Unsupervised Strategies for Information Extraction by Text Segmentation

Eli Cortez, Altigran S. da Silva
Universidade Federal do Amazonas
Departamento de Ciência da Computação
Manaus, AM, Brazil
{eccv,alti}@dcc.ufam.edu.br

ABSTRACT

Information extraction by text segmentation (IETS) applies to cases in which data values of interest are organized in implicit semi-structured records available in textual sources (e.g. postal addresses, bibliographic information, ads). It is an important practical problem that has been frequently addressed in the recent literature. We report here partial results from a PhD thesis work in which we introduce ONDUX (On Demand Unsupervised Information Extraction), a new unsupervised probabilistic approach for IETS. As other unsupervised IETS approaches, ONDUX relies on information available on pre-existing data to associate segments in the input string with attributes of a given domain. Unlike other approaches, we rely on very effective matching strategies instead of explicit learning strategies. The effectiveness of this matching strategy is also exploited to disambiguate the extraction of certain attributes through a reinforcement step that explores sequencing and positioning of attribute values directly learned *on-demand* from test data, with no previous human-driven training, a feature unique to ONDUX. This assigns to ONDUX a high degree of flexibility and results in superior effectiveness, as demonstrated by experimental evaluation we have carried out with textual sources from different domains, in which ONDUX is compared with a state-of-art IETS approach.

Categories and Subject Descriptors

H.2 [**Database Management**]: Miscellaneous ; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Performance, Experimentation

Keywords

Data Management, Information Extraction, Text Segmentation

1. INTRODUCTION

The abundance of on-line sources of text documents containing implicit semi-structured data records in the form of continuous text, such as product descriptions, bibliographic citations, postal addresses, classified ads, etc., has attracted a number of research efforts towards automatically extracting their data values by segmenting the text containing them [1, 3, 11, 16]. This interest is motivated by the necessity of having these data stored in some structured format as relational databases or XML, so that it can be further queried, processed and analyzed.

For instance, an article from "The Washington Post" reports that the revenues by Newspapers from classified ads, which was $17 billion in 2006, has been declining since 2000, while the revenues from *on-line* classified ads grew 6 times in the same period, reaching $3.1 billion. Empowering users with services such as sophisticated searching, dissemination, comparison, personalization on top of this content, can have a significant impact on this business. Extracting and structuring these data is a crucial step towards this goal.

As an example of the information extraction task performed by a typical text segmentation system, consider the input ad *"Regent Square $228,900 1028 Mifflin Ave.; 6 Bedrooms; 2 Bathrooms. 412-638-7273"*. A suitable text segmentation over this string would generate a structured record such as:

⟨neighborhood,"*Regent Square*"⟩,
⟨price,"*$228,900*"⟩,
⟨number,"*1028*"⟩,
⟨street,"*Mifflin Ave.;*"⟩,
⟨bedrooms,"*6 Bedrooms;*"⟩,
⟨bathrooms,"*2 Bathrooms.*"⟩,
⟨phone,"*412-638-7273*"⟩

The dominant approach in information extraction by text segmentation (IETS) is the deployment of statistical methods such as as Hidden Markov Models (HMM) [3] or Conditional Random Fields models (CRF) [10] to automatically learn a statistical model for each application domain. These methods usually require training data consisting of a set of representative segmented and labeled input strings. Currently, methods based on CRF are state-of-art, outperforming HMM-based methods in experimental evaluations presented in the literature [15, 16].

Obtaining a large amount of training data may be very expensive or even unfeasible in some situations. Recognizing this problem, recent papers proposed the use of pre-existing datasets to alleviate the need for manually labeled training

Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research (IDAR 2010), June 11, 2010, Indianapolis, Indiana, USA.

string segments to associate them with their corresponding attributes [1, 11, 16]. In these methods, the learning process takes advantage of known values of a given attribute to train a model for recognizing values of this attribute occurring in an input textual record.

In our work, we look for alternative methods that demand less user labor without compromising the extraction effectiveness. In this context, we introduce ONDUX (**ON**-**D**emand **U**nsupervised Information **EX**traction), an alternative unsupervised probabilistic IETS approach. Similar to previous unsupervised approaches [1, 11, 16], ONDUX also relies on pre-existing data, more specifically, on sets of attribute values from pre-existing data sources, to associate segments in the input string with a given attribute. Different from previous work, there is not an explicit learning process in this step. Instead, we use simple generic matching functions to compute a score measuring the likelihood that a text segments occurs as a typical value of an attribute.

