Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Structured Information Extraction from On-line Advertisements- A Bayesian Approach

Saani H*, Reghu Raj P C

Dept. of Computer Science and Engineering

Govt. Engg. College Sreekrishnapuram, India

Abstract— Advertisements on websites are largely unstructured text even though individuals would naturally want to perform structured search over certain attributes of interest for purposes such as purchasing a car, a book or job searching. This paper describes a method for building an information extraction system for On-line Advertisements using various natural language processing techniques, machine learning, named entity recognition along with supervised learning algorithm. The information extracted from these advertisements can be used to perform search over certain attributes of interest.

Keywords— Information Extraction, Classification, Advertisements, Bayesian Classifier, Named Entity Recognition, Regular Expression.

I. INTRODUCTION

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. This enables much richer forms of queries on the abundant unstructured sources than what is possible with keyword searches alone. When structured and unstructured data coexist, information extraction makes it possible to integrate the two types of sources and pose queries spanning them. The extraction of structure from noisy, unstructured sources is a challenging task, that has engaged a veritable community of researchers for over two decades now. With roots in the Natural Language Processing (NLP) community, the topic of structure extraction now engages many different communities spanning machine learning, information retrieval, database, web, and document analysis. Early extraction tasks were concentrated around the identification of named entities, like people and company names and relationship among them from natural language text. Applications such as comparison shopping, and other automatic portal creation applications, lead to a frenzy of research and commercial activity on the topic. As society became more data oriented with easy online access to both structured and unstructured data, new applications of structure extraction came around.

The technological advances have brought us the possibility to access large amounts of textual information, either in the Internet or in specialised collections. However, people cannot read and digest this information any faster than before. In order to make it useful, it is often required to put this information in some sort of structured format. The information extraction (IE) technology is concerned with structuring the relevant information from a text of a given domain. In other words, the goal of an IE system is to find and link the relevant information while ignoring the extraneous and irrelevant one[6]. The research and development in IE have been mainly motivated by the Message Understanding Conferences (MUC).The web is a perfect publication forum for advertisements (ads for short), since ads websites (allow sellers to) post ads for potential buyers worldwide who can freely access archived and newly-created ads any time and anywhere, which cannot be provided by any traditional publication media. According to a report from eMarketer(.com), on-line advertising surpassed newspaper marketing in 2010 and the margin is widened in 2011, which indicates that on-line ads are popular and proliferating.

The proposed work is motivated by the fact that advertisements posted in a public bulletin board such as Craiglist, Ebay tend to be largely unstructured. The primary reason is that users are not required to post advertisements in strict structured format. Even with a search system, such heterogeneity of the web advertisements places cognition load for users to find information of their interest. By extracting structured information, it will be able to reduce users search time as well as to improve search performance. The aim of the proposed work is to design, implement and verify a system for structured information extraction from advertisements. For the domain of Craiglist Books advertisements, this work is aimed to choose appropriate attributes for the books and extract these attributes by choosing suitable NLP techniques. This work describes a method for building an information extraction system using various natural language processing techniques including machine learning, named entity recognition along with supervised learning algorithm. The system consists of two modules,

1. The task of the first module, the information extraction module consists of tagging (i.e. labelling) the textual content of the advertisement, in order to identify the information units that have to be extracted. Tagging is achieved by using specialised regular expressions, Named entity recognition and relative position analysis.

2. The task of the second module is to classify advertisements into a priori known classes (Engineering, Science, Management, Story or other). This step is needed to guide the information extraction process. Classification is performed using a naive Bayes classifier.

