Which Who are They? People Attribute Extraction and Disambiguation in Web Search Results*

Man LAN
Institute for Infocomm Research, Singapore†
East China Normal University§
mlan@i2r.a-star.edu.sg†
mlan@cs.ecnu.edu.cn§

Yu Zhe ZHANG
Department of Computer Science and Technology
East China Normal University
zyzesty@gmail.com

Yue LU
Department of Computer Science and Technology
East China Normal University
ylu@cs.ecnu.edu.cn

Jian SU
Institute for Infocomm Research, Singapore
sujian@i2r.a-star.edu.sg

Chew Lim TAN
School of Computing, National University of Singapore
tancl@comp.nus.edu.sg

ABSTRACT

People name search often returns a lot of Web pages containing the strings of personal names. Due to namesake, extracting target person attributes (such as *birthday*, *occupation*, *affiliation*, *nationality*, *contact information*, etc.) is expected to be helpful to differentiate documents related to different people and thus group documents related to the same person. This paper presents the methodology for the two tasks of Web people disambiguation: target person Attribute Extraction (AE) and people Clustering. Specifically, in this paper we address three questions: (1) How to effectively extract target person attribute information from raw Web pages? (2) Is the information of extracted attributes able to lead to better performance than the information of raw Web pages for Web people clustering? (3) Which is important for Web people clustering, feature representation or clustering algorithms? To solve them, we first present an effective method to extract different types of target person attributes from raw Web pages by using deep Web page cleaning and processing pipelines with multiple techniques including traditional named entities recognition (NER), regular expression patterns, gazetteer-based matching and so on. Then we explore the methodology for Web people clustering from two aspects, i.e., feature representations (tokens from raw Web page, information of extracted attributes) and clustering strategy. The comparative experimental results showed that deep Web page cleaning contributes significantly to performance improvements for target person attribute extraction task. For people clustering task, the clustering algorithm contributes more to performance improvement than feature representations.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2 [**Software**]: Software Engineering; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Keywords

Web People Search, people disambiguation, clustering, people attribute extraction, deep Web cleaning

1. INTRODUCTION

People name search in the World Wide Web is one of the most common activities of Internet users: around 30% of search engine queries include person names [1]. Due to namesake, a lot of returned Web pages containing the strings of personal names may not refer to the same person. In order to differentiate documents related to different people and thus group documents related to the same person, WePS 2007 [2] was the first competitive evaluation which focused on this people clustering problem. In fact, every person has his own attributes, such as *birthday*, *occupation*, *affiliation*, *nationality*, *contact information*, etc. Intuitively, these attributes are very important clues for people disambiguation. Therefore, it is expected that extracting people attributes from each Web page containing the strings of personal names would be helpful to distinguish documents related to different people and cluster documents related to the same person. For this reason, WePS 2009 proposed two tasks, i.e., target person Attribute Extraction (AE) task [3] which is to extract 16 kinds of attributes values for target individuals on the provided Web pages, and people Clustering task [4] which is to group Web pages to the same person. The Web pages are distributed in their original format (i.e., html).

Generally speaking, the AE task adopts traditional information extraction (IE) and named entities recognition (NER) techniques, but it goes beyond them. The most significant difference comes from the rich resource of formats on raw Web pages, for example, HTML tags, script codes, CSS, and noisy content, for example, Web advertisements (including contextual Ads, banner Ads, Rich Media Ads,

*(Produces the WWW2009-specific release, location and copyright information). For use with www2009-submission.cls V1.4. Supported by ACM.

WWW2009, April 20-24, 2009, Madrid, Spain.
.

etc,), irrelevant links and even fraud anchor words. In most cases, even though the Web pages contain several mentions of these people attributes, they are not related to the target person. On all accounts, we can say that the attribute extraction for target person in general raw Web pages is much more complex than those in plain text. Therefore, the first fundamental question arises here, i.e. "How to effectively extract target person attribute information from raw Web pages?"

At the same time, as we mentioned above, although some people have the same name, they have their own specific traits (or attributes), such as *birthday*, *occupation*, *affiliation*, *nationality*, *contact information*, etc. Consequently, these attributes would be expected to serve as very important clues for people disambiguation and be helpful to group Web pages to different person. Therefore, a second question surfaces here :"Is the information of extracted attributes able to lead to better performance than the information of raw Web pages for Web people clustering?"

Generally, for clustering task, there are two important issues: feature representation and clustering algorithm. Typically, the clustering performance can be improved from these two aspects. However, for Web people clustering, to the best of our knowledge, no comparison studies of the two aspects have been done before. Even given the previous studies in WePS 2007 [2], we could not definitely draw a conclusion as to which dominates the performance of Web people clustering. Thus, a third question emerges here:"Which is important for Web people clustering, feature representation or clustering algorithms?" In this paper, we bridge this gap by performing the comparison study between feature representation and clustering algorithms.

