International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Extract Entities with Iknoweb Framework

Utkarsha Daradmare[1], Bhushan Ugale[2]

[1]Gondwana University, P. G. Student, Department of Computer Science and Engineering,
Ballarpur Institute of Technology, Ballarpur, Maharashtra, India

[2]Gondwana University, Associate Professor, Department of Computer Science and Engineering,
Ballarpur Institute of Technology, Ballarpur, Maharashtra, India

Abstract: ***This review paper presents a study about entity search engine with knowledge mining framework. It describes details about search engine for entity as a lot of information is rapidly growing on the web, extracting this valuable information of real-world entity is most tedious task. Search engine plays a vital role in collecting and understanding this valuable information. In order to reach more accurate resulted information, there is need to develop and utilized unique characteristics of a web. This paper introduces the concept of search engine for entity. It also describes about architecture and iknoweb framework adopted for it. It presents the summary of information about real-world entity.***

Keywords: Entity, Entity disambiguation, Entity extraction, Crawler, Knowledge mining

1. Introduction

The main purpose of developing SEE (Search Engine for Entity) is to present summary of relevant information about the searched entity i.e. person, location, organization etc. instead of navigating through number of web pages. By applying question-answering system in iknoweb framework solves the problem of name disambiguation.

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.Based our entity extraction and search technologies, we have been developing entity search engines to generate summaries of web entities from billions of public web pages and to allow for exploration of their relationships.

Potential applications include information extraction, information retrieval, and knowledge base population. However, this task is challenging due to name variations and entity ambiguity. 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 knowled ge 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.

2. Architecture

An architecture of entity search engine is shown in figure 1.

2.1 A crawler

Crawler is the program which will visits the web pages and fetches the data according to targeted entities. Most of the search engines have such type of program called spider or boat.

Here is the process that a web crawler follows [3]:

- Using the available training data, machine learning model will automatically extracts the information about the entity.
- Extract all the links on that page.
- Follow each of those links to find new pages.
- Extract all the links from all of the new pages found.
- Follow each of those links to find new pages.
- Extract all the links from all of the new pages found.

2.2 Classification

The crawled data is classified into different entity types, such as papers, authors, products, and locations and for each type, a specific entity extractor is built to extract structured entity information from the web data ([1], [2]).

2.3 Aggregation

The data about the same entity will be aggregated.

2.4 Entity Linking and Disambiguation

Once the entity information is extracted and integrated, it is put into the web entity store, and search engines for entity can be constructed based on the structured information in the entity store ([1], [2], [4]).

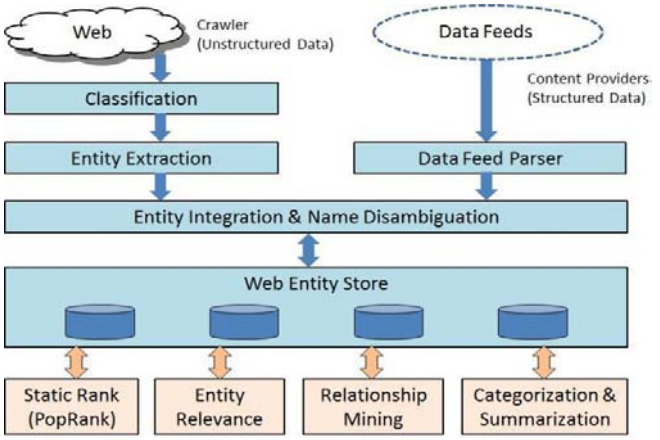

Figure 1: Architecture of Entity Search Engines [1]

The characteristics of entity search engine are as follows ([1], [2]):
Here is the process that a web crawler follows [3]:

- Entity search engines can return a ranked list of entities most relevant for a user query ([1], [9]).
- Entity search engines enable users to explore highly relevant information during searches to discover interesting relationships/facts about the entities associated with their queries.
- Entity search engines detect the popularity of an entity and enable users to browse entities in different categories ranked by their prominence during a given time period.
- Entity search engines rank text blocks from web pages by the likelihood of their being the entity description blocks.

Advanced entity search techniques and frameworks (such as iknoweb) are applied to make search more accurate.

An information about the single entity may distributed over more number of web pages. Entity extraction is the most tedious task because of the name disambiguation problem. Name disambiguation is also the major problem in entity integration. This is the challenging task to improve the search quality of search engine.To overcome the problem of name disambiguation,we propose a novel entity disambiguation framework (called iKnoweb) to add people into the knowledge mining loop and to interactively solve the name disambiguation problem with users ([1], [2], [4]).