Although this simple greedy matching-based strategy is effective (as shown in our experimental results), it may fail for ambiguous attributes with similar domains. This is the case of attributes such as Title and Keywords, found on bibliographic information extracted from paper headings. To solve this, we rely on positioning and sequencing probabilities of the attribute values. While in traditional methods, such as HMM and CRF, these probabilities are assumed as fixed [1, 16] or are learned through a manual labeling process [3, 14, 11], our method can automatically adapt to variable attribute values positioning and sequencing in an unsupervised way. In other words, it does not rely on the explicit association between unsegmented input strings and the corresponding segmented strings (labeled data) that supervised systems require for training, i.e., the labels "come for free" with the attributes of our pre-existing data source. More importantly, as in some unsupervised learning and transductive methods [9], we take advantage of information about the own records we are trying to extract (the test set) by exploiting the high certainty of the matching step in order to incorporate, on demand, information about the positioning and sequencing of attribute values in these records within the extraction model we generate.

To corroborate our claims regarding the high-quality and flexibility of our approach, we present results of preliminary experiments with textual sources from different domains. In these experiments ONDUX is compared with CRF, the state-of-art method in probabilistic information extraction [10, 15], in its unsupervised version [16]. Results of these experiments reveal that ONDUX was able to correctly identify attribute values in all different datasets, outperforming CRF in most of the cases.

In sum, we regard ONDUX as a very effective unsupervised information extraction method that:(1) instead of requiring explicit learning of a model for identifying attributes values on the input texts, uses a simple but very effective greedy strategy based on matching; (2) exploits the high accuracy of this matching strategy to learn from the test data the probabilities of positioning and sequencing of attributes in an unsupervised manner, making no rigid assumptions about the order of the attribute values, thus being much more robust and flexible to changes in patterns; (3) despite the fact of operating on-demand, has processing time of test instances similar to that of methods that use explicit learning such as CRF.

A full paper on ONDUX, containing detailed description of the method and the full set of experiments carried out with it was accepted recently [7]. We are currently preparing public releases of the method as library and as a tool for other research to use.

This paper is organized as follows. Section 2 discusses the main challenges in IETS and previous approaches in the literature. Section 3 presents an overview and the general ideas of ONDUX. Section 4 presents experiments for verifying the effectiveness of our approach comparing it with a state-of-art IETS approach. Section 5 presents a comparison of ONDUX with previous related IETS approaches in the literature. Section 6 concludes the paper giving directions for future work.

2. CHALLENGES AND APPROACHES

Information extraction by text segmentation (IETS) is the problem of segmenting text inputs to extract implicit data values contained in them. Informally, each text input forms an implicit record [15]. A fairly common approach to solve this problem is the use of machine learning techniques, either supervised, i.e., with human-driven training [8, 3, 14], or unsupervised, i.e., with training provided by some form of pre-existing data source [1, 4, 11, 16].

One of the first approaches in the literature addressing this problem was proposed by Freitag and McCallum in [8]. It consisted in generating independent Hidden Markov Models (HMM) for recognizing values of each attribute. This approach was extended in the DATAMOLD tool [3], in which attribute-driven (or *internal*) HMMs are nested as states of an *external* HMM. This external HMM aims at modeling the sequencing of attribute values on the implicit records. Internal and external HMM are trained with user-labeled text segments. Experiments over two real-life datasets yielded very good results in terms of the accuracy of the extraction process.

Later on, *Conditional Random Fields (CRF)* models were proposed as an alternative to HMM for the IETS task [10]. In comparison with HMM, CRF models are suitable for modeling problems in which state transitions and emissions probabilities may vary across hidden states, depending on the input sequence. In [14], a method for extracting bibliographic data from research papers based on CRF is proposed and experimentally evaluated with good results. Currently, CRF constitutes the state-of-art in information extraction due to its flexibility and the quality of the extraction results achieved [14, 11].

Although effective, these supervised IETS approaches based on graphical models such as HMM and CRF usually require users to label a large amount of training input documents. There are cases in which training data is hard to obtain, particularly when a large number of training instances is necessary to cover several features of the test data.

