An Adaptive Method for User Profile Learning

Rim Zghal Rebaï, Leila Ghorbel, Corinne Amel Zayani, and Ikram Amous

MIRACL-ISIMS, Sfax University, Tunis road Km 10, 3021 Sfax, Tunisia
rim_zghal@yahoo.fr, leila.ghorbel@gmail.com, zayani@irit.fr,
ikram.amous@isecs.rnu.tn

Abstract. The user profile is a key element in several systems which provide adapted result to the user. Thus, for a better quality of response and to satisfy the user, the profile's content must always be pertinent. So, the removal of irrelevant content is necessary. In this way, we propose in this paper a semi-supervised learning based method for automatically identifying irrelevant profile elements. The originality of this method is that it is based on a new co-training algorithm which is adapted to the content of any profile. For this, our method includes a preparation data step and a classification profile elements process. A comparative evaluation by the classical co-training algorithm shows that our method is better.

Keywords: user profile, navigation history, data preparation, semi-supe-rvised learning, co-training technique.

1 Introduction

In order to take into account of user's interests, navigation history, preferences etc. several systems, such as adaptive systems, use user profile. This latter must be automatically and frequently updated by the system after each user-system interaction. By the time, and especially after several addition operations, the profile can be overloaded and contains relevant and irrelevant elements. This can affect the result's relevance. For this, the solution is to remove irrelevant elements which are detected by an automatic classification.

In this way, several methods are proposed such as [15] and [10]. These methods use mainly a learning technique [16] which can be unsupervised or supervised. These techniques are used to identify the relevant new elements to be added to the profile or to identify the irrelevant already existing elements to be removed from the profile. The main objective of these methods is that the profile should always be not-overloaded and pertinent. But, by using one of these learning techniques the result is generally not adapted to any user's profile. Especially because each user has his specific interests, preferences, history, etc. and these techniques are applied in the same way whatever the profile's content.

Our contribution in this paper is to propose an automatic classification method adapted to the content of any user's profile. It is based on the semi-supervised

B. Catania, G. Guerrini, and J. Pokorný (Eds.): ADBIS 2013, LNCS 8133, pp. 126–134, 2013.

learning technique by using a new co-training algorithm. In fact, in the literature, several studies have proposed new version of co-training algorithm to make it adapted to their addressed problem, among them we cite [8].

The remaining of this paper is presented as follows. Section 2 presents briefly the required steps to apply the learning techniques and some user profile elements classification works. In section 3, we present our method. Section 4 depicts the evaluation of our proposal. Section 5 presents the conclusion and future work.

2 State of the Art

Automatic learning has been attracting a significant amount of research fields such as information researches, image processing, etc. Before being applied, the data preparation must be performed [5]. It is generally composed of three steps: selection, preprocessing and transformation step. The first step allows to identify the required data for the classification. Then, these data are copied and displayed in a matrix which describes elements and their attributes. The second step deals with cleaning the data in order to correct any inaccuracies or errors such as duplicates, missing information, etc. The third step is to enrich, normalize and code the data to apply the classification. The enrichment is done by adding new attributes. Normalization and coding are done by regrouping or simplifying attributes (coding of discrete attributes, changing type, etc.).

In the literature, we find several techniques of automatic learning. The most used techniques are: the unsupervised, supervised and semi-supervised techniques. In this paper, we are interested in works that apply these techniques on user's profiles either to classify the already existing profile elements in order to remove the irrelevant ones such as [15], [10] and [3] or to classify the new elements to be added to the profile such as [4], [11] and [14]. These latter can provide pertinent profiles but after several updating operations, the profile can be overloaded. Moreover, the profile cannot contain all the user's various interests for the reason that a new interest can be added only if there is a similarity with the already existing interests. For these reasons, we are interested in the methods that use learning techniques to classify the already existing profile elements. In [3] the profile elements are represented as hierarchical categories. Each category represents the knowledge about a user interest and has an energy value. This value increases when the user shows interest in the category and decreases by a constant value for each period of time. Based on the energy value, the system classifies the categories: categories that have low-energy will be removed and categories that have high-energy will persist. The proposed method in [15] allows to classify the profile elements (here profile concepts) by using association rules and Bayesian networks. The relevant concepts are maintained. In [10] authors are based on the supervised learning technique by using the K-NN algorithm. The classifier uses labeled users preferences pool to classify the preferences of each user.