II. RELATED WORK

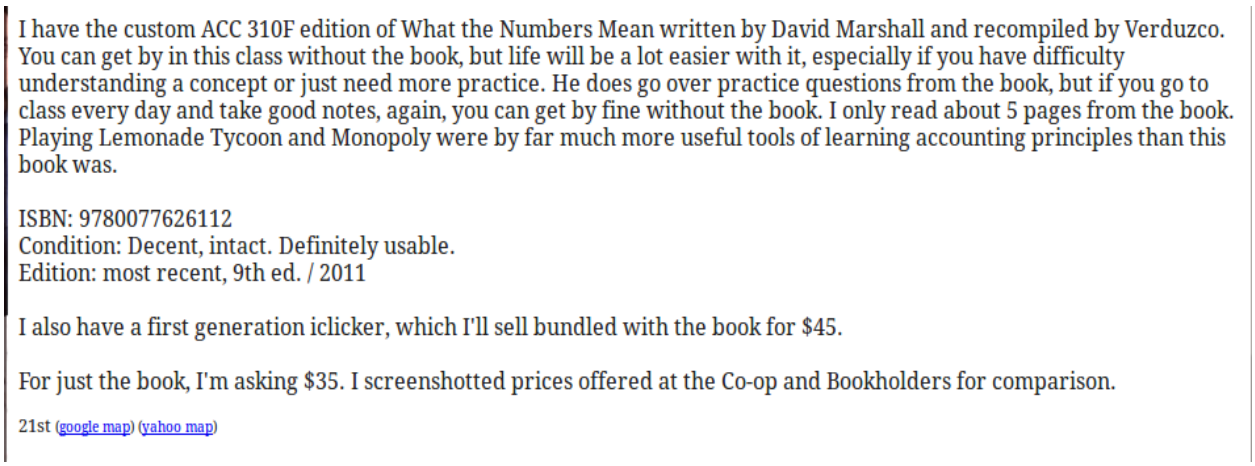
Information Extraction(IE) is an active research area in Natural Language Community. The use of machine learning (ML) methods in Information Extraction applications is mainly focused on the automatic acquisition of the extraction patterns. These patterns are used to extract the information relevant to a particular task from each single document of a given collection. There has been a lot of research works going on in the area of Information Extraction. Advertisement Extraction is one level of Information Extraction. In 2005, Alberto Tellez-Valero et al.[1] proposed a general method for building an information extraction system using regular expressions along with supervised learning algorithms. In this method, the extraction decisions are lead by a set of classifiers instead of sophisticated linguistic analyses. The paper also shows a system called TOPO that allows to extract the information related with natural disasters from newspaper articles in Spanish language. In 2003, Aravind Arasu et al.[2] proposed the comparison between different approaches including statistical, rule based IE. It stated that statistical methods can be useful when the training set is available. For developing rules in rule based IE require creating rules which may require a domain expert to find such rules. Also, the rule based systems are faster and more amenable to optimisations. In 1998, Hsu and Dung [12] classified wrappers into 4 distinct categories, including hand-crafted wrappers using general programming languages, specially designed programming languages or tools, heuristic-based wrappers, and WI approaches. Chang followed this taxonomy and compared WI systems from the user point of view and discriminated IE tools based on the degree of automation. They classified IE tools into four distinct categories, including systems that need programmers, systems that need annotation examples, annotation-free systems and semi supervised systems.

Sarawagi classified HTML wrappers into 3 categories according to the kind of extraction tasks [22]. The first category, record-level wrappers, exploits regularities to discover record boundaries and then extract elements of a single list of homogeneous records from a page. The second category, page-level wrappers, extracts elements of multiple kinds of records. Finally, the site-level wrappers populate a database from pages of a Web site. In 2008, Nipun Bhatia et al.[9] proposed a new approach to extracting information from Craigslist auto mobile ads. It shows that auto mobile ads are somewhat more easier to extract than other advertisements. It is for the reason that named entities in such ads are single words, not multiple words. For example, named entities that they used include 'brand' and 'model' whose instances are 'Toyota', 'Hyundai' etc.

Extracting information from cooking recipes into a machine interpretable format is described in paper [18].They worked with semi-structured XML recipe data. The goal was to reduce the preparation steps of a recipe to ACTION, INGREDIENT, UTENSIL groups. Semantic role labelling was the technique used to identify the relations. In 2012, Xiaoqing Zheng et al.[24] proposed an automated information extraction algorithm that can extract the relevant attribute-value pairs from product descriptions across different sites. A notion, called structural-semantic entropy, is used to locate the data of interest on web pages.