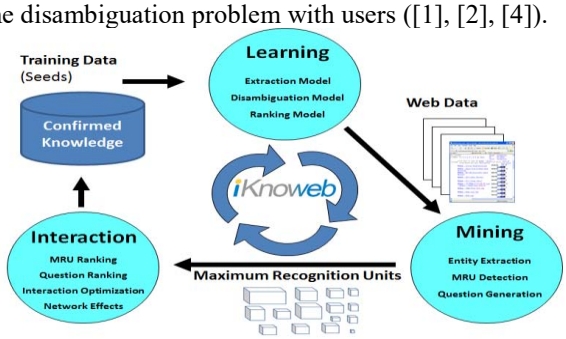

Figure 2: The iknoweb Framework [1]

2.5 Overview of iknoweb

Figure 2 shows the iKnoweb framework and it is explained as follows:

- Using the available training data, machine learning model will automatically extracts the information about the entity.
- The information obtained from extraction process is then merged into MRU's.
- When user enters the query for searching entity, he/she will go through the selecting some MRU or question/answering system to get result more accurately.
- The confirmed knowledge gained through question/answering system is stored into entity store.
- This confirmed knowledge can be used as a seeds for further improvement of entity extraction process.

2.6 Components of iknoweb Framework

Specifically, the iknoweb framework consists of fallowing components [1].

1) Maximum Recognition Unit: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.
2) 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.
3) MRU and Question Re-Ranking: iKnoweb learns from user interactions, and the users will see more and more relevant MRUs and questions after several user interactions.
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.

3. Web Entity Extraction

Web entity extraction is the task of extracting knowledge pieces of an entity from each individual web page within the web corpus and integrating all the pieces of the entity together [1]. Following figure 3 shows the web entity extraction:

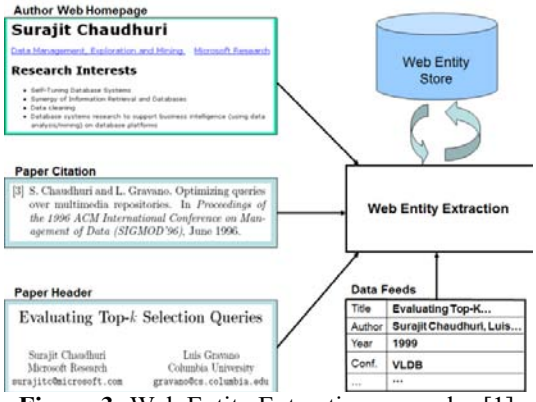

Figure 3: Web Entity Extraction examples [1]

4. 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 lease 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.

5. Conclusion

In search engine for entity which aims and targets to extract valuable and important information as an information unit and solves the problem of name disambiguation with iknoweb framework with more accuracy than traditional searches. By presenting summary of the information about the entity, it removes the necessity to navigate through all the web pages for getting complete view of entity.

6. Acknowledgment

The authors duly acknowledge the support provided by the Management and Principal of BIT Engineering College-Maharashtra by means of providing all the study related facilities.

References

[1] ZaiqingNie, Ji-Rong Wen and Wei-Ying Ma, “Statistical entity extraction from web,” *IEEE*, vol. 100, No.9, Year 2012.
[2] Pinky Paul and Mr. Thomas George, “Entity Search Engine,” IJCSMC, vol. 3, Issue. 2, Feb. 2014, pp. 877–880.
[3] Nileshjain, PriyankaMangal, “An approach to build a web crawler using clustering based k-means algorithm,” journal of global research in computer science, vol. 4, No. 12, Dec. 2013.
[4] ShaikMuneebAhamed, Sd.Afzal Ahmad, and P.Babu, **“**Entity Extraction Using Statistical Methods Using Interactive Knowledge Mining Framework,” IJCSN, ISSN: 2231 – 1882, Vol. 2, Issue. 1, Year 2013.
[5] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, ”Open Information Extraction from theWeb,” IJCAI-07,pp. 2670-2676.
[6] Michal Laclavik, Stefan Dlugolinsky and MarekCiglan.” Discovering Relations by Entity Search in Lightweight Semantic Text Graphs,”Computing and Informatics, Vol. 32, Year 2013, pp. 1001-1028, V 2014-Jul-24.
[7] Renaud Delbru,” Searching Web Data: an Entity Retrieval Model,”Digital Enterprise Research Institute,National University of Ireland, Galway, Sept. 2010.
[8] Manika Nanda,” The Named Entity Recognizer Framework,” IJIRAE, Vol. 1, Issue. 4, May 2014.
[9] AlexandrosKomninosand AviArampatzis, **“**Entity Ranking as a Search Engine Front-End,” IJAIT, Vol. 6, No.1&2,Year2013.Available:http://www.iariajournals.org/internet_technology/

Author Profile

Utkarsha M. Daradmare, received the BE degree in Information Technology from K.D.K College of Engineering in 2010 and 2013 , Now pursuing M-Tech final year in Computer science and engineering from Ballarpur Institute of Technology During the session 2014 to 2016 respectively. This study represents search engine for entity with interactive knowledge mining framework.