The Impact of Query Refinement in the Web People Search Task

Javier Artiles
UNED NLP & IR group
Madrid, Spain
javart@bec.uned.es

Julio Gonzalo
UNED NLP & IR group
Madrid, Spain
julio@lsi.uned.es

Enrique Amigó
UNED NLP & IR group
Madrid, Spain
enrique@lsi.uned.es

Abstract

Searching for a person name in a Web Search Engine usually leads to a number of web pages that refer to several people sharing the same name. In this paper we study whether it is reasonable to assume that pages about the desired person can be filtered by the user by adding query terms. Our results indicate that, although in most occasions there is a query refinement that gives all and only those pages related to an individual, it is unlikely that the user is able to find this expression a priori.

1 Introduction

The Web has now become an essential resource to obtain information about individuals but, at the same time, its growth has made web people search (WePS) a challenging task, because every single name is usually shared by many different people. One of the mainstream approaches to solve this problem is designing meta-search engines that cluster search results, producing one cluster per person which contains all documents referring to this person.

Up to now, two evaluation campaigns – WePS 1 in 2007 (Artiles et al., 2007) and WePS 2 in 2009 (Artiles et al., 2009) – have produced datasets for this clustering task, with over 15 research groups submitting results in each campaign. Since the release of the first datasets, this task is becoming an increasingly popular research topic among Information Retrieval and Natural Language Processing researchers.

For precision oriented queries (for instance, finding the homepage, the email or the phone number of a given person), clustered results might help locating the desired data faster while avoiding confusion with other people sharing the same name. But the utility of clustering is more obvious for recall oriented queries, where the goal is to mine the web for information about a person. In a typical hiring process, for instance, candidates are evaluated not only according to their cv, but also according to their *web profile*, i.e. information about them available in the Web.

One question that naturally arises is whether search results clustering can effectively help users for this task. Eventually, a query refinement made by the user – for instance, adding an affiliation or a location – might have the desired disambiguation effect without compromising recall. The hypothesis underlying most research on Web People Search is that query refinement is risky, because it can enhance precision but it will usually harm recall. Adding the current affiliation of a person, for instance, might make information about previous jobs disappear from search results.

This hypothesis has not, up to now, been empirically confirmed, and it is the goal of this paper. We want to evaluate the actual impact of using query refinements in the Web People Search (WePS) clustering task (as defined in the framework of the WePS evaluation). For this, we have studied to what extent a query refinement can successfully filter relevant results and which type of refinements are the most successful. In our experiments we have considered the search results associated to one individual as a set of relevant documents, and we have tested the ability of different query refinement strategies to retrieve those documents. Our results are conclusive: in most occasions there is a “near-perfect” refinement that filters out most relevant information about a given person, but this refinement is very hard to predict from a user’s perspective.

In Section 2 we describe the datasets that where used for our experiments. The experimental methodology and results are presented in Section 3. Finally we present our conclusions in 4.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 361–364,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

2 Dataset

2.1 The WePS-2 corpus

For our experiments we have used the WePS-2 testbed (Artiles et al., 2009) [1]. It consists of 30 datasets, each one related to one ambiguous name: 10 names were sampled from the US Census, 10 from Wikipedia, and 10 from the Computer Science domain (Programme Committee members of the ACL 2008 Conference). Each dataset consists of, at most, 100 web pages written in English and retrieved as the top search results of a web search engine, using the (quoted) person name as query[2].

Annotators were asked to organize the web pages from each dataset in groups where all documents refer to the same person. For instance, the "James Patterson“ web results were gruped in four clusters according to the four individuals mentioned with that name in the documents. In cases where a web page refers to more than one person using the same ambiguous name (e.g. a web page with search results from Amazon), the document is assigned to as many groups as necessary. Documents were discarded when there wasn't enough information to cluster them correctly.

2.2 Query refinement candidates

In order to generate query refinement candidates, we extracted several types of features from each document. First, we applied a simple preprocessing to the HTML documents in the corpus, converting them to plain text and tokenizing. Then, we extracted tokens and word n-grams for each document (up to four words lenght). A list of English stopwords was used to remove tokens and n-grams beginning or ending with a stopword. Using the Stanford Named Entity Recognition Tool[3] we obtained the lists of persons, locations and organizations mentioned in each document.

