# Named Entity Recognition and Identification for Finding the Owner of a Home Page

Vassilis Plachouras[1,2], Matthieu Rivière[2], and Michalis Vazirgiannis[1,3]

[1] LIX, École Polytechnique, Palaiseau, France
vassilis.plachouras@presans.com
[2] PRESANS, X-TEC, École Polytechnique, Palaiseau, France
matthieu.riviere@presans.com
[3] Dept of Informatics, AUEB, Athens, Greece
mvazirg@aueb.gr

**Abstract.** Entity-based applications, such as expert search or online social networks where users search for persons, require high-quality datasets of named entity references. Obtaining such high-quality datasets can be achieved by automatically extracting metadata from Web pages. In this work, we focus on the identification of the named entity that corresponds to the owner of a particular Web page, for example, a home page or an organizational staff Web page. More specifically, from a set of named entities that have already been extracted from a Web page, we identify the one which corresponds to the owner of the home page. First, we develop a set of features which are combined in a scoring function to select the named entity of the Web page owner. Second, we formulate the problem as a classification problem in which a pair of a Web page and named entity is classified as being *associated* or not. We evaluate the proposed approaches on a set of Web pages in which we have previously identified named entities. Our experimental results show that we can identify the named entity corresponding to the owner of a home page with accuracy over 90%.

**Keywords:** named entity recognition, entity selection.

## 1 Introduction

Developing named entity-based datasets is a central task to applications such as expert search engines and scientific digital library portals, where researchers and organizations are the key entities to index and search for. However, developing such datasets is challenging because information must be extracted from unstructured or semi-structured sources. One approach involves the extraction of information from bibliographic metadata of scientific publications. DBLP[1] is an example of a site offering an index of the literature in Computer Science. CiteSeerX[2] crawls the Web to collect files that correspond to publications and

[1] http://www.informatik.uni-trier.de/~ley/db/
[2] http://citeseerx.ist.psu.edu/

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 554–565, 2012.

from which information is extracted. A complementary approach is to extract information and to identify researchers from crawled Web pages of academic institutions. By exploiting the publicly available Web sites of academic institutions, this approach has the potential to achieve higher coverage when there is no bibliographic metadata available.

In this work, we consider the latter approach to develop an entity-based dataset of researchers, covering several research fields and different countries. We describe a mechanism for identifying with high accuracy the named entity that corresponds to the owner of a Web page. In other words, given a Web page $p$, we identify the person entity $e$ of the Web page's owner. Such Web pages are either home pages of people, or organizational staff Web pages, similar to online business cards. For example, the owner's named entity of a researcher's home page is the researcher's name.

Related works propose to identify the owner of a home page by learning models of the structure of home pages and the position of names on the Web page. For example, Gollapalli *et al.* [8] select the first identified name on the Web page. This simple heuristic is effective because the name of the owner of the home page is likely to appear before any other name. However, the effectiveness of this heuristic is highly dependent on the effectiveness of named entity recognition, because if the first name is not identified correctly, then the selected name is likely to be wrong.

We take a different approach, where we first apply a named entity recognizer to extract all names appearing on a Web page and then, we exploit the likely redundancy of the names' occurrences on a Web page to identify the name of the Web page's owner. For example, the name of the Web page's owner is likely to appear more than once in the Web page. In addition, it may appear both in its full form as well as in abbreviated forms in publication references. We develop weighting functions for the identified named entities and select the top-scoring ones as the named entities of a Web page's owner. The weighting functions are based on the output of the named entity recognizer and exploit similarities between names by constructing a graph whose vertices are the named entities that have been recognized in a Web page. Furthermore, we treat the problem of selecting the named entity as a binary classification problem and train an SVM classifier to identify those named entities.

An important advantage of our approach over existing ones is that it does not depend on a particular named entity recognition model. Instead, it can use any method that detects the named entities with the required granularity, that is, given names, last names and middle names. The experimental results, based on a dataset of 472 home pages manually annotated with the name of their owner, show that our proposed approaches can achieve over 87% precision in identifying the named entity of the Web page owner. When considering only the Web pages in which the correct name has been recognized at least once by the named entity recognition model, the introduced approaches achieve over 95% precision.