To address this problem, recent approaches presented in the literature propose the use of pre-existing data for easing the training process [1, 11, 16]. According to this strategy, models for recognizing values of an attribute are generated from values of this attribute occurring in a database previously available. These approaches take advantage of large amounts of existing structured datasets with little or no user effort.

Following this strategy, recent methods in the literature use *reference tables* in combination with graphical models, that is, HMMs [1] or CRFs [11, 16]. For recognizing values of a given attribute among segments of the input string, a model is trained using values available on the reference table for this attribute. No manually labeled training input strings are required for this. Once attribute values are recognized, records can be extracted. The methods proposed in [1, 16] assume that attributes values in the input text follow a single global order. This order is learned from a sample batch of the test instances. On the other hand, the method proposed in [11] can deal with records bearing different attribute value orders. To accomplish this, the CRF model must be learned using additional manually labeled input strings.

A similar strategy is used in [4]. However, when extracting data from a source in a given domain, this approach may take advantage not only from pre-existing datasets, but also from other sources containing data on the same domain, which is extracted simultaneously from all sources using a 2-state HMM for each attribute. Record extraction is addressed in an unsupervised way by aligning records from the sources being extracted.

As these approaches alleviate or even eliminate the need for users to label segments in training input strings; we regard them as *unsupervised* IETS approaches. Despite this, experimental results reported for these methods reveal extraction quality levels similar to those obtained with traditional supervised IETS methods [8, 3, 14].

Our method ONDUX can also be regarded as unsupervised, since it relies on pre-existing data sources to recognize attribute values on input strings. In a first step, it deploys effective generic similarity functions to label text segments based on matching scores between these segments and known values of a given attribute. Next, assigned labels are revised based on a reinforcement step that takes into account sequencing and positioning of attribute values directly learned *on-demand* from test data, with no previous human-driven training. As demonstrated by experimental results, in which ONDUX is compared with a state-of-art IETS approach, these features yield highly accurate results which are in most cases superior to the state-of-the-art.

3. THE ONDUX METHOD

In this section, we present an overview of ONDUX, our unsupervised probabilistic approach for IETS. Given a text input T containing a set of implicit textual records, ONDUX identifies data values available in these records and associates these values with proper attributes.

Consider an input string I representing a real classified ad such as the one presented in Figure 1(a). Informally, the IETS problem consists in segmenting I in a way such that each segment s receives a label ℓ corresponding to an attribute a_ℓ, where s represents a value in the domain of a_ℓ. This is illustrated in Figure 1(d), which is an example of the outcome produced by ONDUX.

Similar to previous approaches [1, 11, 16], in ONDUX, we use attribute values that come from pre-existing data sources from each domain (e.g. addresses, bibliographic data, etc.) to label segments in the input text. These values are used to form domain-specific *Knowledge Bases*(*KBs*).

A Knowledge Base is a set of pairs $K = \{\langle a_1, O_1\rangle, \ldots, \langle a_n, O_n\rangle\}$ in which each a_i is a distinct attribute, and O_i is a set of strings $\{o_{i,1}, \ldots, o_{i,n_i}\}$ called *occurrences*. Intuitively, O_i is a set of strings representing plausible or typical values for attribute a_i.

Given a data source on a certain domain which includes values associated with fields or attributes, building a Knowledge Base is a simple process that consists in creating pairs of attributes and sets of occurrences. Example of possible data sources are: databases, reference tables, ontologies, etc.

In Figure 2 we present a very simple example of a KB-which includes only four attributes: *Neighborhood*, *Street*, *Bathrooms*, and *Phone*.

The first step in ONDUX operation is called *Blocking*. In this step, the input string is roughly segmented into units we call *blocks*. Blocks are simply sequences of terms (words) that are likely to form a value of an attribute. Thus, although terms in a block must all belong to a same value, a single attribute value may have terms split among two or more blocks. This concept is illustrated in Figure 1(c). Observe that the blocks containing terms "Mifflin" and "Ave" are parts of the same value of attribute Street.

Next, in the *Matching* step, blocks are matched against known attribute values, which are available in the Knowledge Base, using a small set of specific matching functions. By the end of the matching step, each block is *pre-labeled* with the name of the attribute for which the best match was found.

We notice that Blocking and Matching steps alone are enough to correctly label the large majority of the segments in the input string. Indeed, experiments with different domains, which we have performed and reported here, show that blocks are correctly pre-labeled in more than 80% of the cases. This is illustrated in Figure 1(d) in which the Matching was able to successfully label all blocks except for the ones containing the terms "Regent Square" and "Mifflin".

