A PATTERN-MATCHING METHOD FOR EXTRACTING PERSONAL INFORMATION IN FARSI CONTENT

Hojjat EMAMI[1], Hossein SHIRAZI[2], Ahmad ABDOLLAHZADEH BARFOROUSH[3], Maryam HOURALI[4]

Personal information extraction as a subtask of entity profiling is the process of identifying and extracting useful and structured information of the focus person named-entity from the content related to it. This is a demanding task, so it is of great importance of how to extract the personal information in a given text. In this paper, we present an approach to extract the relevant information of the focus persons in Farsi content. Our approach contains three steps, pre-processing, attribute extraction and cross-document profile fusion. In pre-processing step, we prepare the input text as system's desired format using some existing text processing tools. In attribute extraction phase, we identify and extract all the related attributes of the target persons in a given text. To fulfill this aim, we use a pattern-matching approach. This method contains a set of keyword-based patterns and manually collected pre-compiled attribute-value candidate lists. In cross-document profile fusion phase, we integrate the distributed information of the focus persons from multiple documents and resolve ambiguity between person names. We evaluate our approach on a sample Farsi textual corpus drawn from news Websites and Wikipedia articles. The experimental results show that our approach is capable to extract personal information, thereby laying the foundation for person named-entity profiling in Farsi.

Keywords: information extraction, named-entity profiling, attribute extraction, named-entity disambiguation, pattern-matching method, Farsi language

1. Introduction

Personal information extraction or in other words, personal name profiling is the process of identifying, extracting and fusing information about the focus person named-entity from the data related to it [1]. Personal name profiling is a well studied problem in the context of Web mining, natural language processing, information extraction and personalized systems. There are wide varieties of applications to which personal name profiling can be helpful, such as personalized

[1] Social Network and Intelligent Systems Laboratory, Faculty of Artificial Intelligence, Malek-Ashtar University of Technology, Tehran, Iran, e-mail: hojjatemami@yahoo.com

[2] Social Network and Intelligent Systems Laboratory, Faculty of Artificial Intelligence, Malek-Ashtar University of Technology, Tehran, Iran, e-mail: shirazi@mut.ac.ir

[3] Intelligent Systems Laboratory, Computer Engineering Department, Amir Kabir University of Technology; Tehran, Iran; e-mail: ahmadaku@aut.ac.ir

[4] Department of Information, Communication & Security, Malek-Ashtar University of Technology, Tehran, Iran, e-mail: mhourali@mut.ac.ir

search [2], [3], information filtering [4], [5], recommender systems [5]–[8], semantic search engines [6], [9], [10], adaptive e-learning systems [11], intelligent tutoring systems [11]–[13], e-commerce applications [14], intelligent information retrieval [15], ads generation [16], active and passive help systems [17]–[19], knowledge management systems [20], [21], etc.

In the literature, various approaches have been developed for the problem of personal name profiling. The bulk of research focused on the extraction of personal information in English or other well-known languages and for some languages such as Farsi there have not been any effort yet. Farsi is one of the rich, less-studied and less resourced languages [22]. Two well-known varieties of Farsi are Tajik and Dari. The majority of the people in Iran, Tajikistan and Afghanistan speak Farsi and its varieties including Tajik and Dari. In total, the Farsi speaking people constituents the 1.5% of the world's population. In spite of the fact that in recent years, the amount of available content in Farsi has increased considerably, systems to analyze Farsi content automatically are not as easily available as they are for other well-known languages. Therefore, in this paper, we develop a personal name profiling system to extract the persons in question and their related information in Farsi content. The problem of personal name profiling is closely related to the slot filling problem studied in information extraction field [23]. However, there are several differences. The problem of personal name profiling contains additional subtasks such as document-internal and cross-document profile fusion. So, many of the methods were developed for slot filling task, are not directly applicable for entity profiling problem.