The remainder of this paper is organized as follows. In Section 2 we present related works from the literature. In Section 3 we briefly describe the named

entity recognition method and features we employ to identify the named entities from which we select the Web page owner named entities. Section 4 introduces a framework for the selection of named entities and describes two baseline approaches and one based the construction of a graph from named entities that are similar, exploiting the redundancy in the named entity occurrences. In Section 5 we describe an approach based on supervised machine learning and, more specifically, a binary SVM classifier, which is trained to select the named entities from a home page. Section 6 describes our dataset and the experimental results we have obtained. Finally, Section 7 closes this work with some concluding remarks.

## 2 Related Work

The approach to identify the named entity of the home page owner is primarily related to named entity recognition, metadata extraction from Web pages, and to coreference resolution.

*Named Entity Recognition.* Supervised machine learning techniques are typically used to identify named entities in texts and Web pages. An example of a generative model is a Hidden Markov Model (HMM), in which the hidden states are used to model the tag classes of words. Bikel *et al.* [1] develop a named entity recognizer based on a HMM, where the hidden states of the HMM correspond to a number of name classes, such as person names, or organization names, and the features involve checking for capitalization, whether a word contains only letters, or digits. Chieu and Ng[4] employ a Maximum Entropy Classifier, which classifies each word in a text as the beginning of a named entity, the continuation, or the last word of a named entity. The features employed by the Maximum Entropy Classifier are binary. Takeuchi and Collier [14] explore the use of Support Vector Machines for named entity recognition, computing features from a context of the three previous and three following tokens. An approach that has been commonly used and results in state-of-the-art performance is Conditional Random Fields (CRF) [10], which is an undirected graphical model or a Markov Random Field. Culota *et al.* [5] extract contact information from the home pages of persons identified in email corpora. Minkow *et al.* [11] apply a CRF model to recognize names in emails, using features which are primarily based on gazetteers for person first and last names, names of organizations and locations, but not using deep natural language processing. Zhu *et al.* [17] propose two-dimensional CRF, which take into account not only the sequence of information objects in a Web page, but also the dependencies between neighboring blocks. Shi and Wang have proposed a dual-layer CRF, which aims to process more accurately cascades of subtasks in Natural Language Processing [13]. For example, one such cascade of tasks is the identification of the full person names in a text, as well as the given and last names. In our work, we employ an approach similar to cascaded CRFs where the predicted labels for the person names are used as a feature in the prediction of the first and last names of persons. We discuss in more detail the CRF model we employ in Section 3.

*Metadata Extraction.* While we use named entity recognition to identify person names on a Web page, the focus of our work is on selecting the name of the owner of a professional or academic home page. Hence, our work is more closely related to [9][3][8]. Kato et al. [9] employ the concept of information sender to identify the author of a Web page or the organization to which the Web page belongs. They treat the problem as a ranking problem evaluating at the top-5 results. The reported precision at ranks 1 and 5 is 0.586 and 0.752, respectively. Changuel et. al [3] extract the author of Web pages, not necessarily home pages, by building a decision tree with the C4.5 algorithm and employing a small set of features. They report a precision of approximately 0.812 in identifying the author of a web page. Gollapalli *et al.* [8] identify the owner of a home page by applying a standard named entity recognition model and selecting the first identified name. Zheng et al. [16] describe an approach based on Conditional Random Fields to identify the metadata about authors from their home pages using visual features, such as the position of DOM nodes on the rendered Web page. Finally, Tang et al. [15] start from a dataset of bibliographic metadata and create Web search engine queries to retrieve the home page of a user. Their setting is different from ours where we aim to extract the names of persons, without assuming that we have any information about the names *a priori*.

*Coreference Resolution.* The task of selecting the main entity from the set of entities identified on a home page is related to coreference resolution, which determines whether two textual expressions refer to the same entity or not. Typical coreference resolution methods employ supervised learning [12][6] and rely on the linguistic analysis of text to extract features. The task of identifying the owner of a home page does not require the full resolution of all references, and hence, it is not necessary to apply coreference resolution at a first step.

## 3 Named Entity Recognition

Before presenting our approach to named entity selection, we describe the Named Entity Recognition (NER) system we first apply to extract names from home pages. We have developed a NER system based on supervised learning of a Conditional Random Field (CRF) to learn to recognize the full names of persons, as well as their first, middle and last names. We did not employ an existing NER system such as the Stanford Named Entity Recognizer[3] [7] for two main reasons. First, we require better granularity in identifying first, last and middle names in addition to full names. Second, our objective is to process input from Web pages. Hence, we develop features that exploit term frequency statistics in the anchor text of incoming hyperlinks of Web pages.