As we said at the beginning of this section, all these presented methods should perform the data preparation before applying the most appropriate learning technique to their contexts. By using these techniques, these methods [15], [10] and [3] are able to identify irrelevant profile elements. However, all these methods are based on the supervised learning technique which provides a prediction model that is not usually adapted on any user profile.

In this paper, we propose a method of profile elements classification adapted to the content of any profile. This method is based on a semi-supervised learning technique and uses a new co-training algorithm.

3 New Method Based on Co-training Algorithm

To classify profile elements into relevant/irrelevant, we propose an automatic method which can be applied to overloaded users' profiles. These profiles respect the profile model proposed in [20] and are composed of several parts. In this work, we are interested on the navigation history part which contains mainly the visited domains. Each visited domain is composed of the visited sub-domains, if exist, the visited documents and the visited links. These profiles are obtained after several navigation sessions in the INEX 2007[1] corpus which is part of the collection WIKIPEDIA XML. The 110000 XML documents in this corpus are related to one or more domains and interconnected by XLINK simple links. The used navigation method and the updating process are detailed in our previous work [19].

To apply the appropriate learning technique on these profiles, the data preparation is required. In our work, it consists of two steps: (i) selection, and (ii) transformation of data.

3.1 Data Preparation

The purpose of the data preparation is to provide a set of labeled user profiles on which the classification will be based. It consists of two steps: the selection of data related to each element (visited domains, visited sub-domains, visited documents and visited links) and the transformation of these data (cf. Fig. 1).

The data selection step consists in extracting the four user profile elements and their attributes. Table 1 presents all the extracted attributes.

These attributes describe mainly the different identifiers, the number of visits and clicks, the date of visits and clicks and the duration of visits.

After this step, we obtain four databases for each profile. These databases are the input of the data transformation step. This step allows to: (i) change the coding of some attributes, (ii) enrich the databases by adding new attributes and (iii) filter some not-discriminates attributes. The main objectives of this step are to facilitate the semi-automatic labeling of the profiles and improve the precision rates of the classification. For the coding of attributes, we have changed the coding of the attributes related to the date of visit in order to differentiate the recent dates and not-recent dates. So, each date will be replaced either by

[1] http://www-connex.lip6.fr/~denoyer/wikipediaXML

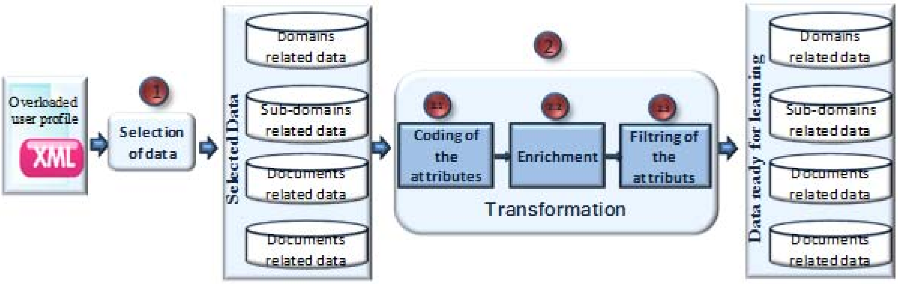
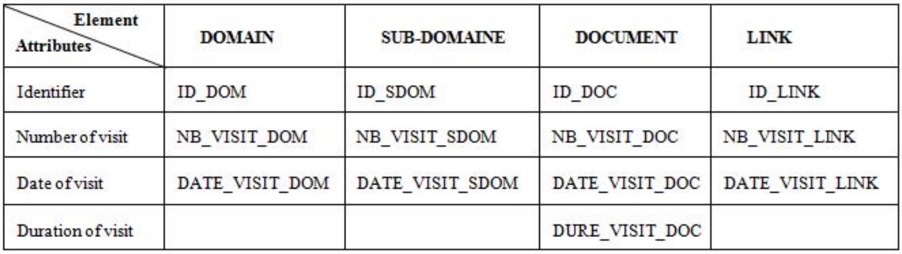