Additionally, we used attributes manually annotated for the WePS-2 Attribute Extraction Task (Sekine and Artiles, 2009). These are person attributes (affiliation, occupation, variations of name, date of birth, etc.) for each individual sharing the name searched. These attributes emulate the kind of query refinements that a user might try in a typical people search scenario.

field	F	prec.	recall	cover.
ae_affiliation	0.99	0.98	1.00	0.46
ae_award	1.00	1.00	1.00	0.04
ae_birthplace	1.00	1.00	1.00	0.09
ae_degree	0.85	0.80	1.00	0.10
ae_email	1.00	1.00	1.00	0.11
ae_fax	1.00	1.00	1.00	0.06
ae_location	0.99	0.99	1.00	0.27
ae_major	1.00	1.00	1.00	0.07
ae_mentor	1.00	1.00	1.00	0.03
ae_nationality	1.00	1.00	1.00	0.01
ae_occupation	0.95	0.93	1.00	0.48
ae_phone	0.99	0.99	1.00	0.13
ae_relatives	0.99	0.98	1.00	0.15
ae_school	0.99	0.99	1.00	0.15
ae_work	0.96	0.95	1.00	0.07
stf_location	0.96	0.95	1.00	0.93
stf_organization	1.00	1.00	1.00	0.98
stf_person	0.98	0.97	1.00	0.82
tokens	1.00	1.00	1.00	1.00
bigrams	1.00	1.00	1.00	0.98
trigrams	1.00	1.00	1.00	1.00
fourgrams	1.00	1.00	1.00	0.98
fivegrams	1.00	1.00	1.00	0.98

Table 1: Results for clusters of size 1

field	F	prec.	recall	cover.
ae_affiliation	0.76	0.99	0.65	0.40
ae_award	0.67	1.00	0.50	0.02
ae_birthplace	0.67	1.00	0.50	0.10
ae_degree	0.63	0.87	0.54	0.15
ae_email	0.74	1.00	0.60	0.16
ae_fax	0.67	1.00	0.50	0.09
ae_location	0.77	1.00	0.66	0.32
ae_major	0.71	1.00	0.56	0.09
ae_mentor	0.75	1.00	0.63	0.04
ae_nationality	0.67	1.00	0.50	0.01
ae_occupation	0.76	0.98	0.65	0.52
ae_phone	0.75	1.00	0.63	0.13
ae_relatives	0.78	0.96	0.68	0.15
ae_school	0.68	0.96	0.56	0.17
ae_work	0.81	1.00	0.72	0.17
stf_location	0.83	0.97	0.77	0.98
stf_organization	0.89	1.00	0.83	1.00
stf_person	0.83	0.99	0.74	0.98
tokens	0.96	0.99	0.94	1.00
bigrams	0.95	1.00	0.92	1.00
trigrams	0.94	1.00	0.92	1.00
fourgrams	0.91	1.00	0.86	0.99
fivegrams	0.89	1.00	0.84	0.99

Table 2: Results for clusters of size 2

field	F	prec.	recall	cover.
ae_affiliation	0.51	0.96	0.39	0.81
ae_award	0.26	1.00	0.16	0.20
ae_birthplace	0.33	0.99	0.24	0.28
ae_degree	0.37	0.90	0.26	0.36
ae_email	0.35	0.96	0.23	0.33
ae_fax	0.30	1.00	0.19	0.15
ae_location	0.34	0.96	0.23	0.64
ae_major	0.30	0.97	0.20	0.22
ae_mentor	0.23	0.95	0.15	0.22
ae_nationality	0.36	0.88	0.26	0.16
ae_occupation	0.52	0.93	0.40	0.80
ae_phone	0.34	0.96	0.23	0.33
ae_relatives	0.32	0.95	0.22	0.16
ae_school	0.40	0.95	0.29	0.43
ae_work	0.45	0.94	0.34	0.38
stf_location	0.62	0.87	0.53	1.00
stf_organization	0.67	0.96	0.56	1.00
stf_person	0.59	0.95	0.47	1.00
tokens	0.87	0.90	0.86	1.00
bigrams	0.79	0.95	0.70	1.00
trigrams	0.75	0.96	0.65	1.00
fourgrams	0.67	0.97	0.55	1.00
fivegrams	0.62	0.96	0.50	1.00