To train the CRF model, we have manually annotated all the names in 95 Web pages. We first split the textual content of a Web page in sentences using the DOM tree and regular expressions. Next, we tokenize each sentence by splitting tokens at non-alphanumeric characters, and we annotate the tokens. We use

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

the *begin, inside, outside* (BIO) convention for labels. For example the sentence *"Chris Bishop is a Distinguished Scientist at ..."*[4] is tokenized and labeled as follows:

| Chris | Bishop | is | a | Distinguished | Scientist | at | ... |
|---|---|---|---|---|---|---|---|
| BPERSON | IPERSON | O | O | O | O | O | |
| BFNAME | BLNAME | O | O | O | O | O | |

where BPERSON denotes the beginning of a full name, BFNAME denotes the beginning of a first name, and BLNAME denotes the beginning of a last name. The label IPERSON denotes that the corresponding token is inside a person name. The label O denotes that the corresponding token does not belong to any of the classes we consider.

Next, we train a CRF model using five types of features. The first type of features corresponds to the tokens themselves. The second type corresponds to two features, whose value depends on the form of the examined token. The first feature indicates whether the token contains only numerical digits, or it is a single upper case letter, or a punctuation symbol, or a capitalized word, *etc.*). The second feature indicates whether the token is an alphanumeric string. The third type of features relies on two gazetteers for first names and geographic locations, respectively, and comprises two binary features indicating whether the token is a first name, and whether the token is a geographic location. The fourth type of features is based on a full-text index of Web pages and comprises 4 features. More specifically, two features correspond to the logarithm of the number of documents in which the term occurs in the body, and the anchor text of incoming links respectively. The two next features correspond to flags indicating whether the term occurs in the title or the anchor text of incoming links of the currently processed document. The fifth type of features comprises one feature indicating whether the token occurs in the anchor text of an outgoing hyperlink in the currently processed document, differentiating between links to Web pages in the same or different domains. Note that the last two types of features depend on the distribution of terms in a full text index of Web pages, and the text associated with the link structure of Web pages. We employ the implementation of CRF++[5].

We learn the CRF model and apply it to unseen Web pages in the following way. First, we train a CRF to recognize full names and assign labels BPERSON and IPERSON . The assigned labels are then used to learn a second model where the assigned labels constitute an additional twelfth feature used in the recognition of first, middle and last names, assigning labels BFNAME, IFNAME, BMNAME, IMNAME, and BLNAME and ILNAME, respectively. After applying the CRF models to label tokens, we aggregate consecutive tokens with B and I labels in an entity $e$. For each entity $e$, $t(e)$ is the type of the entity where $t(e) \in \{\texttt{PERSON}, \texttt{FNAME}, \texttt{MNAME}, \texttt{LNAME}\}$, $c(e)$ is the average confidence of the label assignment over the entity's tokens, and $s(e)$ is the concatenation of the tokens to form the string representation of $e$.

[4] Quoted from http://research.microsoft.com/en-us/um/people/cmbishop/

[5] Available from http://crfpp.sourceforge.net/

The accuracy of the named entity recognition could potentially be higher if we employed language-specific features, such as Part Of Speech (POS) tags. However, our aim is to apply the developed approaches to a wide range of input Web pages, irrespectively of the language they are written in. We offset the potentially lower accuracy of the CRF named entity recognition by weighting the different occurrences of names, as described in the following section.

## 4 Finding Named Entities of Web Page Owner

In this section, we study the problem of selecting the named entity corresponding to the owner of a home page. We operate on the output of the named entity recognition process described in Section 3 to select the entity $e$ with type $t(e) = PERSON$. First, we describe the framework for weighting the identified entities (Section 4.1). Then, we introduce two baseline weighting functions for entities based on the features used by the NER system (Section 4.2), and a third weighting function based on a graph representation of the named entities (Section 4.3).

### 4.1 Entity Selection Framework

We perform entity selection in the following framework. $S(t, str)$ is the set of all entities of type $t(e) = t$ and string representation $s(e) = str$:

$$S(str) = \{e | t(e) = t \wedge s(e) = str\} \tag{1}$$

When $t =$ `PERSON` we write $S(str) = S(\texttt{PERSON}, str)$. For each set $S(str)$, we compute a weight $w_{str}$ and rank $S(str)$ in descending order of $w_{str}$. The selected named entities of the processed Web page are the ones belonging to the top ranked $S(str)$.

