Feature Selection on 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. Searching people on the Web is one of the most common activities carried out by Internet users. However, search engine results are usually generated without taking into account the inherent ambiguity of people names. Homonymy is the state of one of a group of words of sharing the same spelling and the same pronunciation but having different meanings. In this paper we have focused our research on the task of discriminating people that share the same name but have different occupation. We compare two different approaches for selecting features on documents of homonyms from a supervised and unsupervised viewpoint. We have used a high scalable clustering method based on fingerprinting in order to discriminate a set of homonyms taken from a testbed corpus used in one international competition. The proposed system performed well in comparison with other results reported in literature.

1 Introduction

The homonym discrimination on the web is a task that requires a special attention by the natural language engineering community. Searching people in Internet is one of the most common activities performed by the World Wide Web users [1]. The main challenge consists on discriminating people with presence in the Web that share the same name but have different occupation. In Figure 1 we can see a snapshot of Google[1] when searching people on the Web. The results retrieved are neither considered in any way belonging to people nor to be classified or clustered. This an undesirable behaviour of a wise search engine.

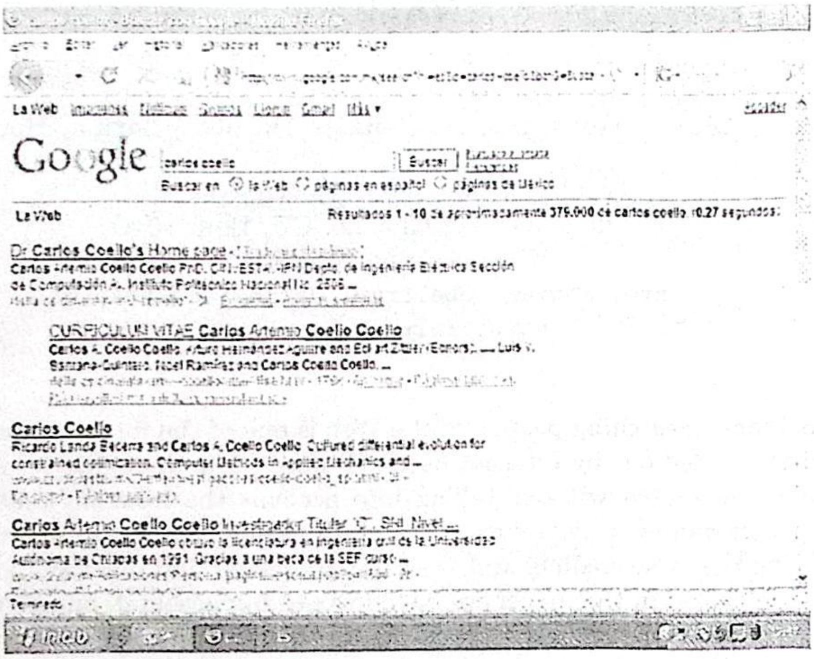
Spock.com[2] instead assumes by default that the user is searching people on Internet and, therefore, it brings together all the results (people webpages) that the system considers that share the same occupation (see Figure 2). In theory, an information retrieval system which is able to recognize queries related to people should have similar behaviour to the Spock system; otherwise, it should have a google-like behaviour.

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

[1] http://www.google.com

[2] http://www.spock.com

© J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad and H. Sossa (Eds.)
Special Issue in Advances in Pattern Recognition
Research in Computing Science 44, 2009, pp. 77-85


La Web
Google
Buscar
Buscar en: ○ la Web ○ páginas en español ○ páginas de México
Terminado
Inicio

carlos coello

La Web — Resultados 1 - 10 de aproximadamente 379.000 de carlos coello (0.27 segundos)

