Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis

Jorge J. García Flores[1], Pierre Zweigenbaum[1], Zhao Yue[2], and William Turner[1]

[1] LIMSI - CNRS, B.P. 133, F-91403, Orsay Cedex, France
[2] Université Paul Valéry, Route de Mende, F-34199, Montpellier, France
{jgflores,pz,turner}@limsi.fr, yue.zhao@etu.univ-montp3.fr

Abstract. This paper presents the *Unoporuno* system: an application of natural language processing methods to the sociology of migration. Our approach extracts names of people from a scientific publications database, refines Web search queries using bibliographical data and decides of the international mobility category of a person according to the location analysis of those snippets classified as mobility traces. In order to identify mobility traces, snippets are filtered with a name validation grammar, analyzed with mobility related semantic features and classified with a support vector machine. This classification method is completed by a semi-automatic one, where *Unoporuno* selects 5 snippets to help a sociologist decide upon the mobility status of authors. Empirical evidence for the automatic person classification task suggest that *Unoporuno* classified 78% of the mobile persons in the right mobility category, with F=0.71. We also present empirical evidence for the semi-automatic task: in 80% of the cases sociologist are able to choose the right category with a moderate level of inter-rater agreement (0.60) based on the 5 snippet selection.

1 Introduction

Among the Latin-American authors who published scientific articles about biotechnology during 2011, how many of them are living abroad? And how many of them have studied in foreign universities before coming back to work in their home countries? Sociologists of migration, and in particular those working on the "brain drain" issue—that is, the idea that talent mobility is a serious problem affecting developing countries—find it hard to answer such fine-grained questions using traditional data sources such as demographic registers, labour surveys or population census, which require a great deal of field work and are carried out too infrequently to provide a constantly updated picture of talent mobility [1]. They have been experimenting other methods such as using browsers to search the Web for biographical evidence of mobility but are faced with the "needle in the haystack" problem [2]: it takes them a great deal of time to wade through the results of a browser search (hundreds of snippets) to find the precise evidence they need (a CV, a personal Web Page, etc.) to classify a person in one of the following categories:

H. Isahara and K. Kanzaki (Eds.): JapTAL 2012, LNAI 7614, pp. 180–191, 2012.

- Mobile: has gone abroad for professional or academic reasons and has lived away from the country of origin for at least one year.
- Local: has only spent short periods of time abroad (less than one year).

Web People Search (WePS) systems [3] are concerned with clustering the results of ambiguous name queries in order to distinguish between people with the same names. Our system also aims at finding people on the Web, however, it differs from WePS because its starting point is not a user query, but a publication record. Information can consequently be extracted on, for example, an author's geographical location, his or her affiliation or topics from the publication's title in order to refine name-only queries. We call this the *Mobility Traces Classification* (MTC) task. This paper presents *Unoporuno*: an NLP system for carrying out the MTC task. Its main contribution consists in implementing a metasearch engine based upon bibliographical query refinement and multilingual Web search. The resulting snippets are first filtered using a personal name grammar that recognizes valid name variations; then classified on the basis of the mobility-related features they contain; and finally ranked statistically according to their calculated relevance for deciding on the mobility status of a person. We present two variants of the MTC task: an automatic one, where *Unoporuno* decides on the mobility status of a person based on a location analysis of the top ranked snippets, and a semi-automatic one, where only the "top five" snippets in the ranking are presented to a sociologist, who manually attributes a mobility status. The article is organized as follows: Section 2 presents related work; Section 3 provides methodological details about the *Unoporuno* pipeline: Section 4 presents an overall evaluation of each step of the pipeline; Section 5 presents the results, and Section 6 discusses these results and outlines future work.