### 4.2 Baseline Entity Selection

A simple way to weight a set $S(str)$ of `PERSON` entities with the same string representation is to sum the confidence $c(e)$ of the label assignment for each entity $e \in S(str)$:

$$w_{str} = \sum_{e \in S(str)} c(e) \tag{2}$$

The intuition for defining the weight $w_{str}$ as the sum of the confidences is that it reflects both the number of times the same string has been identified as a `PERSON` entity as well as the confidence in the recognition.

The weighting of $S(str)$ from Eq. 2 is only based on the average confidence of the label assignment to each token of the entities in $S(str)$. We can improve the weighting by incorporating more information regarding the position of the occurrences of entities.

$$w_{str} = \sum_{e \in S(str)} (w_a anchor(e) + w_t title(e) + w_c c(e)) \tag{3}$$

where $anchor(e) = 1$ if $s(e)$ occurs in the anchor text of incoming hyperlinks of the processed Web page, otherwise $anchor(e) = 0$. Similarly, $title(e) = 1$ if $s(e)$ occurs in the title of the processed Web page, otherwise $title(e) = 0$. The parameters $w_a, w_t, w_c$ control the importance of each of the three features and are set during training.

### 4.3 Graph-Based Entity Selection

The baseline weighting of set $S(str)$ according to Eq. 2 and 3 only consider the entities of type `PERSON` with the same string representation. However, they ignore any similarities between the identified entities in order to compute an improved weight. Suppose that on a Web page the full name of a researcher appears only twice at the top of the Web page, and the name of the most frequent co-author appears in abbreviated form once for each publication of the researcher[6]. In such a setting, the baseline weighting functions may select the abbreviated name of the co-author as the named entity of the URL's owner, instead of the full name of the researcher.

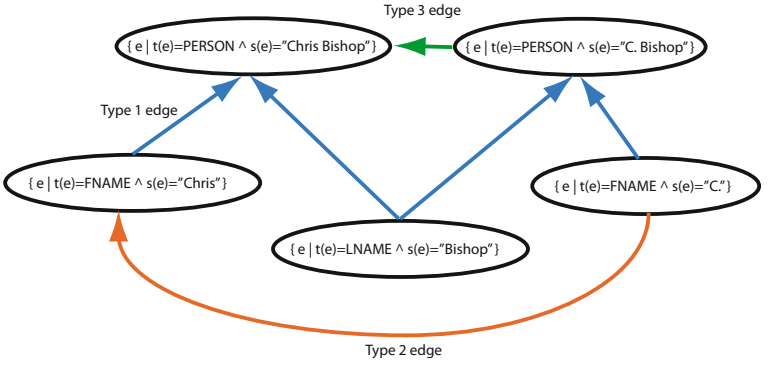


**Fig. 1.** The graph constructed from the sets of identified named entities in a Web page

We overcome the limitations of the baseline weightings by introducing a novel graph-based weighting for sets of entities. We define a directed graph $G = \{V, E\}$ where $V$ is the set of vertexes and $E$ is the set of edges. Each set $S(t, str) = \{e | t(e) = t \wedge s(e) = str\}$ of entities with given type $t$ and string representation $str$, corresponds to a vertex of $V$. Hence, the graph is constructed from all identified names in the Web page.

We define three types of directed edges in graph $G$. The set of vertices having an edge of type $i$ to $S(t, str)$ is denoted by $in_i(t, str)$. When $t =$ `PERSON`, we can write $in_i(str)$. The three types of directed edges are defined as follows:

[6] For example, `http://www.cs.washington.edu/homes/pedrod/`

- A type 1 edge connects sets of `FNAME, MNAME, LNAME` entities to the corresponding sets of `PERSON` entities in which they occur.
- A type 2 edge connects a set $S(t, str1)$ to $S(t, str2)$ when string $str1$ is an abbreviated form of $str2$ and $t \in \{\texttt{FNAME, MNAME, LNAME}\}$.
- A type 3 edge connects a set $S(str1)$ to $S(str2)$ when the name $str1$ is an abbreviated form of the name $str2$. Formally, $S(str1) \in in_3(S(str2))$ if there exists $S(t, str3) \in in_1(\texttt{PERSON}, str1) \cap in_1(\texttt{PERSON}, str2)$ and there exist $S(t', str4) \in in_1(\texttt{PERSON}, str1))$, $S(t'', str5) \in in_1(\texttt{PERSON}, str2)$ where $S(t', str4) \in in_2(t'', str5)$.