Problems such as this are likely to occur in two cases. First, *Mismatching*, happens when two distinct attributes have domains with a large intersection. For instance, when extracting from scientific paper headings, values from attributes Title and Keywords have usually several terms (words) in common. In our running example, as shown in Figure 1(c), "Regent Square" was mistakenly labeled with *Street* instead of *Neighborhood*. Second, *Unmatching*, happens when no matching was found for the block in the Knowledge Base, as the case of the block containing the term "Mifflin" in Figure 1(c).

To deal with such problems, our method deploys a third step we call *Reinforcement* in which the pre-labeling resulting from the Matching step is reinforced by taking into consideration the positioning and the sequencing of labeled blocks in the input texts.

To accomplish this, first, a probabilistic HMM-like graph model we call PSM(Positioning and Sequencing Model) is built. This model captures (*i*) the probability of a block labeled with ℓ appear in position p in the input text, and (*ii*) the probability of a block labeled with ℓ appear before a block labeled with m in the input text. Next, these probabilities are used to reinforce the pre-labeling resulting from the Labeling step, assigning labels to previous unmatched blocks and changing labels for blocks found to be mismatched so far.

One important point to highlight regarding ONDUX is that PSM is built without manual training, using the pre-labeling resulting from the Matching step. This implies that the model is learned *on-demand* from test instances, with no *a priori* training, relying on the very effective matching strategies of the Matching step.

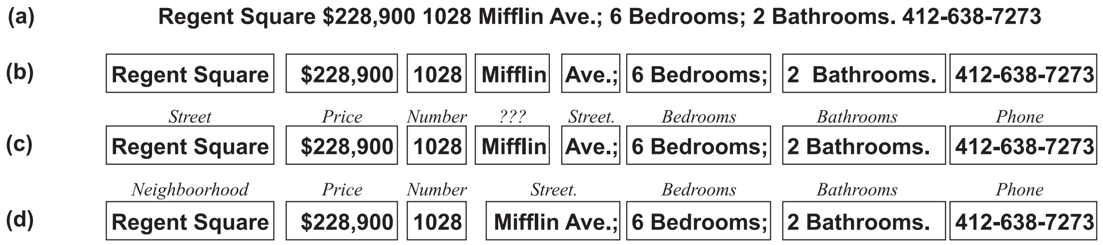

Figure 1: Example of an extraction process on a classified ad using ONDUX.

$$\begin{aligned} K =& \{\langle Neighborhood, O_{Neighborhood}\rangle,\ \langle Street, O_{Street}\rangle,\ \langle Bathrooms, O_{Bathrooms}, Phone, O_{Phone}\rangle\} \\ O_{Neighborhood} =& \{\text{“Regent Square”,“Milenight Park”}\} \\ O_{Street} =& \{\text{“Regent St.”,“Morewood Ave.”,“Square Ave. Park”}\} \\ O_{Bathrooms} =& \{\text{“Two Bathrooms”,“5 Bathrooms”}\} \\ O_{Phone} =& \{\text{“(323) 462-6252”,“171 289-7527”}\} \end{aligned}$$

Figure 2: A simple example of a Knowledge Base.

4. EXPERIMENTAL RESULTS

In this section, we report an experimental evaluation we have carried out with ONDUX using a real dataset to show that our method is a robust, accurate, and efficient unsupervised approach for IETS. We first describe the experimental setup and metrics used. Then, we report results on extraction quality and performance. Due to lack of space, we present here results for only one dataset. In [7] we present a larger set of experiments with domains and datasets.

4.1 Setup

Baselines

In the experiments, we compare ONDUX with an unsupervised version of CRF, a state-of-art IETS approach. This version was developed by adapting the publicly available implementation of CRF by Sunita Sarawagi [1], according to what is described in [16]. We call this version *U-CRF*. We believe that *U-CRF* represents the most suitable baseline for comparing with ONDUX, as it delivers top performance while at the same time does not require user-provided training. However, since this our first baseline assumes, as we shall see in more details later, that the order of the text sequences to be extracted is fixed, we also included the standard CRF model [10] (called *S-CRF*), that does not have this limitation at all but requires manually labeled training data.