2 Related Work

Evaluation campaigns of the Web People Search task [4, 5, 6, 3] generally focus on ways of clustering documents into sets which characterize different persons that share the same name. Quoted-named queries are used as input to the WePS task, but with no query refinement. Artiles et. al. [7] showed that when query refinement is used, in most cases relevant pages are found, but they note that human users seldom know what refinement terms to use in order to produce these positive results. In contrast, our MTC task provides a semantically rich context for Web People Search. Its input is a bibliographical record, from which we extract topics, organizations and locations to enrich multilingual name queries. However, this greatly increases the number of Web queries and search results (an average of 400 snippets per person, compared to 100 for WePS). For that reason, instead of directly processing Web pages, *Unoporuno* implements a common strategy [8] which consists in filtering and classifying the snippets found by the search engine. This requires extracting suitable features directly from these snippets, and implementing a statistical classification of Google snippets based upon the semantic features of mobility. Previous work on the linguistics and semantics of Web search has largely focused on queries [9, 10], but relatively few studies have focused on snippets, even though an eye tracking study has shown that snippets are looked at longer [11] than titles, images or URL address.

3 NLP Pipeline

Unoporuno[1] is a metasearch engine for query refinement and snippet classification (see Figure 1). Its input is a Web of Science (WoS) data extraction: e.g., all the biotechnology publications of Uruguayan researchers in October 2011. The output consists of 5 Web search snippets that are presented to the sociologist in order to classify the person in one of the above mentioned categories. This section describes each of the steps of the *Unoporuno* processing pipeline.

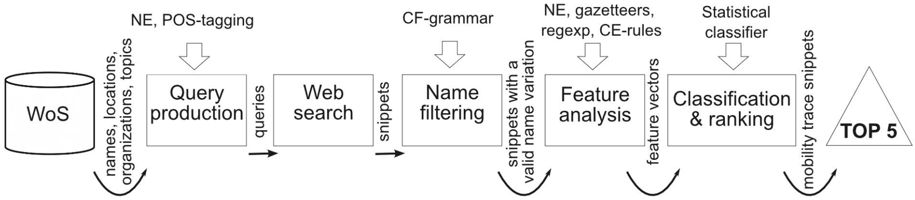

Fig. 1. NLP pipeline for Mobility Traces Classification

3.1 Pre-processing and Query Production

The pre-processor extracts author names, publication titles, organizations and locations from the input file, which is a bibliographical extraction from the ISI Web of Knowledge[2]. The ISI export format separates author, publication title and affiliation in different columns. The first step of the pre-processing consists in extracting geographical locations and organizations from the author's affiliation. Authors are then filtered by affiliation and name. Researchers affiliated to non-Spanish speaking countries are filtered out, except for those with a Spanish first or last name. A Spanish name list was built from census[3] data and geodemographic analysis for that purpose [12]. The second step focuses on query refinement and production. Names of people are combined with noun phrases from the publication's title; these noun phrases are identified using the Freeling bilingual POS-tagger and NE recognizer [13]. Geographical locations are translated into Spanish or English and multilingual queries are generated. When an organization's language is neither of the two (for instance, the *Karolinska Institutet* from Sweden) queries are built using Spanish, English and the organization's language (Swedish in our example). Language detection is made using

[1] Open source available at `https://github.com/unoporuno/unoporuno`

[2] `http://www.isiwebofknowledge.com/`

[3] `http://www.ine.es/daco/daco42/nombyapel/nombyapel.htm`

Google translator; city and country translation with home-made gazetteers. An average of 19 queries per person are produced[4].

3.2 Name Matching Filter

The mass of snippets resulting from Web search queries is then filtered to select those with a valid variation of the person's name. A context-free grammar for Spanish and English names was developed for this purpose[5] (see Table 1). By successive recursive operations on the parse tree of the name, we produce a set of regular expressions which recognize name abbreviations, initial expansions, first and last name inversions or partial name suppression in the snippet. For the test set (see Table 4), the name filtering process retained 3,476 snippets among the 11,266 retrieved by the queries. The grammar has four terminal elements: 1) a common name, 2) a name with a particle (*Ana Ozores* ***de*** *Clarin*), 3) a name with a typographical link (*Ana Ozores***-***Clarin*) and 4) the initial. The syntax requires at least a name (or an initial) and a last name. It takes into account Spanish name formation rules, using father and mother last names, and ensures that one of the last names cannot be compressed as an initial. Table 2 describes valid operations on any branch of the Spanish name grammar tree. Operations are implemented as recursive algorithms that compress, expand or suppress elements from the tree. Regular expressions are produced and then used to validate name variations found in the snippets. When any given regular expression is true, the snippet is considered as containing a valid name variation. A total of 27 possible variations for a name were identified during the acquisition process.