Figure 1 illustrates a graph constructed from a set of identified named entities. The graph has two vertexes of type `PERSON`, one vertex of type `LNAME` for the last name *'Bishop'* and two vertexes of type `FNAME` for the first name *'Chris'* and its abbreviated form *'C.'* There are four edges of type 1, linking the vertexes of type `FNAME` and `LNAME` to the corresponding vertexes of type `PERSON`. There is one edge of type 2 which links the vertex $S(\texttt{FNAME}, \text{'C.'})$ to the vertex $S(\texttt{FNAME}, \text{'Chris'})$. Finally, there is one edge of type 3 from $S(\text{'C. Bishop'})$ to $S(\text{'Chris Bishop'})$ because both vertexes have incoming links from the same vertex $S(\texttt{LNAME}, \text{'Bishop'})$ and there is a type 2 edge between two of their `FNAME` linking vertexes.

The graph $G$, which is constructed as described above, is a directed acyclic graph (DAG). From the definition of type 1 edges, we cannot have a cycle involving vertices of type `PERSON` and any other entity type because type 1 edges always point to vertices of type `PERSON`. Hence, a cycle may involve either type 2 edges exclusively or type 3 edges exclusively. Since a type 3 edge exists only if there is a type 2 edge, and the two edges cannot be in the same path, then there exists a cycle with type 3 edges only if there exists a cycle with type 2 edges. However, there cannot be a cycle with type 2 edges, because type 2 edges link an abbreviated name to its full form. Hence, there cannot be any cycle in the graph $G$.

Once we have constructed the graph from the named entities identified in a Web page, we compute a weight for each vertex $S(str)$, corresponding to the sum of the Baseline 2 score from Eq. 3 plus the sum of the scores of vertices that link to $S(str)$.

$$w_{str} = \sum_{e \in S(str)} (w_a anchor(e) + w_t title(e) + w_c c(e)) + \sum_{S(str') \in in_i(str)} w_{str'} \quad (4)$$

Finally, we select the set $S(str)$ with the highest score $w_{str}$ according to Eq. 4. The intuition is that the scores of abbreviated named entities propagate to the entities corresponding to full names.

## 5 Learning to Select Named Entities

The baseline and the graph-based scoring functions make use of the output of the NER system to score entities found in a Web page and select the ones

which are more likely to refer to the owner of the Web page. However, all three functions will always produce a score for the entities, even when the named entity of the owner of the Web page is not among the identified named entities. For example, a researcher may have a set of Web pages documenting a software he has written and released as open-source. The functions introduced earlier will always select one set of entities as the owner for the considered Web page. Moreover, extending these functions with arbitrary features is not trivial. In this section, we investigate the problem of selecting the named entities as a binary classification problem in a supervised learning setting.

In particular, we formulate the classification problem $y(x) \in \{-1, 1\}$, where $x \in \mathcal{X} = \{(\text{URL}, S(\texttt{PERSON}, str))\}$. The input $x$ is a pair of a URL and a set $S(\texttt{PERSON}, str)$. For the output, $y(x) = 1$ when the named entities in $S(\texttt{PERSON}, str)$ correspond to the owner of Web page with URL, otherwise, $y(x) = -1$. For each input point $x$, we compute 13 features:

- the graph-based score of $S(\texttt{PERSON}, str)$ from Eq. 4
- the rank of $S(\texttt{PERSON}, str)$ when all sets of PERSON entities are ordered in ascending order of the Baseline 1, Baseline 2, graph-based scoring functions, as well as in the order of occurrence (4 features)
- the sum of the cardinalities $|S(t, str')|$ where $S(t, str') \in in_i(\texttt{PERSON}, str)$ for each type of links (3 features)
- the number of edges of type $i$ pointing to $S(\texttt{PERSON}, str)$ for $i = 1, 2, 3$ (3 features)
- 1 if *str* appears in an email address found in the content of URL, otherwise 0
- 1 if *str* appears to be emphasized in the text of home page identified by URL, otherwise 0

The feature values are normalized between -1 and +1 on a per home page basis. We employ an SVM classifier with radial-basis kernel from LIBSVM[7] [2]. For a given home page identified by URL, if the SVM classifies as +1 more than one pairs $(\text{URL}, S(\texttt{PERSON}, str))$, we select the one with the highest estimated probability, as computed by the SVM classifier.