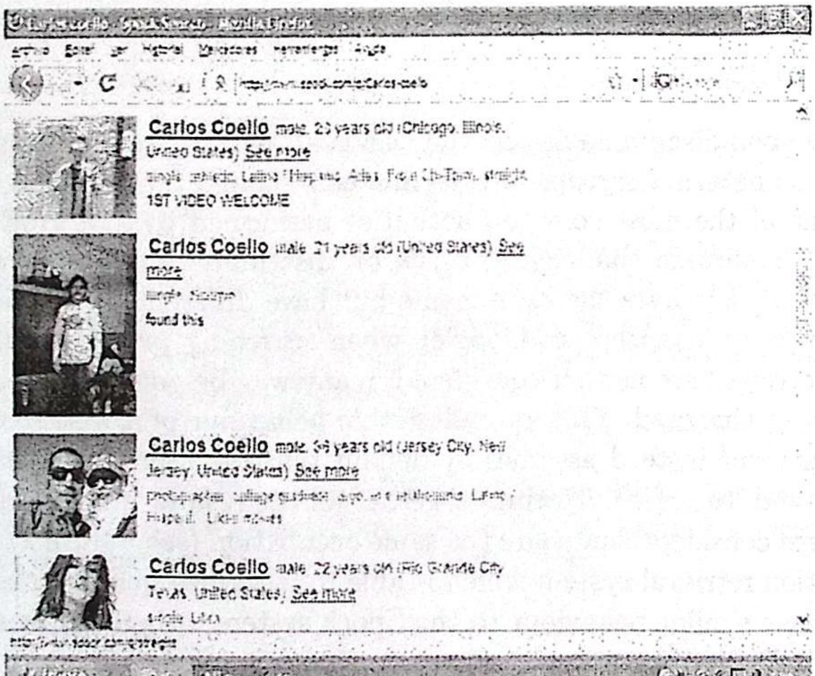
Dr Carlos Coello's Home page
Carlos Artemio Coello Coello PhD. CINVESTAV-IPN Depto. de Ingeniería Eléctrica Sección de Computación Av. Instituto Politécnico Nacional No. 2508 ...

CURRICULUM VITAE Carlos Artemio Coello Coello
Carlos A. Coello Coello, Arturo Hernández Aguirre and Eckart Zitzler (Editors). ... Luis V. Santana-Quintero, Noel Ramírez and Carlos Coello Coello. ...

Carlos Coello
Ricardo Landa Becerra and Carlos A. Coello Coello. Cultured differential evolution for constrained optimization. Computer Methods in Applied Mechanics and ...

Carlos Artemio Coello Coello Investigador Titular "C", SNI Nivel ...
Carlos Artemio Coello Coello obtuvo la licenciatura en ingeniería civil de la Universidad Autónoma de Chiapas en 1991. Gracias a una beca de la SEP cursó ...

Fig. 1. Google results when searching people on the WWW.



Carlos Coello male, 20 years old (Chicago, Illinois, United States) See more
single, athletic, Latino / Hispanic, Ashes From Chi-Town, straight
1ST VIDEO WELCOME

Carlos Coello male, 21 years old (United States) See more
single, Scorpio
found this

Carlos Coello male, 36 years old (Jersey City, New Jersey, United States) See more
photographer, college graduate, ..., Latino / Hispanic, Likes movies

Carlos Coello male, 22 years old (Rio Grande City, Texas, United States) See more
single, Libra

Fig. 2. Spock results when searching people on the WWW.

The aim of this paper is to automatically discover the most salient features for the classification of homonyms. We present a semi-supervised classification method which relies the training phase on the WePS-1 collection [1], which contains 49 ambiguous people names. We evaluate the obtained features on a test dataset in order to see the performance obtained on this task. Moreover, we report the performance of the classifier when using the WePS-2 collection, which is a standard corpus used at the Web People Search Task [2] 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].

The classifier used on the experiments was constructed on the basis of a system which relies its text retrieval techniques on hash functions [3]. In particular, we have constructed a new vectorial coordinate 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.

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 amount of information available in Internet, we consider that an important 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. In Section 2 we describe the components of the implemented system. Section 3 describes into detail the two feature selection techniques used in the experiments. The document fingerprint technique used in the document indexing and clustering process in the Web People Search framework is explained in Section 4. The experimental results are discussed in Section 5, whereas the conclusions are given in Section 6.

2 Description of the implemented system

The system architecture follows the classical approach of supervised classification (see Figure 3) and it 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.

Feature selection: The process of document feature selection is performed in two steps as presented in Figure 4, and it is described in Section 3. In a brief, all the categories of different people names are clustered in order to conform a general set of features which hopefully will describe a certain occupation.

Document representation: We see each document as a vector of features and, therefore, we may formally express the representation of each webpage as follows.

$$\overrightarrow{d} = [w(f_1), w(f_2), \cdots, w(f_n)], \tag{1}$$

where $w(f_i)$ is the weigth of the i-th feature recognized in the document d.

Indexing/clustering: The indexing process was carried out by using the formula expressed in Equation (5). 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.