III. ADVERTISEMENT EXTRACTION

If one looks into advertisement websites, users are filling most of the important data into body of the advertisements. As the body of the advertisement is just a plain text (see figure 1), it is impossible to filter by such data. Information extraction comes to be helpful with this. Here the task is to create and verify extraction algorithm to gain as many attributes (fields) as possible from the book advertisements.



I have the custom ACC 310F edition of What the Numbers Mean written by David Marshall and recompiled by Verduzco. You can get by in this class without the book, but life will be a lot easier with it, especially if you have difficulty understanding a concept or just need more practice. He does go over practice questions from the book, but if you go to class every day and take good notes, again, you can get by fine without the book. I only read about 5 pages from the book. Playing Lemonade Tycoon and Monopoly were by far much more useful tools of learning accounting principles than this book was.

ISBN: 9780077626112
Condition: Decent, intact. Definitely usable.
Edition: most recent, 9th ed. / 2011

I also have a first generation iclicker, which I'll sell bundled with the book for $45.

For just the book, I'm asking $35. I screenshotted prices offered at the Co-op and Bookholders for comparison.

21st (google map) (yahoo map)

Figure 1. Advertisement

The problem can be characterised as an information extraction task, one where to populate rows in a relational database with values for certain attributes of interest. An information task may be defined as extracting segments of the

text to populate the columns of a database but in this case it would be ideal if we could go beyond this to understand the semantics of some of the attributes; for example, we would like to be able to extract an actual numeric value for the price of a book rather than just the raw text that represents the price in the advertisement. While at first it may seem like this is an easy task, on closer inspection of the advertisements it becomes clear that there are many ambiguities to deal with since the advertisements are not well-formatted and may short forms and acronyms are used.

IV. SYSTEM DESIGN

The proposed system is a information extraction system, that accepts request for book information as user query. The query may be title of a book or part of a book title. Output of the system will be collection of book details from document class library. The query processing is performed with the help of a Bayesian Classifier. The system flow is shown in figure (2 and 3).

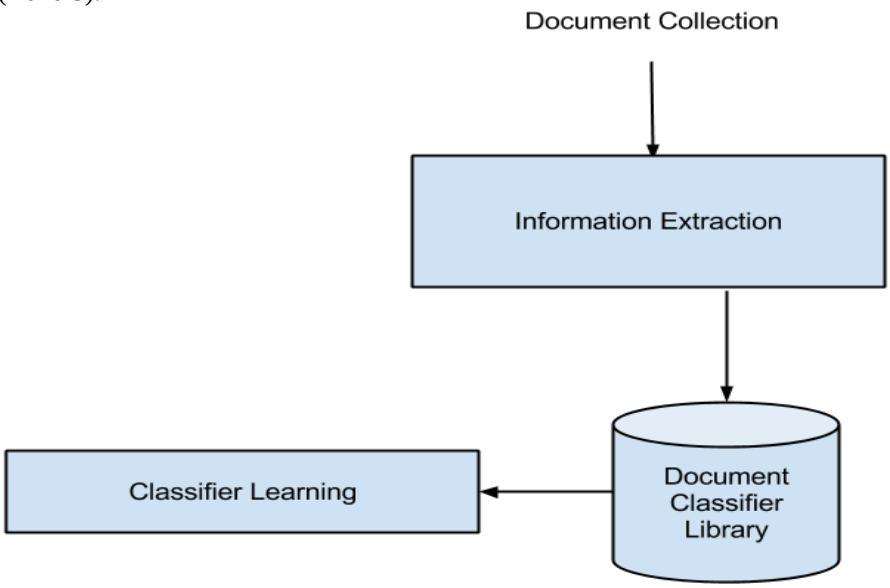


Figure 2: Work Flow diagram1

The system has following modules:

- Document Collection Module
- Pre processing Module
- Information Extraction Module
- Classifier Learning Module
- Query Processing Module

A. Document Collection and Pre processing

Data set for this work is the book advertisements from one of the famous On-line marketplace Craiglist. 200 ads in HTML file format are downloaded and saved separately. These forms the document collection for the proposed work. Each of these postings were unstructured and the relevant information is contained in title and main body. These HTML pages contains many design elements and also parts of the web page which are not connected to the particular advertisement - commercials, menus, footer text etc. In order to annotate these advertisements,the HTML postings are converted to text. Converting them to text resulted in the loss of HTML tags, however care was maintained to ensure that tags that help recognise the title in these advertisements were not lost. The free tool html2text was used for this purpose.

