Entity name extraction from faculty directories

João Mateus de Freitas Veneroso
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
jmfveneroso@gmail.com

Berthier Ribeiro-Neto
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
berthier.dcc.ufmg@gmail.com

ABSTRACT

Reliable researcher affiliation data is necessary to allow enquiring about international research group productivity and publication patterns. Public bibliographic databases such as DBLP and Google Scholar hold invaluable data about the academic environment. However, the researcher affiliation information is frequently missing or outdated. We propose a statistical data extraction method to acquire affiliation information directly from university websites and solve the name extraction task in general. Previous approaches to web data extraction either lack in flexibility, because wrappers do not generalize well on cross website tasks, or they lack in precision, because domain agnostic methods neglect useful properties of this particular application domain. Our statistical approach solves the name extraction task with a framework that incorporates both textual and structural features to yield an outstanding tradeoff between generality and precision. We conducted experiments over a collection of 152 faculty web pages in multiple languages from universities in 49 countries and obtained 94.40% precision, 97.61% recall and 0.9597 F-measure at the extraction task.

1 INTRODUCTION

Web data extraction is the task of automatically extracting structured information from unstructured or semi-structured web documents. Typically, Information Extraction tasks consist of mapping unstructured or poorly structured data to a semantically well defined structure. The input is most commonly composed of a set of documents that describe a group of entities in a similar manner. The Information Extraction task consists of identifying these entities and organizing them according to a template.

HTML documents most often lie in between the structured / unstructured data paradigm, which means that authors take a rather relaxed approach in regard to formal structure. Hierarchy, element disposition, class names, and other features related to the document structure and indirectly associated with the data itself are valuable information in the task of identifying entities and determining relationships, yet we cannot expect these features to be completely constrained by any underlying pattern. Like natural language, organization patterns tend to follow some guidelines but are in no way subject to strict rules.

We are interested in the task of name extraction, particularly extracting researcher names from university websites to complement data from public databases such as the DBLP repository (http://dblp.uni-trier.de/) [1]. The DBLP database has sparse information about author affiliation and only contains information about computer science researchers. We propose a probabilistic method

[1] This is useful, for instance, if one needs to compare the research output of deparments in Germany, or study international publication patterns.

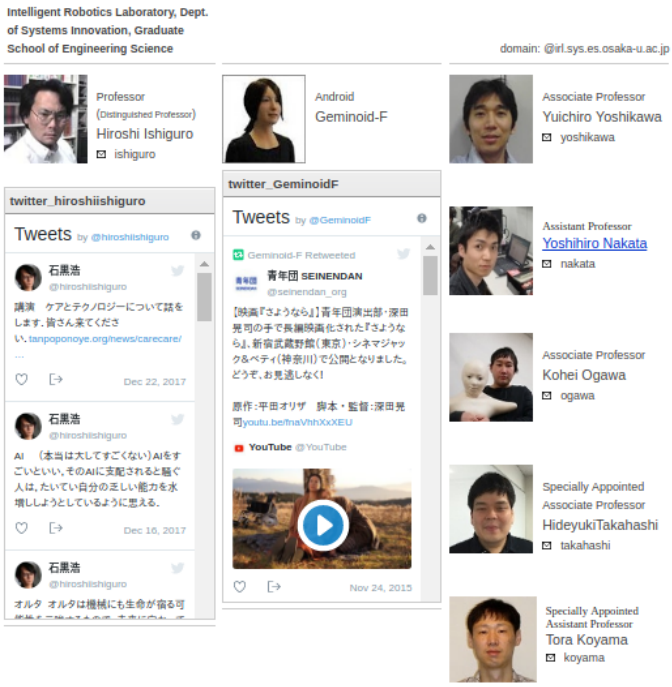

Figure 1: Example of a faculty directory

to handle the name extraction problem in general, without having to rely on metadata from PDF papers.

To acknowledge the complexity of this extraction task, take for example a snippet of the staff page for the intelligent robotics laboratory from Osaka University shown in figure 1. There is some structure to the way member profiles are arranged, but the organization is rather flexible even considering this single website. Other websites can show very different patterns, ranging from tables and lists to free form.

Researcher names can appear inside plain text, similar to typical named entity recognition scenarios or in a tabular structure. Names may be part of larger sentences such as in "Michael Johnson Chair" and "John Doe Avenue" yielding false positives. Names can be composed of common words (e.g. Summer Hall) yielding false negatives. There is no rule that fits all cases.

