An Unsupervised Approach based on Fingerprinting to the Web People Search task

David Pinto, Mireya Tovar, Darnes Vilariño & Héctor Díaz
Faculty of Computer Science, BUAP
14 & Av. San Claudio, CU
Puebla, Mexico, 72570
{dpinto, mtovar, darnes}@cs.buap.mx

Héctor Jiménez-Salazar
Department of Information Technologies
Autonomous Metropolitan University
Cuajimalpa, DF, Mexico, 11850
hgimenezs@gmail.com

ABSTRACT
In this paper we present the results obtained in the Web People Search task (WePS-2) when using an unsupervised approach based on document fingerprinting techniques. In the context of document indexing/retrieval, we consider a document fingerprint to be a specific code which may be used to uniquely identify this document from the rest of the text collection. In terms of implementation, we consider a document fingerprint may be obtained through hash-based indexing. The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WePS framework needs to be further investigated.

Categories and Subject Descriptors
H.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval—Clustering, Retrieval models

General Terms
Algorithms

Keywords
Clustering, information retrieval, document fingerprinting

1. INTRODUCTION
Classifying people names on the web is a task that requires a special attention by the classification societies community. Searching people in Internet is one of the most common activities performed by the World Wide Web users [2]. The main challenge consists in bringing together all the results (of a given search engine) that share the same occupation/profession (which very often are ambiguous) by using a highly scalable classification method.

In this paper, we report the results obtained in the Web People Search task [3] when using a system based on hash-based text retrieval techniques [6]. In particular, we have constructed a new vectorial coordinates system for the representation of the original data and, thereafter, we calculate the distance of the vectorial representation of each input dataset by means of a hash function.

The experiments were carried out by using the WePS-2 collection. In summary, it is made up of 30 ambiguous names of people, each name with a number of html pages with information related with that people name. The complete description of the evaluated corpus is given into detail in [3].

We must take into account that the fingerprinting technique may allow indexing and classifying of documents in a one single step. Therefore, given the huge amount of information available in Internet, we consider that the main contribution of this research work consists of providing a very fast way of classifying people names on the World Wide Web.

The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WePS framework needs to be further investigated.

The remainder of this document is structured as follows. Section 2 presents the document fingerprint technique used in the process of indexing and clustering of documents in the Web People Search framework. In Section 3 we describe the components of the implemented system. The experimental results are discussed in Section 4. Finally in Section 5 the conclusions are given.

2. DOCUMENT FINGERPRINTING
Document indexing based on fingerprinting is a powerful technology for similarity search in huge volumes of documents. The goal is to provide a proper hash function which quasi-uniquely identifies each document, so that the hash collisions may be interpreted as similarity indication.

Formally, given two documents $d_1$ and $d_2$, and the fingerprint of the two documents $h(d_1)$ and $h(d_2)$, respectively. We consider $d_1$ and $d_2$ to be $\epsilon$-similar iff $|h(d_1) - h(d_2)| < \epsilon$.

In the context of document indexing/clustering/retrieval a fingerprint $h(d)$ of a document $d$ may be considered as a set of encoded substrings taken from $d$, which serve to identify $d$ uniquely.

Defining the specific hash function to encode the substrings of the documents is the main challenge of the fingerprinting technique. In particular, in the implementation of the BUAP Web People Search system we defined a small set of term-frequency vectors (which are used as reference for a new system coordinates) in order to be considered as the new reference for the vectorial representation of each document of the WePS-2 collection. In Figure 1 we may see an overview of the proposed approach.
Formally, given a set of $k$ reference vectors, $\{r_1, r_2, \ldots, r_k\}$, and the vectorial representation of a document $d$, we defined the fingerprint of $d$ as shown in equation (1).

$$h(d) = \sum_{i=1}^{k} r_i \cdot d^i$$  

The specific features used in the vectorial representation of the documents are explained in the following section.

3. THE BUAP SYSTEM

The system comprises the following components:

Pre-processing: We have programmed two implementations in order to perform the HTML to text conversion. The first HTML to text converter was programmed with Java, whereas the second was implemented with AWK. No HTML tags nor url’s were considered in the text extraction.