We propose a three stage approach to automatically extract person named-entity profile in Farsi content. In the first step, we prepare the input text as system's desired format using some existing text processing tools. In the second step, we identify and extract attributes of the focus person named-entities and integrate them into discourse profiles. To extract the personal attributes, we use a pattern-matching approach. This approach is a combination of keyword-based patterns and pre-compiled attribute-value candidate lists. Each keyword-based pattern is designed to extract a particular kind of personal attribute. In the third step, we use two simple but efficient heuristics to resolve ambiguity between person named-entities and integrate the discourse profiles to form corpus-level profiles. The first heuristic is based on a simple string-matching of the focus persons' names or aliases, and the second uses profile similarity method to merge discourse profiles and resolve ambiguity between entities. To summary, our contributions in this article include:

(1) formalizing the problem of personal name profiling as a three step combination approach

(2) assessing a pattern-matching approach to extract entity-centric information in Farsi content

(3) developing two simple but efficient heuristics for cross-document profile fusion and name disambiguation

(4) evaluation of the effectiveness of the proposed approach

Our approach for the problem of personal name profiling is the first effort designed for Farsi. In order to assess the effectiveness of our method we built a sample Farsi textual corpus. Experimental results on the benchmark corpus show that our approach is capable to extract personal information.

After this short introduction, which describes the problem of personal name profiling, its usefulness and the current status for Farsi, the rest of this paper is organized as follows. Section 2 presents a brief survey of state of the art on the topic. Section 3 describes our proposed personal name profiling method for Farsi. Section 4 describes the experimental results on a small Farsi corpus. Finally, Section 5 draws some conclusions, and identifies future work.

2. Related Work

Named-entity profiling is a central task in information extraction. This task aims to distill structured and useful information for the entities in question from a large textual document collection. The task of extracting profiles of various entities such as person, organization and location from textual corpora has been the objective of various researches and attracted researchers from various fields, e.g., relation extraction, question answering and Web people search.

Early named-entity profiling systems are often rule-based system. Rule-based approaches use linguistic patterns developed by humans to match text and extract entity-centric attributes. For example, Li et al. [24] presented a multi-level modular approach to entity profile extraction from English textual content. Their approach relies on various linguistic and information extraction rules. Watanabe et al. [25] proposed a rule-based approach to extract attributes of the people from Web documents. Their approach contains two stages, attribute identification and relevant attribute-value selection. In the attribute identification stage, all potential attribute strings in a given text are marked. In the relevant attribute-value selection stage, the attribute values relevant to a person named-entity are extracted. The advantage of rule-based approaches is that they can achieve good performance on a specific target domain. Rule-based approaches encounter several limitations. They are domain-specific and require human expertise to design appropriate attribute extraction rules.

To overcome the limitations of the rule-based approaches, researchers decomposed the entity profiling problem into several components such as named entity recognition, relation extraction, and profile fusion. To solve these subtasks, researchers have developed statistical, probabilistic, machine learning, data mining and other specific information extraction and retrieval approaches. For example, Tang et al. [26] described a combination approach for Web user profiling. This approach consists of three subtasks including profile extraction, profile integration, and user interest discovery. Specifically, they have proposed a Tree-structured Conditional Random Field (TCRF) to extract the profile information from the Web pages and proposed a probabilistic model to solve the ambiguity between entities. Kristjansson et al. [27] proposed a method to extract the contact information from emails. They developed an interactive information extraction system to assist the user to populate a contact database from emails. Yu et al. [28] proposed a cascade information extraction framework for identifying personal information from resumes. This method contains two steps: a) segmenting the resume into some blocks and assigning labels to them, b) extracting detailed personal information such as name, family, address and email from blocks.

Few named-entity profiling approaches have relaying on the combination approaches. For example, Chen et al. [29] presented a combination method to extract personal information from Web documents. They used a combination of rule-based and machine learning approaches to extract personal attributes. Furthermore, for name disambiguation, they employed a clustering approach, which uses lightweight features to resolve ambiguity between named-entities.

Although each of the proposed methods in the literature aims to extract useful, structured, and valuable information about various entities from textual data, but most of them have focused on a specific entity and on some information about that entity. Existing methods developed for named-entity profiling are restricted to limited number of languages, and for many other languages including Farsi, there has not been any effort yet. Currently, the performance of named entity profiling systems is far from ideal state. Top-score person named-entity profiling systems designed for English textual corpora rarely exceeding 50% F1-score [29]. This indicates that personal named-entity profiling is a challenging task. In this paper, we introduce an approach for person named-entity profiling in Farsi, which is described in the following sections.