Table 1. Context-free grammar for Spanish and English name parsing

PersonName	→	FirstName LastName
FirstName	→	FirstName Initial
FirstName	→	FirstName CommonName
FirstName	→	Initial
FirstName	→	CommonName
FirstName	→	TypoName
FirstName	→	CommonName ParticleName
LastName	→	MainLastName
LastName	→	MainLastName CommonName
LastName	→	MainLastName ParticleName
LastName	→	MainLastName TypoName
LastName	→	MainLastName Initial
MainLastName	→	CommonName
MainLastName	→	ParticleName
MainLastName	→	TypoName

[4] Google querying using P. Krumins Python Library, `http://bit.ly/EUizu`

[5] It was implemented with the NLTK formal grammar library [14].

Table 2. Regular expressions generated from the name grammar to check valid name variations

Operation	**Description**
n: name	$n \rightarrow N$
a: surname	$a \rightarrow A$
C: compression	$CnLa$(Noe Lopez) → N\.?Lopez
E: expansion	$EnLa$(Eva M Perez) → Eva M[$a - z$]+ Perez
L: literal	$LnLa$(Noe Lopez) → Noe Lopez
X: extra element	$LnXLa$(Eva M Perez) → Eva M [$A - Z$][$a - z$]+ Perez
V: inversion	$VCnLa$(Eva M Perez) → Perez,? +E[\.]?[−]?M \ .?
SI:suppress initial	$SInSIa$(Noe J Lopez F) → Noe Lopez

3.3 Semantic Features Analysis

Feature analysis consists in searching in the snippet content for mobility related information. The rationale is that the snippet contents might give clues about mobility traces not directly visible in the snippet, but which are contained in the referred to document. Feature analysis is performed by means of regular expression and gazetteers. To design the multilingual rules, an extensive n-gram based analysis of the 58,220 snippets from the training set (see Section 4.1) and NE's from the JRC base [15] was performed. Acronyms received special treatment: a list of uppercase sequences were extracted from all snippets of the test set and transformed into content-specific *Unoporuno* queries whose results were then analyzed to find significant acronyms for mobility. Most of the features are binary. The underlying idea is to represent a snippet as a vector of binary features. Table 3 shows semantic features used to analyze snippets. Features 1 to 8 convey very simple information, while features 9 to 14 capture more complex phrases (biography, profession, academic background) that needed a deeper linguistic analysis. Regular expressions and gazetteer rules were preferred to deeper linguistic techniques because of the multilingual character of snippets.

3.4 Snippet Classification and Ranking

The last step statistically classifies and ranks the snippets. The ranking is used both by the automatic Mobility Traces Classification (MTC) task (to attribute a mobility status to a person) and by the semi-automatic MTC task (to select the top-5 snippets that will be presented to the sociologist). Four classifiers from the Weka toolkit (Decision trees, Naive Bayes, NBTrees and SVM) were trained on the training set and tested on the test set snippets (see Table 4). The classification process takes all the snippets of a person, classifies them and then ranks those classified as mobile to select the top-5. Classifiers were trained on three categories: strong mobility trace, weak mobility trace and no trace. Mobility traces are considered strong if both points of the movement (origin and destination) are visible in the document referred to the snippet. Traces are considered as weak if only one point of the potential movement is visible. We use

Table 3. Semantic features for snippet analysis (*cities have more than 100,000 hab)