## 6 Experimental Results

In this section, we describe the experimental setting in which we evaluate the introduced methods. First, we evaluate the CRF-based named entity recognition (Section 6.1). Next, we describe the dataset we use for entity selection and we present the obtained results (Section 6.2).

### 6.1 Named Entity Recognition Evaluation

In this section, we present evaluation results for the NER system we describe in Section 3. Starting from a set of 95 annotated Web pages, we randomize their

[7] Available from `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

order and split them in three folds. We use each fold once to test the CRF model we learn on the other two folds. Table 1 reports the micro-averaged precision, recall and F-measure for each of the labels we assign during the first and the second passes of the CRF-based NER system, respectively.

The NER system assigns `BPERSON` and `IPERSON` labels with high precision and recall. This is consistent with results reported for NER systems trained on much larger corpora [7]. First and last names are identified with an accuracy of more than 0.80. The obtained precision for middle names is significantly lower, mainly due to the small number of training examples available.

**Table 1.** Number of annotated tokens, micro-averaged precision, recall, and F-measure for each of the labels assigned in the first and second passes, respectively

| Label | # of Annotated Tokens | Precision | Recall | F-Measure |
|---|---|---|---|---|
| | Pass 1 | | | |
| `BPERSON` | 3326 | 0.931 | 0.918 | 0.924 |
| `IPERSON` | 6552 | 0.955 | 0.913 | 0.934 |
| `O` | 35364 | 0.978 | 0.988 | 0.983 |
| | Pass 2 | | | |
| `BFIRST` | 2942 | 0.820 | 0.900 | 0.858 |
| `BLAST` | 2956 | 0.818 | 0.879 | 0.848 |
| `BMIDDLE` | 131 | 0.290 | 0.344 | 0.315 |
| `IFIRST` | 1783 | 0.820 | 0.871 | 0.844 |
| `ILAST` | 777 | 0.832 | 0.793 | 0.812 |
| `IMIDDLE` | 120 | 0.542 | 0.375 | 0.443 |
| `O` | 36533 | 0.975 | 0.960 | 0.967 |

### 6.2 Evaluation and Experimental Results

We have evaluated the introduced approaches using a dataset of home pages, for which we have manually identified the full name, as well as the first, middle and last names of the home page owners. We have sampled a total of 472 home pages from a large crawl of university and research organization Web sites.

The NER model, described in Section 3, has identified the correct name at least once in 432 out of the 472 home pages. Out of the 432 pages, 66% of the pages are written in English, 27% are written in French and 3% of the pages are written in German. The remaining 4% of the pages are written in Danish, Italian, Polish, Portuguese and Swedish. We distinguish between perfect and partial matches of names. We have a perfect match when the entity weighting ranks first a name matching perfectly the correct one. A partial match occurs when the entity weighting ranks first an abbreviated version of the correct name.

Table 2 reports the accuracy of perfect and partial identifications over the 432 home pages for which the correct answer is among the identified named entities. We also report results computed over the total number of home pages. The first approach (Order) is a naïve heuristic where the first identified person name is

selected as the owner's name for the corresponding page. The effectiveness of this heuristic depends on the accuracy of the underlying NER system because any wrong identification of names will lead to an error in the selection [8]. The two next approaches, Baseline 1 and Baseline 2, correspond to the selection of entities using Eq. 2 and 3, respectively. The fourth and fifth rows in Table 2 display the results obtained with the graph-based and the SVM-based entity selection approaches, respectively.