State-of-the-art named entity recognition approaches such as Conditional Random Fields do not perform so well in information extraction scenarios such as the one presented in figure 1, because the text is insufficient to provide proper contextual information about the semantic category of a word.

A key problem in the task of entity name extraction is accounting for all possible name combinations. Many names share their

spelling with common words, or are missing from our database because they are uncommon or unique. Since we lack a database with all possible name combinations in every language, we propose a holistic statistical method that accounts for discrepancies in data organization by assigning probabilities to sequences of tokens without relying on per-website training. Our method of label assignement resembles a Naive Bayesian classifier without the assumption of token independence. We also rely as little as possible on contextual information to avoid the mistakes of other classifiers. The base model already achieves recall and precision rates above 90%, but the performance can be increased even further by incorporating HTML structural features to estimate better probabilities, especially to avoid filtering out false negatives.

2 RELATED WORK

In the last 20 years, the astonishing growth of public information in the web has led to the development of a number of different approaches to the problem of web data extraction. Traditionally, the task was solved by designing special purpose programs called wrappers to recognize relevant data and store records in a structured format. These early tools varied wildly relative to their degree of automation.

It was readily perceived that manual wrapper generation was a rather tedious and error prone process, unsuited for large scale operations. Wrappers tend to break frequently because they rely heavily on web page features that can change often. So, in the late nineties, several authors advocated for wrapper induction, a technique that consists of automatically constructing wrappers from a small set of examples by identifying delimiters or context tokens that single out the desired attributes. Some remarkable wrapper induction methods are WIEN [1], Soft Mealy [2] and STALKER [3].

Despite being better than constructing wrappers manually, wrapper induction methods still suffered from a lack of expressive power and flexibility. These methods had trouble handling records with missing attributes or unusual structures because patterns could only be identified if they happened at least once in the examples.

Other approaches such as NoDoSE ([4]) and Debye ([5]) brought greater flexibility to wrapper induction methods by requiring a greater level of human interaction through graphical user interfaces. Web data extraction techniques often require some sort of assistance from human experts to boost accuracy. One of the main challenges in the field lies in determining an adequate tradeoff between the degree of automation and the precision and recall of the data extraction tool.

To automate the task of web data extraction completely some approaches, such as Road Runner [6], removed entirely the need for data examples. Road Runner parses documents belonging to a same class (e.g. books on Amazon) and generates wrappers based on their similarities and differences, yielding comparable results to those obtained by wrapper induction methods. However like previous approaches, it was unsuited for cross site extraction tasks because the learned rules were not general enough.

NLP based approaches aimed at extracting more general rules that could possibly be employed over multiple websites. RAPIER [7] is a method of rule extraction that uses information such as part-of-speech tags and semantic classes from a lexicon to derive patterns from a set of training examples. This approach is more flexible than the wrapper induction methods, however it achieves much lower rates of recall and precision.

In 2002, a survey by Laender et al. [8] made a thorough classification of the early approaches with a taxonomy based on their main technology, being them: languages for wrapper development, HTML-aware tools, NLP-based tools, Wrapper Induction Tools, Modeling-based tools and Ontology-based tools. Some noteworthy examples from this era are:

- TSIMMIS [9] and WebOQL [10], which are special purpose languages for building wrappers.
- Road Runner [6], XWRAP [11] and W4F [12], which are HTML-aware tools that infer meaningful patterns from the HTML structure.
- RAPIER [7], SRV [13], WHISK [14], which are NLP-based tools.
- WIEN [1], Soft Mealy [2] and STALKER [3] which are wrapper induction methods.
- NoDoSE [4] and Debye [5], which are semi supervised modeling based tools that require some interaction with the user by means of a graphical user interface.

In 2006, Chang et al. [15] complemented the previous surveys with semisupervised technologies such as Thresher [16], IEPAD [17] and OLERA [18]. They differed from supervised and unsupervised methods because they either needed only a rough description of data from users for extraction rule generation or some level of post processing that needed user attention. The survey also mentioned newer unsupervised methods such as DeLa [19], Exalg [20] and Depta [21].

Most of the early information extraction systems were rule-based with either manual rule description or automatic rule learning from examples, thus they suffered from a lack of flexibility when dealing with noisy and unstructured data. Huge progress in the field of statistical learning led to the development of statistical models that tried to solve this problem.

