

Efficient and enhancement Prediction of Data based on Ontology using AF crawler (Academic-focused crawler)

M.S. VinayagaMoorthy*, J.JennetInfantia[#]

*Professor, A.V.C College of Engineering, Mannampandal – 609 305, Tamilnadu, India

[#]M.E CSE Student, A.V.C College of Engineering, Mannampandal – 609 305, Tamilnadu, India

ssvmoorthy@gmail.com, jenet.rjcraj@gmail.com

ABSTRACT

The Internet is a global system of interconnected computer networks that use the standard Internet protocol suite. The Internet carries an extensive range of information resources and services, such as the inter-linked hypertext documents and applications of the World Wide Web. Web-scale search engines (e.g. Google Image Search, Bing Image Search) mostly rely on surrounding text mining features. It is difficult for them to interpret users. Online advertising is very popular with numerous industries, including the traditional mining service industry where mining service advertisements are effective carriers of mining service information. Users may encounter three major issues heterogeneity, ubiquity, and ambiguity, when searching for information in mining in an Internet. In this paper, we focus the poetic based on AF crawler (Academic-focused crawler), this framework incorporates the technologies of rule based semantic focused crawling and ontology learning, in order to maintain the performance of this crawler. The innovations of this research lie in the design of an unsupervised framework for vocabulary-based ontology learning, and a hybrid algorithm for matching semantically relevant concepts and metadata. Set of visual features which are both effective and efficient in Internet search are designed. Experimental evaluation shows that our approach significantly improves the precision of top ranked ontology and also the user experience.

Keyword: ontology, Entity Resolution, Rule based entity

1.

In many applications, a real-world entity may appear in multiple data sources so that the entity may have quite different descriptions. For example, there are several ways to represent a person's name or a

INTRODUCTION

mailing address. Thus, it is necessary to identify the records referring to the same real-world entity, which is called Entity Resolution (ER). ER is one of the most important problems in data cleaning and

arises in many applications such as information integration and information retrieval. Because of its importance, it has attracted much attention in the literature. Traditional ER approaches obtain a result based on similarity comparison among records, assuming that records referring to the same entity are more similar to each other (compact set property). However, such property may not hold so traditional ER approaches cannot identify records correctly in some cases.

Observation 1. The existence of some attribute-value pairs are useful to identify records.

Observation 2. The nonexistence of some attribute-value pairs are also useful to identify records.

TABLE 1
Paper-Author Records

| | id | name | coauthors | title |
|-------|----------|----------|------------------|----------------|
| e_1 | o_{11} | wei wang | zhang | inferring... |
| | o_{12} | wei wang | duncan, kum, pei | social... |
| | o_{13} | wei wang | cheng, li, kum | measuring... |
| e_2 | o_{21} | wei wang | lin, pei | threshold... |
| | o_{22} | wei wang | lin, hua, pei | ranking... |
| e_3 | o_{31} | wei wang | shi, zhang | picturebook... |
| | o_{32} | wei wang | pei, shi, xu | utility... |

Based on the observations, we are able to develop the following rules to identify records in Table 1.

- R1: $\forall oi$, if $oi[name]$ is “wei wang” and $oi[coauthors]$ includes “kum”, then oi refers to entity e_1 ;
- R2: $\forall oi$, if $oi[name]$ is “wei wang” and $oi[coauthors]$ includes “lin”, then oi refers to entity e_2 ;
- R3: $\forall oi$, if $oi[name]$ is “wei wang” and $oi[coauthors]$ includes “shi”, then oi refers to entity e_3 ;
- R4: $\forall oi$, if $oi[name]$ is “wei wang” and $oi[coauthors]$ includes “zhang” and excludes “shi”, then oi refers to entity e_1 .

This example shows that the disadvantages of the traditional ER methods can be overcome by employing rules generated from the entities’ information.

2. RELATED WORK

In this rule system for entity resolution, called ER rule, is defined in it. (1) The If clause includes constraints on attributes of records, such as “including zhang in coauthors”, and (2) the Then clause indicates the real world entity referred by the records that satisfy the first clause of the rule, such as “refers to entity e_1 ”. Therefore $A \Rightarrow B$ to express their rules. If o satisfies A Then o refers to B ” for ER. It denote the left-hand side and the right-hand side of a rule r as $LHS(r)$ and $RHS(r)$ respectively.