3. Our Proposed Method

In this paper, we study a specific variant of the general entity profiling problem, namely person named-entity profiling in Farsi content. There are three subtasks in the proposed procedure of extracting the personal named-entity profile, including pre-processing, entity-centric attribute extraction and cross-document profile fusion. Pre-processing aims to prepare the input text as system’s desired format. Attributes extraction phase attempts to identify entity-centric information in the form of attributes and valid values for the focus entity. To fulfill this aim, we propose a pattern-matching method, which contains a set of keyword-based patterns and pre-compiled attribute-value candidates. The list of pre-compiled attribute-value candidates is manually drawn from knowledge resources on the Web such as Wikipedia[5]. Cross-document profile fusion task aims to resolve *polysemy* and *synonymy* problems. In the parlance of person named-entity profiling, the *polysemy* means the same name can refer to multiple discourse named-entity profile, and the synonymy means the same person can have multiple profiles that should be aggregate into an integrated corpus-level profile. To integrate entity-centric information from multiple documents and resolve ambiguity between named-entities, we propose two simple heuristics: string-matching method and profile similarity method. In the string-matching approach, different profiles belong to a same person are integrated using a simple string matching of the persons' names or aliases. In profile similarity method, merging different discourse profiles is based on the information constituents the personal profiles. Fig. 1 shows an overview of our approach and following sections explain the components of our proposed method in more details.

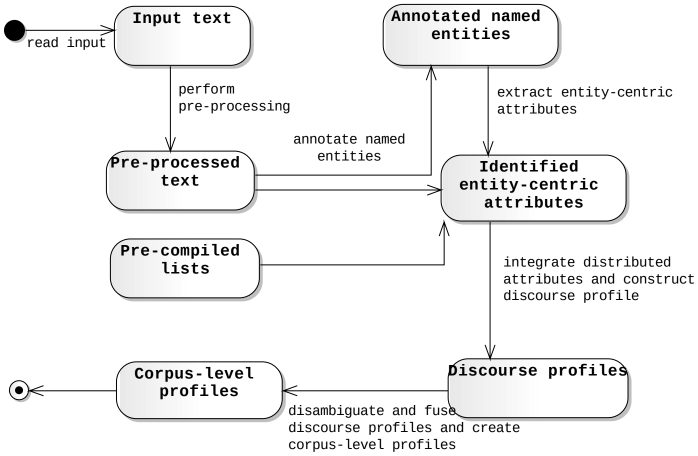

Fig. 1. Overview of our proposed approach for the problem of person named-entity profiling

A. Pre-processing

Pre-processing is the starting point of the entity profiling system. The task of pre-processing is to prepare the input text according to the system's desired format. In pre-processing phase, the system, determines boundaries of sentences, extracts part of speech tags, phrase chunks, and lemmas of the sentences' tokens, dependency labels and other necessary information from input documents using existing tools. We then use a named entity tagger to annotate named entities from the text and classify them into sets of predefined types of named entities such as person, place and organization. A high performance named entity tagger is crucial for entity profiling system since accurately identifying named-entities is anchoring point for extracting entity profiles. We adopt an existing named entity

[5] www.wikipedia.org

tagger system [30] with performance around 73% precision/recall combined F1-score.

B. Attribute extraction

We developed a pattern-matching method to extract some attributes of person named-entities. Our method attempts to extract 10 kinds of personal information from the text. The list of attributes is given in Table 1. Our pattern-matching method for extracting personal attributes generally works as follows: first a set of keyword-based rules are defined. These rules are used to match the text and locate personal information unites. Taking several rules, our algorithm iteratively extracts more attributes. Each rule is designed to extract one particular type of personal attribute. Each rule consists of a pattern and an action. The pattern is a regular expression defined to identify personal attributes. Each pattern to extract a specific attribute of the focus person uses a set of pre-compiled attribute value candidates. The list of candidate keywords are manually created using resources available on the Web such as Wikipedia.