In 2008, Sarawagi [22] produced a survey that classified wrappers in rule-based methods, statistical methods and hybrid models, bringing together the fields of named entity recognition, relationship extraction and information extraction. The rule based methods encompass most of the previous models. The statistical methods convert the extraction task into a token labeling task, identifying the target entities through the assignment of labels. Any classifiers such as a Support Vector Machines, Logistic Classifiers or Neural Networks could be employed to perform this task. However Hidden Markov Models, Maximum Entropy Taggers and Conditional Random Fields tend to perform better at most extraction tasks because of the way they model token and label dependencies. Hybrid models incorporate both rule-based and statistical methods.

Statistical models have proven to be reliable tools for performing numerous NLP tasks. However, information from web documents relevant to data extraction tasks is usually arranged in a tabular form rather than in a plain text format. Therefore, typical state of the art classifiers can yield poor results if they rely exclusively on textual information. In the web information extraction task, the document structure must be incorporated as a feature in an effective classifier.

More recently, surveys by Ferrara et al. [23] and Schulz et al. [24] updated the previous surveys with some interesting innovations. Some examples are: the Visual Box Model [25], a data extraction system that produces a visualization of the web page to exploit visual cues to identify data presented in a tabular form; automatic wrapper adaptation [26], a technique that tries to reduce the cost of wrapper maintenance by measuring the similarity of HTML trees and adapting wrappers to the new page structure; AutoRM [27], a method to mine records from a single web page by identifying similar data regions through DOM tree analysis; and Knowledge Vault [28], a method that combines different extraction approaches to feed a probabilistic knowledge base.

In 2016, Varlamov et al. [29] argued that the degree of automation can no longer be the main classification criterion for the data extraction systems because unsupervised methods which were widely considered to be the state of the art when dealing with individual websites performed poorly or were innapropriate on cross site extraction tasks. The authors proposed a classification of methods by the extent of their application. The competing approaches were separated into two groups: methods for individual websites and methods that are applicable to whole application domains.

The first group contains most of the earlier approaches, including the supervised approaches: SRV [13], RAPIER [7], WHISK [14], WIEN [1] SoftMealy [2] and STALKER [3]; and the unsupervised approaches: RoadRunner [6] and EXALG [20].

The second group is divided between domain specific methods and domain agnostic methods. Domain specific methods are designed for extracting data about a particular application domain across multiple websites. Domain specific methods integrate information about the particular application domain in the course of its development and thus are able to achieve superior performance in comparison to domain agnostic methods. One example is the method of comment extraction from blog posts described by Kao et al. [30]. By incorporating multiple techniques and domain specific features they are able to build a classifier that differentiates comment-blocks and non-comment blocks. This is only one of many methods tuned for various domains. Our name extractor method also belongs to this category.

Domain agnostic methods are the most general extraction methods. They can extract information from any application domain from multiple websites. They pose the hardest challenge because the tool must infer data relevance without any prior training in that particular application domain. Some examples are: ODE ([31]), ObjectRunner ([32]), and AMBER ([33]). These approaches are more flexible but yield worse results than domain specific methods.

3 IMPLEMENTATION OVERVIEW

The name extraction problem is no different from a Named Entity Recognition problem, however approaches that typically achieve high accuracy on NER tasks like Conditional Random Fields, Hidden Markov Models or Maximum Entropy Models do not necessarily perform so well when handling tabular data, as is most common in web data extraction tasks. For example, in narrative text, a classifier could learn patterns such as "X spoke to Y" and then discover from the sentence "Alice spoke to Bob" that Alice and Bob are names. However in the name extraction task, the lack of context words prevents these methods from detecting useful patterns over sequences of tokens. For instance, tabular data such as the one found in most faculty directories provides minimal textual context.

We are also interested in developing a method to extract researcher names regardless of the website's main language, even though at this moment we will not be considering non extended-ASCII encodings such as Chinese and Farsi characters. Training classifiers over multiple corpora in different languages could be in itself a very challenging task. Instead we rely on a priori probabilities and structural HTML features to attribute label probabilities to tokens. When manually scanning through faculty pages in search of researcher names, the intuitive method to extract them is usually identifying a few cases for which we are certain (e.g. Dr. John Smith), then identifying structural patterns to classify other cases for which we may be not so sure, either because they lie outside our knowledge base (e.g. T.J. Bok) or because they sound like common words (e.g. Summer Hall).