2.1 Syntax

An ER-rule is syntactically defined as $T_1 \wedge \dots \wedge T_m \Rightarrow e$, where $T_i (1 \leq i \leq m)$ is a clause with the form of $(A_i op_i v_i), (v_i op_i A_i), (A_i op_i v_i)$ or $(v_i op_i A_i)$, where A_i is an attribute, v_i is a constant in the domain of A_i and op_i can be any domain- dependent operator defined by users, such as exact match operator $=$, fuzzy

match operator~ [16] for string value, for numeric value, or \in for set value. The clause with form (A_i, op_i, v_i) or (v_i, op_i, A_i) is called positive clause, and the clause with form (A_i, op_i, v_i) or (v_i, op_i, A_i) is called negative clause.

Each ER-rule r can be assigned a weight $w(r)$ in $[0:1]$ to reflect the level of confidence that r is correct. Intuitively, the more records are identified by an ER-rule r , the more possible r is correct. Therefore, given a data set S , we define the weight of each ER-rule r as:

$$W(r) = |S(r)| / |S(RHS(r))|$$

; Where $S(r)$ denotes the records in S that are identified by r and $S(RHS(r))$ denotes the records in S that refer to entity $RHS(r)$.

2.2 Semantics

In the following definitions, let o be a record, S be a data set, r be an ER-rule and R be an ER-rule set. For the convenience of discussion, assume the mapping from each record in S to its actual entity is given. Since an ER-rule does not include disjoint clauses,

Definition 1. o matches the LHS of r if o satisfies all the clauses in $LHS(r)$. o matches the RHS of r if o refers to entity $RHS(r)$.

Definition 2. o satisfies r , denoted by $o \models r$, if o does not match $LHS(r)$ or matches $RHS(r)$

Definition 3. o is identified by r , if o matches both $LHS(r)$ and $RHS(r)$. Note that, if o is identified by r , o must satisfy r . If o satisfies r , o might not be identified by r .

2.3 Properties of ER-Rule Set

Definition 4 (Validity). R is valid for S if each ER-rule in R is valid for S .

Definition 5 (Consistency). R is consistent for S if o matches both $LHS(r_1)$ and $LHS(r_2)$ then $RHS(r_1) = RHS(r_2)$ for all $o \in S$ and $r_1, r_2 \in R$.

Definition 6 (Completeness). R is complete for S if for $\forall o \in S$ such that o matches $LHS(r)$. The following proposition shows that an ER-rule set will contain redundant rules if it is not independent.

Proposition 1. If both r and R are valid for S and $R \models r$, then for any record o in S , if o is identified by r , o must be identified by R .

Proposition 2. If R is valid for S , R is consistent for S .

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Web search engines and some other sites use Web crawling or speeding software to update their web content or indexes of

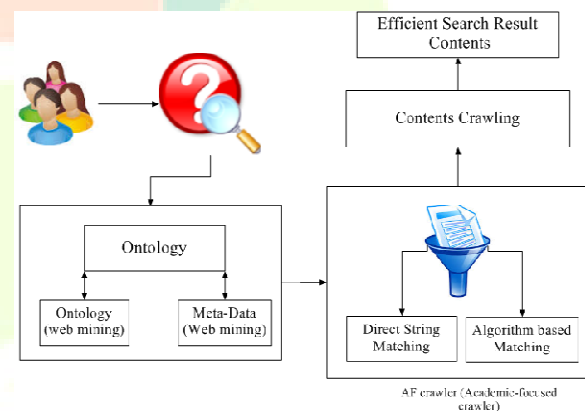
others sites web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping. The pages will be crawled through Implicit Navigation path to lead users from entry pages to thread pages. In this project, the index page is identified. Through which the Index or Thread URL detection is happening. A clear segregation of page identification is happening. Page identification is to classify the page into Index pages or Thread pages. Based on the nature of the page, the pages were segregated. The pages will be navigated with the concept of page flipping. The flow is little bit superficial flow due to automation process. Index Thread Page Flipping (ITF) Regex Learning is utilized to backtrack the threads in the website.

DISADVANTAGES

1. The Mining Service is used to store the manually generated and indexed mining service.
2. The intention of users can be highly diverse.

3. Does not contain the exact results and processing speed kept low.
4. Crawling based on when update the web content or indexes.
5. It visit all the pages. So the performance is low.