Table 3: Results for clusters of size >=3

3 Experiments

In our experiments we consider each set of documents (cluster) related to one individual in the WePS corpus as a set of relevant documents for a person search. For instance the James Patter-

[1] http://nlp.uned.es/weps

[2] We used the Yahoo! search service API.

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

field	F	prec.	recall	cover.
best-ae	1.00	0.99	1.00	**0.74**
best-all	1.00	1.00	1.00	1.00
best-ner	1.00	1.00	1.00	0.99
best-nl	1.00	1.00	1.00	1.00

Table 4: Results for clusters of size 1

field	F	prec.	recall	cover.
best-ae	0.77	1.00	0.65	**0.79**
best-all	0.95	1.00	0.93	1.00
best-ner	0.92	0.99	0.88	1.00
best-nl	0.96	1.00	0.94	1.00

Table 5: Results for clusters of size 2

field	F	prec.	recall	cover.
best-ae	0.60	0.97	0.47	**0.92**
best-all	0.89	0.96	0.85	1.00
best-ner	0.74	0.95	0.63	1.00
best-nl	0.89	0.95	0.85	1.00

Table 6: Results for clusters of size >=3

son dataset in the WePS corpus contains a total of 100 documents, and 10 of them belong to a British politician named James Patterson. The WePS-2 corpus contains a total of 552 clusters that were used to evaluate the different types of QRs.

For each person cluster, our goal is to find the best query refinements; in an ideal case, an expression that is present in all documents in the cluster, and not present in documents outside the cluster. For each QR type (affiliation, e-mail, n-grams of various sizes, etc.) we consider all candidates found in at least one document from the cluster, and pick up the one that leads to the best harmonic mean ($F_{\alpha=.5}$) of precision and recall on the cluster documents (there might be more than one).

For instance, when we evaluate a set of *token* QR candidates for the politician in the James Patterson dataset we find that among all the tokens that appear in the documents of its cluster, "republican" gives us a perfect score, while “politician“ obtains a low precision (we retrieve documents of other politicians named James Patterson).

In some cases a cluster might not have any candidate for a particular type of QR. For instance, manual person attributes like phone number are sparse and won't be available for every individual, whereas tokens and ngrams are always present. We exclude those cases when computing F, and instead we report a *coverage* measure which represents the number of clusters which have at least one candidate of this type of QR. This way we know how often we can use an attribute (coverage) and how useful it is when available (F measure).

field	1	2	>=3
ae_affiliation	20.96	17.88	29.41
ae_occupation	20.25	21.79	24.60
ae_work	3.23	8.38	8.56
ae_location	12.66	12.29	8.02
ae_school	7.03	6.70	6.42
ae_degree	3.23	3.91	5.35
ae_email	5.34	6.15	4.28
ae_phone	6.19	5.03	3.21
ae_nationality	0.28	0.00	3.21
ae_relatives	7.03	5.03	2.67
ae_birthplace	4.22	5.03	1.60
ae_fax	2.95	1.68	1.60
ae_major	3.52	3.91	1.07
ae_mentor	1.41	2.23	0.00
ae_award	1.69	0.00	0.00

Table 7: Distribution of the person attributes used for the "best-ae“ strategy

These figures represent a ceiling for each type of query refinement: they represent the efficiency of the query when the user selects the best possible refinement for a given QR type.

We have split the results in three groups depending on the size of the target cluster: (i) rare people, mentioned in only one document (335 clusters of size 1); (ii)people that appear in two documents (92 clusters of size 2), often these documents belong to the same domain, or are very similar; and (iii) all other cases (125 clusters of size >=3).