3.1 Pre-processing

Before applying the classifier over the web page content, we must run the input web page through a series of pre-processing steps to clean the data as much as possible. The HTML document is first parsed producing a DOM tree. At this stage, malformed HTML is converted into a valid DOM tree as it is commonly done by most browsers nowadays. Header details, script tags, style tags, hidden tags and comments are removed. Emails, URLs, and honorifics are removed through the use of regular expressions. Spaces, tabs, newlines and special characters are used as separators in a tokenization stage, producing a list of tokens. Each token is saved as an object with the following attributes:

Value: the token's text in lowercase with converted accented characters.
DOM Element: a pointer to the DOM Element that contains this token in the DOM Tree.
Previous Token: the previous token object.
Next Token: the next token object.

With this simple structure we can extract various features associated with entity names, with the benefit of having a list of tokens such as the one we would obtain from a plain text format. Through the DOM Element pointer we can figure out the parent elements, nesting depth, siblings and other information that can be used to improve estimates in the probabilistic method we adopt.

3.2 Name Extraction

Let $t = (t_1, t_2, ..., t_n)$ be a sequence of token objects obtained on the pre-processing stage, and $y = (y_1, y_2, ..., y_n)$ be a sequence of labels attributed to these tokens where y_i can be either a "Name Label" (N), meaning that token t_i is a person's name, or a "Word Label" (W), meaning that token t_i is a common word (not a person's name). Then, the problem of extracting names from a sequence of tokens is just a series of binary classification problems. Considering that each token t_i has a probability $P(t_i = y_i)$ of having label y_i, the problem of finding an optimal sequence of labels y^* for a sequence of tokens t can be written as:

$$y^* = \underset{y}{\operatorname{argmax}}\, P(t_1 = y_1, t_2 = y_2, ..., t_n = y_n) \quad (1)$$

We may employ the chain rule to explore the relationship between joint and conditional probabilities. For ease of exposure, consider that $P(Y_i) \equiv P(t_i = y_i)$ yielding:

$$P(Y_1, Y_2, ..., Y_n) = P(Y_1)P(Y_2|Y_1)...P(Y_n|Y_{n-1}, Y_{n-2}, ...) \quad (2)$$

A k-gram model could approximate the probabilities $P(Y_1, Y_2, ..., Y_n)$ by looking just at the first k tokens and sliding a window of fixed size k over the token sequence. However, the conditional probabilities $P(Y_i|Y_{i-1}, ...)$ are hard to estimate, because the joint distribution depends both on the previous labels and the previous tokens. If we express them in terms of joint probabilities the problem becomes more evident:

$$P(Y_i|Y_{i-1}, Y_{i-2}, ...) \equiv P(t_i, y_i|t_{i-1}, y_{i-1}, t_{i-2}, y_{i-2}, ...) \quad (3)$$

To simplify our modeling, let us assume that the probability that token t_i has label y_i depends on the values of previous labels but is independent of the previous tokens. That is, we assume that, given a sequence of tokens {"$John$", "$Smith$"}, the conditional probability $P(\text{"}Smith\text{"}|\text{"}John\text{"} = Name)$ is equivalent to $P(\text{"}Smith\text{"}|Any\ name)$. In other words, this means that the probability of Smith being a last name is the same regardless of a person's first name, as long as we can make sure that the previous token is a name. By taking this assumption, Equation 3 becomes:

$$P(Y_i|Y_{i-1}, Y_{i-2}, ...) = P(t_i, y_i|y_{i-1}, y_{i-2}, ...) \quad (4)$$

Once again we can employ the chain rule to obtain:

$$P(Y_i|Y_{i-1}, Y_{i-2}, ...) = P(t_i|y_i, y_{i-1}, ...)P(y_i|y_{i-1}, y_{i-2}, ...) \quad (5)$$

In Equation 5, the probability $P(t_i|y_i, y_{i-1}, ...)$ depends on the current and previous labels. We make a simplifying assumption that $P(t_i|y_i, y_{i-1}, ...)$ can be approximated by $P(t_i|y_i)$. The reasoning behind this is that previous labels have neglible influence over the particular value of token t_i. For example, we assume that $P(t_i = \text{"}John\text{"}|t_i = Name)$ is a good aproximation for $P(t_i = \text{"}John\text{"}|y_i = Name, y_{i-1} = Word, ...)$. With this assumption, Equation 5 becomes:

$$P(Y_i|Y_{i-1}, Y_{i-2}, ...) = P(t_i|y_i)P(y_i|y_{i-1}, y_{i-2}, ...) \quad (6)$$