4. SYSTEM ARCHITECTURE



5. PROPOSED SYSTEM

In this method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. Scanning the entire web pages through rule based or data mining is used. In this project, we are trying to create an automation engine which will take care of traversing the contents dynamically. Moving towards the

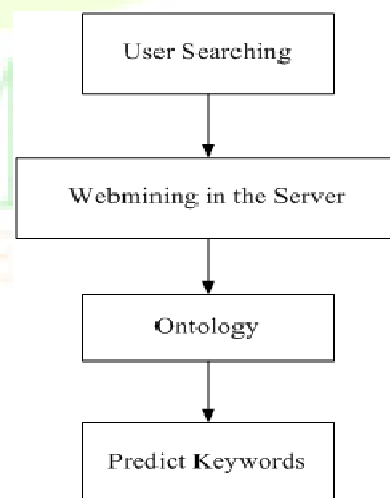
hyperlinks related to the forum and cleanup the related links integrating the missed out data pages in future were considered as the core proposed approaches included in the system. In our proposed system, the features of differential rule based ontology using content extraction instead of an inefficient entire system scanning. Ontology summarizes basic information about categories of data depend on the items (e.g. Rose color, rose flower. This option will enhance the performance of the system very much. Moving towards the hyperlinks associated with the forum and cleanup the connectedlinks emerge Integrating the left out information pages with ontology method predict efficient based on AF crawler (Academic-focused crawler) algorithm.

ADVANTAGES

- 1.This is used to semantically index the generated mining service metadata in the mining service ontology.
- 2.The Mining Service Metadata Base is used to store the automatically generated and indexed mining service metadata.
- 3.High performance and accuracy. Less time consuming.

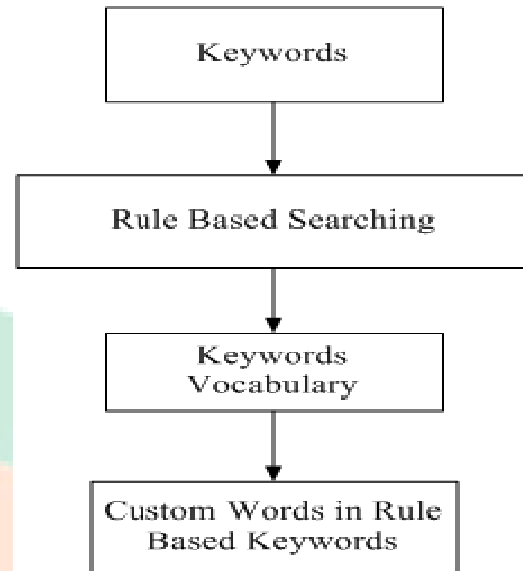
5.1 ONTOLOGY CONSTRUCTION

- In this module the user ontology search in the web mining based on the queries.
- The ontology predict based on the collected keywords in various web mining.
- The keywords are collected from various group websites, based on the query and able to migrate or get the priority.
- A priorities-based method to enrich the decisions made by the user as they complete the tasks. At the beginning of the simulation, each node starts with the same priority. An example will illustrate the significance of this technique. Based on the priority the user is able to view the related keywords information.



5.2 RULE BASED SEARCHING

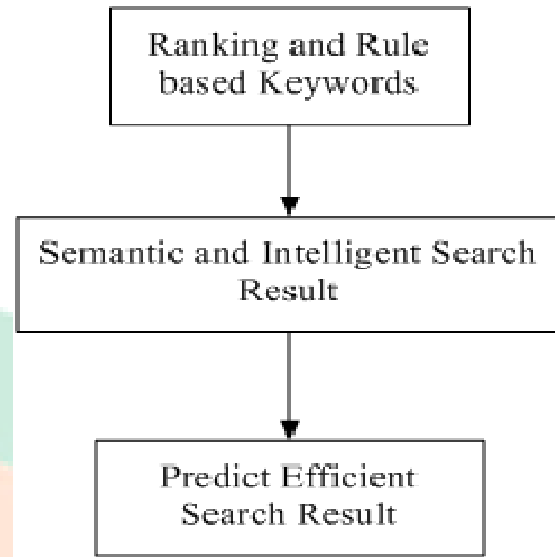
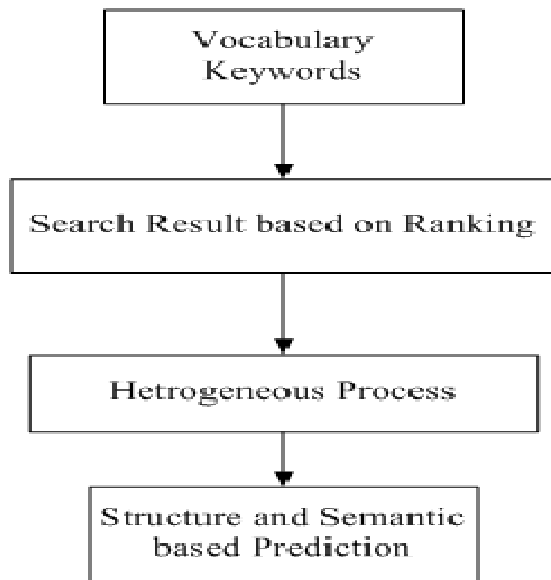
- In this the authorized user is able to search and get results based on rule in the web Browser.
- As a result the User is used to have a relevant search and avoid all irrelevant details and save the time by increasing the performance.
- User's search query is affected by the group membership and previous experience in many other factor. In Proposed Solution, a search engine that understands the user's vocabulary. Extend the reach of resulting by user custom words, with the aid of user profiling.
- If 80% of Actions are caused by a rule based search, by capturing user's Intention, the Search Experience could improve.