B. Information Extraction

There are two kinds of approaches usually adopted for the task of information extraction. The first approach is a rule based approach where rules are carefully crafted manually to identify instances of attributes. The second approach is to provide a statistical model with enough training data so that it can learn the rules that are most effective; usually this

approach is successful when ample training data is available. In this work the rule based method is used. In the proposed system, the aim is to extract the features or attributes of a book. The attributes considered are Book Title, Author, Price, Publication, Contact information etc. Here the attributes or candidates are extracted by nltks' regular expressions.

Regular expressions (RE) are well known way how to describe regular language. In all modern programming languages there is an implementation of RE used for pattern matching in operations like searching or replacing the text. In my case I am using RE to find candidate fields for information extraction. The candidates to be extracted may be present in body of the HTML file or may be at Title. 'Title' of the Book commonly present in the advertisements' title. 'Author' extraction is made by use of both rule based and Named Entity Recognition. The 'Author' may be specified separately in the body part of the ads, in such cases here used Regular Expression to extract the 'Author' information. In other case NLTK's Named Entity Recogniser will extract the details. To use NER, first the sentence is tokenised. After tokenising, POS tagging for each token is performed. Then using nltk.ne_chunk module named entities are identified.

'Publisher', 'Edition' are identify using nltk.re module. 'Contact' is another parameter where RE is used. The pattern is more complicated (contact can be phone number or e-mail).

$[0 - 9]\{9, 12\}|$

$(\backslash\backslash+)\{0, 1\}([0 - 9]\{2, 3\})\{3, 5\}|$

$[\wedge \backslash\backslash s@.] + ([\backslash\backslash.] \wedge \backslash\backslash s@\backslash\backslash.])+)\{1, 3\}$

The RE consists of three parts joined by(|) junction, the first part for the phone number with no spaces, the second part for the phone number with spaces between the groups of two or three numbers and finally the last part for the e-mail address.

Regular expression for price is:

$\backslash\$\backslash d + (\backslash.\backslash d+)?$

Extraction of price is a challenging task. For example, Figure 4. shows two prices mentioned in the advertisement related to different options for buying but only one of these is the true price of the book for sale. This is an example of the ambiguities that may arise while extracting the price. Figure 4 is another example showing another challenge in the price.

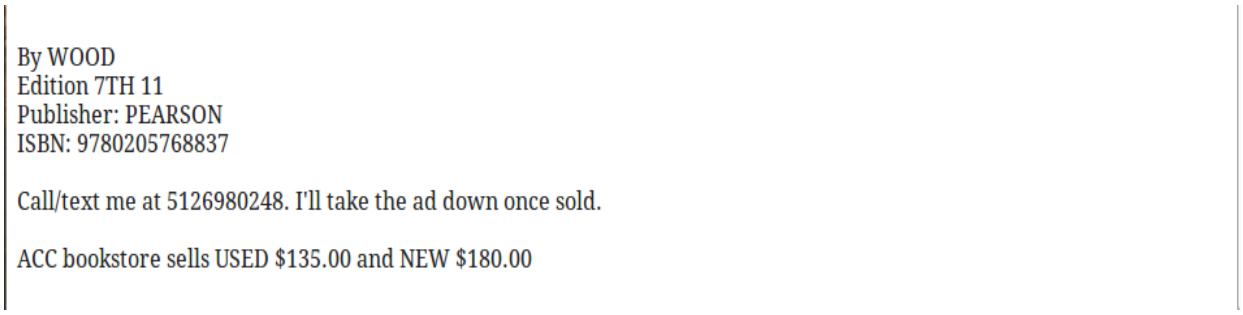


Figure 4. ambiguity in prices

The RE identifies all price values presented in the advertisement. Sometimes there may be more than one price in the advertisement. In such cases the actual price is identified by taking the previous two tokens of the identified price.