Therefore, the purpose of this study is to address the above three questions. To solve them, in this paper we present the methodology for the two tasks of WePS 2009 Evaluation: target person Attribute Extraction (AE) and people Clustering. Specifically, to address the first question, we extract different types of attributes by adopting processing pipelines with multiple techniques including named entities recognition (NER), regular expression patterns, gazetteer-based matching, manually-constructed rules based on shallow and deep Web pages cleaning. Regarding the last two questions, we examine the performance of people clustering in terms of different feature representations (i.e. tokens from raw Web pages, clean tokens from AE results and their integration) and different clustering strategies on two benchmark corpora, i.e. WePS 2007 corpus and WePS 2009 corpus.

The rest of the paper is structured as follows. Section 2 presents the methodology of the two tasks. Section 3 reports the experimental results and discussion. Finally, Section 4 summarizes our concluding remarks and suggests the future work.

2. METHODOLOGY

2.1 Web Page Cleaning

Traditional information extraction (IE) and named entities recognition (NER) techniques have performed well in recognition of Names, Organization, Location entities, etc. Intuitively, explorations would be done to port existing IE or NER systems into this people attribution extraction (AE) task. However, few of these traditional techniques achieved satisfactory performance on raw Web pages. Unlike conventional data or plain text, Web pages typically have a large amount of information that is not part of the main contents of the pages, including: (1) rich resource of formats and functional codes, e.g., HTML tags, script codes, CSS, etc, (2) irrelevant and noisy information, e.g., contextual Ads, navigation banner, Rich Media Ads, copyright notices, and even fraud anchor words with links, etc, (3) confusing information, e.g., in most cases, even though the Web pages contain many mentions of people attributes, they are not relevant to the target person, for example, the *email* addresses of the web masters, friends, colleagues, or even other person who make comments in this Web page, etc. Therefore, attribute extraction for target person in general raw Web pages is much more complex than those in traditional domain, eg newswire. All such irrelevant information in Web pages, i.e. *Web page noise*, can seriously harm the two tasks of WePS 2009, i.e. AE and Clustering. Therefore, the first and important step is to do *Web page cleaning*, which is very challenging since it is to decide which content of the page is meaningful and relevant to the target person and which is noisy.

In recent years, several similar research studies have been done on Web page cleaning, including detection of informative blocks in Web pages [5], detection of a frequent template or patterns of Web pages [6], and assignment of different weights to different blocks in Web pages [7], etc. However, these methods have different limitations on their own. For example, [5] hold two strong assumptions: (1) the system knows a *priori* how a Web page can be partitioned into coherent content blocks; and (2) the system knows a *priori* which blocks are the same blocks in different Web pages. In [6], the partitioning of a Web page is pre-fixed. The work in [7] is based on the observation that in commercial web site Web pages follow some fixed layouts and presentation styles. All these methods are not suitable for this AE task since our purpose is to do an automatical Web page cleaning without any *priori*. As a result, we implement the Web page cleaning work by ourselves as follows.

Firstly, like most commonly-used HTML cleaning tools which only concentrates on data extraction from Web pages, we do a shallow Web page cleaning as follows:

- repair missing or non-closed tags;
- strip away all HTML tags and script codes as many as possible;
- remove the content between several pairs of tags, such as <select> and </select>, <style> and </style>, etc.;
- extract the content between the *title*, the *body* and the anchor tags for each page.

However, the resulting Web pages still contain a lot of noise. Therefore, besides the above shallow cleaning operations, we do a further deep cleaning work as follows:

- some HTML tags are replaced by white space (such as <p>, <td>), while others are converted into line separators (such as <li>,
, <tr>);
- remove content between a pair of tags and controllers, such as ListView, ListBox, ComboBox, etc.;

If all the Web pages are written form the same template, it is easy to do the structure analysis and remove the irrelevant content from the header, the left navigation bar and the footer of the Web pages. However, these Web pages with various templates are returned from the current search engine and downloaded from the Internet directly. It is not practical to manually clean these Web pages. In order to simplify the recognition of these noisy content, we also include the deep cleaning work as follows:

- remove all non-*url* textual contents between a pair of anchors for the purpose of removing advertisements;
- remove all textual content after "*copyright*" keywords for the purpose of removing footer information.

Note that the above Web page cleaning is the first step of the succedent two tasks of WePS 2009, i.e., the AE and Clustering tasks. That is, these two tasks are performed on the resulting documents of Web page cleaning as mentioned above.