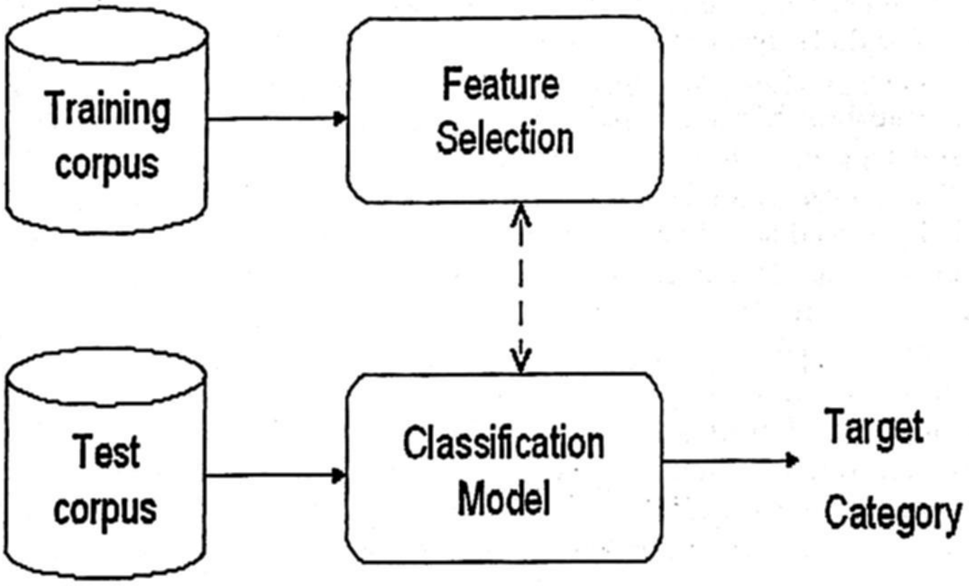


Fig. 3. The classification approach.

3 Feature Selection Techniques

We have considered that each person webpage should be classified according to the abilities of that person with respect to a specific occupation. Therefore, we focus our investigation on discovering such features that identify the role of a given person on his occupation. However, we do not expect that all the occupations will be represented on the training corpus nor that each person will completely describe the features of some occupation.

Our approach extract the most salient features of each occupation. This process is carried out within all the person webpages by means of the two following feature selection techniques.

Entropy: We calculate the entropy (H) of each word (w_i) within each cluster/category for every person. Formally, given a cluster (occupation) of one person C_j, which is made up of a document set ($\{d_1, d_2, \cdots, d_n\}$). The most salient features of C_j are those words w_i that are obtained by Equation (2).

$$Features(C_j) = \{w_i | H(w_i, C_j) > \beta\} \quad (2)$$

with

$$H(w_i, C_j) = p(w_i, C_j) * log(p(w_i, C_j)) \quad (3)$$

where $p(w_i, C_j)$ is the probability of word w_i in cluster C_j, and β is a real value which is used as a threshold, whose value range between cero and one.

Conditional probability: We selected those words that represent better each category by using the conditional probability of word w_i given the category C_j, i.e., $p(w_i|C_j)$. We sort these probabilities and, thereafter, we just selected those that are above of a given threshold γ ($\gamma \in \mathbb{R}$ and $\gamma \in [0, 1]$) as shown in Equation (4).

$$Features(C_j) = \{w_i | p(w_i|C_j) > \gamma\} \quad (4)$$

Once the most salient features of each category are obtained, we cluster these features in order to bring together all that features that belong to the same occupation by using the K-Star clustering method [4]. In summary, given the k categories we are using for extracting the most salient features, we firstly extract k sets of features (one per category) and, thereafter, we consider merging these feature sets by using a clustering method in order to get k' new feature sets ($k' \leq k$). The complete process is presented in Figure 4.

4 Document Fingerprinting and classification

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.

4.1 Document classification

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 $|h(d_1) - h(d_2)| < \epsilon$.