Table 1

The list of 10 attributes in the person named-entity profile

No.	Attribute	Possible value
1	Birth place	NE(Location)
2	Date of birth	NE(Time)
3	Death place	NE(Location)
4	Date of death	NE(Time)
5	Nationality	NE(Location)
6	Affiliation	Noun Phrase
7	Relatives	NE(Person)
8	Occupation	Noun Phrase
9	Email	NE(Email)
10	URL	NE(URL)

The text is then compared against the patterns and when a pattern matches a sequence of words in a sentence of the text (i.e., when one of keywords regarding the target attribute appears in a sentence), the specified action fires and extracts appropriate value to the target attribute. For example, to extract the value for attribute “occupation”, we create a list of occupations using Wikipedia[6]. We then compare the text against the entries of occupation list. There is a relationship between an occupation attribute value and a focus person named-entity only when one of the entries in occupation list and that person named-entity co-occur in a sentence of the text. For example, given the following sentence:

“Alê (Ali) ostade-e (professor) daneshgah-e Tehran (University of Tehran) ast (is)

Translation: Ali is the professor at university of Tehran.

[6] https://en.wikipedia.org/wiki/Lists_of_occupations

The following sample rule concludes that the noun phrase “*ostade (professor)*” is a possible value for occupation attribute of the entity *Alê (Ali).*

$$NE1(Person) + NP1(Occupation) + NE2(Organization)$$
$$\rightarrow Occupation(NE1, NP1)$$

Where *NE* and *NP* are abbreviates for *named-entity* and *noun phrase*. This simple pattern connects a person named-entity (*NE1*) with a noun phrase (*NP1*) with the “occupation” relationship. It also extracts an *“affiliation”* attribute between *NE1 (Person)* and *NE2 (Organization).*

When extracting attributes, we aggregate the extracted attributes into discourse profiles. In the profile integration process, for constructing discourse profile for each person named-entity, his/her related attributes are assembling while merging redundant attributes.

C. Cross-document profile fusion

A vital problem, which increases the accuracy of profiling system, is cross-document profile fusion. The purpose of cross-document profile fusion is to determine if multiple extracted profiles belong to the same person, resolve ambiguous named-entities, and makes an exactly one to one correspondence between entities and their profiles. In the literature, various approaches have been proposed for solving cross-document profile fusion. Some of the most popular methods are supervised classification and various unsupervised clustering techniques. Here, we used to simple but efficient heuristics for solving cross-document profile fusion problem.

- In the first approach, different profiles belong to a same person are aggregated using a simple string matching of the persons' names or aliases. In this case, there are a set of person names and the goal is to separate irrelevant names and conversely merging relatively same names. To fulfill this aim, a similarity matrix M_d is defined. Each element of the matrix shows the degree of similarity between any two persons' names. Edit distance metric is used to compute the similarity between names. While merging process, if the similarity value between any two names in M_d matrix is greater than an empirical stop-threshold, then these two names are merging. The stop threshold directly affects the robustness of name disambiguation task. In order to find the optimal stop threshold, we run the grouping process with all possible stop thresholds, and choose the one, which has the best performance as the optimal stop threshold. In our implementations, we set the threshold value to 0.70.
- In the second approach, merging different discourse profiles is based on the information constituents the profiles. In this approach, all extracted attributes of an entity can be used to distinguish or merge different profiles. For

instance, given the profile of two entities *"Mr. John Minca"* and *"Mrs. Minca"*, if the "gender" property was extracted for these two entities, then these entities would not be merged. Although, the success of this method is strictly depends on the existing important attributes in the text. To compute the similarity of two discourse profiles, we define the profile similarity measure as follows:

$$S(P_i, P_j) = \frac{1}{|A|} \times \sum_{k \in A} \begin{cases} W_k(P_i, P_j) & if \quad (C(P_i, P_j) > \alpha) \\ 0 & otherwise \end{cases} \tag{1}$$