2.2 Attribute Extraction for Target Person

The AE task is to extract attributes for target person from raw Web pages. These attributes are empirically defined and selected by organizers so that they have to be general enough to cover most people, useful for the disambiguation, and meaningful for the evaluation. Table 1 lists these 16 types of people attributes at WePS 2009.

Table 1: Definition of 16 attributes of person at WePS 2009

Attribute Class	Examples of Attribute Value
Date of birth	4 February 1888, 7th August
Birth place	Brookline, Massachusetts
Other name	JFK
Occupation	Politician, Editor
Affiliation	University of California, Los Angeles
Award	Pulitzer Prize
School	Stanford University
Major	Mathematics
Degree	Ph.D.
Mentor	Tony Visconti
Nationality	Amercian
Relatives	mum's name
Phone	+1(111)111-1111
FAX	(111)111-1111
Email	xxx@yyy.com
Web site	http://nlp.cs.nyu.edu

Roughly, these attributes can be grouped into four categories according to the type and the number of their values. Table 2 lists the different types of these attributes and their corresponding potential extraction method.

Generally, we extract attribute candidates by using processing pipelines with multiple techniques including traditional NER, regular expression patterns, gazetteer-based matching, manually-constructed rules and so on. However, we have to keep in our mind that the amount of attribute candidates extracted using the above methods is huge, which means the recall measure of these extracted attributes is high. Therefore, for different attributes, we may need different strategies to make filtering and disambiguation for the purpose of improving precision measure. For example, we may build a stop list for *email* filtering and remove the noisy email address, such as "*webmaster@xxx*", etc. Specifically, for different types of attributes, different extraction and disambiguation methods have been adopted in our work as follows.

Table 2: The attribute value type and corresponding potential extraction method

Category	Attributes	Method
typical pattern	Phone FAX Email Web site Date of birth	regular expression patterns
NER with limited candidates	Degree Nationality Major Occupation	Gazetteer-based matching
traditional NER (including location, person, organization)	Birth place Affiliation School Other name Mentor Relatives	traditional NER tool
Freely NER	Award	

Regarding the four enumerable attributes, i.e., *Occupation*, *Major*, *Degree*, *Nationality*, a gazetteer (dictionary) is constructed from public resources (such as *wikipedia*). Then we use a simple dictionary matching algorithm to extract these attributes values.

With regard to the 7 attributes, i.e., *Birth place*, *Affiliation*, *School*, *Mentor*, *Relatives*, *Email*, *Web site*, we first adopt a NER tool named ESpotter [8] to extract names, organizations, locations, emails and urls from Web pages. Then for each attribute, we examine if the corresponding attribute trigger keywords are available in the same sentences. For example, to check if the location entity is a *birth place* for the target person, we examine if there is a attribute trigger keyword, such as "born", "birth place", etc, in the same sentence with the location entity. Specifically, for *Email* and *Web site*, we also construct a stop list for filtering, including the very common values such as "*webmaster@xxx*", "*wikipedia*","*wiki*", and so on.

Regarding *Date of birth*, we first construct several regular expression patterns to recognize textual or numerical date patterns. Then we check if the current sentence has attribute trigger keywords, such as "*born*", "*birthday*", "*birth date*", "*birth*", etc, and has a referent for the target person as well. If yes, then this date is recognized as a *Date of birth* for the target person.

With respect to *Other name* attribute, we generate several name regular expression patterns to extract different types of other names for the target person, such as first name alone, last name alone, capitalized first letter from the first name or last name in combination with capitalized letters and the names, etc.

For *Phone* and *FAX* attributes, it is easy to extract them by using regular expression patterns. Then we further identify if it is a phone or fax number by examining the relative locations of these trigger keywords including *phone*, *fax*, *con-*

tact, *dial*, and so on, in the same sentence.

Recognizing the *Award* attribute is the most challenging task due to the great diversity of expressions in natural language. We explore a lot of methods and no one performs well on the extraction of *Award* attribute.

2.3 People Clustering

The people Clustering task is to distinguish documents related to different people and cluster documents related to the same person. For this task, there are two issues, i.e., feature representation and clustering algorithm.

2.3.1 Feature Representation

Based on resulting documents after deep Web page cleaning, we adopt two kinds of token representations. The first set of tokens are extracted from the raw Web pages, which we call "*Raw token*" set. The second set of tokens are obtained only from the extracted attributes for target person, which we call "*AE token*" set.

