A Testbed for People Searching Strategies in the WWW

Javier Artiles
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
javart@bec.uned.es

Julio Gonzalo
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
julio@lsi.uned.es

Felisa Verdejo
Dpto. Lenguajes y
Sistemas Informáticos
UNED, Spain
felisa@lsi.uned.es

ABSTRACT

This paper describes the creation of a testbed to evaluate *people searching* strategies on the World-Wide-Web. This task involves resolving person names' ambiguity and locating relevant information characterising every individual under the same name.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Clustering; H.3.4 [**Systems and Software**]: Performance evaluation

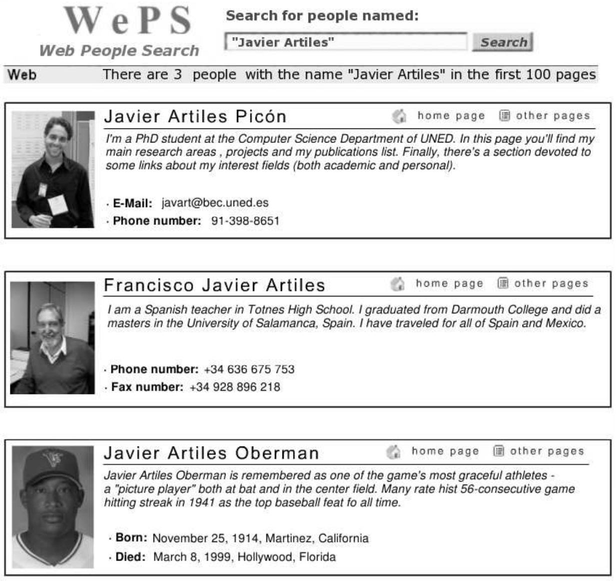
General Terms

Information Retrieval, Information Extraction, Web search

1. MOTIVATION

Finding people -information about people- in the World-Wide-Web is one of the most common activities of Internet users: around 30% of search engine queries include person names [2]. Person names, however, are highly ambiguous: for instance, only 90,000 different names are shared by 100 million people according to the U.S. Census Bureau [2]. In most cases, therefore, the results for this type of searches are a mixture of pages about different people that share the same name. Instead of a ranked list of results, an ideal search engine would return a list of people descriptions, from which the user might select the person she is looking for, and directly access all relevant information for this person. Figure 1 illustrates this idea.

In this paper, we describe the creation of a testbed to evaluate strategies addressing this *people searching* task on web documents. We provide: **(i)** a corpus of web pages retrieved using person names as queries to web search engines; **(ii)** a classification of pages according to the different people (with the same name) they refer to; **(iii)** manual annotations of relevant information -found in the web pages-describing them (e-mail, image, profession, phone number, etc.); **(iv)** the results of applying a general purpose clustering algorithm to that annotated data, which serve as a baseline for the ambiguity resolution problem.



There are 3 people with the name "Javier Artiles" in the first 100 pages

Javier Artiles Picón

I'm a PhD student at the Computer Science Department of UNED. In this page you'll find my main research areas , projects and my publications list. Finally, there's a section devoted to some links about my interest fields (both academic and personal).

- **E-Mail:** javart@bec.uned.es
- **Phone number:** 91-398-8651

Francisco Javier Artiles

I am a Spanish teacher in Totnes High School. I graduated from Darmouth College and did a masters in the University of Salamanca, Spain. I have traveled for all of Spain and Mexico.

- **Phone number:** +34 636 675 753
- **Fax number:** +34 928 896 218

Javier Artiles Oberman

Javier Artiles Oberman is remembered as one of the game's most graceful athletes - a "picture player" both at bat and in the center field. Many rate hist 56-consecutive game hitting streak in 1941 as the top baseball feat fo all time.

- **Born:** November 25, 1914, Martinez, California
- **Died:** March 8, 1999, Hollywood, Florida

Figure 1: Mock-up interface of a search engine able to resolve person names ambiguity

2. THE *WEPS* CORPUS TESTBED

The creation of the WePS (*Web People Search*) corpus consisted of the following steps:

1. Generating ten English person names, using random combinations of the most frequent first and last names in the U.S. Census 1990[1].

2. Collecting the first 100 web pages retrieved by the Google search engine for every (quoted) person name.

3. Grouping documents according to the person they refer to, for every person name.

4. Classifying every web document in the collection as a (i) *homepage entry* (ii) *part of a homepage* (iii) *reference page* (exclusively containing information about the person) and (iv) *other.*

5. Annotating all the occurrences of certain types of

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

[1] http://www.census.gov/genealogy/www/freqnames.html

Table 1: Corpus annotation statistics

category	instances	per name	per person
home page	28	2.8	0.06
part of h.p.	15	1.5	0.03
reference p.	412	41.2	1
other	532	53.2	1.2
tags	**instances**	**per name**	**per person**
name	5,374	537.4	11.60
job	2,105	210.5	4.55
author_of	1,823	182.3	3.94
definition	438	43.8	0.95
date birth	387	38.7	0.84
date death	256	25.6	0.55
image	232	23.2	0.50
place birth	386	38.6	0.83
email	282	28.2	0.61
location	185	18.5	0.40
phone num	136	13.6	0.29
address	86	8.6	0.19
place death	85	8.5	0.18
fax num	37	3.7	0.08
Total	**11,812**	**1,181.2**	**25.51**

descriptive information: name, job, person image, date of birth/death, place of birth/death, email address, postal address, fax/phone number, location (where the person lives in), author_of (e.g. books, paintings, patents...) and description (a brief definition of the person).

