words to make predictions. The reality here is that the labelled words from pages annotated by each user is very limited while we have large number of data unlabelled. The semi-supervised learning method is especially efficient under this kind of circumstance. It combines the information from unlabelled words in learning while the supervised learning method only used labelled words in training. As the supervised learning method uses only very limited training data, it won't be very efficient to explore the information from large amount of unlabelled data. The out-performance of the semi-supervised method has been shown in [14]. Therefore, we construct pro-

file (annotated data) for each user and apply the semi-supervised method to predict its current information need based on its own behaviors.

3 THE USER STUDY – LILAC

Our approach uses the data from a user study, named "LILAC" (Learn from the Internet: Log, Annotation, Content) [18], which attempts to evaluate the browsing behavior models by actual users working on their day-to-day tasks.

All LILAC participants were required to install a customized browser (i.e. *WebIC*) on their own computer, and they were instructed to use *WebIC* to browse their own choice of web pages. Each participant was instructed to make annotation by clicking "MarkIC" button whenever s/he found a page s/he considered to be an ICpage, and s/he was also required to evaluate the usefulness of the suggested page by *WebIC*. As part of this evaluation, the subjects were instructed to "Tell us what you feel about the suggested page", to indicate whether the information provided on the suggested page was relevant to his/her search task. They could choose one of five options as follows:

- *Fully* answered my question
- *Somewhat* relevant, not answering my question fully
- *Interesting*, but not so relevant
- *Remotely* related, but still in left field
- *Irrelevant*, not related at all.

LILAC considered four models: the three behavior models and "Followed Hyperlink Word" (FHW) [6], which is used as a baseline.

These browsing behavior models describe how users locate relevant pages on the Web as general, which have been trained based on the "browsing features" [15] extracted from the annotated data (e.g. "ratio of the pages in the session that contain word w", "latest relative location of the page that contained w", etc.). We have developed 35 *browsing features* to describe how each word w relates to the sequence of pages visited. We trained the model on the data previously annotated by all subjects. That is, the users initially used the models, which were based on a model obtained prior to the study. During the 2^{nd} week, they used the models, based on the training data obtained from week 1, as well as the prior model, and so forth. Note that the models used in LILAC were trained based on the entire population.

To suggest a page, *WebIC* first computes browsing features for all stemmed non-stopwords that appear in the current session. It then determines which of these words to submit as a query to a search engine, using one of the models learned previously. *WebIC* then recommends the top page returned, and also required the subject to give an evaluation on the suggested page.

In order to train these behavior models, the study participants must actively make annotations while browsing the Web; this is both inconvenient for the user,

and unrealistic in a production version of Web recommender system. To partially avoid this problem, we have tried to train a behavior model based on previous evaluation results in LILAC. We found that the results of such a model are similar to the results of training the model directly on the original ICpages. This observation is significant as it will allow us to continuously refine the model without requiring the user's annotation.

4 PERSONALIZED WEB RECOMMENDER SYSTEM

Here, we propose to build a personalized Web recommender system based on semi-supervised learning, using a user's previous evaluations on the suggested pages. Since the semi-supervised learning method can still predict fairly well even with very few labelled data, the user may be relaxed by making evaluations as less as possible. The user can also help improve the performance the semi-supervised learner by providing more evaluations.

Imaging the system watched the user's browsing on the Web. Whenever the user asks for recommendation, the system will collect the observed page sequence, and suggest a Web page which is expected to satisfy the user's current information need.

Extracting Browsing Features. The system will first identify all the non-stop stemmed words in the current browsing session, and extract a *browsing feature* vector for each word

$$w = (bf_1, bf_2, \ldots, bf_i, \ldots, bf_n),$$

in which bf_i is the i^{th} value of w's browsing features. This results in a big matrix, where each row corresponds to a word encountered, and each column, to the value of a particular browsing feature. The matrix will be fed into the semi-supervised predictive model as unlabelled data points.

Predicting by combining labelled and unlabelled data. Based on the observation from LILAC, we can make prediction by using previous evaluation results. The predictive model records every query (i.e. a list of keywords) sent to search engine to retrieve the suggested page, associated with the user's evaluation. Here, we only consider the two extreme evaluation options: "Fully" and "Irrelevant", and label each word in the query as *Fully* (+) or *Irrelevant* (-), corresponding to the evaluation on the page that returned by sending the query to a search engine. Note that we only use the browsing features of each word, not the words themselves. Table 1 shows one example of such labelled data set.