where, A is the total number of attributes between two target profiles P_i and P_j, and $W_k(P_i, P_j)$ is the similarity of kth attribute-value pair. We only compute the similarity if profiles' coverage (common seen patterns between two profiles, $C(P_i, P_j)$), is at least α patterns. This can emphasize the precision of name disambiguation and improve performance of the cross-document profile fusion task. We set α to 2 in order to ignore poor-content profiles. The attributes on the profiles may be single-value attribute (e.g., gender) or multiple-value attribute (e.g., education). In order to compute $W_k(P_i, P_j)$, we first compute single-value similarities for all possible patterns between two target profiles and pick the maximum valued ones. $W_k(P_i, P_j)$ is calculated as follows:

$$W_k(P_i, P_j) = \frac{1}{\min\left(\left|\vartheta(P_i)\right|, \left|\vartheta(P_j)\right|\right)} \times \sum_{s \in \vartheta(P_i)} \max\left(\delta\left(s, \vartheta(P_j)\right)\right) \tag{2}$$

where, δ is the set of single-value pattern similarities calculated between an element $s \in \vartheta(P_i)$ and all elements in $\vartheta(P_j)$. $\delta(s, \vartheta(P_j))$ is defined as follows:

$$\delta\left(s, \vartheta(P_j)\right) = \left\{\varphi(s, l) \middle| l \in \vartheta(P_j)\right\}, \quad \begin{cases} 1 & if\ (s == l) \\ 0 & otherwise \end{cases} \tag{3}$$

In Eq. (3) $\varphi(s, l)$ gives 1 for two identical elements and 0 for non-identical elements.

4. Experiments and Results

A. Benchmark dataset

For evaluating the effectiveness of our proposed approach, we built a suitable Farsi textual corpus with annotated profiles. Documents in the corpus randomly have been drawn from Tabnak[7] and the Entekhab[8] news Websites with the date range 1 Dec. 2013 through 20 Feb. 2014, and some articles about well-known persons (such as popular athletes and politicians) have been taken from

[7] http://www.tabnak.ir/

[8] http://www.entekhab.ir/

Wikipedia. The documents in the corpus have been chosen from diverse topics including politic (50 news instances), economic (50 news instances), sport (50 news instances), and social (50 news instances). The corpus is saved collectively in a relational database.

To evaluate the performance of our approach, four human annotators are conducted to build an annotated gold-standard set based on the benchmark corpus. For disagreements in the annotation, we conducted "majority voting". The gold-standard set consists of identified person profiles associated with their attribute (slots) and fillers. The gold-standard set contains 68 profiles each of them related to a unique person, with the average ~4.28 slots per each profile. The structure of gold-standard set is a matrix $M_{S\times 2}$, where S indicates the total number of attributes in the gold- standard set, that each slot consists of two parameters, the first for slot name and the latter for slot filler.

B. Performance measures

The quality of our approach is measured by the following three criteria: *precision*, *recall* and *F1-measure* [31]. We defined these metrics in the following.

1) Precision

This metric is defined as follows [32]:

$$precision = \frac{\#correct}{\#correct + \#incorrect} \tag{4}$$

where *#correct* is the number of correct slots extracted by the system and *#incorrect* is the number of incorrect slots among all extracted slots. A slot is said to be filled correctly, if it is align with a slot in the annotated gold-standard set and if it is assigned a valid value.

2) Recall

Let *#total* denote the total number of slots expected to be filled according to the gold-standard set, then the recall measure can be defined as follows [32].

$$recall = \frac{\#correct}{\#total} \tag{5}$$

3) F1-measure

This metric is another well-known evaluation measure, which uses the ideas of precision and recall from information retrieval. This metric is simply is defined as follow.

$$F_{\beta} - measure = \left(\beta^2 + 1\right) \times \frac{precision \times recall}{\left(\beta^2 \times precision\right) + recall} \tag{6}$$

where β is a non-negative value, which used to adjust precision and recall relative weighting. In our work, β is set to 1, which gives equal weight for both precision and recall.