Finally, by replacing Equation 6 into Equation 2 and grouping together the second terms of Equation 6 to form the joint probability $P(y_1, y_2, \ldots, y_n)$, we obtain:

$$P(Y_1, \ldots, Y_n) = P(y_1, \ldots, y_n)P(t_1|y_1)P(t_2|y_2)\ldots P(t_n|y_n) \quad (7)$$

Equation 7 can be split into two parts: the prior $P(y_1, y_2, \ldots, y_n)$ and the conditional token probabilities $P(t_1|y_1)P(t_2|y_2)\ldots P(t_n|y_n)$.

```
<div>
  <b>John Smith</b>
  <span>Professor Emeritus</span>
</div>
```

Figure 2: Tag transition example

3.2.1 Prior probabilities. The prior probabilities are given by the expression $P(y_1, y_2, \ldots, y_n)$, which represents the probability of a series of tokens assuming labels $\{y_1, y_2, ..., y_n\}$ without any prior knowledge about the actual token values.

By assuming a window of size $k \leq n$ we may approximate the priors by calculating the joint probability $P(y_1, y_2, \ldots, y_k)$. We observed empirically that when a name token is found, it is very likely that the next token will also be a name. However, arbitrary sequences of non-name tokens do not alter significantly the probabilities for the next token's label. Put in other words, $P(y_i|y_{i-1} \neq Name) \approx P(y_i)$. We can use this property to slide the window of k tokens over the entire sequence of tokens and estimate probabilities without deviating too much from the real distribution. We start at the beginning of the token list and calculate probabilities for each possible sequence of labels, taking the most likely one. If the first token is not a name we slide the window right by one token and repeat the process. If the first token is a name, then we extract all the tokens that are part of that name starting from the first one and slide the window right by the number of name tokens extracted. It is important that the window is at least the number of tokens in the longest name (5 seems to be enough on most cases), so we can extract it as a single name.

Let N be a name label and W be a word label, then for a window of size k we would need to estimate prior probabilities for all 2^k possible sequences of labels. For example, taking a window of size 4, the sequences would be: $\{W, W, W, W\}$, $\{W, W, W, N\}$, $\{W, W, N, W\}$, $\{W, W, N, N\}$, ..., $\{N, N, N, N\}$.

When names occur next to each other there is no way to tell where the first name ends and the second one starts. To delimit name boundaries we need to estimate priors for different sequences of name labels in addition to our previous priors. Let the first and second name labels be N_1 and N_2, respectively. Then, taking a window of size 4, the prior $\{N, N, N, N\}$, for example, would expand into four different cases: $\{N_1, N_1, N_1, N_1\}$, $\{N_1, N_1, N_1, N_2\}$, $\{N_1, N_1, N_2, N_2\}$, $\{N_1, N_2, N_2, N_2\}$.

Most of the times when names happen inside a list they tend to be contained inside a single HTML element. Even though this is not always the case, this knowledge can be incorporated as an additional piece of evidence in our model. This evidence becomes especially useful when we are trying to delimit name boundaries.

A tag transition occurs when a token occurs inside an HTML element and the next token occurs inside a different HTML element. Take for example the HTML snippet in figure 2 with the sequence of tokens { John, Smith, Professor, Emeritus }. There is a tag transition between the tokens "Smith" and "Professor" because these consecutive tokens are inside different HTML tags. Let $*$ indicate a tag transition in a sequence of tokens, then the label sequence $\{y_1, y_2*, y_3, y_4\}$, means that the first two tokens are contained inside a single HTML element, while the remaining tokens are inside a different HTML element. For a window of size 4, we would

need to estimate prior probabilities with 4 possible tag transitions: $\{y_1, y_2, y_3, y_4\}$, $\{y_1*, y_2, y_3, y_4\}$, $\{y_1, y_2*, y_3, y_4\}$, $\{y_1, y_2, y_3*, y_4\}$. We could estimate sequences with multiple tag transitions inside a window of tokens, however estimates with a single transition have shown good experimental results.

3.2.2 Conditional token probabilities. The conditional probabilities are given by the second part of equation 7, that is $P(t_1|y_1)P(t_2|y_2)\ldots P(t_n|y_n)$. We need to estimate conditional token probabilities for both labels: names and words. So we need to know $P(t_i|N)$, the probability that a name is t_i and $P(t_i|W)$, the probability that a common word is t_i.

