# ENTITY EXTRACTION USING STATISTICAL METHODS USING INTERACTIVE KNOWLEDGE MINING FRAMEWORK

**SHAIK MUNEEB AHAMED[1], SD.AFZAL AHMAD[2], P.BABU[3]**

[1]PG Student, QCET
[2,3]Associate Professor, QCET, Nellore

**Abstract-** There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases. 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. In this paper, we introduce our recent work on statistical extraction of structured entities, named entities, entity facts and relations from Web. We also briefly introduce iKnoweb, an interactive knowledge mining framework for entity information integration. We will use two novel web applications, Microsoft Academic Search (aka Libra) and EntityCube, as working examples.

## 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 will discuss the recent results and trends in web entity extraction, in the context of two novel web applications.

Entity Cube (http://www.entitycube.com) for users to search and browse summaries of entities including people, organizations, and locations. The Chinese version of EntityCube is called Renlifang (http://renlifang.msra.cn);

Microsoft Academic Search (aka Libra Academic, http://academic.research.microsoft.com) for users to search and browse information about academic entities including papers, authors, organizations, conferences, and journals.

The entities and their relationships in Entity Cube and Libra are automatically mined from billions of crawled web pages and integrated with existing structured knowledge from content providers. For each crawled web page, we extract entity information and detect relationships, covering a spectrum of everyday individuals and well-known people, locations, conferences, journals, and organizations.

**Existing System:**
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.

**Proposed System:**
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.

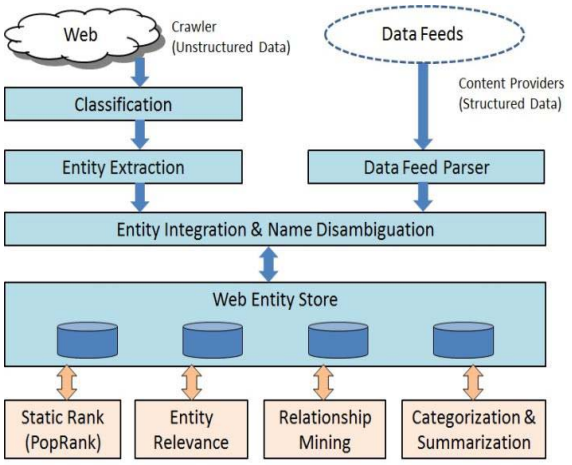


**Fig:- System Architecture of Entity Search Engines**

Modules:

- Admin

It receives the request from clients and processes it, sends the requested video corresponding to client bandwidth.

- Client

It send the request to the server, request contains requested file, available bandwidth of client and receives the video based on its available bandwidth.

**IMPLEMENTATION**

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

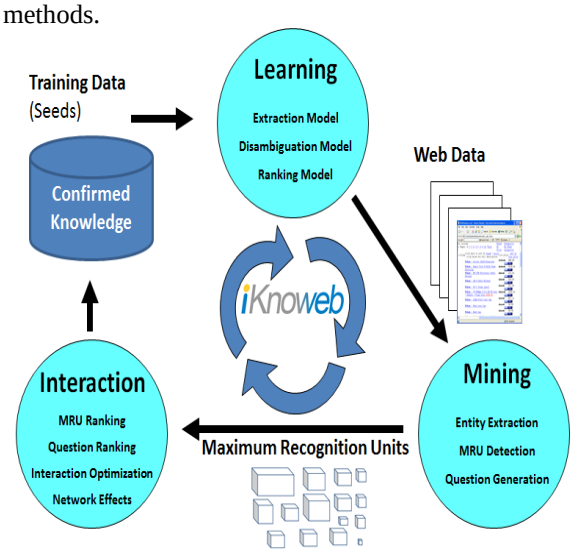


**Figure 13. The iKnoweb Framework**

Main Modules:-

**MODULE DESCRIPTION:**

1. Web Entity Extraction
2. Detecting Maximum Recognition Units
3. Question Generation
4. Network Effects
5. Interaction Optimization

Modules Description

**1.** Web Entity Extraction

➢ Visual Layout Features

- Web pages usually contain many explicit or implicit visual separators such as lines, blank area, image, font size and color, 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.
- Visual information itself can also produce powerful features to assist the extraction. For example, if an element has the maximal font-size and centered at the top of a paper header, it will be the title with high probability.

➢ Text Features

- Text content is the most natural feature to use for entity extraction.In web pages, there are a lot of HTML elements which only contain very short text fragments (which are not natural sentences). We do not further segment these short text fragments into individual words.
- Instead, we consider them as the atomic labeling units for web entity extraction. For long text sentences/paragraphs within web pages, however, we further segment them into text fragments using algorithms like Semi-CRF .

➢ Knowledge Base Features

- 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. In this way we can build more robust feature functions based on the element emission probabilities than those on the word emission probabilities.

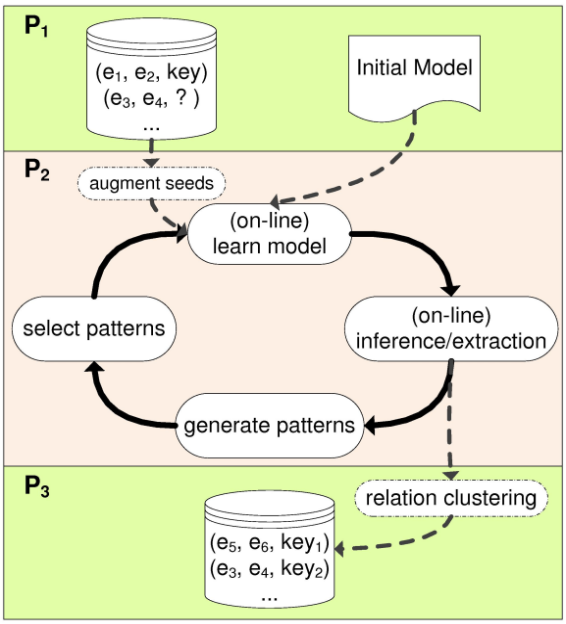


**Figure 11. The StatSnowball Framework, with three parts: P1 (input), P2 (statistical extraction model), and P3 (output).**

- The knowledge base can be used to see if there are some matches between the current text fragment and stored attributes. We can apply the set of domain-independent string transformations to compute the matching degrees between them.

**2.** Detecting Maximum Recognition Units
We need to automatically detect highly accurate knowledge units, and the key here is to ensure that the precision is higher than or equal to that of human performance.

**3.** Question Generation
By asking easy questions, iKnoweb can gain broad knowledge about the targeted entity. An example question could be: "Is the person a researcher? (Yes or No)", the answer can help the system find the topic of the web appearances of the entity.

**4.** Network Effects
A new User will directly benefit from the knowledge contributed by others, and our learning algorithm will be improved through users participation.

**5.** Interaction Optimization
This component is used to determine when to ask questions, and when to invite users to initiate the interaction and to provide more signals.

**CONCLUSION:**

How to accurately extract structured information about real-world entities from the Web has led to significant interest recently. This paper summarizes our recent research work on statistical web entity extraction, 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. Specifically, 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.

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] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 2670–2676.

[4] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In Proceedings of the Association for Computational Linguistics (ACL), 2008, pp. 28-36.

[5] S. Brin. Extraction patterns and relations from the World Wide Web. In International Workshop on the Web and Databases (WebDB), 1998, pp. 172—183.

[6] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma. Block-Level Link Analysis. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK (SIGIR), 2004, pp. 440-447.

[7] 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.

[8] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. Block-based Web Search. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Manuscript