Named entity recognition: We used the Stanford Named Entity Recognizer [4] in order to extract names of places, organizations and people names from the target documents.

Document representation: The features used to represent each document comprised all the named entities recognized by the Stanford NER. Thus, the vectorial representation of a document $d$ was:

$$d = \{tf(ne_1), tf(ne_2), \ldots, tf(ne_m)\}$$  

where $tf(ne_i)$ is the frequency of the $i$-th named entity recognized in the document $d$.

Reference vector generation: In order to generate the $k$ reference vectors, we calculated the named entity vocabulary, $VN_E$, of the entire target collection. We sorted this vocabulary in a non-increased order according to the named entity frequency (over the complete collection) and, thereafter, we selected only those named entities whose frequency were between the so called “transition” range [5]. The transition range allows to obtain the mid-frequency terms of a given vocabulary. A typical formula used to obtain the center of the transition range (transition point) is given in Equation (3).

$$TP_V = \frac{\sqrt{8 \times I_1 + 1} - 1}{2}$$  

where $I_1$ represents the number of terms (in our case, named entities) with frequency equal to 1.

Once the transition point has been found, we may extract the mid-frequency named entities, which are those which obtain the closest frequency values to $TP_V$, i.e.,

$$V_{TP} = \{ne_i | ne_i \in VN_E, U_1 \leq tf(ne_i) \leq U_2\}$$  

where $tf(ne_i)$ is the frequency of the $i$-th entity over the complete document collection and $U_1$ is a lower threshold obtained by a given neighbourhood value of the TP: $U_1 = (1 - NTP) \times TP_V$, where $0 \leq NTP < 1$. $U_2$ is the upper threshold and it is calculated in a similar way: $U_2 = (1 + NTP) \times TP_V$.

Thus, the representation of the $j$-th reference vector $r_j$ is given as follows:

$$r_j = \{tf(ne_1), tf(ne_2), \ldots, tf(ne_m)\}$$  

Indexing/clustering: The indexing process was carried out by using the formula expressed in Equation (1). We used a specific threshold ($\epsilon$) in order to determine a range of hash-based values (documents) that should belong to the same cluster. The overlapping of clusters was not considered but it may be easily implemented.

4. EXPERIMENTAL RESULTS

Besides the evaluation of the WePS-2 collection, we performed a set of experiments over the training and test dataset of the WePS-1 collection (see [2] for a complete description of these datasets). The obtained results are presented in Tables 1, 2 and 3, respectively.

In these tables we may see the following set of metrics used to evaluate the performance of the implemented system:

- **BEP**: BCubed Precision
- **BER**: BCubed Recall
- **FMeasure**$_{0.5}$**.BEP-BER**: F-measure of B-Cubed P/R with alpha set to 0.5
- **FMeasure**$_{0.2}$**.BEP-BER**: F-measure of B-Cubed P/R with alpha set to 0.2
- **P**: Purity
- **IP**: Inverse Purity
- **FMeasure**$_{0.5}$**.P-IP**: F-measure of Purity and Inverse Purity with alpha set to 0.5
- **FMeasure**$_{0.2}$**.P-IP**: F-measure of Purity and Inverse Purity with alpha set to 0.2
For more details about the evaluation metrics please refer to [1]. The baselines and the rationale for $F$-measures with alpha 0.2 are explained in the WePS-1 task description paper [2].

We have tested six different approaches varying the document similarity threshold ($\epsilon$), the HTML to text converter and the use or not of named entities. The name of each approach as well as their description is given as follows:

BUAP_1 ($\epsilon = 0.0004$): Java-based HTML to text converter without NER, i.e., all the document terms are used.

BUAP_2 ($\epsilon = 0.0004$): Java-based HTML to text converter with NER, i.e., all the document named entities are used.

BUAP_3 ($\epsilon = 0.0004$): AWK-based HTML to text converter without NER, i.e., all the document terms are used.

BUAP_4 ($\epsilon = 0.0004$): AWK-based HTML to text converter with NER, i.e., all the document named entities are used.

BUAP_5 ($\epsilon = 0.3$): Java-based HTML to text converter without NER, i.e., all the document terms are used.

