Chapter 2
Related Work

Abstract In the literature, different approaches have been proposed to address the problem of extracting valuable data from the Web. In this chapter is presented an overview of such approaches. It begins by presenting a broad set of Web extraction methods and tools. Following a taxonomy previously used in the literature (Laender et al. 2002), they are divided into distinct groups according to their main approach. These groups are: *Languages for Wrapper Development*, *Wrapper Induction Methods*, *NLP-based Methods*, *Ontology-based Methods*, and *HTML-aware Methods*. Next, it is specifically presented probabilistic graph-based methods, *supervised* and *unsupervised*, and discusses their main characteristics in comparison to the unsupervised approach presented in this book.

Keywords Information extraction · Wrappers · NLP · HTML · Probabilistic methods · CRF

2.1 Web Extraction Methods and Tools

By the early 2000s, several tools and methods had been discussed in the literature for extracting valuable data from the Web. A survey on this early work is presented in Laender et al. (2002), where the authors proposed a *taxonomy* for grouping different Web extraction methods and tools based on the main approach used by each method. Here, the same taxonomy was adopted. In what follows, is described the main characteristics of the methods and tools belonging to each group.

2.1.1 Languages for Wrapper Development

One of the first initiatives for addressing the problem of extracting valuable data from the Web was the use of specialized programs able to identify data of interest and map them to some suitable format as, for instance, XML or relational tables.

E. Cortez and A. S. da Silva, *Unsupervised Information Extraction by Text Segmentation*, SpringerBriefs in Computer Science, DOI: 10.1007/978-3-319-02597-1_2,

These programs are called *wrappers*. Different *languages* were specially designed to assist users in developing wrappers. Such languages were proposed as alternatives to general purpose languages such as Perl and Java, which were prevalent at that time for this task.

Some of the best known tools that adopt this approach are Minerva (Crescenzi and Mecca 1998), TSIMMIS (Hammer et al. 1997), and Web-OQL (Arocena and Mendelzon 1998). Although such languages provide effective approaches for wrapper generation, their main drawback is that they required manual wrapper development. Due to such a limitation, efforts have been made to automate the wrapper generation process.

2.1.2 Wrapper Induction Methods

There were also efforts to use machine-learning techniques to semi-automatically induce wrappers (Hsu and Dung 1998; Kushmerick 2000; Muslea et al. 2001). In general, these approaches consist of using training examples to generate automata that recognize instances in contexts similar to the ones of the given examples.

The approach proposed by Kushmerick (2000) and adopted in the WEIN system relies on examples from the source to be wrapped. The main drawbacks of this work are: (1) it does not deal with missing or out-of-order components and (2) although it identifies the need for extraction of complex objects present in nested structures, the solution provided is computationally intractable and has not been implemented.

These two features of semi-structured data extraction are addressed in SoftMealy (Hsu and Dung 1998) and Stalker (Muslea et al. 2001). Both systems also generate wrappers, generalizing the given examples through machine-learning techniques, and are very effective in wrapping several types of Web pages. The main problem with SoftMealy is that every possible absence of a component and every different ordering of the components must be represented beforehand by an example. Stalker (Muslea et al. 2001) can deal with such variations in a much more flexible way since each object component is extracted independently through a top-down decomposition procedure.

The main drawback to all these approaches is that the extraction process relies on the knowledge of the structure of HTML pages. In WEIN and SoftMealy, for example, pages are assumed to have a defined structure (e.g., a head, then a body with a set of tuples, and then a tail) that must be flat. This prevents the exclusive extraction of the objects (or subobjects) of interest and might generate extraction difficulties if unwanted text portions (such as advertisements) occur between tuples or tuple components in the page body. In Stalker, the extraction of nested objects is possible but the approach also relies on a previous description of the entire source page.