For our experiments, the conditional token probabilities were obtained by maximum likelihood estimation with Laplace smoothing to account for tokens that didn't occur in the corpus. The $P(t_i|N)$ probabilities were estimated over a collection of approximately 1.5 million author names from the DBLP database. The $P(t_i|W)$ probabilities were estimated over a corpus of approximately a hundred thousand documents obtained by a web crawler that collected university pages. In the latter case all names were removed from the corpus in order to keep only non-name tokens.

Conditional token probabilities can be made more precise by incorporating features in equation 7. We do that by changing the token conditional probabilities to:

$$P(t_i, f_1, f_2, \ldots, f_n|y_i) = P(t_i|y_i)P(f_1|y_i)\ldots P(f_n|y_i) \qquad (8)$$

where f_i are features, which are assumed to be independent. Features can be textual or structural. Textual features take textual clues from context words and the current token value. An example of textual feature is the token's length, because names tend to be shorter than common words. Structural features infer token probabilities based on the surrounding HTML structure. An example of structural feature is the HTML tag, because names tend to occur inside the same HTML tags in a given document.

3.2.3 Secondary estimates. Structural features that were estimated over the entire corpus ended up being too general. HTML structure varies wildly between different websites, so we cannot extract useful probabilities looking at the entire collection. For example, if all names appear inside a <tr> tag in a given document it does not mean that names tend to appear inside <tr> tags rather than any other HTML tag in other documents. However for that particular document we may be able to identify other names and exclude non name tokens by knowing that tokens that occur inside <tr> tags have a higher probability of being names.

Given that the basic algorithm (without structural features) was able to extract a satisfactory number of names from a web page with sufficient precision on a first run, we may estimate probabilities for a structural feature $P(f_j|y_i)$ by looking at the extracted names. In fact, since we already attributed a label to every token in the web page, we can access the originating DOM element through the token object and estimate feature probabilities by maximum likelihood, as we did for the token conditional probabilities. For example, we could easily estimate the probability that a name will occur inside a tag $P(div|Name)$ with this method and incorporate this information as a feature in Equation 8 on a next run. Since estimates should get better after a second run, this method can be repeated multiple times to further increase the algorithm's performance. This method boosted the algorithm's performance considerably in our experiments, even with a single secondary run.

Id	Feature	Description
1	Token incidence	How often a token happens in a document
2	Token length	The token's character length
3	First+Second parents	First and second HTML parents combined
4	Third parent	Third HTML parent
5	CSS class name	Innermost valid CSS class
6	Child number	The child number relative to the HTML parent
7	Nesting depth	The number of HTML parents up to root

Table 1: Features

3.2.4 Features. Table 1 describes the list of features tested in our experiments. The third feature "First + Second parents" is the combined tag names of the first and second HTML elements from leaf to root in the DOM tree, starting from the text node. This combination yields better results than separate features. The fourth feature is the same thing for the third parent. The meaning of the remaining features is self evident.

Only features that showed some evidence of improvement are being shown here. The missing features were usually thought to handle specific problems such as street names that could be confused with a person's name. They involved looking at context words, word capitalization, honorifics, etc. However, these trials were not very productive.

By design, the features were chosen with improvements to recall rather than precision in mind. The reason is that when a data extractor misses useful results, it becomes hard to obtain that data. However cleaning bad results is a much simpler task to be carried manually.

3.3 Experiments

Features	Precision	Recall	F-measure
None	91.71%	91.39%	0.9155
1	91.28% (0.3880)	93.79% (0.1836)	0.9252 (0.3026)
2	92.00% (0.4189)	92.03% (0.4086)	0.9201 (0.4042)
1+2	91.83% (0.4660)	93.90% (0.1586)	0.9285 (0.2336)

Table 2: Textual features experiment

The test collection was a set of 152 manually labeled faculty directory web pages from 49 different countries with 11782 researcher names. These names are the target of our extraction procedure. We calculated the precision, recall and f-measures for each modeling variant. All measures were tested for statistical significance with a one tailed paired T-test in comparison with the base model, the t-values are presented inside parenthesis next to each measure. Results were considered to be significant when the t-value was smaller than 0.05. Statistically significant results are boldened on all tables. All measures were obtained by the averaged results of a 5 fold cross validation run.

The base model uses only priors with the details already described in the last section and token conditional probabilities, but no textual or structural features. Extracted names were only considered to be correct when they resulted on an exact match with the test data.

