

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IJCSMC, Vol. 3, Issue. 2, February 2014, pg.877 – 880

RESEARCH ARTICLE

ENTITY SEARCH ENGINES

Pinky Paul

Jyothi Engineering College, Thrissur

Mr. Thomas George

Assistant Professor, JECC, Thrissur

Abstract— *This review paper presents a study about entity search engine .It describes details about entity search engine architecture and it's working or the different methods adopted for entity extraction on it. All these methods are described based on the working examples such as entity cube and Microsoft academic search. Search engine for searching the summaries of an entity that make the user searching easy. In entity search engine, it extracts all the entities and relationships from heterogeneous web pages through different techniques. And finally integrate all these extracted information into a single unit.*

Key terms - *Entity Extraction, Markov Logic Networks, Shallow Natural Language processing Entity Relationship Mining, semi CRF, Web Page Segmentation, Interactive Knowledge Mining*

I. INTRODUCTION

Entity search engine extract the all the information about a particular entity and entity relationships from heterogeneous web pages and integrate it with the information from knowledge base about that entity and act as a single unit. So for that introduce a vision based entity extraction model. This model is created by considering the visual features of the web.

For retrieving the entities and relationships from the web different patterns have to be generated. For that Statistical Snowball (StatSnowball)[2] approach is introduced. This approach iteratively discovers extraction patterns in a bootstrapping manner. Starting with a handful set of initial seeds, it iteratively generates new extraction patterns and extracts new entity facts. The discovered extraction patterns can be used as the text features for web entity extraction in general.

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, search engine need knowledge in users' minds to help connect knowledge pieces automatically mined by algorithms. So 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.

Microsoft developed two entity search engine Entity cube and Microsoft Academic Search (aka Libra Academic).But they are not fully implemented. Now it is under the development. China developed an entity search engine called Renlifang is a Chinese version and it's English version is Entity cube. Entity cube is for users to search and browse summaries of entities including

people, organizations, and locations. Entity cube that is an automatically generated entity relationship graph based on knowledge extracted from billions of web pages.

Microsoft Academic Search (aka Libra Academic) is for users to search and browse information about academic entities including papers, authors, organizations, conferences, and journals.

II. ARCHITECTURE

First, a crawler fetches web data related to the targeted entities, and the crawled data is classified into different entity types, such as papers, authors, products, and locations. For each type, a specific entity extractor is built to extract structured entity information from the web data.

At the same time, information about the same entity is aggregated from different data sources including both unstructured web pages and the structured data feeds from content providers.

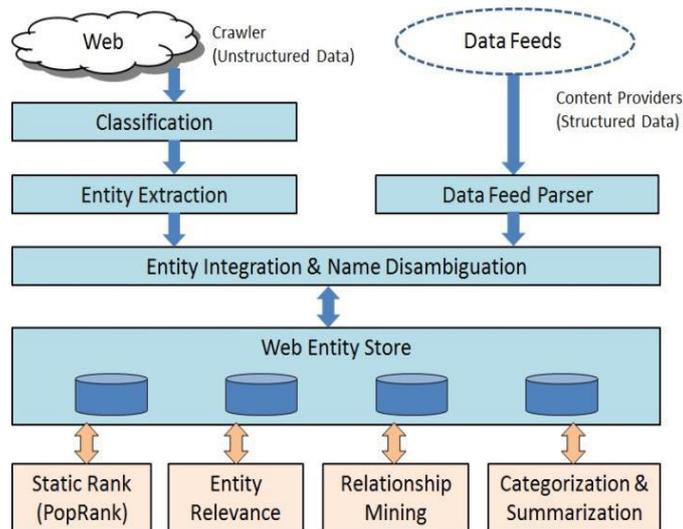


Figure 1. System Architecture of Entity Search Engines

Once the entity information is extracted and integrated, it is put into the web entity store, and entity search engines can be constructed based on the structured information in the entity store. Moreover, advanced entity ranking and mining techniques can be applied to make search more accurate and intelligent.

III. VISION BASED WEB ENTITY EXTRACTION

In the vision based web entity extraction model, three types of features are used. They are visual layout features, text patterns, and knowledge base features. These statistical Extraction models jointly optimize both page layout understanding and text understanding for web entity extraction leveraging these three types of features.

VIPS (VIsion-based Page Segmentation) algorithm [3] to extract the semantic structure for a web page. Such semantic structure is a hierarchical structure in which each node will correspond to a block. Each node will be assigned a value (Degree of Coherence) to indicate how coherent of the content in the block based on visual perception. The VIPS algorithm makes full use of page layout feature. It first extracts all the suitable blocks from the html DOM tree, and then it tries to find the separators between these extracted blocks. Here, separators denote the horizontal or vertical lines in a web page that visually cross with no blocks. Finally, based on these separators, the semantic structure for the web page is constructed. VIPS algorithm employs a top-down approach, which is very effective.