We also report on the aggregated results for certain subsets of QR types. For instance, if we want to know what results will get a user that picks the best person attribute, we consider all types of attributes (e-mail, affiliation, etc.) for every cluster, and pick up the ones that lead to the best results.

We consider four groups: (i) *best-all* selects the best QR among all the available QR types (ii) *best-ae* considers all manually annotated attributes (iii) *best-ner* considers automatically annotated NEs; and (iv) *best-ng* uses only tokens and ngrams.

3.1 Results

The results of the evaluation for each cluster size (one, two, more than two) are presented in Tables 1, 2 and 3. These tables display results for each QR type. Then Tables 4, 5 and 6 show the results for aggregated QR types.

Two main results can be highlighted: (i) The best overall refinement is, in average, very good ($F = .89$ for clusters of size ≥ 3). In other words, there is usually at least one QR that leads to (approximately) the desired set of results; (ii) this best refinement, however, is not necessarily an intuitive choice for the user. One would expect users to refine the query with a person's attribute, such as his affiliation or location. But the results for the best (manually extracted) attribute are significantly worse ($F = .60$ for clusters of size ≥ 3), and they cannot always be used (coverage is .74, .79 and .92 for clusters of size 1, 2 and ≥ 3).

The manually tagged attributes from WePS-2 are very precise, although their individual coverage over the different person clusters is generally low. Affiliation and occupation, which are the most frequent, obtain the largest coverage (0.81 and 0.80 for sizes ≥ 3). Also the recall of this type of QRs is low in clusters of two, three or more documents. When evaluating the "best-ae" strategy we found that in many clusters there is at least one manual attribute that can be used as QR with high precision. This is the case mostly for clusters of three or more documents (0.92 coverage) and it decreases with smaller clusters, probably because there is less information about the person and thus less biographical attributes are to be found.

In Table 7 we show the distribution of the actual QR types selected by the "best-ae" strategy. The best type is affiliation, which is selected in 29% of the cases. Affiliation and occupation together cover around half of the cases (54%), and the rest is a long tail where each attribute makes a small contribution to the total. Again, this is a strong indication that the best refinement is probably very difficult to predict a priori for the user.

Automatically recognized named entities in the documents obtain better results, in general, than manually tagged attributes. This is probably due to the fact that they can capture all kinds of related entities, or simply entities that happen to coocur with the person name. For instance, the pages of a university professor that is usually mentioned together with his PhD students could be refined with any of their names. This goes to show that a good QR can be any information related to the person, and that we might need to know the person very well in advance in order to choose this QR.

Tokens and ngrams give us a kind of "upper boundary" of what is possible to achieve using QRs. They include almost anything that is found in the manual attributes and the named entities. They also frequently include QRs that are not realistic for a human refinement. For instance, in clusters of only two documents it is not uncommon that both pages belong to the same domain or that they are near duplicates. In those cases tokens and ngram QR will probably include non informative strings. In some cases the QRs found are neither directly biographical or related NEs, but topical information (e.g. the term "soccer" in the pages of a football player or the ngram "alignment via structured multilabel" that is the title of a paper written by a Computer Science researcher). These cases widen even more the range of effective QRs. The overall results of using tokens and ngrams are almost perfect for all clusters, but at the cost of considering every possible bit of information about the person or even unrelated text.

4 Conclusions

In this paper we have studied the potential effects of using query refinements to perform the Web People Search task. We have shown that although in theory there are query refinements that perform well to retrieve the documents of most individuals, the nature of these ideal refinements varies widely in the studied dataset, and there is no single intuitive strategy leading to robust results. Even if the attributes of the person are well known beforehand (which is hardly realistic, given that in most cases this is precisely the information needed by the user), there is no way of anticipating which expression will lead to good results for a particular person. These results confirm that search results clustering might indeed be of practical help for users in Web people search.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. 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.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2009. Weps 2 evaluation campaign: overview of the web people search clustering task. In *WePS 2 Evaluation Workshop. WWW Conference 2009*.

Satoshi Sekine and Javier Artiles. 2009. Weps2 attribute extraction task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*.