C. Numerical example

The proposed method is implemented using Java programming language on a Pentium® Core TM 2 Quad CPU 2.66 GHz with 2.00 GB RAM. Obviously, the quality of profiling system is directly affected by its related components including pre-processing, entity-centric attribute extraction and cross-document profile fusion. This means that the higher the performance of the system's components is, the higher the total efficiency of the profiling system is. For the sake of evaluating our approach, we devise two evaluation levels, component-level evaluation and system-level evaluation. In component-level, we evaluate the performance of attribute extraction component, and in system-level evaluation, we evaluate the total efficiency of the system.

1) Evaluation for attribute extraction

The performance of attribute extraction phase determines the performance in the following phases of the profiling system. We randomly pick *n* (*n=5*) instances of each of the attribute class for manual checking the performance of the attribute extraction component. This process is repeated three times. Finally, the performance measures are calculated as the average of three times. We note that the target attributes at this stage are those collected in discourse named-entity profiles. Table 2 shows attributes and the obtained performance scores by the attribute extraction phase.

Table 2

Performance of attribute extraction component

Attribute Class	Precision (%)	Recall (%)	F1-score (%)
Email	34.1	30.5	32.2
URL	32.5	29.7	31
Birth place	27.1	20.3	23.2
Date of birth	25.6	17.35	20.7
Death place	23.4	19.7	21.4
Date of death	18.4	17.2	17.8
Nationality	15.8	15.5	15.6
Relatives	13.9	12.2	13
Occupation	11.8	10.6	11.2
Affiliation	6.1	5.4	5.7
Overall	20.9	17.8	19.2

Using the pattern-matching algorithm for attribute extraction, the F-score is reached to [5.7%-32.2%], the precision is reached to [6.1%-34.1%]; nevertheless, the recall is reached to [5.4%-30.5%]. As shown in Table 2, the attributes of *"email"*, *"URL"*, *"birth place"*, *"date of birth"*, *"date of death"*, *"death place"* have achieved relatively good performance, because their candidate values are expressed by fix and easily predictable patterns. In contrast the attributes of *"affiliation"*, *"relative"*, *"occupation"*, and *"nationality"* have not achieved good results, because their instances can take various forms in both vocabulary and structure in different domains. Such attributes cannot easily be extracted by our pattern-matching method. The low score for some attributes such as *affiliation and occupation* is partially due to the requirement of better development of patterns. With defining patterns that are more precise, more attributes can be discovered; subsequently the system performance can improve.

2) System performance

Table 3 gives system performance obtained for each of the personal attributes based on the different cross-document profile fusion techniques. An automatic scorer module is used to measures performance metrics based on a gold-standard set. From the table we can say that using the name matching method, the F1-score can reach to [6.3%-33.2%].

Table 3

Performance of our profiling approach (in terms of F1-score) with different cross-document profile fusion techniques

Attribute Class	Profile fusion using string-matching	Profile fusion using profile similarity
Email	33.2	35.3
URL	32.3	33.2
Birth place	23.6	25.3
Death place	22	26.1
Nationality	18.8	16.5
Date of birth	17.5	19.8
Date of death	16.3	17.8
Relatives	12.6	13.7
Occupation	11.3	12.4
Affiliation	6.3	6.9
Overall	19.4	20.7

Given this result we can conclude the name variation phenomenon is more serious than that of name ambiguity. Using profile similarity in profile fusion process, the performance is improved significantly. In this case, the precision/recall combined F1-score is reached to [6.9%-35.3%]. This means that profile fusion based on profile's content resolves the problem of name variation by addressing the whole profile attributes in fusion process. It can be seen that profile similarity approach is more reliable than name string-matching approach.

Personal information extraction has proven very challenging. Our proposed approach can achieve as high as 35.3% F1-score, considering our method as being the first effort for personal named-entity profiling in Farsi. However, the system performance is far from ideal state, and further efforts are need to improve the results. Our approach decomposes the problem of person named-entity profiling into three subtasks include pre-processing, attribute extraction and cross-document profile fusion/discrimination. It is obvious that the quality of personal name profiling system is directly affected by its relevant components. This means that how much pre-requisite components to be implemented with higher performance, the total efficiency of the system increases greatly. Even with the low F1-score obtained for personal attributes, the profiling system demonstrates enormous value in gathering information about entities. One of the main challenges that decrease the system performance is that the filler values for the majority of attributes in person profile can be expressed in various forms and structures in the text. The keyword-based pattern matching approach that we used to extract personal attributes cannot completely identify various forms and structures of attributes. In our approach, we used pre-compiled attribute value lists to mark potential attributes in text. These lists are not complete, because the values that attributes such as affiliation and occupations can take are open sets and we cannot enumerate all the candidate attribute-values using pre-compiled lists. Moreover, using pre-compiled lists brings name ambiguity problem. In order to increase the system performance, it is need to spend more time on the development of more precise attribute extraction patterns, or use another set of patterns that do not depend to pre-compiled lists.