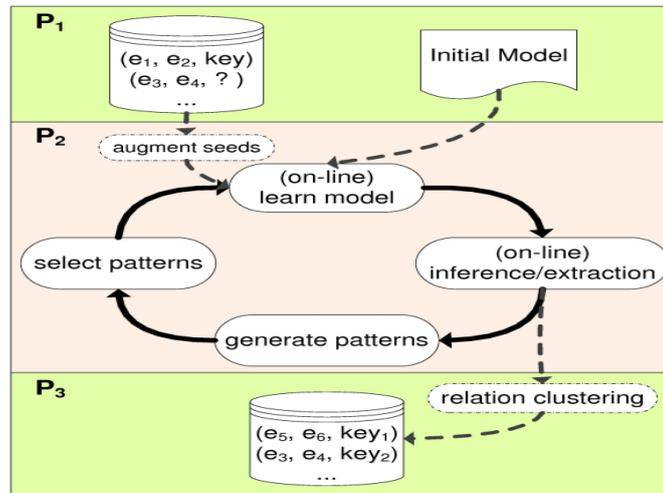
A well-defined joint statistical model can integrate both the visual layout understanding and the web text understanding (considering visual layout features, text patterns, and knowledge base features) together, so that the labelling results of the HTML elements and page layout can give a priori for further understanding the texts within the HTML elements, while the

understanding of the text fragments with the HTML elements can also give semantic suggestions to improve page layout understanding. Different models of vision based web entity extraction s are Vision-based Page Layout Understanding, Web Page Text Segmentation and Labelling, Joint Optimization of Layout and Text Understanding.

IV. STATSNOVBALL

StatSnowball automatically generates and selects the extraction patterns. It uses the information redundancy property of the web. StatSnowball [2] iteratively discovers both new facts/relations of an entity .In addition to the entity relation/fact extraction task, the discovered text patterns can also be used as text features. Statistical Snowball (StatSnowball), which is a bootstrapping system and can perform both traditional relation extraction and Open IE.

Generally, StatSnowball has three parts. The first part P1 is the input, which contains a set of seeds and an initial model. The seeds are not required to contain relation keywords that indicate the relationship. Thus, there are two types of seeds, i.e., seeds with relation keywords like (e1, e2, key) or seeds without relation keywords like (e3, e4, ?). If the initial model is empty, then will first use the seeds to generate extraction patterns in order to start the process.



The second part P2 is the statistical extraction model. To start the iterative extraction process, StatSnowball takes the input seeds and the initial model (can be empty) in P1 to learn an extractor. The third step in P2 is to generate extraction patterns with the newly identified relation tuples.

The patterns are generated based on the keyword matching patterns and general patterns. In keyword matching patterns, there are two parts. The first part is from the initial seeds. Users can provide seeds with some keywords to indicate the relationships. We take these keywords to define candidate patterns, e.g., a candidate pattern should contain at least one of these keywords. The second parts of the keywords are those that are automatically discovered during the Stat-Snowball extraction process. General extraction patterns are all based on a shallow natural language processing (NLP) technique—part-of speech tagging (POS). Much work has been done to investigate the usability of shallow or deep linguistic structures for various application tasks such as named entity extraction, and relationship identification

Selecting patterns is a feature induction problem of Markov random fields or Markov networks. In MLN [4], the problem is called structure learning. Alchemy uses a generative approach to learning the structure of MLN by using beam search to generate candidate formulae and selecting good candidates according to the gain in (weighted) pseudo-likelihood. In StatSnowball, we apply the ℓ_1 -norm regularized MLE as defined in the problem P and do discriminative structure learning. First use the generated patterns to formulate a set of candidate formulae of MLN. Then, we apply the algorithm to optimize the ℓ_1 -norm penalized conditional likelihood function as in the problem P, which yields a sparse model by setting some formulae's weights to zeros. The zero-weighted formulae are discarded and the resultant model is passed to the next step for re-training.

V. RELATED WORK

After extracting all the entities and their relationships from unstructured and structured data all these information has to integrate it into a single unit that is interactive entity information integration. This is the last phase of the entity search engine construction. The information about a single entity may be distributed in diverse web sources. So entity information integration is unavoidable one. The most challenging problem in entity information integration is name disambiguation. For solving this name disambiguation propose a novel knowledge mining framework (called iKnoweb). This adds people into the knowledge mining loop and to interactively solve the name disambiguation problem with users. Because the same knowledge may be represented using different text patterns in different web pages, this motivates us to use bootstrapping methods to interactively discover new patterns through some popular seed knowledge.

One important concept in iKnoweb[1] is Maximum Recognition Units (MRU), which serves as atomic units in the interactive name disambiguation process. A Maximum Recognition Unit is a group of knowledge pieces (such as web appearances, scientific papers, entity facts, or data records), which are fully automatically assigned to the same entity identifier with 100% confidence that they refer to the same entity (or at least with accuracy equal to or higher than that of human performance), and each Maximum Recognition Unit contains the maximal number of knowledge pieces which could be automatically assigned to the entity given the available technology and information.

VI. CONCLUSION

In Entity Search Engine, which targets to extract and integrate all the related web information about the same entity together as an information unit. In web entity extraction, it is important to take advantage of the following unique characteristics of the web: visual layout, information redundancy, information fragmentation, and the availability of a knowledge base. 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. Statistical snowball work to automatically discover text patterns from billions of web pages leveraging the information redundancy property of the Web. 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.

REFERENCES

1. Zaiqing Nie, Ji - Rong Wen, and Wei-Ying Ma, *Fellow, IEEE*. Statistical Entity Extraction from Web. Proceedings of IEEE vol.100 no.9 year 2012.
2. Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, Ji- Rong Wen. StatSnowball: A Statistical Approach to Extracting Entity Relationships. In Proceedings of the 18th international conference on World Wide Web.
3. Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report, MSR-TR-2003-79, 2003.
4. M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.
5. Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma. Web Object Retrieval. In Proceedings of the 16th international conference on World Wide Web (WWW).
6. S. Brin. Extraction patterns and relations from the World Wide Web. In International Workshop on the Web and Databases.