In order to obtain *Raw token* set, we first remove stop words (512 stop words), numerals and punctuations from clean Web pages. Then the Porter's stemming [9] is performed to reduce words to their base forms. We also set minimal term length as 3 (ie, each token has 3 letters at least). Finally, by using χ^2 metric, the top $200 - 400$ features are selected.

Apparently, the feature size of the *AE token* set is much smaller than that of *Raw token* set since a lot of irrelevant content has been removed from raw Web pages and only attributes information extracted has been retained. Therefore, there is no necessary to perform removing stop words and feature selection steps since their purpose is to decrease the feature set size. We thus only extract stemmed words (tokens) from the extracted attributes for target person and adopted them to represent the content of Web pages. Consequently, some Web pages would be converted into null vectors if there is no person attribute extracted from them.

2.3.2 Clustering Algorithm

The clustering algorithms have been widely studied by a lot of researchers for several decades. In this paper, we simply adopt the most widely-used *K-means* algorithm and usually the k value is set whether to be the minimal number which makes the clustering convergent if larger than 2 or to be 2.

2.4 Performance Evaluation

Regarding the AE task, we adopt the traditional *precision*, *recall* and F_1 measure for each individual attribute and for the overall answers to evaluate the performance of attribute extraction. Specifically, *recall* is defined as the number of correctly identified attribute values by system divided by the total number of attribute values in golden data. *Precision* is defined as the number of correctly identified attribute values by system divided by the total number of attribute values the system produced. F_1 function attributes equal importance to *precision* (p) and *recall* (r) and it is computed as:

$$F_1 = \frac{2 * p * r}{p + r} \quad (1)$$

For the general clustering task, there are several different measures to evaluate system performance. In order to overcome the limitations of standard clustering measures when dealing with overlapped clusters, WePS 2009 adopts the extended B-Cubed clustering measure [10] according to feedback from WePS 2007.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this paper, we conduct three series of experiments under various experimental circumstances to address the three questions raised in Section 1.

The main purpose of the first series of experiments is to address the first question: "How to effectively extract target person attribute information from raw Web pages?" A Gold Standard training corpus for AE task is provided by WePS 2009 [11], which consists of 17 names from the WePS 2007 dataset and has been carefully annotated by organizers. To address the first question, we first compare the results on this AE golden standard corpus by using shallow and deep Web pages cleaning. Then we perform attribute extraction according to the methods described in Section 2.2 on the WePS 2009 corpus based on deep Web pages cleaning.

The aim of the second series of experiments is to address the second question, i.e. "Is the information of extracted attributes able to lead to better performance than the information of raw Web pages for Web people clustering?" To accomplish it, we first compare the clustering results on the AE golden standard corpus using different feature representations, i.e. *Raw tokens* from raw Web page, *AE tokens* from extracted attributes information, and their integration. Then we perform the comparative experiment on the WePS 2009 corpus.

In order to address the third question, i.e., to compare the contributions of feature representation and algorithms for clustering task, we conduct the third series of experiments. To make the comparison reasonable and meaningful, three simple baseline approaches are applied to the data, i.e. All-In-One baseline, One-In-One baseline and heuristical approach. The third series of experiments is to compare the performance of the above three feature representations and different clustering strategies.

3.1 Experiment 1: Explore Methods for Attribute Extraction for Target Person

Table 3 lists the comparative result of AE on the golden truth file with 17 annotated names using shallow and deep Web cleaning. It is clearly observed that deep Web page cleaning significantly improved the AE performance rather than the shallow Web page cleaning. The averaged F_1 measure over all 17 annotated names improved more than +417% by using deep cleaning. This promising result indicates that only simply removing HTML tags and script codes is not quite enough. To get rid of a lot of noises from Web page, it is crucial to do deep Web page cleaning. In addition, it is also interesting to observed that the performance improved more on the Wikipedia Names than on the ECDL names. The possible reason may be that the ECDL names are selected from one academic domain and thus many Web pages are more concentrated on the same person.

Table 4 depicts the results of two official runs of AE task on the WePS 2009 test data, which has been released on early January of 2009. The results of ECNU_1 and ECNU_2 are obtained based on shallow Web page cleaning and deep Web page cleaning as described in Section 2.1, respectively.