5. Conclusions

In this paper, we have addressed the problem of extracting person named-entity profiles in Farsi content. We formalized the problem as three subtasks including pre-processing, entity-centric attribute extraction and cross-document profile fusion. After pre-processing phase and preparing the input text as system's desired format, the attributes of entities are extracted using a pattern-matching technique. Then, identified attributes are fused into several discourse profiles. Finally, related (distinct) person named-entities' profiles are fused (separated) by cross-document profile fusion heuristics. Experiments on Farsi textual corpora show that our method is able to find and extract structured personal information.

One of the main drawbacks of our proposed method is that it requires more human labor to design attribute extraction patterns. Another drawback is that our method ignores semantic information in the text, and don't consider dependency between attributes. These problems decrease system’s performance. For future work and to overcome these problems, we intent to use semantic-based machine learning techniques for extracting personal named-entity profiles.

REFERENCES

[1] I. Zukerman and D. W. Albrecht, “Predictive statistical models for user modeling,” User Model. User-adapt. Interact., vol. 11, no. 1–2, pp. 5–18, 2001.

[2] S. Calegari and G. Pasi, “Definition of User Profiles based on the YAGO Ontology,” in In M. Melucci, S. Mizzaro & G. Pasi (Eds.), IIR, CEUR Workshop Proceedings, CEUR-WS.org, 2011.

[3] M. Daoud, L. Tamine, and M. Boughanem, “A personalized graph-based document ranking model using a semantic user profile,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6075 LNCS, pp. 171–182.

[4] E. Michlmayr, S. Cayzer, and P. Shabajee, “Add-A-Tag: Learning adaptive user profiles from bookmark collections,” in 1st International Conference on Weblogs and Social Media (ICWSM’2007), Boulder, Colorado (USA), 2007.

[5] H. N. Kim, I. Ha, K. S. Lee, G. S. Jo, and A. El-Saddik, “Collaborative user modeling for enhanced content filtering in recommender systems,” Decis. Support Syst., vol. 51, no. 4, pp. 772–781, 2011.

[6] X. Zhou, W. Wang, and Q. Jin, “Multi-dimensional attributes and measures for dynamical user profiling in social networking environments,” Multimed. Tools Appl., pp. 1–14, 2014.

[7] C.-C. Hung, Y.-C. Huang, J. Y. Hsu, and D. K.-C. Wu, “Tag-based user profiling for social media recommendation,” in Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI, 2008, pp. 49–55.

[8] X. Zhou, Y. Xu, Y. Li, A. Josang, and C. Cox, “The state-of-the-art in personalized recommender systems for social networking,” Artif. Intell. Rev., vol. 37, no. 2, pp. 119–132, 2012.

[9] K. Sugiyama, K. Hatano, and M. Yoshikawa, “Adaptive web search based on user profile constructed without any effort from users,” in Proceedings of the 13th conference on World Wide Web - WWW ’04, 2004, pp. 675–684.

[10] A. Sieg, B. Mobasher, and R. Burke, “Web search personalization with ontological user profiles,” in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007, pp. 525–534.

[11] C. Froschl, “User modeling and user profiling in adaptive e-learning systems,” Graz University of Technology, A-8010 Graz, Austria, 2005.

[12] A. C. Martins, L. Faria, and C. V. De Carvalho, “User Modeling in Adaptive Hypermedia Educational Systems,” Educ. Technol. Soc., vol. 11, pp. 194–207, 2008.