Name	Type	Description	Type
PhD thesis	regex	The snippet links to a PhD thesis	bool
LinkedIn	gazet	The snippet links to a LinkedIn Web page	bool
Publication	gazet	The snippet links to a scientific publication	bool
e-mail	regex	The snippet contains an email	bool
Non Latin-American nationality	gazet	The snippet contains a nationality from a non Latin-American country	bool
Latin-American nationality	gazet	The snippet contains a nationality from a Latin-American country	bool
Person name found in *URL*	regex	Personal first or last name in the *http* address	bool
CV	regex	The snippets links to a CV	bool
Profession	regex	The snippet contains a profession name	bool
Degree	regex	The snippet contains academic information	bool
Biographical sentence	regex	The snippet contains a biographical sentence	bool
Organization acronym	gazet	The snippet contains an organization acronym	bool
City & region	gazet	The snippet contains a city or region name	bool
Country	gazet	The snippet contains a country name	bool
Organization	regex	The snippet contains an organization name	bool
Feature count	-	Number of features found in the snippet	int

a geographical heuristic to select the top-5 mobility snippets. First, we extract all the geographical locations from snippets classified as strong and weak mobility traces. Second, we calculate frequent countries from those locations. Finally, we include in the top-5 three snippets containing locations outside Latin-America, and two containing Latin-American locations. If one of both locations is missing, snippets are sorted in decreasing order of their feature count.

3.5 Person Classification

In the automatic MTC task, persons are classified using geographical data found in the snippets, with no sociologist annotation at all. First, the title and the description of those snippets classified as mobility traces are parsed to extract locations. Second, locations are associated to a country. For this experiment, we consider only cities and countries as locations. Neither organizations nor nationalities nor any element of the URL are associated to a country yet. The relation between a city and a country is obtained through a qualified gazetteer of 3,545 world cities with more than 100,000 inhabitants extracted from Wikipedia. Finally, the person is classified according to the most frequent countries in the snippet selection: if Latin-American and non Latin-American countries are found in the frequent countries list, the person is classified as mobile; otherwise, the person is classified as local. If no locations are found in the snippet list, the person is not classified.

4 Evaluation

4.1 Data

Table 4 summarises the data collected for mobility trace classification. Training and test datasets have no overlap. The training set comes from two sources. First, 102 researchers from Argentina, Colombia and Uruguay extracted from WoS, and whose mobility traces were annotated manually by sociologists. Second, 646 Latin-American researchers from WoS who were treated by *Unoporuno* using the baseline top-5 classification criteria. For each of these 646 researchers, the top-5 snippets were manually annotated. The test set was created using information collected from an on-line survey of Uruguayan researchers: 25 of these researchers answered that they live abroad or have been abroad for longer than a year for professional or academic purposes (mobile category). The other 25 answered that they had only spent short periods abroad (local category). Each document that a snippet pointed to was manually annotated as:

Table 4. Two hand annotated corpora for the MTC task.

Gold standard	Training set	Testing set
Researchers	102+646	50
Home country	Argentina, Uruguay, Colombia	Uruguay
Queries	782+10471	609
Filtered snippets	5544+52676	3476
Mobility traces	397+214	134
Home/destination country traces	921+770	1091 (home) 252 (dest)
No trace	4226+51692	1999

- Mobility trace: the snippet links to a document containing clear evidence of international mobility.
- Destination country trace: the snippet links to a document containing partial evidence of mobility (e.g., an affiliation to a foreign university).
- Home country trace: the linked document shows no international mobility, but an affiliation of the researcher to his home country.
- No trace: none of the above.

4.2 Name Matching Filter Evaluation

As written above, the name matching filter selects those snippets containing valid variations of a person name (see Section 3.2). To evaluate the grammar and regular expressions used for this step, we first selected on a random basis 100 snippets containing a valid variation of 10 persons names (positives) and 100 snippets containing no valid variation of 10 person names (negatives). Then we manually annotated false positives from the first set and false negatives from the second.