C. Classifier Learning

The proposed system uses a Bayesian classifier to perform the classification of the user query. As is the case with all degenerative classifiers, the classifier is first trained on a corpus. The downloaded HTML training corpus was first converted to text files. It was then annotated using the selected attribute values. Care was taken to extract the attributes from body and title. Using the title and body training data the the classifier was trained.

Bayesian classifier, as discussed earlier is a general class of feature based classifier. The important aspect of these classifier is the independent assumption. Bag of words assumption was taken to collect features from the training data. The classifier takes the structured information extracted from the downloaded advertisements to classify the user query to one of previously defined document classes. To achieve the goal, the title of the each extracted document is tokenised using a word punctuation tokeniser. Then bi-gram collection is calculated by using a nltk collocation bi-gram finder. Most commonly occurring bi-grams are taken as the feature.

D. User Query Processing

The user request for book details are entered through the GUI which is created using python Tkinter module. the query is fed to the Bayesian classifier as test data. The classifier will collect the features from the query and classifies the query to one of the predefined document classes. The class to which the query belongs in the document class library is searched and corresponding book details are displayed through the GUI.

V. RESULT AND OBSERVATION

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. In practice queries may be ill-posed and there may be different shades of relevancy.

The proposed system is an attempt to implement information extraction from On-line advertisements. Data set contains 200 advertisements on Book are downloaded from Craiglist site. To evaluate this system, Precision and Recall are used as evaluation metrics.

Precision-is the fraction of the documents retrieved that are relevant to the user´ information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall-is the fraction of the documents that are relevant to the query that are successfully retrieved

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

To evaluate the Extractor ,200 documents are given to the extractor as input, 90% correct output was produced for all the attributes except 'Author' information. For the author information the extractor provides 80% correct output. To evaluate the Classifier the document collection is divided into two sets, test set and training set. 100 documents are selected as test set and remaining is as training set. Initially 20 queries are taken as test set and training set contains 50 documents. 42 queries are classified correctly. The evaluation is performed many times by increasing the size of the document collection. The table 1: shows the performance of the overall system in terms of Recall.

TABLE 1. RESULT

Number of documents	**Number of queries**	**Correct answers**	**Recall in%**
20	50	42	84
30	50	41	82
80	50	41	82

VI. CONCLUSION AND FUTURE WORKS

The proposed system is an attempt to designing a system for structured information extraction from On-line advertisements. The system is a working system that applies NLP techniques to the real world problem of extracting meaningful features from Craigslist Book postings. The information present in the advertisements are unstructured or semi structured. This work uses regular expressions to find candidates for extraction instead of generalisation patterns. Named Entity Recognition is also used to extract some of the features. Bayesian classifier approach is used to classify the user request to one of the predefined document classes and the system will try search the document class to which the request belongs to extract book details. The system explained here could extract information from advertisements on singe book. The system not tried to extract GENERIC TITLE, that is advertisement that meant for more than one book for example Children´ story books are available for sale.

It would be interesting to use deeper semantic elements of the postings to discover the useful information. An obvious future work would be using large training data. The system presented here only focuses on Craiglists Books advertisements. Another future work is Scaling the system to a generic domain.

REFERENCES

[1] Alberto Tellez-Valero, Manuel Montes -y-GÃ¸mez, Luis VillaseÃ´or -Pineda, ”A Machine Learning Approach to Information Extraction”.Computational intelligence and Linguistics text processing,Springer,539-547,2005

[2] Aravind Arasu, Hector Garcia-Molina ”Extracting Structured Data from Web Pages”.International conference on Management of Data, Proceedings of the 2003 ACM, 337-348, 2003

[3] Bird S, Klein E, and Loper E, Natural Language Processing with Python, OâAZReilly Media, 2009.