5.3 RANKING BASED ON RULE

- The extracting the rule based keywords from a heterogeneous web mining, which are generated from multiple ranking predicted.
- In this problem, clustering of web mining such that the keywords in the same template is required, and thus, the correctness of extracted templates depends on the quality of clustering.
- A web mining technique for rule based keywords ontology searching is proposed in this paper.
- The proposed technique is based on rule, so the structure and semantic similarity keywords and forwarding

similarity in ontology based on the ranking.



6. SCREENSHOT

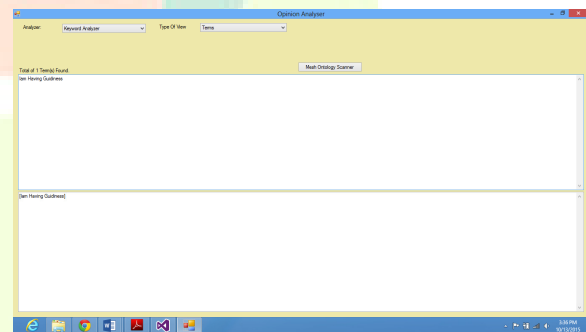


Figure 6.1 opinion Analyser

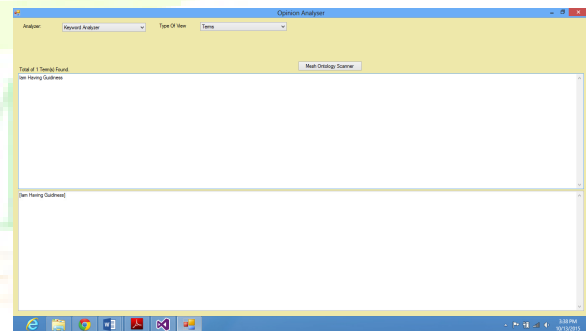


Figure 6.2 KeywordAnalyser

5.4 PREDICT EFFICIENT SEARCH BASED ON AF CRAWLER (ACADEMIC-FOCUSED CRAWLER)

- Mapping web pages or objects to a knowledge base relative to an ontology in general search.
- Make current search engines more semantic and intelligent based on AF crawler (Academic-focused crawler) algorithm.
- Ontology-based query expansion outperformed over keyword based search in rule base kinds of query.
- Predict the efficient search result based on the rule based keywords search.