2.1.3 NLP-Based Methods

Besides wrapper induction, there were other approaches for learning extraction patterns that were more suitable for extracting data from semi-structured texts such as newspaper classified advertisements, seminar announcements, and job posting, which present grammatical elements. In general, these approaches use techniques typical of Natural Language Processing (i.e., semantic class, part-of-the-speech tagging, etc.) sometimes combined with the recognition of syntactic elements (delimiters). This is the case of Rapier (Mooney 1999) and SRV (Freitag and McCallum 2000). WHISK (Soderland 1999) goes beyond and addresses a large spectrum of types of documents ranging from rigidly formatted to free text. For formatted text, this system has a behavior that is closer to wrapper induction systems like WEIN (Kushmerick 2000).

Recently, several new methods that also explore Natural Language Processing techniques have been proposed to deal with the *Open Information Extraction* (Etzioni et al. 2008) problem. In this context, the goal is to perform Web scale extraction from all types of textual documents available on the Web. The system makes a single data-driven pass over its dataset and extracts a large set of relational tuples without requiring any human input. Banko et al. (2007, 2009) introduced a system called TEXTRUNNER, an open information extraction system that is able to extract tuples from large datasets and also allow their exploration via user queries. Different from the presented unsupervised approach, these open information extraction approaches rely heavily on linguistic information requiring the presence of grammatical elements.

2.1.4 Ontology-Based Methods

An ontology-based approach to extracting data from Web sources was proposed by Embley et al. (1999a). This approach uses a semantic data model to provide an ontology that describes the data of interest, including relationships, lexical appearances, and context keywords. By parsing this ontology, a relational database schema and a constant/keyword recognizer are automatically generated, which are then used to extract the data that will populate the database. Prior to the application of the ontology, the approach requires the application of an automatic procedure to extract chunks of text containing data "items" (or records) of interest (Embley et al. 1999b). Then, the extraction process proceeds from the set of records extracted. Not only does this approach require the user to provide a conceptual description of the data to be extracted, but relies mainly on the expected contents of the pages, which is anticipated by the prespecified ontology. Further, this approach requires a specialist to build the ontology using a notation specially designed for this task.

2.1.5 HTML-Aware Methods

Crescenzi et al. (2001) proposed RoadRunner, a method that heavily explores the inherent features of HTML documents to automatically generate wrappers. RoadRunner works by comparing the HTML structure of two (or more) given sample pages belonging to a same "page class," generating as a result a schema for the data contained in the pages. To accurately capture all possible structural variations occurring on pages of a same page class, it is possible to provide more than two sample pages. The extraction process is based on an algorithm that compares the tag structure of the sample pages and generates regular expressions that handle structural mismatches found between the two structures. It should be noted that the process is fully automatic and no user intervention is required, a feature that was unique to RoadRunner by that time. Although very effective, RoadRunner relies on specific HTML features to uncover the structure of the objects to be extracted. In such cases, fully automated tools tend to make a lot of misinterpretations, in the sense that they can extract several unwanted data.

There are also methods that rely on the representation of the HTML documents as *DOM* trees. Reis et al. (2004) and Dalvi et al. (2009) propose techniques based on tree edit distance to perform the extraction task. In Zhao et al. (2005) the authors propose the use of both the visual content of the HTML pages as displayed on a browser and the HTML DOM tree to perform the extraction.

More recently, a set of methods has been proposed for detecting and extracting information available on HTML tables. A system that is able to explore tabular information available within HTML pages is described by Cafarella et al. (2008). For this, the *Webtables* system relies on the HTML markup to automatically detect the occurrence of tables and then extract attribute-value pairs. Following the same idea of exploring HTML structures, such as tables and lists, Elmeleegy et al. (2009) propose a technique that is able to not only extract information from HTML tables, but also lists, thus combining HTML markup characteristics with string alignment.

As it can be noticed, all of these approaches rely on the regularity of HTML documents and depend heavily on the HTML tags (document structure) to extract information of interest. In some cases, this assigned to these approaches good extraction results, however, precludes their usage in a large number of textual sources that are available on the Web. As seen in Fig. 1.1, the scenario that information extraction approaches faces nowadays includes textual sources in different formats and styles, and more specifically, free texts without any tag to explicitly indicate data of interest. In order to deal with these general textual sources the use of probabilistic graph-based approaches has been proposed, as described below.