4.3 Semantic Features Evaluation

Two tests were used for semantic feature evaluation. First, a detailed evaluation of individual feature performance, and second, an evaluation of the impact of each feature on the whole automatic MTC task. Then, we performed ablation tests of the automatic MTC task in order to evaluate the impact of each feature on the main task. For each of the 15 features, we made a random selection of 50 snippets with the feature on (positives) and 50 snippet with the feature off (negatives). Then we annotated false positives from the first set and false negatives from the second. For the feature impact evaluation on the overall task, we trained 15 SVM ablated classifiers. An ablated classifier is trained by removing one feature from the original 15 feature set. The automatic person classification process was run 15 times with a 14 feature-set classifier, and the results compared to the full 15 feature-set run.

4.4 Snippet Classifiers Evaluation

We selected the best classifier by evaluating the top-5 snippets in two ways. First, we measured P@5 (precision at the fifth snippet), R@5 (recall at the fifth snippet) and F@5 based on the observed category value of each snippet. Second, we simulated whether a sociologist would be able to make a decision based on the top-5 snippets. This would be the case if at least one snippet allowed the sociologist to classify a person in the right category. Table 5 shows how to decide whether a snippet has this property given how it was manually annotated and the person's true mobility status. Based on this we defined the Oracle Decision Rate (ODR) of a classifier as the proportion of persons for which the top-5 has this property. We also computed that rate for the mobile persons only (mODR).

4.5 User Evaluation of the Semi-automatic MTC Task

We evaluated the ability of sociologists to classify persons given the top-5 snippets produced by the classifier that obtained the best ODR. Three pairs of sociologists classified subsets of 10 persons (5 mobile, 5 local) of the test set. A seventh sociologist was asked to classify the entire test set (50 persons: 25 mobile, 25 local). Precision, Recall and F-measure were computed using the true

Table 5. Criteria for a decision enabling snippet

Person class	Snippet class	Relevant iff
Mobile	Mobility trace	Always (no exception)
Mobile	Destination country trace	There is also a home country trace in the top-5
Mobile	Home country trace	There is also a destination country trace in the top-5
Local	Home country trace	Always (no exception)

mobility status of the people. Kappa was computed for each pair of users sharing the same dataset. A first experiment on automatic person classification is presented as well.

4.6 Automatic MTC Task Evaluation

We performed an automatic person classification test on a set of 25 mobile and 25 local persons. The test consisted in classifying automatically a person as being mobile or local based on geographical criteria, and comparing *Unoporuno* results with the real mobility classes.

5 Results

From results in Table 6 we can observe an F=0.93 for the name filtering process, and an F>0.80 for all the semantic features. While we can expect to get a very high F when simply controlling for snippet links to a LinkedIn page, more complex features, like biographical sentences, academic degree or organization get a fairly good score. However, further evaluation is needed to measure the impact of the semantic feature analysis on the overall task (see Table 9). Table 7 presents the results of the compared evaluation of statistical classifiers of snippets. The tests were performed on binary trained classifiers (a snippet can be a mobility trace or no trace at all) that selected the top-5 according to the confidence of the prediction. The best score was obtained by the SVM classifier, whose difference with the baseline score is statistically significant ($p < 0.05$). Table 8 presents the

Table 6. Name matching and semantic features evaluation (tp=true postivies; fp=false positives; fn=false negatives; P=precision; R=recall; F=F-measure)

Feature	**tp**	**fp**	**fn**	**P**	**R**	**F**
Name matching filter	99	1.00	13	0.99	0.88	**0.93**
PhD tesis	47	3	5	0.94	1.00	**0.97**
LinkedIn	50	0	0	1.00	1.00	**1.00**
Publication	50	0	8	1.00	0.86	**0.93**
e-mail	46	4	0	0.92	1.00	**0.96**
Non Latin-American nat.	33	17	0	0.66	1.00	**0.80**
Latin-American nat.	48	2	1	0.96	0.98	**0.97**
Person name in URL	47	3	0	0.94	0.90	**0.92**
CV	48	2	3	0.96	0.94	**0.95**
Academic degree	47	3	1	0.94	0.98	**0.96**
Profession	48	2	4	0.96	0.92	**0.94**
Biographical sentence	49	1	1	0.98	0.98	**0.98**
Organization acronym	44	6	5	0.88	0.90	**0.89**
City	40	10	3	0.80	0.93	**0.86**
Country	44	6	2	0.88	0.96	**0.92**
Organization	47	3	10	0.94	0.82	**0.88**
Feature count	-	-	-	-	-	-

Table 7. Top-5 automatic evaluation on the testing set. Classifiers were trained on a binary basis; no geographical data was used for top-5 selection.