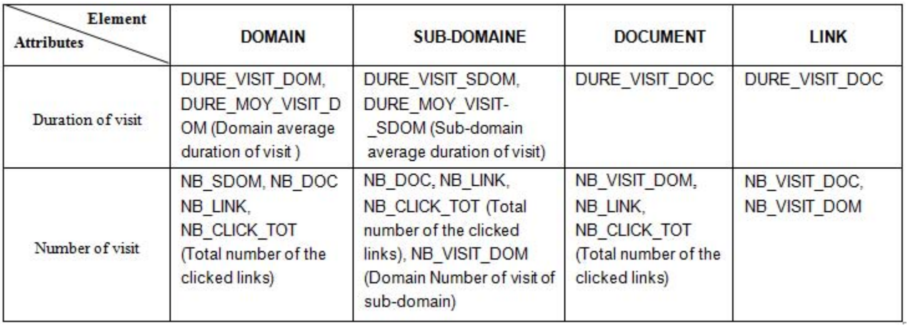
Fig. 1. Data preparation

Table 1. Selected elements and attributes from profile



Element / Attributes	DOMAIN	SUB-DOMAINE	DOCUMENT	LINK
Identifier	ID_DOM	ID_SDOM	ID_DOC	ID_LINK
Number of visit	NB_VISIT_DOM	NB_VISIT_SDOM	NB_VISIT_DOC	NB_VISIT_LINK
Date of visit	DATE_VISIT_DOM	DATE_VISIT_SDOM	DATE_VISIT_DOC	DATE_VISIT_LINK
Duration of visit			DURE_VISIT_DOC	

Table 2. Added attributes of each element

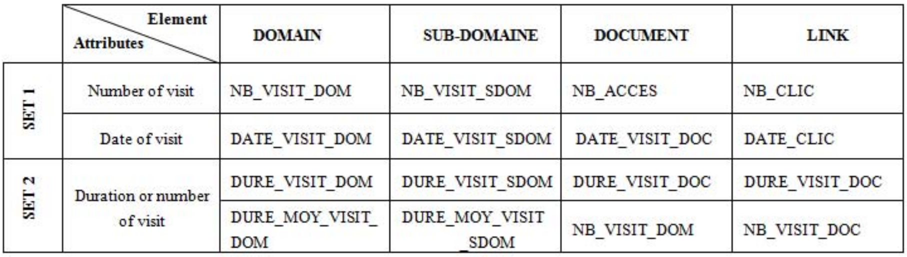


Element / Attributes	DOMAIN	SUB-DOMAINE	DOCUMENT	LINK
Duration of visit	DURE_VISIT_DOM, DURE_MOY_VISIT_DOM (Domain average duration of visit)	DURE_VISIT_SDOM, DURE_MOY_VISIT-_SDOM (Sub-domain average duration of visit)	DURE_VISIT_DOC	DURE_VISIT_DOC
Number of visit	NB_SDOM, NB_DOC NB_LINK, NB_CLICK_TOT (Total number of the clicked links)	NB_DOC, NB_LINK, NB_CLICK_TOT (Total number of the clicked links), NB_VISIT_DOM (Domain Number of visit of sub-domain)	NB_VISIT_DOM, NB_LINK, NB_CLICK_TOT (Total number of the clicked links)	NB_VISIT_DOC, NB_VISIT_DOM

”R” (ie. Recent date) or by ”NR” (ie. Not Recent date). As for the enrichment of the databases, we firstly added to each element some attributes related to the other elements. Table 2 illustrates the added attributes of each element.