**Table 2.** Fraction of Web pages for which there is a perfect or partial match, when using Order, Baseline 1, Baseline 2, Graph and SVM-based entity selection

| | Perfect | Partial | Perfect+Partial | (Perfect+Partial)/All Pages |
|---|---|---|---|---|
| Order | 0.847 | 0.035 | 0.882 | 0.807 |
| Baseline 1 | 0.789 | 0.090 | 0.880 | 0.805 |
| Baseline 2 | 0.875 | 0.039 | 0.914 | 0.837 |
| Graph-based | 0.944 | 0.014 | 0.958 | 0.877 |
| SVM-based | 0.954 | 0.009 | 0.963 | 0.881 |

The best-performing approach is the SVM-based one, which achieves perfect matches in 95.4% of the home pages when the named entity recognition identifies the correct name at least once. If we consider both perfect and partial matches, then we have a match in 96.3% of the home pages. When we calculate the results on all the home pages, including the ones in which named entity recognition did not identify the correct named entity, we achieve a precision of 88.1%.

## 7 Conclusions

In this work, we have introduced a novel method to select among recognized named entities in a home page the one corresponds to the owner. Our method uses the output of a named entity recognition system and exploits the redundancy and the similarities between names to select the correct one. The introduced methods are developed independently of the employed named entity recognition approach. Indeed, they can be used with any NER approach that identifies person names, but also first, middle and last names. In a dataset of more than 400 home pages, our methods identify the correct name for more than 90% of the home pages in which a NER system identifies at least once the correct name in the processed page. The comparison of our methods with a heuristic based on the order of names shows that our approaches achieve important improvements in effectiveness because they are more robust with respect to the accuracy of the employed NER system.

We have applied the developed methods in the context of researchers' home pages. In the future, we will evaluate it in the context of different applications, such as the automatic creation of online social networks, or people search. We also aim to apply the developed methods for identifying the name of the owner of a Web page as a feature to improve the classification of Web pages.

## References


1. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Procs. of the 5th ANLC, pp. 194–201 (1997)
2. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27 (2011)
3. Changuel, S., Labroche, N., Bouchon-Meunier, B.: Automatic web pages author extraction. In: Procs. of the 8th FQAS, pp. 300–311 (2009)
4. Chieu, H.L., Ng, H.T.: Named entity recognition with a maximum entropy approach. In: Procs. of the 7th Conference on Natural Language Learning at HLT-NAACL 2003, CONLL 2003, vol. 4, pp. 160–163 (2003)
5. Culotta, A., Bekkerman, R., McCallum, A.: Extracting social networks and contact information from email and the web. In: CEAS (2004)
6. Culotta, A., Wick, M., Hall, R., McCallum, A.: First-order probabilistic models for coreference resolution. In: Procs. of HLT/NAACL, pp. 81–88 (2007)
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Procs. of the 43rd Annual Meeting on ACL, pp. 363–370 (2005)
8. Gollapalli, S.D., Giles, C.L., Mitra, P., Caragea, C.: On identifying academic homepages for digital libraries. In: Procs. of the 11th JCDL, pp. 123–132 (2011)
9. Kato, Y., Kawahara, D., Inui, K., Kurohashi, S., Shibata, T.: Extracting the author of web pages. In: Procs. of the 2nd ACM WICOW, pp. 35–42 (2008)
10. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Procs. of the 18th ICML, pp. 282–289 (2001)
11. Minkov, E., Wang, R.C., Cohen, W.W.: Extracting personal names from email: applying named entity recognition to informal text. In: Procs. of the Conf. on HLT and EMNLP, HLT 2005, pp. 443–450 (2005)
12. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Procs. of the 40th Annual Meeting on ACL, ACL 2002, pp. 104–111 (2002)
13. Shi, Y., Wang, M.: A dual-layer crfs based joint decoding method for cascaded segmentation and labeling tasks. In: Procs. of the 20th IJCAI, pp. 1707–1712 (2007)
14. Takeuchi, K., Collier, N.: Use of support vector machines in extended named entity recognition. In: Procs. of the 6th Conference on Natural Language Learning, COLING 2002, vol. 20, pp. 1–7 (2002)
15. Tang, J., Zhang, D., Yao, L.: Social network extraction of academic researchers. In: Procs. of the 7th ICDM, pp. 292–301 (2007)
16. Zheng, S., Zhou, D., Li, J., Giles, C.L.: Extracting author meta-data from web using visual features. In: Procs. of the 7th ICDMW, pp. 33–40 (2007)
17. Zhu, J., Nie, Z., Wen, J.R., Zhang, B., Ma, W.Y.: 2d conditional random fields for web information extraction. In: Procs. of the 22nd ICML, pp. 1044–1051 (2005)