	Top-5 snippets			**Persons**	
Classifier	**P@5**	**R@5**	**F@5**	**ODR**	**mODR**
Baseline	0.46	0.08	0.14	0.82	0.72
Dsc. trees	0.32	0.09	0.15	0.76	0.68
Naive Bayes	0.37	0.11	0.17	0.82	0.76
NBtree	0.32	0.11	0.16	0.78	0.72
SVM	**0.48**	**0.13**	**0.20**	**0.88**	**0.84**

Table 8. MTC task evaluation on the test set with 7 sociologist users (SVM classifier). Average **F=0.79** for the first six evaluators (sets A,B,C). Pers. = number of persons in dataset. *Set D corresponds to the full test set.

			First user			**Second user**			
Data	**Pers.**	**Users**	**P**	**R**	**F**	**P**	**R**	**F**	κ
set A	10	E1, E2	0.83	0.56	0.67	0.89	0.89	0.89	0.31
set B	10	E3, E4	0.75	0.75	0.75	0.88	0.78	0.82	0.81
set C	10	E5, E6	0.75	0.75	0.75	0.89	0.89	0.89	0.68
set D*	50	E7	**0.80**	**0.88**	**0.83**	**avg. kappa:**			**0.60**

Table 9. Automatic person classification on the Testing set (50 researchers, SVM classifier)

Id	**Ablated feature**	**P**	**R**	**F**	F variation
1	ALL FEATURES (no ablation)	**0.64**	**0.78**	**0.71**	-
2	PhD thesis	0.62	0.72	0.67	**-0.04**
3	LinkedIn	0.64	0.75	0.69	**-0.02**
4	Publication	0.64	0.72	0.68	**-0.03**
5	e-mail	0.62	0.78	0.69	**-0.02**
6	Non Latin-American nat.	0.62	0.75	0.68	**-0.03**
7	Latin-American nat.	0.64	0.82	0.72	**+0.01**
8	Person name in URL	0.61	0.83	0.7	**-0.01**
9	CV	0.59	0.86	0.7	**-0.01**
10	Academic degree	0.64	0.78	0.71	**0**
11	Profession	0.64	0.72	0.68	**-0.03**
12	Biographical sentence	0.61	0.79	0.69	**-0.02**
13	Organization acronym	0.64	0.72	0.68	**-0.03**
14	City	0.62	0.75	0.68	**-0.03**
15	Country	0.64	0.78	0.71	**0**
16	Organization	0.64	0.78	0.71	**0**
17	Feature count	0,6	0,78	0,68	**-0,03**

results of the semi-automatic MTC task carried out by sociologists. An average F=0.79 was obtained for the first six evaluators with an average inter-evaluator agreement kappa=0.60 (considered as moderate to substantial). The seventh evaluator annotated the entire test set with an F=0.83. From the analysis of the results we observe that a) in approximately 80% of the cases a sociologist received the right evidence to decide on mobility status; b) the annotator disagreement was higher for the local than for the mobile category (66% agreement for mobile, only 53% for local); c) moderate inter-annotator agreement might be related to low P@5 and R@5: improvements in the snippet classifier could have an impact on this agreement. Finally, an ablation test was made to estimate the impact of each feature on the automatic MTC task. Table 9 shows the result of the automatic classification of 50 persons (25 mobile, 25 local) using all 15 features, which gets an F=0.71, with a mobile recall of 0.78. Further runs were performed, each ablating one feature, in order to estimate the impact of each feature on precision and recall (see Table 9). The features whose absence impacts the overall performance were PhD thesis, Publication, Organization acronym, City, Feature count, Profession and City. In contrast, Organization, Academic degree and Country do not contribute.