2.2 Probabilistic Graph-Based Methods

Due to limitations of the extraction methods that are based on the HTML structure of Web pages, new methods, based on probabilistic graph-based approaches such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF), were created to tackle the problem of extracting valuable data from textual sources. A fairly common approach to solve this problem is the use of machine-learning techniques, either supervised, i.e., with human-driven training, or unsupervised, i.e., with training provided by some form of pre-existing data source.

2.2.1 Supervised Probabilistic Graph-Based Methods

One of the first approaches in the literature addressing the extraction problem with a probabilistic graph-based approach was proposed by Freitag and McCallum (2000). It consisted in generating independent Hidden Markov Models (HMM) for recognizing values of each attribute. This approach was extended in the DATAMOLD tool (Borkar et al. 2001), in which attribute-driven (or *internal*) HMM are nested as states of *external* HMM. These external HMM aim at modeling the sequencing of attribute values on the implicit records. Internal and external HMM are manually 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, *Conditional Random Fields (CRF)* models were proposed as an alternative to HMM for the extraction of valuable information from text (Lafferty et al. 2001). 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. Peng and McCallum (2006) proposed a supervised method for extracting bibliographic data from research papers based on CRF that showed good results in the experimental evaluation they conducted.

Kristjansson et al. (2004) also proposed the use of CRF to the task of filling Web forms with values available in unstructured texts. In this context, it is needed to extract valuable data from these texts and submit them to a predefined Web form with different form fields. Their interactive information extraction system assists the user in filling in form fields while giving the user confidence in the integrity of the data. The user is presented with an interactive interface that allows both the rapid verification of automatic field assignments and the correction of errors.

Although effective, these supervised information extraction 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.

2.2.2 Unsupervised Probabilistic Graph-Based Methods

To address the problem of requiring large amounts of manually created training sets, recent approaches presented in the literature propose the use of pre-existing data for easing the training process (Agichtein and Ganti 2004; Cortez et al. 2007; Mansuri and Sarawagi 2006; Zhao et al. 2008). These approaches take advantage of the existence of large amounts of structured datasets that can be used with little or no user effort.

According to the strategy of relying on pre-existing data, models for recognizing values of an attribute are generated from values of this attribute occurring in a dataset previously available. Mansuri and Sarawagi (2006) proposed a method based on Conditional Random Fields to extract valuable data from unstructured textual portions. The proposed method relies on pre-existing data to learn content-based features and hand-labeled training sets to learn structure-related features.

Agichtein and Ganti (2004) and Zhao et al. (2008) proposed methods that are able to train a model relying only on a pre-existing dataset and, then, use it for recognizing values of attributes among segments of the input string. No manually labeled training input strings are required for this. Once attribute values are recognized, records can be extracted. These methods assume that attributes values in the input text follow a single global order, which is learned from a sample batch of the test instances. The difference between the methods proposed by Agichtein and Ganti and the one proposed by Zhao et al. is that the first relies on Hidden Markov Models and the second relies on Conditional Random Fields. Despite this, both follow the same assumptions regarding a global attribute order in the input text.

The main difference between the presented approach and the ones presented by Agichtein and Ganti, Mansuri and Sarawagi, and Zhao et al., is the way that structure-related features (Sarawagi 2008) are learned. In the presented unsupervised approach these features, when necessary, are captured by a specific model, which, as demonstrated in the experiments, is flexible enough to assimilate and represent variations in the order of attributes in the input texts and can be learned without user-provided training. The methods proposed by Agichtein and Ganti (2004) and Zhao et al. (2008) are also capable of automatically learning structure-related features, but they cannot handle distinct orderings on the input, since they assume a single total order for the input texts. These make the application of these methods difficult to a range of practical situations. Thus, in practical applications, the presented unsupervised approach can be seen as the best alternative. The method proposed in Mansuri and Sarawagi (2006) can handle distinct ordering, but user-provided training is needed to learn the structure-related features, similar to what happens with the standard supervised CRF model, thus increasing the user dependency and the cost to apply the method in several practical situations.

A similar strategy is used by Chuang et al. (2007). 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 two-state HMM

for each attribute. Record extraction is addressed in an unsupervised way by aligning records from the sources being extracted.