Table 1 summarises the results of this exhaustive annotation process. A total of 11,812 text fragments have been semantically annotated, with an average of 25.51 annotations per person. The ambiguity of our set of person names is very high, with an average of 41 different people sharing each person name (see Table 2). This indicates that they are very common names, and also that, in general, none of them corresponds to any web celebrity (i.e. a person dominating top-ranked web hits). The most common tags are *name* (which includes name variants), *job* and *author_of* (mostly titles of books and other written materials). Note that there are few pages classified as *home page*, and even less pages tagged as a *part of a home page*. Nevertheless, each identified person has, in average, one explicit description (*reference page*).

3. BASELINE AMBIGUITY RESOLUTION USING CLUSTERING TECHNIQUES

How difficult is the ambiguity resolution task? Does it demand strategies specifically designed for it, or will generic clustering techniques suffice? Is it necessary to consider the full content of web pages, or the snippets provided by search engines provide enough information for an accurate grouping of results?

To provide initial answers to these questions, we have implemented and tested the *Agglomerative Vector Space* clustering algorithm, which has been previously used to evaluate similar tasks [1], and does not require fixing the number of clusters (K) a priori. In our experiments, terms have been weighted with a logarithmic tf-idf criteria.

The clustering method has been tested using two approaches to build the vector representation of the documents: the first one (*full text*) uses all the textual contents of the web page as input for the algorithm, and the second one (*snippets*) only considers the snippets in the ranked lists provided by Google. Roughly speaking, they consist of a window of approximately 18 terms around one or more occurrences of the person name in the web page. In both cases, we only take into account text inside the html <body> tag, removing stopwords and html tags. Words were stemmed using Porter's algorithm. The similarity threshold for the clustering algorithm has been empirically adjusted to 0.1.

Table 2: Clustering using full text/snippets as a bag of terms

	#	Full Text/Snippets		
name	**people**	**K**	$F_{\alpha=0.5}$	$F_{\alpha=0.2}$
Ann Hill	55	51/38	.88/.81	.88/.88
Angela Thomas	36	34/37	.81/.88	.82/.88
Brenda Clark	23	30/27	.88/.87	.85/.84
Christine King	29	33/44	.67/.74	.70/.70
Helen Miller	38	46/64	.62/.65	.60/.57
Lisa Harris	30	33/36	.83/.79	.83/.76
Mary Johnson	54	40/41	.75/.77	.83/.83
Nancy Thompson	47	33/42	.81/.78	.81/.77
Samuel Baker	38	26/31	.79/.84	.87/.87
Sarah Wilson	62	35/47	.70/.86	.81/.86
Mean	41	36/40	.77/.79	.80/.79

Table 2 shows the results of the experiment. Two different evaluation measures are reported: $F_{\alpha=0.5}$ is a harmonic mean of purity and inverse purity, and $F_{\alpha=0.2}$ is a version of F that gives more importance to inverse purity. Rationale for using $F_{\alpha=0.2}$ is that, from a user's point of view, it is easier to discard a few incorrect web pages in a cluster which has all the information needed, than having to collect the relevant information across many different clusters. In average, clustering with full text obtains $F_{\alpha=0.2} = .80$, which can be seen as a reasonably strong baseline for the task. In addition, using only snippets (which can be much more efficient when searching online) gives $F_{\alpha=0.2} = .79$ (-1.3%). This small difference suggests that snippets can be useful for clustering. But, of course, the next step would be extracting all descriptive features of each person, a task for which the full web page content is necessary.

4. CONCLUSIONS

The WePS corpus provides an initial testbed to test people search strategies over the web. We are currently working to expand and balance the corpus, including two additional types of person names: less frequent names, on one hand (for which ambiguity should be lower), and "celebrity" names, where one person dominates the top-ranked results of search engine results. The expanded corpus will be available at http://nlp.uned.es.

5. REFERENCES

[1] C. H. Gooi and J. Allan. Cross-Document Coreference on a Large Scale Corpus. Technical report, Center for Intelligent Information Retrieval, Department of Computer Science. Univ. of Massachusetts, 2004.

[2] R. V. Guha and A. Garg. Disambiguating People in Search. In *Proceedings of the 13th World Wide Web Conference (WWW 2004), ACM Press, 2004.*