[4] Chia Hui Chang, ”Automatic Information Extraction from Semi Structure Web Pages by Pattern Discovery”.Dept of CS and IT. National Central University, Taiwan ,2008

[5] Claire Cardie, ”Empirical methods in information extraction”.AI magazine, 18:65âAS79, 1997.

[6] Cowie, J., Lehnert, W, ”Information Extraction”.Communications of the ACM, Vol. 39,No. 1 (1996) 80-91

[7] Hsu, C.N. and Dung M, ” Generating finite-state transducers for semi-structured data extraction from the web”. Journal of Information Systems, 23(8): 521-538, 1998.

[8] J. Cordeiro and P. Brazdil, ”Learning text extraction rules, without ignoring stop words”,In 4th International Workshop on Pattern Recognition in Information Systems âA¸PRIS, pages 128âAS138. Cl âAS DBLP, 2004.

[9] Jiawei Han and Micheline Kamber, ”Data Mining: Concepts and Techniques”.Morgan Kaufmann Publishers Inc., San Francisco, CA, USA,2005.

[10] JosÃl’ Kahan and Marja-Ritta Koivunen, ”Annotea: an open rdf infrastructure for shared web annotations. In Proceedings of the 10th international conference on World Wide Web”, WWW 2001, pages 623âAS632,New York, NY,USA, 2001. ACM.

[11] Josh Herbach, Rohan Jain, ”Information Extraction from Housing Advertisements by David Murray”.Department of Computer Science, Stanford University,2009

[12] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan, ”Data mining: a knowledge discovery approach”.Springer Science+Business Media, Boston,- 2007.

[13] Maria S. Pera, Rani Qumsiyeh, Yiu-Kai Ng, ”Web-based Closed-Domain Data Extraction on Online Advertisements”.Journal of Information Systems, vol 38, Issue 2, pp. 183-197, Elsevier, April 2013

[14] Rubens, Mathew and Agarwal, Puneet, ”Information Extraction from Online Automotive Classifieds”.Dept. Of Computer Science ,Stanford University,2002

[15] Michele Banko, Michael J Cafarella, Stephen Soderl, Matt Broadhead,and Oren Etzioni, ”Open information extraction from the web”.In Proceedings of the IJCAI, pages 2670âAS2676, 2007.

[16] Nipun Bhatia, Rakshit Kumar, Shashank Senapaty, ”Extraction of Structured Information From Online Auto mobile Advertisements”.Department of Computer Science, Stanford University -2008

[17] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, AnaMaria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates, ”Web-scale information extraction in know-it-all: (preliminary results). In Proceedings of the 13th international conference on World Wide Web”, WWW âAZ04, pages S110, New York, NY, USA, 2004. ACM.

[18] Rahul Agarwal, Kevin Miller, ”Information Extraction from Recipes”.Department of Computer Science, Stanford University -2008

[19] Ram n Arag s Peleato, Jean-C dric Chappelier, ”Automated Information Extraction out of Classified Advertisements”.International Conference on Applications of Natural Language to Information Systems, France) – june 2000.

[20] Romain Loth, Delphine Battistelli , FranÃ˘ois -RÃl’gis Chaumartin, Hugues de Mazancourt, Jean-Luc Minel, Axelle Vinckx, ”Linguistic information extraction for job ads (SIRE project)”,Adaptivity, Personalisation and Fusion of Heterogeneous Information, 222-224,2010

[21] Siegfried Handschuh, Steffen Staab, and Fabio Ciravegn, ”semi-automatic creation of metadata. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web”, EKAW 2002, pages 358âAS372,London, UK, 2002.

[22] Sarawagi, S., ” Automation in information extraction and integration”, Tutorial of The 28th International Conference on Very Large Data Bases (VLDB), 2002.

[23] Tom Michael Mitchell, ”Machine Learning”.McGraw-Hill Series in Computer Science. WCB/McGraw-Hill, Boston, MA-1997

[24] Xiaoqing Zheng, Yiling Gu, Yinsheng Li, ”Data Extraction from Web Pages Based on Structural-Semantic Entropy”.WWW Companion volume, 2012