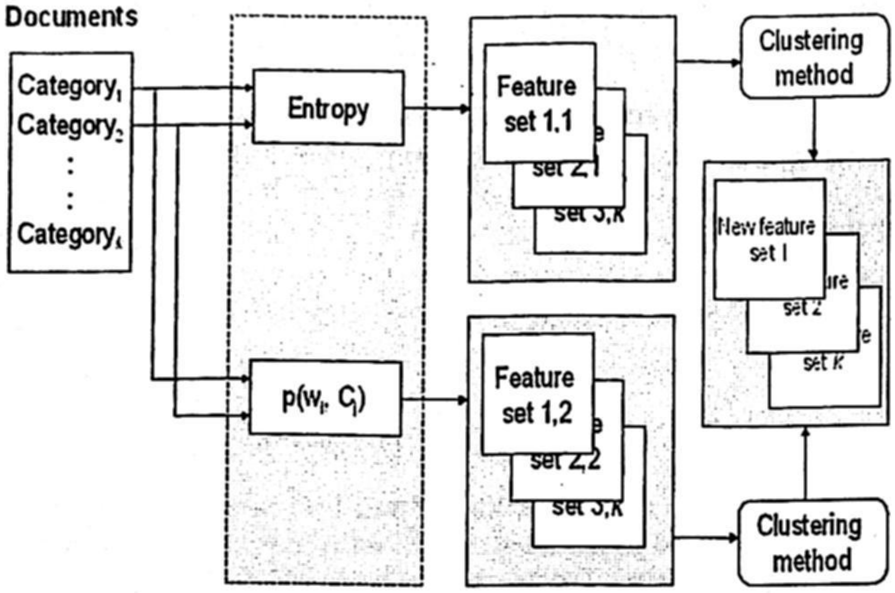


Fig. 4. The process of discovering general features for the homonymy descrimination task.

4.2 Document fingerprinting

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 proposed Web People Search system we defined a small set k of term-frequency vectors $(\{\overrightarrow{r_1}, \overrightarrow{r_2}, \cdots, \overrightarrow{r_k}\})$ (which are used as reference for a new coordinate system) in order to be considered as the new reference for the vectorial representation of each document of the WePS-2 collection. In Figure 5 we may see an overview of the proposed approach.

Formally, given a set of k reference vectors, $\{\overrightarrow{r_1}, \overrightarrow{r_2}, \cdots, \overrightarrow{r_k}\}$, and the vectorial representation of a document d $(\overrightarrow{d})$. We defined the fingerprint of $\overrightarrow{d}$ as shown in equation (5).

$$h(\overrightarrow{d}) = \sum_{i=1}^{k} \overrightarrow{r_1} \cdot \overrightarrow{d}^t \qquad (5)$$

The specific features used in the vectorial representation of the documents are the ones explained in Section 3.

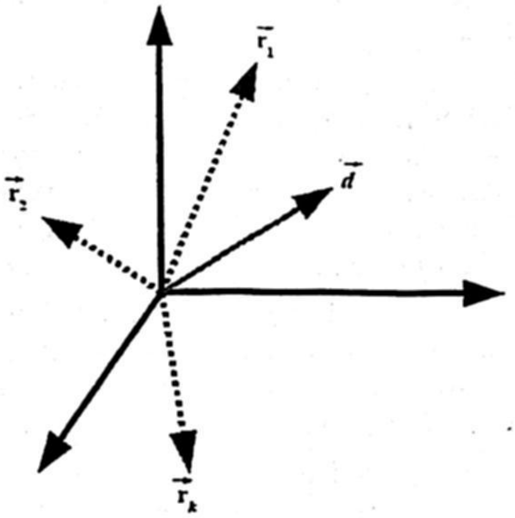


Fig. 5. The new vectorial coordinate system used on the implemented hash-based function for fingerprinting.

5 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
FMesure_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 [5]. 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 the two approaches that result from selecting features by using entropy and conditional probability on each category. 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 0.4.

Although, some of the implemented approaches obtained acceptable results in comparison with the baselines, in the case of the WePS-1 test collection, we could not outperform one of 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 shows 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. We are considering the use of other supervised classifiers in order to extract the most important features of the WePS collection.

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

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

			FMeasure_0.5	FMeasure_0.5		
run	BEP	BER	BEP-BER	P-IP	IP	P
ALL_IN_ONE_BASELINE	0.43	1.0	0.53	0.67	1.0	0.56
COMBINED_BASELINE	0.43	1.0	0.52	0.87	1.0	0.78
ONE_IN_ONE_BASELINE	1.0	0.24	0.34	0.34	0.24	1.0
Entropy	0,46	0,9	0,54	0,65	0,94	0,57
Conditional probability	0.44	1.00	0.53	0.67	1.00	0.56

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

			FMeasure_0.5	FMeasure_0.5		
run	BEP	BER	BEP-BER	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
Entropy	0.57	0.93	0.65	0.61	0.96	0.50
Conditional probability	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
Entropy	0.21	0.92	0.30	0.36	0.96	0.25
Conditional probability	0.18	0.98	0.25	0.36	1.00	0.25

6 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 WePS-1 test and training datasets showed an acceptable performance of the proposed algorithm. However, the proposed reference vector for the fingerprinting-based model did not help too much when evaluating with the WePS datasets. 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. Shin, K., Han, S.Y.: Fast clustering algorithm for information organization. In: CICLing. Volume 2588 of LNCS., Springer-Verlang (2003) 619–622
5. 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