Table 3: Comparative results on the golden standard corpus using shallow and deep Web cleaning

Name	Shallow Cleaning			Deep Cleaning			Improved
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	(%)
Wikipedia Names							
Alexander_Macomb	14.93	3.02	5.02	57.36	23.12	32.95	+557
David_Lodge	19.50	7.07	10.37	59.75	43.30	50.21	+384
George_Clinton	10.91	5.25	7.09	36.22	24.33	29.11	+311
John_Kennedy	11.79	4.99	7.01	43.83	29.15	35.01	+399
Michael_Howard	12.85	5.45	7.66	48.76	35.18	40.88	+434
Paul_Collins	15.30	2.73	4.63	59.53	22.22	32.36	+599
Tony_Abbott	13.88	3.26	5.28	61.28	32.05	42.09	+697
Average	14.16	4.54	6.87	52.39	29.91	38.08	+454
ECDL-06 Names							
Allan_Hanbury	14.12	1.24	2.28	68.94	16.72	26.91	+108
Andrew_Powell	17.49	6.20	9.15	37.97	17.89	24.32	+166
Anita_Coleman	26.70	5.36	8.93	60.32	22.38	32.65	+266
Christine_Borgman	11.13	3.92	5.80	44.33	24.93	31.91	+450
Donna_Harman	14.05	1.46	2.65	52.94	10.08	16.93	+538
Edward_Fox	17.55	6.67	9.66	46.18	26.87	33.97	+252
Gregory_Crane	17.67	3.17	5.37	57.35	19.83	29.47	+449
Jane_Hunter	6.88	2.60	3.78	26.36	12.61	17.06	+352
Paul_Clough	14.51	2.95	4.90	58.02	24.03	33.98	+593
Thomas_Baker	11.08	3.81	5.67	48.57	28.86	36.20	+539
Average	15.12	3.74	5.99	50.10	20.42	29.01	+384
Total Average	14.73	4.07	6.37	51.04	24.33	32.95	+417

Table 4: Results on the WePS 2009 test corpus

Official Run	Name	TP(Match)	FP(Over-generate)	FN(Miss)	P (%)	R (%)	F1 (%)
ECNU_1	Benjamin_Snyder	38	1168	272	3.15	12.26	5.01
	Hao_Zhang	83	819	262	9.20	24.06	13.31
	Amanda_Lentz	50	2168	262	2.25	16.03	3.95
	Otis_Lee	38	413	292	8.426	11.52	9.73
	Bertram_Brooker	101	1567	335	6.06	23.16	9.60
	Jason_Hart	81	891	685	8.33	10.57	9.32
	Average				**6.24**	**16.27**	**9.02**
ECNU_2	Benjamin_Snyder	37	520	273	6.64	11.94	8.54
	Hao_Zhang	82	702	263	10.46	23.77	14.53
	Amanda_Lentz	34	683	278	4.74	10.90	6.61
	Otis_Lee	38	354	292	9.69	11.52	10.53
	Bertram_Brooker	103	821	333	11.15	23.62	15.15
	Jason_Hart	81	744	685	9.82	10.57	10.18
	Average				**8.75**	**15.39**	**11.16**

Although the absolute values of *precision* and *recall* both are quite low on the whole, they are on the top of the participated systems. This indicates that the WePS 2009 test data is more noisy than the WePS 2007 training data and thus the AE task of WePS 2009 is more challenging. From this table, we can find that once again deep Web cleaning performs better than shallow Web cleaning. However, when we take a close look at the two results, it is observed that deep Web cleaning significantly decreases the size of attribute candidates and thus improves the *precision* measure rather than the *recall* measure. A large amount of true attributes have not been extracted from the Web pages. Therefore, to significantly improve the overall performance, only adopting Web page cleaning is not enough and it is extremely necessary to explore other strategies of improving *recall* measure.

Furthermore, for most attributes, such as *email*, *web site*, *phone*, *FAX*, *date of birth*, *birth place*, *degree*, *nationality*, *major*, it may be relatively easy to extract these information from the Web pages, but it is not quite easy to identify which are really related to the target person. Therefore, for these attributes, more good filtering strategies are still needed to improve the *precision*. On the other hand, for some attributes, such as *other name*, *relatives*, *mentor*, *award*, *affiliation*, *school*, *occupation*, it is quite difficult not only to extract them from Web pages but also to disambiguate them. Although manually-constructed rules using regular express patterns, traditional NER tools and gazetteer may capture many patterns and named entities, they still cannot capture all expressions due to the huge diversity of human natural language. Among them, the *award* attribute is most chal-

lenging due to wide varieties of award names, we explored many methods and all failed. Therefore, for these attributes, we still need to explore more good techniques to extract all potential attribute candidates and good filtering techniques to disambiguate them.

3.2 Experiment 2: Is AE Helpful to Improve the Clustering Performance?

To address the second question, we first conduct comparative experiment on the AE golden standard corpus using three feature representations, i.e. i.e. *Raw tokens* from raw Web page, *AE tokens* from extracted attributes information, and their integration. Our consideration is that since the *AE tokens* extracted from carefully annotated attribute values on the gold standard corpus precisely contain correct attributes information about the target person and would not contain noisy or irrelevant information that *Raw tokens* would capture from the raw Web pages, they are expected to perform better than *Raw tokens*. The reason for using the integration of these two representations is that we would like to examine if the integration would capture some information left out from the *AE tokens* alone.