Secondly, we have semi-automatically labeled the profiles elements and added the labels ”+” for the relevant elements and ”-” for the irrelevant ones. The labeling starts with domains, sub-domains, documents and links. It is performed as follows; if an element is irrelevant then all its child elements are automatically labeled as irrelevant, otherwise all its child elements must be manually labeled.

Table 3. Discriminate attributes



	Element Attributes	DOMAIN	SUB-DOMAINE	DOCUMENT	LINK
SET 1	Number of visit	NB_VISIT_DOM	NB_VISIT_SDOM	NB_ACCES	NB_CLIC
	Date of visit	DATE_VISIT_DOM	DATE_VISIT_SDOM	DATE_VISIT_DOC	DATE_CLIC
SET 2	Duration or number of visit	DURE_VISIT_DOM	DURE_VISIT_SDOM	DURE_VISIT_DOC	DURE_VISIT_DOC
		DURE_MOY_VISIT_ DOM	DURE_MOY_VISIT _SDOM	NB_VISIT_DOM	NB_VISIT_DOC

Finally, based on the obtained labeled profiles, we proceeded to the filtering of the most discriminate attributes for classification. For this, we used ReliefF [9] algorithm. In the literature, there are many different algorithms for attributes selection such as Fisher filtering, Feature ranking, etc.; we are interested in ReliefF algorithm because it is unaffected by attributes interaction [9].The selected attributes are illustrated in table 3.

At the end of the data preparation, we obtain a set of labeled user profiles that represent labeled pool on which our proposed classification process will be based.

3.2 Profile Elements Classification

In our work, the data preparation and especially the semi-automatic labeling of a large number of users' profiles is a considerable work and the classification task must be adapted to the content of any user's profile. For these reasons, the semi-supervised learning technique has been chosen to classify automatically user's profile elements.

In the literature, several techniques for semi-supervised learning are proposed such as the self-training, the co-training, S3VM and T-SVM. As it is simple, able to provide better adapted result to classification problems of data, we are more interested in co-training.

Based on [17], the idea of co-training is that two separated classifiers are trained using the data of the Labeled Pool (LP) having two sub-attribute sets respectively. For the reason that co-training assumes that attributes for training must be split into two sets. Each sub-attribute set is sufficient to train a classifier and the two sets are conditionally independent given the class. Then, each classifier generates a prediction model. Based on this model, the classifier assigns labels to unlabelled data given as input. After that, the most confident predicted ones (obtained by the two classifiers), are selected and added to LP and the process repeats. When training is completed, after n rounds, the labels of the data to be classified are predicted. This process of co-training has been successfully applied to several classification fields. In our case, this process cannot robustly classify the elements of any profile. In fact, the content of any profile varies from one user to another according to the history of each one (durations of visits, total duration of sessions, etc.). So, two generated prediction models (by the two classifiers) based on one LP (the set of the labeled profiles result of the data preparation), which contains a mixture labeled data from several various users' profiles, cannot usually be applied to any profile and cannot provide good classification result. Therefore, we propose a new method based on a new co-training algorithm which can be adapted to the content of any user's profile. This method is based on N Adapted Labeled Pools (N-ALP). The initial content of N-ALP is similar to the content of the N-LP. Each LP consists in an overloaded labeled user profile. That means that we considered that each profile represents one LP.

