ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

A Survey on Entity Extraction from the Web Pages Methodology and Application

K.B.Anusha

Assistant Professor, Department of CSE, AITAM, Tekkali, Andhra Pradesh, India

ABSTRACT: The Current Development of Web applications such as Blogs and Wiki enable users to easily create and disseminate their contents in the web. As the contents on the Web are rapidly growing, the Quantity of information is recently becoming more important in the web. Our approach defines a log-linear model over latent extraction predicates, which select lists of entities from the web page. The main challenge is to define features on widely varying candidate entity lists. Extracting and integrating these entity information from the Web is of great significance. Comparing to traditional information extraction problems, web entity extraction needs to solve several new challenges to fully take advantage of the unique characteristic of the Web. We also briefly introduce iKnoweb, an interactive knowledge mining framework for entity information integration.

KEYWORDS*:* Entity Extraction, latent Extraction, Entity Search, Entity Relationship Mining, Web Page Segmentation, Interactive Knowledge Mining.

I. INTRODUCTION

The need for collecting and understanding Web information about a real-world entity (such as a person or a product) is currently fulfilled manually through search engines. However, information about a single entity might appear in thousands of Web pages. Even if a search engine could find all the relevant Web pages about an entity, the user would need to sift through all these pages to get a complete view of the entity. Some basic understanding of the structure and the semantics of the web pages could significantly improve people's browsing and searching experience. In this paper we propose the information about a single entity may be distributed in diverse web sources, entity information integration is required. The most challenging problem in entity information integration is name disambiguation. This is because we simply don't have enough signals on the Web to make automated disambiguation decisions with high confidence. In many cases, we need knowledge in users' minds to help connect knowledge pieces automatically mined by algorithms. We propose a novel knowledge mining framework (called iKnoweb) to add people into the knowledge mining loop and to interactively solve the name disambiguation problem with users.

The main objective of this paper is to introduce the web entity extraction problem and to summarize the solutions for this problem.

Web entity extraction is different from traditional information extraction in the following ways:

- **Visual Layout**: In a web page, there is much visual structure which could be very useful in segmenting the web pages into a set of appropriate atomic elements instead of a set of words and in tagging the atomic elements using the attribute names.
- **Information Redundancy**: The same knowledge/fact about an entity may redundantly exist in multiple heterogeneous web pages with different text or layout patterns, and this redundancy could be very useful in statistical pattern discovery.
- **Information Fragmentation**: Information about a single entity is distributed in diverse web sources, each source may only have a small piece of its information, and the format of web pages across heterogeneous data sources is very different.
- **Knowledge Base**: The existing structured information about an entity in the knowledge databases could be very useful in extracting knowledge from other sources about this entity.

DOI: 10.15680/IJIRCCE.2018.0605087

II.RELATED WORK

Traditional “closed” IE work was discussed in above. Recent efforts seeking to undertake large scale extraction indicate a growing interest in the problem this year, Sekine in 2006] proposed a paradigm for on-demand information extraction,” which aims to eliminate customization involved with adapting IE systems to new topics. using unsupervised learning methods, the system automatically creates patterns and performs extraction based on a topic that has been specified by a user. Also this year, Shinyama and Sekine described an approach to “unrestricted relation discovery” that was developed independently of our work, and tested on a collection of 28,000 newswire articles.

This work contains the important idea of avoiding relation specificity, but does not scale to the Web as explained below. Given a collection of documents, their system first performs clustering of the entire set of articles, partitioning the corpus into sets of articles believed to discuss similar topics. Within each cluster, named-entity recognition, co-reference resolution and deep linguistic parse structures are computed and then used to automatically identify relations between sets of entities. This use of “heavy” linguistic machinery would

be problematic if applied to the Web. Shinyama and Sekine’s system, which uses pair wise vector-space clustering, initially requires an O(D2) effort where D is the number of documents. Each document assigned to a cluster is then subject to linguistic processing, potentially resulting in another pass through the set of input documents. This is far more expensive for large document collections than TEXTRUNNER’s O(D+T log T) runtime as presented earlier. From a collection of 28,000 newswire articles, Shinyama and Sekine were able to discover 101 relations. While it is difficult to measure the exact number of relations found by TEXTRUNNER on its 9,000,000Web page corpus, it is at least two or three orders of magnitude greater than 101.

III . EXPERIMENTAL METHODOLOGY

In this section we describe the data, metrics, and method used to test experimentally.

3.1 Data: Knowledge Base

The knowledge base was created from citations of scientific publications in Computer Science and from author home pages identified by the University of Trier's Home Page Search Web site. At the time of the experiments it contained 93,989 author names and 9,292 organization names. Simple heuristics were used to identify and conflate names that differed due to minor spelling errors, common abbreviations, or minor naming variations, but there remain many multiple entries for single individuals (e.g.,\John Smith", \J. Smith", \John Q. Smith") and single organizations.

3.2 Data: Web Pages

Two sets of pages (training and test) were downloaded from the Web. The pages were chosen because they fell into one of three categories: i) listed a journal editorial board, ii) listed a conference program committee, or iii) listed a conference program (e.g., paper titles and author names). The training pages were available for examination and used during system development. This set was augmented occasionally during development, to provide new pages for blind testing. This set eventually consisted of 19 pages. The set of 26 testing pages was kept separate from the training data, and was not used or examined during system development.