The predictive model then identifies the unlabelled data that inputed after the browsing feature extraction, to make prediction by combining these labelled and unlabelled data. To do so, the predictive model first combines the labelled browsing feature vectors with the unlabelled ones as

$$\mathcal{W} = \{w_1, \ldots, w_l, w_{l+1}, \ldots w_n\}$$

Query	Browsing Features	Evaluation
⋮		
data	...0.15 ...0.23 ...	+
+mining	...0.32 ...0.8 ...	
+software	...0.15 ...0.62 ...	
+free	...0.80 ...0.34 ...	
⋮		

Table 1. The labelled data in semi-supervised predictive model

and a label set

$$\mathcal{Y} = \{-1, 1\}.$$

The first l words w_i are labelled as $y_i \in \mathcal{Y}$ according to user's previous evaluation and the remaining words are unlabelled. Each word w_i is associated with a browsing feature vector. We want to make predictions for the remaining unlabelled words. The main steps of the semi-supervised algorithm are described according to [13] in Algorithm 1.

Algorithm 1 Semi-supervised learning algorithm

INPUT: word features set $\mathcal{W} = \{w_1, \ldots, w_l, w_{l+1}, \ldots, w_n\}$ and a vector $y = \{y_1, \ldots, y_l, 0, \ldots, 0\}$.

1. Construct a Gaussian weight matrix such that

$$G_{ij} = e^{(-||w_i - w_j||^2/2\sigma^2)}$$

2. Construct the matrix

$$S = D^{-1/2} G D^{-1/2}$$

in which D is a diagonal matrix that

$$D(i,i) = \sum_j G(i,j)$$

3. Compute

$$f = (I - \alpha S)^{-1} y$$

where $\alpha \in [0, 1]$

OUTPUT: a label vector $y_i = f_i$ for each word w_i.

We then rank the words with positive outcomes (i.e. $y_i > 0$) and rank these words according to y_i, select top $m = 4$ words as query to retrieve a page from search engine.

Incorporating evaluations. As we stated above, the user's evaluation on the suggested page can also help predict the query which will trigger search engine to

return a relevant Web page. We encourage the user to evaluate the recommendations as possibly as s/he can. The more the labelled data, the better performance of the predictive model. If the user would like to make the evaluation, we will label the keywords in the query according to the evaluation outcome, and append these new labelled feature vectors into the model (i.e. Table 1). Thus, the newly labelled data can be taken into account for the next time's prediction.

We fulfil personalization in the recommender system by maintaining different predictive model (i.e. labelled feature vectors) for each individual, which can also catch up with the user's browsing behavior shifting. We can also suggest page for a group of users if the predictive model can obtain labelled data (browsing feature vectors with evaluation label) from each of its members.

5 EXPERIMENT

To measure the performance of the semi-supervised predictive model, we run the off-line testing that simulates the subjects' behavior in the "LILAC" user study.

We assume that the user's evaluation of the suggested page is based on the similarity between his/her own ICpage and the suggested page [17]. For each "MarkIC" session in LILAC, the user annotated the ICpage only if it satisfies his/her current information need. In such cases, the ICpage can equivalently be described as the page that can fully answer his/her question, which means that the "Fully" suggested page contains very similar content as the ICpage. Alternatively, a page that is evaluated as "Irrelevant" contains unrelated information. Intuitively, the more a suggested page is similar to ICpage, the higher probability it will be evaluated as "Fully". Thus the higher the similarity score of the page suggested by a model, the better the model.

Several "similarity functions" ($f(p_{ic}, p_S)$ on the ICpage p_{ic}, and the suggested page p_S) have been proposed and then verified by LILAC data [17]. Among them, Information Theoretic Measure (ITM) has the simplest format, and it performs promisingly on LILAC data. ITM is a simplified version of the measure that was proposed in [10].

$$f_{ITM}(p_{IC}, p_S) = \frac{|W_{IC} \cap W_S|}{|W_{IC} \cup W_S|}$$

in which W_{IC} are the words in the ICpage, while W_S denotes the words in the suggested page, after removing stop-word and stemming. ITM returns a large number if p_S was a "Fully" page, and a small number if p_S was an "Irrelevant" page. Figure 1 shows the whole process for testing the personalized web recommendation on LILAC data.