Table 5 depicts the comparative results of clustering on the golden standard corpus using the three representations. Here, "*Raw tokens*" denotes the tokens extracted from the raw Web pages, "*AE tokens*" denotes the tokens extracted from the annotated attribute values extracted from the gold standard, "*integration*" denotes their integration. From the results in Table 5, we observed that *AE tokens* performs better than *Raw tokens* in terms of *precision* measure while *Raw tokens* performs better than *AE tokens* in terms of *recall* measure. This is reasonable since the *AE tokens* contain more clean information than *Raw tokens* from raw Web pages, the former would provide a better performance in terms of *precision*. On the other hand, this result indicates that *Raw tokens* would capture some information left by using *AE tokens* only and thus *Raw tokens* representation has a better performance than *AE tokens* in terms of *recall* measure. Moreover, for their *integration* representation, since the number of *Raw tokens* is large and it dominates the quantity of features, the performance of *integration* is dominated by the performance of *Raw tokens*. With respect to the F_1 measure, the *Raw tokens* representation and the *integration* perform better than *AE tokens* representation. This result is beyond our expectation. It shows that although the *Raw tokens* from raw Web pages contain more noises than *AE tokens*, they also carry a lot of useful information for identifying different person.

Table 6 depicts the comparative results of clustering on the WePS 2009 test corpus using the three representations. The result from Table 6 is similar to that from Table 5. Again *Raw tokens* and the *integration* representations perform better than *AE tokens* representation.

However, so far it is too early for us to draw a definite conclusion that *AE tokens* is not helpful to improve the performance of clustering. The reason may lie in the insufficient or incorrect information of extracted attributes and the inappropriate choosing of clustering algorithms as well. So far, we only adopt the simple *K-means* algorithm and the choosing of k value is arbitrary. The performance of these representations would be different given other density-based or hierarchical-based clustering algorithms. Therefore, we still need a lot of exploration of the usage of these extracted attributes for clustering in our future work.

3.3 Experiment 3: Which Makes More Contribution to Clustering Performance?

Given the above experimental results, it is not clear that which contributes more to the Web people clustering performance, feature representation or clustering algorithms. There are three simple baseline approaches applied to the people clustering task, i.e. *All-In-One* baseline, *One-In-One* baseline and *heuristical* baseline. The *All-In-One* baseline provides a clustering solution where all the documents are assigned to a single cluster. On the other hand, the *One-In-One* baseline gives another extreme clustering solution, where each document is assigned to a different cluster. Besides the above two baselines, we also present an *heuristical* baseline based on simple sampling. That is, for each individual name, we randomly select a very small number of Web pages (roughly less than 10 samples) from whole corpus (approximately 100 Web pages for each name) and manually browse through them. If the number of samples which are referred to one same person is more than 3[1], then we apply *All-In-One* strategy on this name. Otherwise, we use *One-In-One* strategy. Note that the above three baseline approaches are not related to the feature representations. That is, no matter what kinds of feature representations adopted, the result of these three baseline on clustering is consistent.

Table 7 depicts the results of AE task using different clustering strategies on WePS 2009 test data. The first three are results from three clustering strategy baselines, namely, *All-In-One*, *One-In-One* and *heuristical*. The last three are results from *K-means* method with different feature representations as mentioned before.

Table 7: Results of different clustering strategies on the WePS 2009 test corpus

Strategy	BEP	BER	F1
All-In-One	0.43	1	0.60
One-In-One	1	0.24	0.39
Heuristic	**0.78**	**0.74**	**0.76**
Raw Tokens	0.53	0.66	0.59
AE Tokens	0.5	0.55	0.52
integration	0.56	0.59	0.57

The result is quite interesting and several observations from this result are worth discussion. First, it is clearly that clustering strategy contributes more to the clustering performance than feature representation. Therefore, only simply using the AE information without choosing appropriate clustering algorithm would not result in a significant performance improvement. Second, among these clustering strategies, the *heuristical* approach performs the best. It indicates that sampling is an important and useful technique for clustering. However, to improve the performance of clustering, we still need to explore more clustering strategies in our future work.

4. CONCLUDING REMARKS

[1]Typically, this threshold is manually set based on experience.