FLUX-CiM (Cortez et al. 2007, 2009) is an unsupervised approach for extracting metadata from bibliographic citations that rely on the same ideas adopted by the unsupervised approach presented here. While FLUX-CiM also relies on content-based features learned from pre-existing data, it uses a set of domain-specific heuristics based on assumptions regarding bibliographic metadata to perform the extraction task. This includes the use of punctuation as attribute value delimiters, the occurrence of single values for attributes other than author names, etc. Thus, the presented unsupervised approach can be seen as a generalization of FLUX-CiM.

Michelson and Knoblock (2007) presented an unsupervised approach to exploit pre-existing data for extraction. To accomplish this, initially the user has to specify a large repository with distinct sets of pre-existing data. Once this repository is chosen, using simple vector-space model similarities between the input text and the available sets of pre-existing data, the system automatically finds the most suitable set for the given extraction task. Now that a set of pre-existing data was chosen, the system relies on predefined string distance metrics such as Jaro-Winkler and Smith-Waterman, and fine-tuned thresholds to perform the extraction of valuable data. This work differs from the presented unsupervised approach in the sense that it relies on the use of predefined string similarity functions other than content-based features based on vocabulary. Moreover, the proposed system requires the availability of large pre-existing datasets in order to perform the extraction task. In the unsupervised approach presented here, this is alleviated since, when possible, it is able to automatically induce structure-related features from content-based features, helping the extraction process.

In order to support these unsupervised extraction methods that have been recently proposed in the literature, Chiang et al. (2012) developed a system called *AutoDict* that is able to automatically discover dictionaries to support unsupervised probabilistic graph-based methods. Moreover, Serra et al. (2011) show that Wikipedia can be used to support information extraction methods. Thus, these works show that it is feasible to acquire pre-existing structured datasets in order to create unsupervised extraction methods.

References

Agichtein, E., & Ganti, V. (2004). Mining reference tables for automatic text segmentation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 20–29). USA: Seattle.

Arocena, G., & Mendelzon, A. (1998). Weboql: Restructuring documents, databases and webs. *Proceedings of the IEEE ICDE International Conference on Data Engineering* (pp. 24–33). USA: Orlando.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., & Etzioni, O. (2009). Open information extraction for the web. PhD thesis, University of Washington, Washington.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence* (pp. 2670–2676). India: Hyderabad.

Borkar, V., Deshmukh, K., & Sarawagi, S. (2001). Automatic segmentation of text into structured records. *Proceedings of the ACM SIGMOD International Conference on Management of Data Conference* (pp. 175–186). USA: Santa Barbara.

Cafarella, M., Halevy, A., Wang, D., Wu, E., & Zhang, Y. (2008). Webtables: Exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, *1*(1), 538–549.

Chiang, F., Andritsos, P., Zhu, E., & Miller, R. (2012). Autodict: Automated dictionary discovery. *Proceedings of the IEEE ICDE International Conference on Data Engineering* (pp. 1277–1280). USA: Washington.

Chuang, S., Chang, K., & Zhai, C. (2007). Context-aware wrapping: synchronized data extraction. *Proceedings of the VLDB International Conference on Very Large Data Bases* (pp. 699–710). Austria: Viena.

Cortez, E., da Silva, A., Gonçalves, M., Mesquita, F., & de Moura, E. (2007). FLUX-CIM: flexible unsupervised extraction of citation metadata. *Proceedings of the ACM/IEEE JCDL Joint Conference on Digital Libraries* (pp. 215–224). Canada: Vancouver.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., & de Moura, E. S. (2009). A flexible approach for extracting metadata from bibliographic citations. *Journal of the American Society for Information Science and Technology*, *60*(6), 1144–1158.

Crescenzi, V., & Mecca, G. (1998). Grammars have exceptions. *Information Systems*, *23*(8), 539–565.

Crescenzi, V., Mecca, G., & Merialdo, P. (2001). Roadrunner: Towards automatic data extraction from large web sites. *Proceedings of the VLDB International Conference on Very Large Data Bases* (pp. 109–118). Italy: Rome.