As required by U-CRF, a batch of the input strings is used to infer the order of the attribute values. Based on the information provided in [16], this batch is composed by 10% of the input strings in all cases.

Experimental Data

The sources of previous known data, used to generated the KB for ONDUX ,the references tables for U-CRF, the training data for S-CRF, and the test datasets used in the experiments are summarized in Table 1.

We tried to use the same datasets and sources explored by our baselines, when these were publicly available. In the case of restricted sources/datasets, we tried to obtain public versions of similar ones in the same domains.

Metrics for Evaluation

In the experiments we evaluated the extraction results obtained after the Matching and Reinforcement steps discussed in Section 3. We aim at verifying how each step contributes to the overall effectiveness of ONDUX. In the evaluation we used the well known precision, recall, and F-measure metrics, but all tables report F-measure values.

Let B_i be a reference set and S_i be a test set to be compared with B_i. We define precision (P_i), recall (R_i) and F-measure (F_i) as: $P_i = \frac{|B_i \cap S_i|}{|S_i|}$, $R_i = \frac{|B_i \cap S_i|}{|B_i|}$, $F_i = \frac{2(R_i.P_i)}{(R_i+P_i)}$.

For all the reported comparisons with U-CRF, we used the Student's T-test for determining if the difference in performance was statistically significant. In all cases, we only draw conclusions from results that were significant in, at least, 5% level for both tests. Non-significant values are omitted.

4.2 Extraction Quality

4.2.1 Attribute-Level Results

To demonstrate the effectiveness of the whole extraction process with our method, we evaluate its extraction quality by analyzing, for each attribute, if the (complete) values assigned by our method to this attribute are correct.

Bibliographic Data Domain

This set of experiment was performed using the *CORA* test dataset. This dataset includes bibliographic citations in a variety of styles, including citations for journal papers, conference papers, books, technical reports, etc. Thus, it does not follow the single total attribute order assumption made by [16]. The availability of manually labeled data allowed us to include the S-CRF method in this comparison. A similar experiment is reported in [14]. Because of this, we have to generate our KBand the reference tables for U-CRF using the same data available on the unstructured labeled records we use to train the standard CRF, also from the *CORA* collection. As always, this training data is disjoint from the test dataset. The results for this experiment are presented in Table 2.

First, notice that the high results obtained with the supervised CRF (S-CRF) are similar to those reported in the original experiment [14]. In the case of ONDUX, even though it

[1] http://crf.sourceforge.net/

Domain	Source	Attributes	Records	Dataset	Attributes to be extracted	Text Inputs
Bibliographic Data	CORA	13	350	*CORA*	13	150
	PersonalBib	7	395			

Table 1: Data sources and test datasets used in the experiments.

			ONDUX	
Attribute	*S-CRF*	*U-CRF*	Matching	Reinforc.
Author	0.936	0.906	0.911	**0.960**
Booktitle	0.915	0.768	0.900	0.922
Date	0.900	0.626	0.934	**0.935**
Editor	0.870	0.171	0.779	**0.899**
Institution	**0.933**	0.350	0.821	0.884
Journal	0.906	0.709	0.918	**0.939**
Location	0.887	0.333	0.902	0.915
Note	0.832	0.541	0.908	**0.921**
Pages	**0.985**	0.822	0.934	0.949
Publisher	0.785	0.398	0.892	**0.913**
Tech	0.832	0.166	0.753	0.827
Title	**0.962**	0.775	0.900	0.914
Volume	0.972	0.706	0.983	0.993
Average	0.901	0.559	0.887	**0.921**

Table 2: Extraction over the *CORA* dataset using data from the *CORA* source.

is an unsupervised method, superior results were achieved. Statistically superior results were obtained in 6 out of 13 attributes (results in boldface) and statistical ties were observed in other 4 attributes. The results for U-CRF were rather low; this is explained by heterogeneity of the citations in the collections. While the manual training performed for S-CRF was able to capture this heterogeneity, U-CRF assumed a fixed attribute order. On the other hand, ONDUX was able to capture this heterogeneity through the PSM model, without any manual training.

Still on the Bibliographic data domain, we repeated the extraction task over the *CORA* test dataset, but this time, the previously known data came from the *PersonalBib* dataset. This dataset was used in a similar experiment reported in [11]. Again, our aim is demonstrate the source independent nature of unsupervised IETS methods. Notice that not all attributes from *CORA* were present in *PersonalBib* entries. Thus, we only extracted attribute available on both of them. The results for this experiment are presented in Table 3. Notice that in this case we could not perform manual training , since the previously known data came directly from a structured source.