Table 5: Comparative results of clustering on the golden standard corpus using *Raw tokens*, *AE tokens* and their *integration*

Name	*Raw tokens*			*AE tokens*			*integration*		
	BEP	**BER**	**F1**	**BEP**	**BER**	**F1**	**BEP**	**BER**	**F1**
Wikipedia Names									
Alexander_Macomb	0.5	0.56	0.53	0.96	0.16	0.27	0.53	0.52	0.53
David_Lodge	0.57	0.62	0.6	0.88	0.38	0.53	0.49	0.96	0.65
George_Clinton	0.46	0.64	0.53	0.8	0.33	0.46	0.45	0.55	0.49
John_Kennedy	0.46	0.74	0.56	0.78	0.4	0.53	0.49	0.51	0.5
Michael_Howard	0.24	0.77	0.37	0.69	0.66	0.68	0.3	0.79	0.43
Paul_Collins	0.22	0.72	0.34	0.65	0.4	0.49	0.3	0.63	0.41
Tony_Abbott	0.54	0.66	0.6	0.85	0.27	0.41	0.67	0.51	0.58
Average	0.43	0.67	0.52	0.8	0.37	0.51	0.46	0.64	0.51
ECDL-06 Names									
Allan_Hanbury	0.96	0.84	0.89	1	0.19	0.31	0.95	0.53	0.68
Andrew_Powell	0.13	0.77	0.23	0.77	0.55	0.64	0.13	0.79	0.23
Anita_Coleman	0.7	0.59	0.64	0.83	0.36	0.5	0.68	0.62	0.65
Christine_Borgman	0.81	0.53	0.64	0.94	0.22	0.36	0.81	0.56	0.67
Donna_Harman	0.72	0.64	0.68	0.83	0.21	0.34	0.72	0.64	0.68
Edward_Fox	0.42	0.76	0.54	0.82	0.37	0.51	0.44	0.69	0.54
Gregory_Crane	0.81	0.58	0.68	1	0.33	0.49	0.81	0.58	0.68
Jane_Hunter	0.23	0.76	0.36	0.59	0.57	0.58	0.37	0.86	0.52
Paul_Clough	0.26	0.76	0.39	0.79	0.29	0.42	0.3	0.74	0.43
Thomas_Baker	0.08	0.8	0.15	0.61	0.77	0.68	0.07	0.8	0.13
Average	0.51	0.71	0.59	0.82	0.39	0.53	0.53	0.68	0.52
Total Average	0.48	0.69	0.57	0.81	0.38	0.52	0.50	0.66	0.57

The following conclusions with empirical evidence address the three questions raised in Section 1.

Regarding the first question, the controlled experimental results show that deep Web page cleaning contributes significantly to performance improvements for target person attribute extraction task rather than shallow Web page cleaning. Furthermore, we adopt processing pipelines with multiple techniques to extract attribute candidates, including traditional NER, regular expression patterns, gazetteer-based matching, manually-constructed rules and so on. The difficulty of extraction and disambiguation varies with different person attributes.

The answer to the second question is: not always. It is too early to draw a definite conclusion that AE is not helpful to improve the clustering performance. We should point out that the observations in this paper are made under the controlled experimental settings as indicated in this paper. Once the experimental settings change, for example, different clustering algorithms, different observations would be made.

Regarding the third question, for people clustering task, the clustering algorithm contributes more to performance improvement than feature representations.

We should point out that the observations above are made based on the controlled experiments and therefore, it will be interesting to see in our future work if we can observe the similar results on other more general clustering algorithms, such as density-based, distance-based, hierarchy-based clustering methods and etc. In addition, since the accuracy of extracted attribute values also has a large influence on the performance of consequent clustering task. We believe more advanced NLP techniques and advanced ways of incorporating NLP output could further improve the accuracy performance of attribute extraction for target person, for example, high performance coreference resolution to normalize the person names through different variations, nominal or pronominal expressions could generate more occurrences of the same person names to facilitate the further attribute disambiguation.

5. REFERENCES

[1] R. Guha, and A. Garg. Disambiguation in Web People Search. In *Proc. of the 13th WWW*, pages 148–155. ACM Press, 2004.

[2] Artiles, Javier and Gonzalo, Julio and Sekine, Satoshi. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *Proceedings of SemEval-2007, ACL*, June, 2007.

[3] Satoshi Sekine and Javier Artiles. WePS 2 Evaluation Campaign: overview of the Web People Search Attribute Extraction Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, April, 2009.

[4] Javier Artiles, Julio Gonzalo and Satoshi Sekine. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, April, 2009.

[5] Shian-Hua Lin and Jan-Ming Ho. Discovering informative content blocks from Web documents. In *Preceedings of SIGKDD-2002*, 2002.