Dalvi, N., Bohannon, P., & Sha, F. (2009). Robust web extraction: an approach based on a probabilistic tree-edit model. *Proceedings of the ACM SIGMOD International Conference on Management of Data Conference* (pp. 335–348). Rhode Island, USA: Providence.

Elmeleegy, H., Madhavan, J., & Halevy, A. (2009). Harvesting relational tables from lists on the web. *Proceedings of the VLDB Endowment*, *2*(1), 1078–1089.

Embley, D., Campbell, D., Jiang, Y., Liddle, S., Lonsdale, D., Ng, Y., et al. (1999a). Conceptual-model-based data extraction from multiple-record web pages. *Data and Knowledge Engineering*, *31*(3), 227–251.

Embley, D., Jiang, Y., & Ng, Y. (1999b). Record-boundary discovery in web documents. *ACM SIGMOD Record*, *28*(2), 467–478.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. (2008). Open information extraction from the web. *Communications of the ACM*, *51*(12), 68–74.

Freitag, D., & McCallum, A. (2000). Information extraction with HMM structures learned by Stochastic optimization. *Proceedings of the National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence* (pp. 584–589). USA: Austin.

Hammer, J., McHugh, J., & Garcia-Molina, H. (1997). Semistructured data: The tsimmis experience. *Proceedings of the East-European Symposium on Advances in Databases and Information Systems* (pp. 1–8). Russia: St. Petersburg.

Hsu, C., & Dung, M. (1998). Generating finite-state transducers for semi-structured data extraction from the web. *Information systems*, *23*(8), 521–538.

Kristjansson, T., Culotta, A., Viola, P., & McCallum, A. (2004). Interactive information extraction with constrained conditional random fields. *Proceedings of the AAAI National Conference on Artificial Inteligence* (pp. 412–418). San Jose: USA.

Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, *118*(1–2), 15–68.

Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *SIGMOD Record*, *31*(2), 84–93.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the ICML International Conference on Machine Learning* (pp. 282–289). USA: Williamstown.

Mansuri, I. R., & Sarawagi, S. (2006). Integrating unstructured data into relational databases. *Proceedings of the IEEE ICDE International Conference on Data Engineering* (pp. 29–41). USA: Atlanta.

Michelson, M., & Knoblock, C. (2007). Unsupervised information extraction from unstructured, ungrammatical data sources on the world wide web. *International Journal on Document Analysis and Recognition*, *10*(3), 211–226.

Mooney, R. (1999). Relational learning of pattern-match rules for information extraction. *Proceedings of the National Conference on Artificial Intelligence* (pp. 328–334). USA: Orlando.

Muslea, I., Minton, S., & Knoblock, C. A. (2001). Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, *4*(1–2), 93–114.

Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing and Management*, *42*(4), 963–979.

Reis, D. C., Golgher, P. B., Silva, A. S., & Laender, A. F. (2004). Automatic web news extraction using tree edit distance. *Proceedings of the WWW International World Wide Web Conferences* (pp. 502–511). USA: New York.

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

Serra, E., Cortez, E., da Silva, A., & de Moura, E. (2011). On using wikipedia to build knowledge bases for information extraction by text segmentation. *Journal of Information and Data Management*, *2*(3), 259.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, *34*(1), 233–272.

Zhao, C., Mahmud, J., & Ramakrishnan, I. (2008). Exploiting structured reference data for unsupervised text segmentation with conditional random fields. *Proceedings of the SIAM International Conference on Data Mining* (pp. 420–431). USA: Atlanta.

Zhao, H., Meng, W., Wu, Z., Raghavan, V., & Yu, C. (2005). Fully automatic wrapper generation for search engines. *Proceedings of the WWW International World Wide Web Conferences* (pp. 66–75). Japan: Chiba.

http://www.springer.com/978-3-319-02596-4

Unsupervised Information Extraction by Text Segmentation

Cortez, E.; da Silva, A.

2013, XV, 94 p. 25 illus., Softcover
ISBN: 978-3-319-02596-4