We run the testing independently on each subject's data, which means we will reset the semi-supervised predictive model (i.e. clear the label data) before we start testing on a new subject. For each MarkIC session of a subject in LILAC, ICpage (p_{IC}) and suggested page ($p_{\text{suggest}}^{LILAC}$) generated by $WebIC$ can be collected, and based

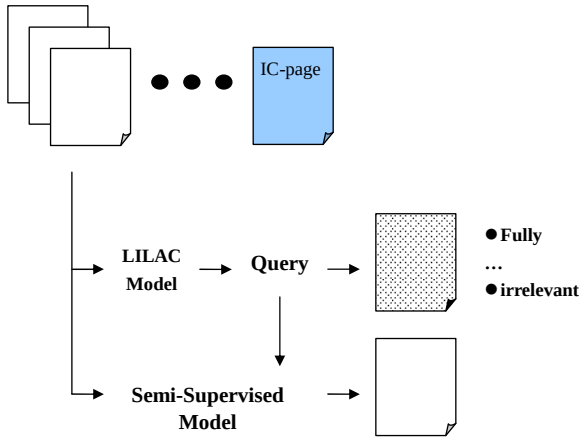

Fig. 1. Testing semi-supervised learning for personalized web recommendation on LILAC data

on the same session, we apply semi-supervised predictive model to generate another suggested page $p^{\text{semi}}_{\text{suggest}}$.

For one-quarter of the MarkIC sessions within the LILAC study data, *WebIC* selected the baseline FHW model. We can now compute the similarity between the user's ICpage p_{IC} and this proposed p^{FHW}_{suggest} page – $f(p_{IC}, p^{FHW}_{\text{suggest}})$ – using ITM, and also the similarity between the ICpage and the suggested page by semi-supervised model, $f(p_{IC}, p^{\text{semi}}_{\text{suggest}})$.

To validate the hypothesis that semi-supervised model is better than FHW, we perform a statistical test (i.e. Wilcoxon) on the correlated samples. If the p value is less than 0.05, then we can conclude that it is better than the FHW model. Table 2 shows several pairs of similarities from "MarkIC" sessions in LILAC.

FHW	Sem-Supervised Model
⋮	
0.100156	0.139225
0.101887	0.131188
0.034173	0.036641
0.05	0.072685
0.146667	0.166667
0.078205	0.116919
⋮	

Table 2. Paired similarities on LILAC data

We then run Wilcoxon test on the paired similarities, with the hypothesis

$$f(p_{IC}, p^{FHW}_{\text{suggest}}) < f(p_{IC}, p^{\text{semi}}_{\text{suggest}}).$$

The p value equals to 0.0137. We can make the conclusion that there exists a significant difference between FHW and semi-supervised model; in other words, semi-supervised model performs better than FHW. Note that the conclusion is qualitative, which means we do know that semi-supervised model works better than FHW, but we do not know how semi-supervised model performs over FHW.

After we have completed computing similarities on one "MarkIC" session, we will incorporate the query generated in LILAC (i.e. keywords' browsing features on the current "MarkIC" session and the evaluation label) into the semi-supervised model, which will be used for the next time's prediction.

6 CONCLUSION AND FUTURE WORK

It is more and more important for a Web recommender system to be able to efficiently help people locate relevant information on the Web, especially to provide personalized web recommendations to save people a lot of time for search information on the web. The standard approach applied supervised learning to build Web user model for producing recommendations. However, the labelled data is limited, and sometimes is very difficult to acquire.

We propose a novel method to generate personalized web recommendation using semi-supervised learning based on previous evaluations. We have developed the semi-supervised predictive model in our Web recommender system, and the preliminary result indicates that it works better than the baseline model.

We are currently investigating additional similarity functions to verify the performance of web recommender system. In the future, we plan to test the semi-supervised predictive model in another user study.

REFERENCES

[1] Agrawal, R.—Srikant, R.: Fast Algorithms for Mining Association Rules. In Proc. of the 20th International Conference on Very Large Databases (VLDB '94), Santiago, Chile, September 1994.

[2] Anderson, C.—Horvitz, E.: Web Montage: A Dynamic Personalized Start Page. In Proceedings of the 11th World Wide Web Conference (WWW 2002), Hawaii, USA 2002.