6 Conclusion and Further Work

We have shown in this paper how we are using NLP techniques in the sociology of migration field with the *Unoporuno* system. From scientific publication databases, our method produces Web People Search queries refined with bibliographical data, classifies the resulting snippets according to mobility related features and then statistically ranks their relevance. The top-5 snippets are selected for evaluation by a sociologist, and our automatic selection algorithm works in 80% of the cases: using the snippets selected by our system, sociologists can access documents on the Web which allow them to take clear-cut decisions on a person's mobility status with a moderate level of inter-evaluator agreement (avg. kappa=0.60). Furthermore, we presented a first experiment towards automatic annotation of mobility traces. The geographical analysis of locations from snippets classified as mobility traces by an SVM classifier was able to find 78% of the mobile persons of the test set (F=0.71). Further evaluation is necessary to calculate the correlation between sociologists manual annotations and those calculated by *Unoporuno*.

Acknowledgments. Research funded by the CIDESAL Europe-Aid project (MIGR/2008/26). Thanks to Jean-Baptiste Meyer, Martn Koolhaas, Julieta Benegochea, Fernando Esteban, Alejandro Blanco, Jean-Paul Sansonnet and Haydée Lugo for their help.

References

[1] Auriol, L., Felix, B., Schaaper, M.: Mapping Careers and Mobility of Doctorate Holders: Draft Guidelines, Model Questionnaire and Indicators. OECD Science, Technology and Industry Working Papers (2010/01) (2010)

[2] Meyer, J.B., Wattiaux, J.P.: Diaspora Knowledge Networks; Vanishing Doubts and Increasing Evidence. International Journal on Multicultural Societies. UNESCO 8(1), 4–24 (2006)

[3] Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., Amigó, E.: WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. In: Conference on Multilingual and Multimodal Information Access Evaluation, CLEF (2010)

[4] Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007). ACL (2007)

[5] Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In: 18th WWW Conference on 2nd Web People Search Evaluation Workshop, WePS 2009 (2009)

[6] Sekine, S., Artiles, J.: WePS2 Attribute Extraction Task. In: 18th WWW Conference on 2nd Web People Search Evaluation Workshop, WePS 2009 (2009)

[7] Artiles, J., Gonzalo, J., Amigó, E.: The impact of query refinement in the web people search task. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort 2009, pp. 361–364. Association for Computational Linguistics, Stroudsburg (2009)

[8] Liu, J., Birnbaum, L., Pardo, B.: Categorizing blogger's interests based on short snippets of blog posts. In: Shanahan, J.G., Amer-Yahia, S., Manolescu, I., Zhang, Y., Evans, D.A., Kolcz, A., Choi, K.S., Chowdhury, A. (eds.) CIKM, pp. 1525–1526. ACM (2008)

[9] Barr, C., Jones, R., Regelson, M.: The linguistic structure of English Web-search queries. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 1021–1030. Association for Computational Linguistics, Stroudsburg (2008)

[10] Li, X.: Understanding the semantic structure of noun phrase queries. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 1337–1345. Association for Computational Linguistics, Stroudsburg (2010)

[11] Marcos, M.C., Gonzalez-Caro, C.: Comportamiento de los usuarios en la página de resultados de los buscadores. Un estudio basado en eye tracking. El Profesional de la Información 19(4) (July-August 2010)

[12] Mateos, P., Longley, P., Webber, R.: El analisis geodemográfico de apellidos en México. Papeles de Población (65), 73–103 (2010)

[13] Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I.: FreeLing 2.1: Five Years of Open-source Language Processing Tools. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, LREC 2010. European Language Resources Association (ELRA), Valletta (2010)

[14] Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (August 2009)

[15] Steinberger, R., Pouliquen, B., Kabadjov, M.A., Belyaeva, J., der Goot, E.V.: JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. In: Proceedings of the International Conferenece, RANLP 2011, pp. 104–110 (2011)