Table 2 describes the results for the base model incorporating different textual features. We considered the variants: (1) base model with no features (our baseline), (2) base model + feature 1 (token incidence), (3) base model + feature 2 (token length), (4) base model + features 1 and 2. Our base model already achieves good results (91.71% precision, 91.39% recall, and 0.9155 F-measure), because it relies on simple but universal probabilities without making assumptions about data organization or contextual information. Notice that the textual features produced no statistically significant improvements in comparison to the base model with no features (the first row).

Features	Precision	Recall	F-measure
3	93.43% (0.1280)	**96.14% (0.0361)**	**0.9477 (0.0475)**
4	93.04% (0.1965)	95.67% (0.0508)	0.9433 (0.0740)
5	93.04% (0.1781)	95.50% (0.0653)	0.9425 (0.0819)
6	92.70% (0.2307)	**95.60% (0.0451)**	0.9413 (0.0725)
7	92.83% (0.2374)	**95.79% (0.0467)**	0.9429 (0.0789)
3 + 5 + 7	94.23% (0.0664)	**97.14% (0.0163)**	**0.9566 (0.0212)**

Table 3: Structural features experiment

Table 3 describes the results for the base model with secondary estimates calculated over double runs and different combinations of structural features. We considered the variants: (1) base model + feature 3 (first and second parents), (2) base model + feature 4 (third parent), (3) base model + feature 5 (CSS class name), (4) base model + features 6 (child number), (5) base model + features 7 (nesting depth), (6) base model + features 3, 5 and 7. We observe in this experiment that features 3, 6 and 7 produced statistically significant improvements to recall. Though, features 4 and 5 also produced small T-values of 0.0508 and 0.0653 respectively. The secondary estimates strategy was very successful in improving all metrics by incorporating structural features on the base model. The best variant (features 3, 5 and 7), achieved a precision of 94.23%, recall of 97.14% and F-measure 0.9566 with statistical significant improvement to both recall and F-measure.

# runs	Precision	Recall	F-measure
3	94.40% (0.0525)	**97.48% (0.0117)**	**0.9592 (0.0151)**
4	94.37% (0.0544)	**97.58% (0.0104)**	**0.9595 (0.0141)**
5	94.40% (0.0523)	**97.61% (0.0100)**	**0.9597 (0.0135)**
6	94.37% (0.0544)	**97.61% (0.0100)**	**0.9596 (0.0138)**

Table 4: Multiple runs experiment

Table 4 describes the results for the best variant from the previous experiments (features 3, 5 and 7) running secondary estimates 3, 4, 5 and 6 times. We observe that multiple runs on the secondary estimates strategy do not increase performance significantly. In fact, it reaches a plateau after only 3 to 4 runs. Notice that all variants achieved statistical significant improvements to recall and F-measure and also got very close to reaching statistical significant improvements to precision. Further runs were omitted because there was no change after 6 runs.

4 CONCLUSION

Our name extraction method achieved 94.40% precision, 97.61% recall and 0.9597 F-measure on a corpus of 152 faculty directory web pages from 49 different countries with 11782 researcher names. The model is tuned for the particular problem of name extraction, but we believe this result can be generalized to solve other data extraction problems. The algorithm is general enough to handle other types of information extraction tasks without the need for too much tinkering. The secondary estimates strategy was very successful in further improving the base model. It can also be remodeled to fit other statistical classifiers since it does not rely on any particular implementation details of our strategy.

REFERENCES

[1] Nicholas Kushmerick. Wrapper induction: efficiency and expressiveness. *Artificial Intelligence*, 118(1-2):15–68, 2000.

[2] Chun Nan Hsu and Ming Tzung Dung. Generating finite-state transducers for semi-structured data extraction from the Web. *Information Systems*, 23(8):521–538, 1998.

[3] Ion Muslea, Steve Minton, and Craig Knoblock. A Hierarchical Approach to Wrapper Induction. *Proc. of the Third Annual Conf. on Autonomous Agents, ACM*, pages 190–197, 1999.

[4] Brad Adelberg. NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents. *ACM SIGMOD Record*, 27(2):283–294, 1998.

[5] Alberto H F Laender, Berthier Ribeiro-Neto, and Altigran S da Silva. DEByE - Date extraction by example. *Data Knowl. Eng.*, 40(2):121–154, 2002.

[6] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Roadrunner: Towards automatic data extraction from large web sites. *Proceedings of the 27th International Conference on Very Large Data Bases*, pages 109–118, 2001.

[7] Mary Elaine Califf and Raymond J Mooney. Relational learning of pattern-match rules for information extraction. *Computational Linguistics*, 4:9–15, 1999.