BUAP_6 ($\epsilon = 0.3$): AWK-based HTML to text converter without NER, i.e., all the document terms are used.

We may see (Tables 1, 2 and 3) that the implemented approaches obtained a performance comparable with two of the proposed baselines, ALL_IN_ONE and ONE_IN_ONE, with a document similarity threshold ($\epsilon$) equal to 0.3 and 0.0004, respectively. The selection of any of the two HTML to text converters was not important with respect to the obtained results. Moreover, we could not confirm the benefit of using named entities with respect to those approaches that did not use them, because the obtained results did not show significant difference among the different approaches.

Although, some of the implemented approaches obtained acceptable results in comparison with the baselines, in the case of the WebPS-2 collection any of the first four approaches outperformed the proposed baselines. We consider that the expected document distribution over the final clusters has played an important role on the obtained results, since the presented algorithm of fingerprinting usually assumes a uniform distribution of documents over the discovered clusters.

The evaluation of the experiments carried out show that the implemented technique could have a positive impact in the analysis/indexing of huge volumes of information. However, the feature set for all the documents in the WebPS framework needs to be further investigated.

As future work, we would like to experiment on feature selection in order to clearly benefit the construction of the reference vector set. Although we would like to keep the process as unsupervised as possible, we are considering the use of supervised classifiers in order to extract the most important features to tackle the Web People Search task.

Finally, we would like to analyse the use of new hash-based functions and new document representations which consider characteristics other than only term or named entity frequencies.

5. CONCLUSIONS

We implemented a hash-based function in order to uniquely identify each document from a text collection in the framework of the Web People Search task. The hash collisions were interpreted as similarity degree among the target documents. In this way, we constructed an algorithm which only takes into account the local features of each document in order to index/cluster them. The experimental results over the WebPS-1 test and training datasets showed an acceptable performance of the proposed algorithm. However, the proposed reference vector for the fingerprinting-based model were useless when evaluating with the WebPS-2 dataset. The proper construction of reference vectors for the automatic and unsupervised classification of people names in the Web needs to be further investigated.

6. ACKNOWLEDGMENTS

This project has been partially supported by CONACYT.

7. REFERENCES


### Table 1: Evaluation of the WePS-2 test dataset.

<table>
<thead>
<tr>
<th>run</th>
<th>run</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>IP</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL_IN_ONE_BASELINE</td>
<td></td>
<td>0.43</td>
<td>1.0</td>
<td>0.66</td>
<td>0.79</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>COMBINED_BASELINE</td>
<td></td>
<td>0.43</td>
<td>1.0</td>
<td>0.65</td>
<td>0.94</td>
<td>0.52</td>
<td>0.87</td>
</tr>
<tr>
<td>ONE_IN_ONE_BASELINE</td>
<td></td>
<td>1.0</td>
<td>0.24</td>
<td>0.27</td>
<td>0.27</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>BUAP_1</td>
<td></td>
<td>0.89</td>
<td>0.25</td>
<td>0.27</td>
<td>0.30</td>
<td>0.33</td>
<td>0.37</td>
</tr>
<tr>
<td>BUAP_2</td>
<td></td>
<td>0.89</td>
<td>0.24</td>
<td>0.27</td>
<td>0.29</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>BUAP_3</td>
<td></td>
<td>0.89</td>
<td>0.24</td>
<td>0.27</td>
<td>0.29</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td>BUAP_4</td>
<td></td>
<td>0.90</td>
<td>0.24</td>
<td>0.27</td>
<td>0.29</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td>BUAP_5</td>
<td></td>
<td>0.44</td>
<td>1.00</td>
<td>0.67</td>
<td>0.80</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>BUAP_6</td>
<td></td>
<td>0.44</td>
<td>1.00</td>
<td>0.66</td>
<td>0.80</td>
<td>0.53</td>
<td>0.67</td>
</tr>
</tbody>
</table>

