For reliable results, a PCDC system must have access to global information regarding the coreference space.

Rich biographic facts have been shown to improve the accuracy of PCDC (Mann and Yarowsky 2003). Indeed, when available, the birth date, the occupation etc. represent a relevant coreference context because the probability that two different persons have the same name, the same birth date and the same occupation is negligible. However, it is equally unlikely to find this information in a news corpus a sufficient number of times. Even for a web corpus, where the amount of this kind of information is higher than in a news corpus, the extended biographic facts, including e-mail address, phones, etc., contribute only with approximately 3% to the total number of coreferences (Elmacioglu et al. 2007).

In order to improve the performances of the PCDC systems based on VSM, some authors have focused on methods that allow a better analysis of the context by extracting the dependency chains (Ng 2007). The special importance of pieces of context has been exploited by implementing a cascade clustering technique (Wei 2006). Other authors have relied on advanced clustering techniques (among others Han et al. 2005, Chen 2006). However, these techniques rely on the precise analysis of the context, which is a time consuming process. It has been also noted that, in spite of deep analysis, the relevant coreference context is hard to find (Vu 2007).

The technique we present in the next sections is complementary to these approaches. We propose a statistical model designed to offer to the PCDC systems information regarding the distribution of PNMs in the corpus. This information is used to reduce the contextual data variation and to attain a good balance between precision and recall.

3 Data Analysis

In this Section we present the data analysis of the PNMs. We are interested in establishing a relationship between the distribution of the PNMs and the relevant context for coreference. As mentioned in the preceding sections, the amount of the relevant context for coreference cannot be decided prior to the investigation of that particular corpus. The performances of a bag of words VSM with a prior defined context approach will vary greatly from corpus to corpus. We have run the following experiment: we have considered the training and test corpora used in Web People Search-1 (WePS-1), which are web page corpora, and we have implemented a bag of word approach with two variants of clustering: agglomerative (A), and hierarchic (H). We have randomly chosen a set of seven names from training and test (14 names in total) and we have compared the results applying the two systems, A and H, on each set of names. In Figure 1 we present the results obtained. The figures on the vertical axes are computed using $F_{\alpha=0.5}$ formula.

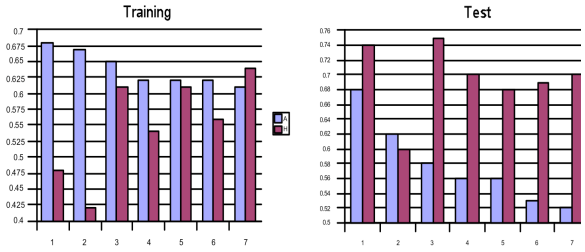

Figure 1. Variation between training and test

We have noticed a great variation in the behavior of the two systems. In order to search for an explanation for this difference we have looked at the distribution in the two corpora of the Named Entities, of the words denoting professions and of the meta-contextual information - e-mails, urls, phones, and addresses. It turns out that these types of contextual information are distributed between training and test approximately evenly. (see Table 1a,b).

Profession	training occ.	test occ.
Doctor	543	668
Lawyer	277	385
Professor	523	490
Researcher	340	166
Teacher	617	569
Coach	467	471
Actor	998	790

Table 1a. Profession words in training and test

Address	training occ.	test occ.
Phone	1,109	1,169
Fax	606	426
e-mail	3,134	2,186

Table 1b. Meta-Context in training and test

By manually investigating the training and test set of our experiment we have reached the conclusion that the reason for the difference is two fold: firstly, while the distribution of the words denoting profession is similar, in the test set the modifiers, for example “internist”, “neurosurgeon” for “doctor”, are more frequent. Secondly, the number of different persons having the same