[8] A.H.F. Laender, B. A. Ribeiro-Neto, and Juliana S.Teixeria. A brief survey of web data extraction tools. *ACM SIGMOD Record 31(2)*, pages 84–93, 2002.

[9] Joachim Hammer, Jason Mchugh, and Hector Garcia-molina. Semistructured Data : The TSIMMIS Experience. *Proceedings of the First East-European Symposium on Advances in Databases and Information Systems*, pages 1–8, 1997.

[10] Gustavo O. Arocena and Alberto O. Mendelzon. WebOQL: Restructuring documents, databases, and webs. *Theory and Practice of Object Systems*, 5(3):127–141, 1999.

[11] L. Liu, C. Pu, and W. Han. XWRAP: an XML-enabled wrapper construction system for Web information sources. *Proceedings of 16th International Conference on Data Engineering*, pages 611–621, 2000.

[12] Arnaud Sahuguet and Fabien Azavant. Building light-weight wrappers for legacy Web data-sources using W4F. *Proceedings of the 25th VLDB Conference*, 99:738–741, 1999.

[13] Dayne Freitag. Information Extraction from HTML: Application of a General Machine Learning Approach. *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pages 517–523, 1998.

[14] Stephen Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1):233–272, 1999.

[15] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006.

[16] Andrew Hogue and David Karger. Thresher : Automating the Unwrapping of Semantic Content from the World Wide Web. *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 86–95, 2005.

[17] C.H. Chang, C.H. Chang, S.C. Lui, and S.C. Lui. IEPAD: information extraction based on pattern discovery. *Proceedings of the 10th international conference on World Wide Web*, pages 681–688, 2001.

[18] Chia Hui Chang and Shih Chien Kuo. OLERA: Semisupervised Web-data extraction with visual support. *IEEE Intelligent Systems*, 19(6):56–64, 2004.

[19] Jiying Wang and Fred H. Lochovsky. Data extraction and label assignment for web databases. *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, page 187, 2003.

[20] A. Arasu, H. Garcia-Molina, Arvind Arasu, and Hector Garcia-Molina. Extracting structured data from Web pages. *2003 ACM SIGMOD International Conference on Management of Data*, pages 337 – 348, 2003.

[21] Yanhong Zhai and Bing Liu. Web data extraction based on partial tree alignment. *Proceedings of the 14th international conference on World Wide Web - WWW '05*, page 76, 2005.

[22] Sunita Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.

[23] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70:301–323, 2014.

[24] Andreas Schulz, Jörg Lässig, and Martin Gaedke. Practical web data extraction: Are we there yet? — A short survey. *IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016*, pages 562—-567, 2016.

[25] Bernhard Krüpl, Marcus Herzog, and Wolfgang Gatterbauer. Using visual cues for extraction of tabular data from arbitrary HTML documents. *Special interest tracks and posters of the 14th international conference on World Wide Web - WWW '05*, pages 1000—-1001, 2005.

[26] Emilio Ferrara and Robert Baumgartner. Automatic wrapper adaptation by tree edit distance matching. *Smart Innovation, Systems and Technologies*, 8:41–54, 2011.

[27] Shengsheng Shi, Chengfei Liu, Yi Shen, Chunfeng Yuan, and Yihua Huang. AutoRM: An effective approach for automatic Web data record mining. *Knowledge-Based Systems*, 89:314–331, 2015.

[28] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 601–610, 2014.

[29] M. I. Varlamov and D. Yu. Turdakov. A survey of methods for the extraction of information from Web resources. *Programming and Computer Software*, 42(5):279–291, 2016.

[30] Huan-An Kao and Hsin-Hsi Chen. Comment Extraction from Blog Posts and Its Applications to Opinion Mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1113–1120, 2010.

[31] Weifeng Su and Frederick H Lochvsky. ODE : Ontology-assisted Data Extraction. 1(212), 2009.

[32] Talel Abdessalem, B Cautis, and Nora Derouiche. ObjectRunner: lightweight, targeted extraction and querying of structured web data. *Proceedings of the VLDB . . .*, 3(2):1585–1588, 2010.

[33] Tim Furche, Georg Gottlob, Giovanni Grasso, Giorgio Orsi, Christian Schallhart, and Cheng Wang. AMBER: Automatic Supervision for Multi-Attribute Extraction. *arXiv preprint*, 1210(5984):1–*22*, 2012.