[3] Billsus, D.—Pazzani, M.: A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling (UM '99), Banff, Canada 1999.

[4] Blum, A.—Mitchell, T.: Combining Labeled and Unlabeled Data With Co-Training. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98).

[5] Budzik, J.—Hammond, K.: WATSON: Anticipating and Contextualizing Information Needs. In Proceedings of 62nd Annual Meeting of the American Society for Information Science, Medford, NJ 1999.

[6] Chi, E.—Pirolli, P.—Chen, K.—Pitkow, J: Using Information Scent to Model User Information Needs and Actions on the Web. In ACM CHI 2001 Conference on Human Factors in Computing Systems, pp. 490–497, Seattle, WA, 2001.

[7] Jansen, B. J.—Spink, A.—Saracevic, T.: Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Information Processing and Management, Vol. 36, 2000, No. 2, pp. 207–227.

[8] Jennings, A.—Higuchi, H.: A User Model Neural Network for a Personal News Service. User Modeling and User-Adapted Interaction, Vol. 3, 1993, No. 1, pp. 1–25.

[9] Liebermann, H.: LETIZIA: An Agent That Assists Web Browsing. In International Joint Conference on Artificial Intelligence, Montreal, Canada, August 1995.

[10] Lin, D. K.: An Information-Theoretic Definition of Similarity. In Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July 1998.

[11] Pirolli, P.—Fu, W.: SNIF-ACT: A Model of Information Foraging on the World Wide Web. In Ninth International Conference on User Modeling, Johnstown, PA 2003.

[12] Spink, A.—Wolfram, D.—Jansen, B.—Saracevic, T.: Searching the Web: The Public and Their Queries. Journal of the American Society of Information Science and Technology, Vol. 52, 2001, No. 3, pp. 226–234.

[13] Zhou, D.—Bousquet, O.—Navin Lal, T.—Weston, J.—Schölkopf, B.: Learning With Local and Global Consistency. In Advances in Neural Information Processing Systems 16, Cambridge, MA, USA.

[14] Zhou, D.—Weston, J.—Gretton, A.—Bousquet, O.—Schölkopf, B.: Ranking on Data Manifolds. In Advances in Neural Information Processing Systems 16, Cambridge, MA, USA.

[15] Zhu, T.—Greiner, R.—-Häubl, G.: An Effective Complete-Web Recommender System. In The Twelfth International World Wide Web Conference (WWW 2003), Budapest (Hungary), May 2003.

[16] Zhu, T.—Greiner, R.—-Häubl, G.: Learning a Model of a Web User's Interests. In The 9th International Conference on User Modeling (UM 2003), Johnstown, USA, June 2003.

[17] Zhu, T.—Greiner, R.—Häubl, G.–Jewell, K.—Price, B.: Off-Line Evaluation of Recommendation Functions. In The 10th International Conference on User Modeling (UM 2005), Edinburgh, Scotland, July 2005.

[18] Zhu, T.—Greiner, R.—Häubl, G.—Jewell, K.—Price, B.: Using Learned Browsing Behavior Models to Recommend Relevant Web Pages. In Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, August 2005.

Tingshao Zhu received his second Ph. D. from the University of Alberta (Canada) in 2006. From 2008, he started working as a Professor at the Graduate University of Chinese Academy of Sciences (CAS) in Beijing. He has extensive experience in data mining and machine learning. The main focus of his current work is in web user behavior modeling, web mining and recommender and data mining.

Bin Hu is Professor, Dean of School of Information Science and Engineering, Lanzhou University, the leader of Intelligent Contextual Computing Group in Pervasive Computing Centre, Reader, Birmingham City University, Visiting Professor in Beijing University of Posts and Telecommunications, China and at ETH Switzerland. He received M. Sc. in computer science from Beijing University of Technology and Ph. D. in computer science from Institute of Computing Technology, Chinese Academy of Science. His research interests include pervasive computing, CSCW and semantic web.

Jingzhi Yan is a lecturer in Lanzhou University. She graduated from Lanzhou University with a Ph. D. degree in mathematics and statistics. Her research interests are in machine learning and graph theory.

Xiaowei Li is a lecturer at Lanzhou University. He graduated from Lanzhou University in 2002 with a M. Sc. degree in computer software and theory. He has many years' research experience in affective learning, ubiquitous computing and data mining.