The results for ONDUX and U-CRF are quite similar to those obtained in the previous experiments, with a large advantage for ONDUX, for the reasons we have already discussed.

5. COMPARISON WITH PREVIOUS APPROACHES

ONDUX falls in the category of methods that apply learning techniques to extract information from data rich input strings. As such, it has several points in common with previous methods that have been successfully applied to such a task, such as HMM [3] and CRF [10]. However, it also has unique characteristics that are worth discussing. As CRF is the current state-of-art method for this problem, we here compare our method to it. More specifically, we compare ONDUX with CRF-based methods in the literature that, like ONDUX, rely on previously known data to generate the extraction model. These are the methods presented in [11] and [16], which we refer to as Extended Semi-CRF (ES-CRF) and Unsupervised CRF (U-CRF, as in the previous section), respectively.

		ONDUX	
Attribute	*U-CRF*	Matching	Reinforcement
Author	0.876	0.733	**0.922**
Booktitle	0.560	0.850	**0.892**
Date	0.488	0.775	**0.895**
Journal	0.553	0.898	**0.908**
Pages	0.503	0.754	**0.849**
Title	0.694	0.682	**0.792**
Volume	0.430	0.914	**0.958**
Average	0.587	0.801	**0.888**

Table 3: Extraction over the *CORA* dataset using data from the *PersonalBib* source.

The first distinction between ONDUX and the other two approaches is the matching step. This step relies on a handful of generic matching functions and does not need to be trained for a specific target source, since it relies only on the known data available on the KB. In the case of text attributes, the matching function is based on the vocabulary of the attribute domain, as represented by terms available in the Knowledge Base, while for the numeric attributes the distribution probability of the known values is used. In CRF models, several distinct state features, i.e., those based only on the properties of each attribute [15], are used for learning the extraction model. In ES-CRF and U-CRF some of these features depend on the previously available data, but other features depend on the specific target source. This is the case of segment length and counting of (previously defined) regular expressions that fire in ES-CRF, and negative examples formed from token sequences taken from the input text in U-CRF.

The main difference between ONDUX and the two similar approaches, ES-CRF and U-CRF, is the way features related to positioning and sequencing, of attributed values (transition features [15]) are learned. In ONDUX these features are captured by the PSM model, which, as demonstrated in our experiments, is flexible enough to assimilate and represent variations in the order of attributes on the input texts and can be learned without user-provided training. U-CRF is also capable of automatically learning the order of attributes, but it cannot handle distinct orderings on the input, since it assumes a single total order for the input texts. This difficult the application of the method to a range of practical situations. For instance, in bibliographic data, it is common to have more than one order in a single dataset. Further, the order may vary when taking information from distinct text input sequences, according to the bibliographic style adopted on each input. The order is even more critical in classified ads, where each announcer adopts its own way of describing the object he/she is trying to sell. Another quite common application is to extract data from online shopping sites to store them in a database. The attributes of the offer, such as price, product, discount and so on, seldom appear in a fixed order. In practical applications like these, ONDUX is the best alternative method. Further, it is as good as the baselines for any other practical application.

In ES-CRF, distinct orderings are handled, but user-provided training is needed to learn the transition features, similarly to what happens with the standard CRF model, thus increasing the user dependency and the cost to apply the method in several practical situations.

Finally, ONDUX is largely influenced by FLUX-CiM [5, 6] an unsupervised approach for extracting metadata from bibliographic citations. While FLUX-CiM also relies on a matching step in which the AF function is also used, it does not include a generic reinforcement step. Instead, it uses a set of domain-specific heuristics based on assumptions regarding bibliographic metadata. This includes the use of punctuation as attribute value delimiters, the occurrence of single values for attributes other than author names, etc. As a consequence, FLUX-CiM could not be adopted as a baseline, since it was not designed for most of the datasets we have in our experiments. ONDUX can thus be seen as a significant improvement over FLUX-CiM, which instead of being applied only to bibliographic metadata, is a general IETS approach whose algorithms do not rely on domain-specific assumptions such as these. Specially, it does not explicitly rely on the use of punctuation.

