The Use of Document Fingerprinting in the Web People Search Task*

David Pinto, Mireya Tovar, Beatriz Beltrán, Darnes Vilariño, Héctor Furlog

Faculty of Computer Science, BUAP
14 Sur & Av. San Claudio, CU, Edif. 104C
Puebla, Mexico, 72570
{dpinto, mtovar, bbeltran, darnes}@cs.buap.mx
http://nlp.cs.buap.mx

Abstract. In the context of document indexing/retrieval, a document fingerprint is considered to be a specific code which may be used to uniquely identify this document from the rest of the text collection. Document fingerprinting is a efficient time-complexity mechanism of indexing data, but issues with respect to precision still being on development. In this paper, we approached the Web People Search task (WePS) by using hash-based document fingerprinting. 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.

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 [1]. 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 [2] when using a system based on hash-based text retrieval techniques [3]. 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 [2].

We must take into account that the fingerprinting technique may allow indexing and clasifying of documents in a one single step. Therefore, given the huge

* This work has been partially supported by the CONACYT project #106625, as well as by the PROMEP/103.5/09/4213 grant.

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 cuasi-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 ϵ-similar iff $|h(d_1) - h(d_2)| < \epsilon$.

In the context of document indexing/clustring/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, $\{\boldsymbol{r_1}, \boldsymbol{r_2}, \cdots, \boldsymbol{r_k}\}$, and the vectorial representation of a document d. We defined the fingerprint of $\boldsymbol{d}$ as shown in equation (1).

$$h(\boldsymbol{d}) = \sum_{i=1}^{k} \boldsymbol{r_1} \cdot \boldsymbol{d} \tag{1}$$

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:

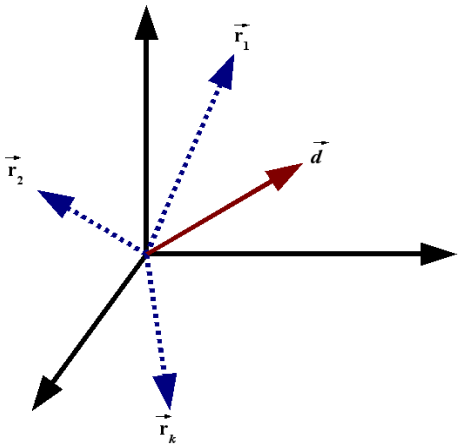

Fig. 1. The new coordinates system used in the implemented hash-based function for fingerprinting.

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:

$$\boldsymbol{d} = \{tf(ne_1), tf(ne_2), \cdots, tf(ne_n)\}, \tag{2}$$

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, V_{NE}, 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).

$$TP_V = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \tag{3}$$

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.,

$$V_{TP} = \{ne_i | ne_i \in V_{NE}, U_1 \leq tf_c(ne_i) \leq U_2\}, \quad (4)$$

where $tf_c(ne_i)$ 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: $U_1 = (1 - NTP) * TP_V$, where $0 \leq NTP < 1$. U_2 is the upper threshold and it is calculated in a similar way: $U_2 = (1 + NTP) * TP_V$. Thus, the representation of the j-th reference vector $\boldsymbol{r_j}$ is given as follows:

$$\boldsymbol{r_j} = \{tf_c(ne_1), tf_c(ne_2), \cdots, tf_c(ne_m)\}. \quad (5)$$

Indexing/clustering: The indexing process was carried out by using the formula expressed in Equation (1). We used a specific threshold (ϵ) 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 [1] 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 [6]. The baselines and the rationale for F-measures with alpha 0.2 are explained in the WePS-1 task description paper [1].

We have tested six differents approaches varying the document similarity threshold (ϵ), the HTML to text converter and the use or not of named entities. The name of each approach as well as their describtion 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 (ϵ) 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 WePS-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 WePS 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 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

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

run	BEP	BER	FMeasure_0.2 BEP-BER	FMeasure_0.2 P-IP	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.43	1.0	0.66	0.79	0.53	0.67	1.0	0.56
COMBINED_BASELINE	0.43	1.0	0.65	0.94	0.52	0.87	1.0	0.78
ONE_IN_ONE_BASELINE	1.0	0.24	0.27	0.27	0.34	0.34	0.24	1.0
BUAP_1	0.89	0.25	0.27	0.30	0.33	0.37	0.27	0.89
BUAP_2	0.89	0.24	0.27	0.29	0.33	0.35	0.26	0.89
BUAP_3	0.89	0.24	0.27	0.29	0.33	0.36	0.26	0.89
BUAP_4	0.90	0.24	0.27	0.29	0.33	0.36	0.26	0.90
BUAP_5	0.44	1.00	0.67	0.80	0.53	0.67	1.00	0.56
BUAP_6	0.44	1.00	0.66	0.80	0.53	0.67	1.00	0.56

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

run	BEP	BER	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.54	1.0	0.64	0.75	1.0	0.65
ONE_IN_ONE_BASELINE	1.0	0.34	0.45	0.46	0.35	1.0
COMBINED_BASELINE	0.48	1.0	0.60	0.9	1.0	0.82
BUAP_1	0.64	0.6	0.56	0.55	0.68	0.54
BUAP_2	0.64	0.6	0.56	0.55	0.68	0.54
BUAP_3	0.6	0.69	0.58	0.56	0.76	0.51
BUAP_4	0.6	0.69	0.58	0.56	0.76	0.51
BUAP_5	0.57	0.93	0.65	0.61	0.96	0.50
BUAP_6	0.54	0.99	0.65	0.61	1.00	0.48

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

run	BEP	BER	FMeasure_0.5 BEP-BER	FMeasure_0.5 P-IP	IP	P
ALL_IN_ONE_BASELINE	0.18	0.98	0.25	0.4	1.0	0.29
COMBINED_BASELINE	0.17	0.99	0.24	0.78	1.0	0.64
ONE_IN_ONE_BASELINE	1.0	0.43	0.57	0.61	0.47	1.0
BUAP_1	0.27	0.61	0.31	0.36	0.71	0.29
BUAP_2	0.27	0.61	0.31	0.36	0.71	0.29
BUAP_3	0.23	0.73	0.28	0.38	0.82	0.29
BUAP_4	0.23	0.73	0.28	0.38	0.82	0.29
BUAP_5	0.21	0.92	0.30	0.36	0.96	0.25
BUAP_6	0.18	0.98	0.25	0.36	1.00	0.25

the local features of each document in order to index/cluster them. The experimental results over the WePS-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 WePS-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.

References

1. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In: Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007, Association for Computational Linguistics (2007) 64–69
2. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: overview of the web people search clustering task. In: Proc. of the 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. (2009)
3. Stein, B.: Principles of hash-based text retrieval. Clarke, Fuhr, Kando, Kraaij, and de Vries, Eds., 30th Annual Int. ACM SIGIR Conf. (2007) 527–534
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the ACL. (2005) 363–370
5. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Proc. of the CICLing 2006 Conference. Volume 3878 of Lecture Notes in Computer Science., Springer-Verlag (2006) 536–546
6. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval **12**(4) (2009) 461–486