To choose the best classifier (learning algorithm), we first made the choice of the supervised learning technique. In literature, there are several techniques of supervised learning. One of the criteria to compare these techniques is the comprehensibility of the generated prediction model. Based on this criterion, we choose the induction of decision trees technique [1]. We have applied to a set of overloaded profiles nine algorithms (ADTRee [18], C4.5 [12], DecisionStump [7], ID3 [13], RandomFoorest [2], and REPTree [13]) and we have obtained the best values of F-Measure and Classification by REPTree. So, we used two REPTree classifiers. Each classifier is based on a set of attributes, set 1 and set 2 (cf. Table 3). A classifier uses attributes related to date and number of visits and the other one uses the duration and number of visit related attributes. At the first round, these classifiers are trained based on the two attributes sets of N-LP and the unlabeled overloaded profile P (the input). So, 2*N prediction models are generated and the most confident labeled elements from each class (relevant/irrelevant) based on these models are added to N-ALP (elements from P). Then, the two classifiers are retrained n-1 rounds on N-ALP and P and after each round 2*N new prediction models are generated and the most confident labeled obtained elements are added to N-ALP. This process is applied to profile elements in the following order: domains, sub-domains, documents and links to obtain a labeled user profile. We only have a filtering step that eliminates the irrelevant elements to obtain a pertinent and not-overloaded user profile.

4 Evaluation

For the evaluation of our proposed method, 20 users have navigated for several sessions in the INEX 2007 corpus until we obtain 20 overloaded profiles. Based on these latter, we carried out a series of experiments. We begin by evaluating the classical co-training algorithm and our proposed co-training algorithm after applying them only to domain elements. Then, we finish by evaluating our algorithm after applying it to all profiles elements (domains, sub-domains, documents and links).

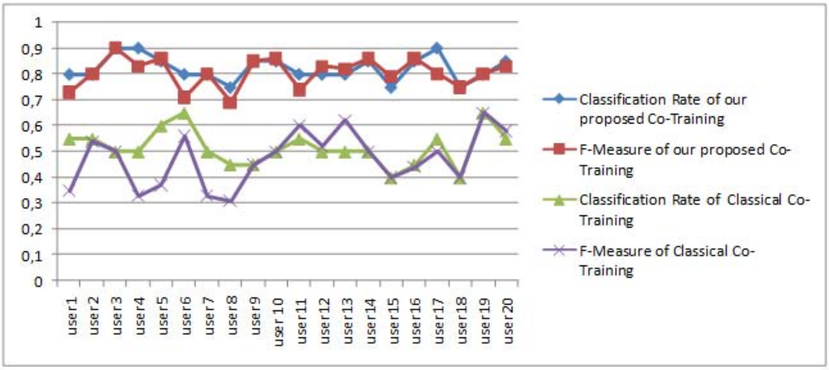


Fig. 2. Comparison between the classical co-training and our proposed co-training algorithm

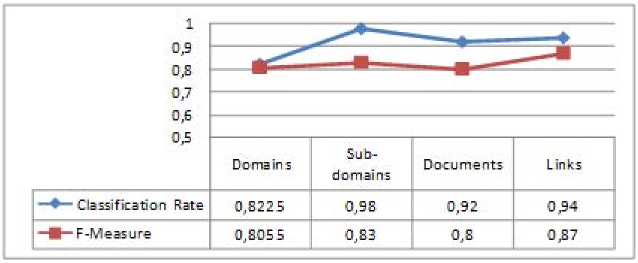
So, for the first evaluation we proceed to the labeling of the profiles' domains elements: (i) manually by the 20 users themselves, (ii) based on the classical co-training algorithm and (iii) based on our proposed method. Then, we compare the profiles obtained by applying the classical and our co-training algorithm with the manual labeled profiles. For this, we used the F-Measure (FM=2*recall*precision/recall+precision) and the classification rates (CA=1-ER, ER is the error rate).

Figure 2 presents the result of the comparison between the classical co-training labeled profiles and our proposed co-training algorithm labeled profiles. We can notice that the best obtained CR and FM by the classical algorithm are 0.65. For example, for user 14 their respective values are: 0.5 and 0.5. In addition, this user labeled manually 11 domains as relevant, while classical co-training algorithm provides only 6 relevant domains. Whereas, by applying our algorithm the CR and FM are improved. For the same user 14, the values of CR and FM are respectively 0.85 and 0.86. Moreover, our method was labeled 9 relevant domains among 11.