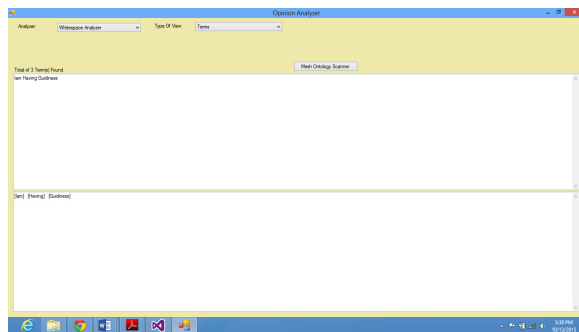


Figure 6.3 WhitespaceAnalyser

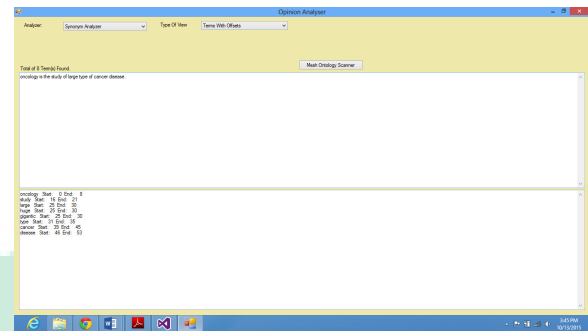


Figure 6.7 Terms with offsets

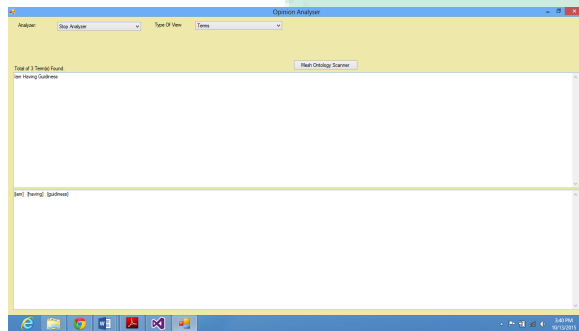


Figure 6.4 StopAnalyser

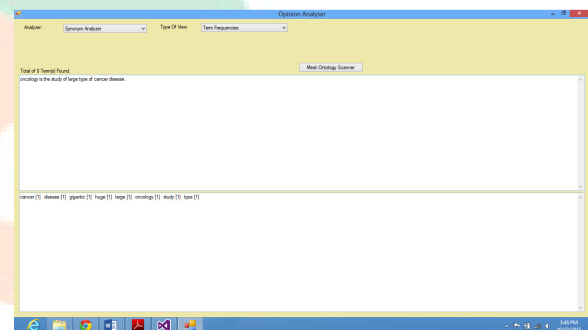


Figure 6.8 Term Frequencies

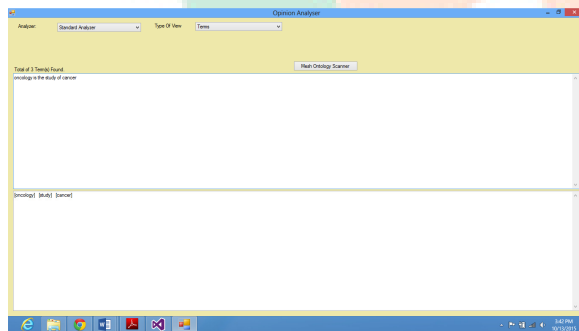


Figure 6.5 StandardAnalyser

Figure 6.9 Mesh ontology Scanner

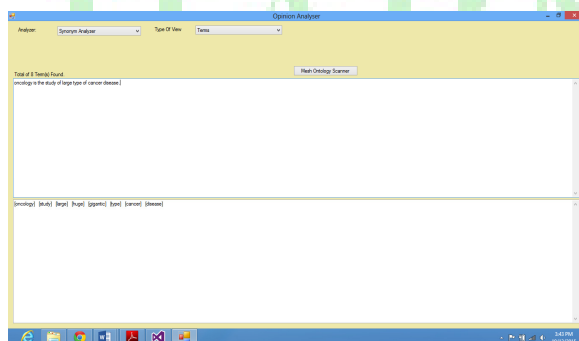


Figure 6.6 Keyword Analyser

7. CONCLUSION

This framework developed a class of ER-rules which are capable to describe the complex matching conditions between records and entities. Based on these rules, we developed an ER algorithm R-ER. The experimentally evaluated our algorithms on real data sets. The experimental results show that our algorithm can achieve a good performance both on efficiency and accuracy.

8.REFERENCES

- 1.P.Christen, “A survey of indexing techniques for scalable record linkage and deduplication,” IEEE Trans. Knowledge Data Eng., vol. 24, no. 9, pp. 1537–1555, Sep. 2012.
2. M. Bilenko, B. Kamath, and R. J. Mooney, “Adaptive blocking: Learning to scale up record linkage,” in Proc. IEEE Int. Conf. Data Mining, 2006, pp. 87–96.
3. S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina, “Entity resolution with iterative blocking,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 219–232.
- 4.O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, “Swoosh: A generic approach to entity resolution,” VLDB J., vol. 18, no. 1, pp. 255–276, 2009.
- 5.C. Xiao, W. Wang, X. Lin, and H. Shang, “Top-k set similarity joins,” in Proc. IEEE Int. Conf. Data Eng., 2009, pp. 916–927.
- 6.A. K. Elmagarmid, G. I. Panagiotis, and S. V. Vassilios, “Duplicate record detection: A survey,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, pp. 1–16, Jan. 2007.



IJARMATE
Your ultra-MATE Research Paper !!!