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented partial results of our research on unsupervised strategies for information extraction by text segmentation. Specifically, we discussed ONDUX (**ON**-**D**emand **U**nsupervised Information E**X**traction), an alternative unsupervised probabilistic approach for IETS. ONDUX also relies on pre-existing data, more specifically, on sets of attributes values from pre-existing data sources to associate segments in the input string with a given attribute. Differently from previous work, there is not an explicit learning process in this step. Instead, we use simple generic matching functions to compute a score measuring the likelihood of text segments to occur as a typical value of an attribute.

To corroborate our claims regarding the high-quality, flexibility and effort-saving features of our approach, we tested our method with several textual sources from different domains and found that it achieved similar or better results than CRF, a state-of-art data extraction model. Our experiments also demonstrate that our approach is able to properly deal with different domains in heterogeneous applications.

We believe that the main contributions of our work are: (1) a very effective unsupervised information extraction method that (2) instead of requiring explicit learning of a model for identifying attributes values in the input texts, uses a simple but very effective greedy strategy based on matching, (3) exploits the high accuracy of this matching strategy to learn from the test data the probabilities of positioning and sequencing of attributes in an unsupervised manner, making no rigid assumptions about the order of the attribute values, thus being much more flexible and robust to changes in patterns, and finally (4) despite the fact it operates on-demand, it has processing time of test instances similar to that of methods that use explicit learning such as CRF.

The work we carried out with ONDUX opens opportunities for several future developments. We intend to investigate the use of alternative matching functions that might better distinguish attribute values. One of the functions we consider is the one proposed in [13], which is based on the commonality of features. In addition, currently ONDUX does not handle nested structures such as lists of values of a same attribute in a record. We also plan to address this issue as future work.

Acknowledgements

This work was supported by a CNPq fellowship grant to Altigran S. Silva and by a CAPES scholarship to Eli Cortez.

7. REFERENCES

[1] E. Agichtein and V. Ganti. Mining reference tables for automatic text segmentation. *Proc. of the ACM SIGKDD* , pages 20–29, Seattle, Washington,USA, 2004.

[2] S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis. Automated ranking of database query results. *Proc. of CIDR*, 2003.

[3] V. R. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. *Proc. of the ACM SIGMOD*, pages 175–186, 2001.

[4] S. Chuang, K. Chang, and C. Zhai. Context-aware wrapping: synchronized data extraction. *Proc. of the 33rd VLDB*, pages 699–710, Viena, Austria, 2007.

[5] E. Cortez, A. da Silva, M. Gonçalves, F. Mesquita, and E. de Moura. FLUX-CIM: flexible unsupervised extraction of citation metadata. *Proc. of the ACM/IEEE JCDL*, pages 215–224, 2007.

[6] E. Cortez, A. da Silva, M. Gonçalves, F. Mesquita, and E. de Moura. A flexible approach for extracting metadata from bibliographic citations. *JASIST*, 60(6):1144-1158, 2009.

[7] E. Cortez, A. S. da Silva, M. A. Gonçalves, and E. S. de Moura. Ondux: On-demand unsupervised learning for information extraction. In *Proc. of the ACM SIGMOD*, 2010.

[8] D. Freitag and A. McCallum. Information extraction with hmm structures learned by stochastic optimization. In *Proc. of the AAAI*, pages 584–589, Austin, Texas, USA, 2000.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. of the ICML*, pages 200–209, Bled, Slovenia, 1999.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the ICML*, pages 282–289, 2001.

[11] I. R. Mansuri and S. Sarawagi. Integrating unstructured data into relational databases. In *Proc. of the ICDE*, pages 29–40, 2006.

[12] F. Mesquita, A. da Silva, E. de Moura, P. Calado, and A. Laender. LABRADOR: Efficiently publishing relational databases on the web by using keyword-based query interfaces. *IPM*, 43(4):983–1004, 2007.

[13] U. Nambiar and S. Kambhampati. Answering imprecise queries over autonomous web databases. In *Proc. of the ICDE*, page 45, 2006.

[14] F. Peng and A. McCallum. Information extraction from research papers using conditional random fields. *IPM*, 42(4):963–979, 2006.

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

[16] C. Zhao, J. Mahmud, and I. V. Ramakrishnan. Exploiting structured reference data for unsupervised text segmentation with conditional random fields. In *Proc. of the SIAM ICDM*, pages 420–431, 2008.