### Table 2: Experimental results with the training dataset of the WePS-1 collection.

<table>
<thead>
<tr>
<th>run</th>
<th>run</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>IP</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL_IN_ONE_BASELINE</td>
<td></td>
<td>0.54</td>
<td>1.0</td>
<td>0.64</td>
<td>0.75</td>
<td>1.0</td>
<td>0.65</td>
</tr>
<tr>
<td>ONE_IN_ONE_BASELINE</td>
<td></td>
<td>1.0</td>
<td>0.34</td>
<td>0.45</td>
<td>0.46</td>
<td>0.35</td>
<td>1.0</td>
</tr>
<tr>
<td>COMBINED_BASELINE</td>
<td></td>
<td>0.48</td>
<td>1.0</td>
<td>0.60</td>
<td>0.9</td>
<td>1.0</td>
<td>0.82</td>
</tr>
<tr>
<td>BUAP_1</td>
<td></td>
<td>0.64</td>
<td>0.6</td>
<td>0.56</td>
<td>0.55</td>
<td>0.68</td>
<td>0.54</td>
</tr>
<tr>
<td>BUAP_2</td>
<td></td>
<td>0.64</td>
<td>0.6</td>
<td>0.56</td>
<td>0.55</td>
<td>0.68</td>
<td>0.54</td>
</tr>
<tr>
<td>BUAP_3</td>
<td></td>
<td>0.6</td>
<td>0.69</td>
<td>0.58</td>
<td>0.56</td>
<td>0.76</td>
<td>0.51</td>
</tr>
<tr>
<td>BUAP_4</td>
<td></td>
<td>0.6</td>
<td>0.69</td>
<td>0.58</td>
<td>0.56</td>
<td>0.76</td>
<td>0.51</td>
</tr>
<tr>
<td>BUAP_5</td>
<td></td>
<td>0.57</td>
<td>0.93</td>
<td>0.65</td>
<td>0.61</td>
<td>0.96</td>
<td>0.50</td>
</tr>
<tr>
<td>BUAP_6</td>
<td></td>
<td>0.54</td>
<td>0.99</td>
<td>0.65</td>
<td>0.61</td>
<td>1.00</td>
<td>0.48</td>
</tr>
</tbody>
</table>

### Table 3: Experimental results with the test dataset of the WePS-1 collection.

<table>
<thead>
<tr>
<th>run</th>
<th>run</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>FMeasure</th>
<th>IP</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL_IN_ONE_BASELINE</td>
<td></td>
<td>0.18</td>
<td>0.98</td>
<td>0.25</td>
<td>0.25</td>
<td>1.0</td>
<td>0.29</td>
</tr>
<tr>
<td>COMBINED_BASELINE</td>
<td></td>
<td>0.17</td>
<td>0.99</td>
<td>0.24</td>
<td>0.28</td>
<td>1.0</td>
<td>0.64</td>
</tr>
<tr>
<td>ONE_IN_ONE_BASELINE</td>
<td></td>
<td>1.0</td>
<td>0.43</td>
<td>0.57</td>
<td>0.61</td>
<td>0.47</td>
<td>1.0</td>
</tr>
<tr>
<td>BUAP_1</td>
<td></td>
<td>0.27</td>
<td>0.61</td>
<td>0.31</td>
<td>0.36</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>BUAP_2</td>
<td></td>
<td>0.27</td>
<td>0.61</td>
<td>0.31</td>
<td>0.36</td>
<td>0.71</td>
<td>0.29</td>
</tr>
<tr>
<td>BUAP_3</td>
<td></td>
<td>0.23</td>
<td>0.73</td>
<td>0.28</td>
<td>0.38</td>
<td>0.82</td>
<td>0.29</td>
</tr>
<tr>
<td>BUAP_4</td>
<td></td>
<td>0.23</td>
<td>0.73</td>
<td>0.28</td>
<td>0.38</td>
<td>0.82</td>
<td>0.29</td>
</tr>
<tr>
<td>BUAP_5</td>
<td></td>
<td>0.21</td>
<td>0.92</td>
<td>0.30</td>
<td>0.36</td>
<td>0.96</td>
<td>0.25</td>
</tr>
<tr>
<td>BUAP_6</td>
<td></td>
<td>0.18</td>
<td>0.98</td>
<td>0.25</td>
<td>0.36</td>
<td>1.00</td>
<td>0.25</td>
</tr>
</tbody>
</table>