[6] Ziv Bar-Yossef, Sridhar Rajagopalan. Template

Table 6: Comparative clustering results on the WePS 2 test corpus using Token, AE and their integration

Name	*Raw tokens*			*AE tokens*			*integration*		
	BEP	**BER**	**F**	**BEP**	**BER**	**F**	**BEP**	**BER**	**F**
AMANDA_LENTZ	0.29	0.59	0.39	0.25	0.73	0.37	0.34	0.73	0.46
BENJAMIN_SNYDER	0.15	0.76	0.25	0.14	0.76	0.24	0.16	0.79	0.27
BERTRAM_BROOKER	1	0.43	0.6	1	0.33	0.49	1	0.25	0.4
CHENG_NIU	0.82	0.53	0.64	0.76	0.38	0.51	0.82	0.52	0.64
DAVID_TUA	1	0.55	0.71	1	0.31	0.47	1	0.21	0.35
DAVID_WEIR	0.29	0.66	0.4	0.32	0.66	0.43	0.44	0.57	0.5
EMILY_BENDER	0.4	0.82	0.54	0.43	0.76	0.55	0.45	0.6	0.51
FRANZ_MASEREEL	0.63	0.65	0.64	0.66	0.38	0.48	0.73	0.52	0.61
GIDEON_MANN	0.83	0.72	0.77	0.61	0.31	0.41	0.9	0.59	0.72
HAO_ZHANG	0.44	0.79	0.56	0.37	0.6	0.45	0.45	0.78	0.57
HELEN_THOMAS	0.96	0.49	0.65	0.96	0.34	0.5	0.96	0.24	0.39
HERB_RITTS	1	0.46	0.63	1	0.31	0.48	1	0.22	0.35
HUI_FANG	0.31	0.67	0.42	0.26	0.66	0.38	0.27	0.63	0.38
IVAN_TITOV	0.78	0.52	0.62	0.58	0.45	0.51	0.79	0.53	0.63
JAMES_PATTERSON	0.93	0.42	0.58	0.92	0.38	0.54	0.92	0.73	0.81
JANELLE_LEE	0.1	0.79	0.18	0.18	0.89	0.3	0.12	0.9	0.21
JASON_HART	0.54	0.65	0.59	0.44	0.48	0.46	0.55	0.45	0.49
JONATHAN_SHAW	0.15	0.72	0.25	0.18	0.59	0.27	0.24	0.6	0.34
JUDITH_SCHWARTZ	0.29	0.9	0.44	0.17	0.61	0.27	0.32	0.74	0.44
LOUIS_LOWE	0.31	0.66	0.42	0.41	0.67	0.51	0.39	0.78	0.52
MIKE_ROBERTSON	0.16	0.75	0.26	0.17	0.74	0.27	0.14	0.96	0.25
MIRELLA_LAPATA	0.98	0.48	0.65	0.98	0.28	0.43	0.98	0.49	0.66
NICHOLAS_MAW	1	0.42	0.59	1	0.27	0.43	1	0.16	0.27
OTIS_LEE	0.38	0.71	0.5	0.41	0.53	0.46	0.41	0.71	0.52
RITA_FISHER	0.58	0.81	0.67	0.5	0.75	0.6	0.62	0.79	0.69
SHARON_CUMMINGS	0.19	0.69	0.3	0.2	0.6	0.3	0.22	0.75	0.34
SUSAN_JONES	0.1	0.93	0.17	0.08	0.82	0.14	0.13	0.81	0.22
TAMER_ELSAYED	0.46	0.65	0.54	0.42	0.39	0.41	0.61	0.29	0.4
THEODORE_SMITH	0.07	0.92	0.13	0.08	0.93	0.14	0.11	0.88	0.2
TOM_LINTON	0.63	0.58	0.6	0.59	0.5	0.54	0.68	0.48	0.56
Average	0.53	0.66	0.59	0.5	0.55	0.52	0.56	0.59	0.57

detection via data mining and its applications. In *Preceedings of WWW-2002*, 2002.

[7] Lan Yi and Bing Liu. Web Page Cleaning for Web Mining through Feature Weighting. In *Preceedings of IJCAI-2003*, 2003.

[8] J. Zhu, V. Uren, and E. Motta. ESpotter: Adaptive Named Entity Recognition for Web Browsing. In *Proc. of Workshop on IT Tools for Know. Manag. Sys. at WM Conference*, Germany, April 11-13, 2005.

[9] M. Porter. An algorithm for suffix stripping. In *Program*, vol. 14, no. 3, pp.130-137, 1980.

[10] E. Amigo, J. Gonzalo, J. Artiles, F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. In *Information Retrieval*, 2008.

[11] http://nlp.uned.es/weps2/