In fact, the average CR increases from 0.515 to 0.8225 and the F-Measure increases from 0.4725 to 0.8055. For example, for the user 14 the CR was 0.40 and becomes 0.85. As for user 4 the FM was improved from 0.33 to 0.83. Thus, the obtained results in figure 2 prove the efficiency of our method.

For the second evaluation, we labeled all the profiles' elements: (i) manually by the 20 users and (iii) based on our proposed algorithm. Then we compare the obtained profiles by using the F-Measure and CA. Figure 3 depicts the obtained average values of the CR and F-Measure.

Based on the values illustrated in figure 3, we obtain 0.9156 as FM average value and 0.8263 as average CR. These values can confirm the effectiveness of our method.



	Domains	Sub-domains	Documents	Links
Classification Rate	0,8225	0,98	0,92	0,94
F-Measure	0,8055	0,83	0,8	0,87

Fig. 3. Evaluation of our method on the four profile elements

5 Conclusion

In this paper, we presented the different steps which provide a pertinent user's profile from an overloaded one: the data preparation and the classification process. This latter is based on a new co-training algorithm adapted to any user's profile. With the comparison of set of users' profiles labeled by the classical co-training algorithm and those labeled by our proposed proves that our algorithm is better.

In the coming works we intend to implement our method in our adaptive navigation architecture [21] and evaluate its reliability on the navigation adaptation.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. CRC Press (1984)
2. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Chen, C., Chen, M., Sun, Y.: A self-adaptive personal view agent. Journal of Intelligent Information Systems 18(2-3), 173–194 (2002)
4. Chunyan, L.: User profile for personalized web search. In: Fuzzy Systems and Knowledge Discovery FSKD, pp. 1847–1850 (2011)
5. Fayyad, M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining Book, pp. 1–34 (1996)
6. Guggenberger, A.: Semi-supervised Learning with Support Vector Machines. Technischen Universität Wien (2008)
7. Iba, W., Langley, P.: Induction of one-level decision trees. In: International Machine Learning Conference. Morgan Kaufmann (1992)
8. Jarraya, S.K., Boukhriss, R.R., Hammami, M., Ben-Abdallah, H.: Cast Shadow Detection Based on Semi-supervised Learning. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part I. LNCS, vol. 7324, pp. 19–26. Springer, Heidelberg (2012)
9. Kononenko, I.: Estimating attributes: Analysis and extensions of relief. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)
10. Montaner, M., Lopez, B., De La Rosa, J.: A Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19(4), 285–330 (2003)

11. Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007)
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
13. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1(1), 81–106 (1986)
14. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: Conference on Information and Knowledge Management, pp. 525–534 (2007)
15. Victoria, E., Analï, A.: Ontology-based user profile learning. Appl. Intell. 36(4), 857–869 (2012)
16. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
17. Xiaojin, Z.: Semi-Supervised Learning Literature Survey. Technical Report (2008)
18. Yoav, F., Llew, M.: The Alternating Decision Tree Learning Algorithm. In: ICML, pp. 124–133 (1999)
19. Zghal Rebaï, R., Zayani, C.A., Amous, I.: An adaptive navigation method in semi-structured data. In: Morzy, T., Härder, T., Wrembel, R. (eds.) Advances in Databases and Information Systems. AISC, vol. 186, pp. 207–215. Springer, Heidelberg (2013)
20. Zghal Rebaï, R., Zayani, C., Amous, I.: A new technology to adapt the navigation. In: International Conference on Internet and Web Applications and Services (in press, 2013)
21. Zghal Rebaï, R., Zayani, C., Amous, I.: MEDI-ADAPT: A distributed architecture for personalized access to heterogeneous semi-structured data. In: Web Information Systems and Technologies, pp. 259–263 (2012)