3.3 Metrics

Three metrics were used to measure the accuracy of named-entity detectors: Precision, Recall, and F-Score.
They are defined as:

$$P = \frac{ee}{ee + eo} \qquad (1)$$

$$R = \frac{ee}{ee + oe} \qquad (2)$$

$$F = \frac{2PR}{R + P} \qquad (3)$$

Where:
ee: number of entities identified correctly;
eo: number of strings mistakenly claimed to be entities; and
oe: number of entities not identified.
These are standard metrics for evaluating named-entity detection.

3.4 Baseline System

We compared a well-known and very effective commercial product based on Hidden Markov Models. IdentiFinder was supplied to us already trained on a large corpus of newswire and similarly well-written text. IdentiFinder was used as it came \out of the box", without any tuning for this task. IdentiFinder is a configurable system. It could have been tuned for this task, by adding features that recognized document structure (e.g., HTML tags), and by training on the set of training documents we examined during KENE system development. However, IdentiFinder is based on a Hidden Markov Model with many parameters. We felt that training on so few documents, and documents with so few similarities to one another, would have reduced its accuracy, and not improved it.

IV. Web Entity Extraction

We summarize our work on web entity extraction. Specifically, we first introduce three types of features we use in web entity extraction: visual layout features, text patterns, and knowledge base features. Then we present a statistical model to jointly optimize both page layout understanding and text understanding for

There exist three types of information that could be utilized for web entity extraction: visual layout features, text patterns, and knowledge base features. In the following, we will discuss them respectively.

4.1 Visual Layout Features

Web pages usually contain many explicit or implicit visual separators such as lines, blank area, image, font size and colour, element size and position. They are very valuable for the extraction process. Specifically, it affects two aspects in our framework: block segmentation and feature function construction.
Using visual information together with delimiters is easy to segment a web page into semantically coherent blocks, and to segment each block of the page into appropriate sequence of elements for web entity extraction.

4.2 Text Features

Text content is the most natural feature to use for entity extraction. Traditionally, the text information is treated as a sequence of words to be labelled. Statistics about word emission probabilities and state transition probabilities are computed on the training dataset, and then these statistics are used to assist labelling the words one by one.

4.3 Knowledge Base Features

For some web entities, there may be some structured information in the knowledge base about them already. This structured information can be used to remarkably improve the extraction accuracy in three ways.

1. First of all, we can treat the information in the knowledge base as additional training examples to compute the *element (i.e. text fragment) emission probability*, which is computed using a linear combination of the emission probability of each word within the element.

2. Secondly, the knowledge base can be used to see if there are some matches between the current text fragment and stored attributes.
3. Thirdly, if we found a good match between the entity information in the web page and the key attributes of an entity in the knowledge base, we can say with high confidence that the information on the web page refers to the same entity in the knowledge base. Then we can use other attributes of this entity in the knowledge base to label the rest elements of the web page or rectify wrong labels.

V.CONCLUSIONS

This paper introduces Open IE from the Web, an unsupervised extraction paradigm that eschews relation-specific extraction in favour of a single extraction pass over the corpus.

During which relations of interest are automatically discovered and efficiently stored. Unlike traditional IE systems that repeatedly incur the cost of corpus analysis with the naming of each new relation, Open IE's one-time relation discovery procedure allows a user to name and explore relationships at interactive speeds. We first introduced our vision-based web entity extraction work, which considers visual layout information and knowledge base features in understanding the page structure and the text content of a web page. We then introduced our statistical snowball work to automatically discover text patterns from billions of web pages leveraging the information redundancy property of the Web. We also introduced iKnoweb, an interactive knowledge mining framework, which collaborates with the end users to connect the extracted knowledge pieces mined from Web and builds an accurate entity knowledge web.

VI.ACKNOWLEDGMENTS

I gratefully acknowledge the support of the Google Natural Language Understanding Focused Program. In addition, I would like to thank anonymous reviewers for their helpful comments.

REFERENCES

[1] Eugene Agichtein, Luis Gravano: *Snowball*: extracting relations from large plain-text collections. In Proceedings of the fifth ACM conference on Digital libraries, pp. 85-94, June 02-07, 2000, San Antonio, Texas, United States. [DOI：10.1145/336597.336644]

[2] G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In Proceedings of International Conference on Machine Learning (ICML), Corvallis, OR, June 2007. [DOI：10.1145/1273496.1273501]

[3] [Rosario and Hearst, 2004] B. Rosario and M. Hearst. Classifying semantic relations in bioscience text. In *Proc. Of ACL*, 2004. [Sekine, 2006] S. Sekine. On-demand information extraction. mIn *Procs. of COLING*, 2006.

[4] A. Arasu and H. Garcia-Molina. 2003. Extracting structured data from web pages. In ACM SIGMOD international conference on Management of data, pages 337–348.

[5]J. Berant, A. Chou, R. Frostig, and P. Liang. 2013 Semantic parsing on Freebase from question-answer pairs. In Empirical Methods in Natural Language Processing (EMNLP).

[6]Hoffmann, C. Zhang, X. Ling, L. S. Zettlemoyer, and D. S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In Association for Computational Linguistics(ACL), pages 541–550.

[7]N. Kushmerick. 1997. Wrapper induction for information extraction. Ph.D. thesis, University ofWashington.

[8] M. L. Mauldin. Information Retrieval by Textn Skimming. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1989.

[9] L. F. Rau. Extracting company names from text. In Proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications, 1991.