[13] P. García, A. Amandi, S. Schiaffino, and M. Campo, “Evaluating Bayesian networks’ precision for detecting students' learning styles,” Comput. Educ., vol. 49, no. 3, pp. 794–808, Nov. 2007.

[14] G. Adomavicius and A. Tuzhilin, “Using data mining methods to build customer profiles,” Computer (Long. Beach. Calif)., vol. 34, no. 2, pp. 74–81, 2001.

[15] P. Mylonas, D. Vallet, P. Castells, M. Fernández, and Y. Avrithis, “Personalized information retrieval based on context and ontological knowledge,” Knowl. Eng. Rev., vol. 23, no. 01, pp. 73–100, 2008.

[16] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola, “Scalable distributed inference of dynamic user interests for behavioral targeting,” in Proceedings of the 17th

ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 114–122.

[17] J. Fink and A. Kobsa, “User modeling for personalized city tours,” Artif. Intell. Rev., vol. 18, no. 1, pp. 33–74, 2002.

[18] F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, and F. Vernero, “Tag-based user modeling for social multi-device adaptive guides,” User Model. User-Adapted Interact., vol. 18, no. 5, pp. 497–538, 2008.

[19] T. Kuflik, C. Callaway, D. Goren-bar, C. Rocchi, O. Stock, and M. Zancanaro, “Non-intrusive User Modeling for a Multimedia Museum Visitors Guide System,” Um 2005, Lnai 3538, pp. 236 – 240, 2005.

[20] Y. Sure, A. Maedche, and S. Staab, “Leveraging Corporate Skill Knowledge-From ProPer to OntoProPer,” in Proceedings of the Third International Conference on Practical Aspects of Knowledge Management (PAKM2000), 2000, pp. 1–9.

[21] L. Razmerita, A. Angehrn, and A. Maedche, “Ontology-based user modeling for knowledge management systems.,” User Model. 2003. Springer Berlin Heidelb., pp. 213–217, 2003.

[22] M. Shamsfard, “Challenges and open problems in Persian text processing,” in Proceedings of 5th Language & Technology Conference (LTC), 2011, pp. 65–69.

[23] B. Min and R. Grishman, “Challenges in the Knowledge Base Population Slot Filling Task,” in Evaluation, 2010, pp. 1137–1142.

[24] W. Li, R. Srihari, C. Niu, and X. Li, “Entity profile extraction from large corpora,” in Pacific Association for Computational Linguistics Conference (PACLING), 2003.

[25] K. Watanabe, D. Bollegala, Y. Matsuo, and M. Ishizuka, “A Two-Step Approach to Extracting Attributes for People on the Web,” in 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[26] J. Tang, L. Yao, D. Zhang, and J. Zhang, “A combination approach to web user profiling,” ACM Trans. Knowl. Discov. from Data, vol. 5, no. 1, pp. 2–44, 2010.

[27] T. Kristjansson, A. Culotta, P. Viola, and A. Mccallum, “Interactive Information Extraction with Constrained Conditional Random Fields,” in In AAAI, 2000, pp. 412–418.

[28] K. Yu, G. Guan, and M. Zhou, “Resume Information Extraction with Cascaded Hybrid Model,” in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL ’05, 2005, no. June, pp. 499–506.

[29] Y. Chen, S. Y. Mei Lee, and C. R. Huang, “A robust web personal name information extraction system,” Expert Syst. Appl., vol. 39, no. 3, pp. 2690–2699, 2012.

[30] P. S. Mortazavi and M. Shamsfard, “Named Entity Recognition in Persian Text,” in 15th Annual Conference of Computer Society of Iran, Tehran, Iran, 2009.

[31] D. M. W. Powers, “Evaluation: From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation, Technical report SIE-07-001, School of Informatics and Engineering, Flinders University, Australia,” 2007.

[32] J. Piskorski and R. Yangarber, “Information Extraction: Past, Present and Future,” in T. Poibeau et al. (eds.), Multi-source, Multilingual Information Extraction and Summarization 11, Theory and Applications of Natural Language Processing, Springer-Verlag Berlin Heidelberg, 2013, pp. 23–50.