ISSN:0975-9646

Avnish Rajput et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 4029-4032

Web Mining: A Survey on Various Annotation Techniques

Avnish Rajput
Department of Computer Science
Truba Institute of Science & IT
Bhopal, India

Prof. **Amit Saxena**
Department of Computer Science
Truba Institute of Science & IT
Bhopal,India

Dr. **Manish Manoria**
Director
Truba Institute of Science & IT
Bhopal, India

***Abstract*— Web Mining is a technique of analyzing the web database so that it can be used for a variety of applications such as text analysis and natural language processing and searching. Analyzing semantic text analysis is one of the efficient techniques of analyzing the text using Annotation based searching. Although there are various techniques implemented for the efficient searching of using annotations. Here in this paper a survey and analysis of various annotations based techniques are analyzed and discussed here so that on the basis of their various advantages and limitations a new and efficient technique is implemented in future.**

***Index Terms*—Component, formatting, style, styling, insert.** *(key words)*

I. INTRODUCTION

Web mining seems to be the part of data mining that is used for the extraction of information from the knowledge databases. Web mining is used for the extraction of web contents and hence works on the basis of Web content and web usage data based and web structure based.

Web content mining is the process of extracting useful information from the contents of web documents and texts. Hence on the basis of Content of the data is the collection of a web page which is designed to hold. These data contents consist of various text and images as well as videos. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision.

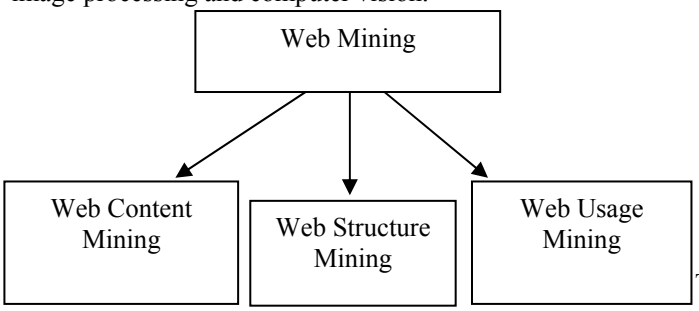

Figure 1 Normal Taxonomy of Web Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications.

In an increasing number of databases have become webs accessible through HTML form-based search interfaces. A recent and emerging trend in data dissemination involves online databases that are hidden behind query forms, thus forming what is referred to as the deep web [1].The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, which is essential for many applications? As compared to the surface web, where the HTML pages are static and data is stored as document files, deep web data is stored in databases. Dynamic HTML pages are generated only after a user submits a query by filling an online form. The emergence of the deep-web is posing many new challenges in data integration. Standard search engines like Google are not able to crawl to these web-sites. At the same time so many domains, manually submitting online queries to numerous queries. A schema is a conceptualization of a domain with a model: Entity-Relationship (ER) model, Object-Oriented (OO) model, XML/XML Schema, or an ontology graph [2]. In each case, there is a natural correspondence between the building blocks of the representation and the notions of element and structure: entities and relationships in ER models; objects and relationships in OO models; elements, sub elements, and IDREFs in XML; and nodes and edges in ontologies [2]

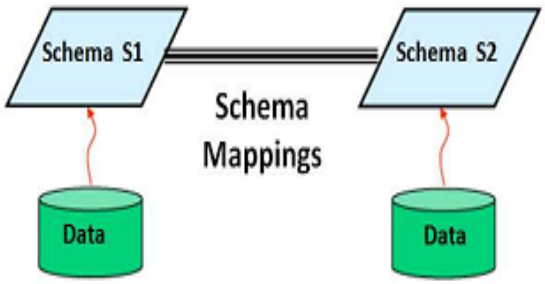

Figure 2: Mappings between two schemas.

II. ANNOTATION AND INFORMATION EXTRACTION

The main focus of this work is to present a concept to make existing web databases accessible and usable for software agents by external semantic annotations.

A. Semantic annotation word set

In a particular part, semantic word can be observing finite. It is distinct as follows.

$$W= \{w_1, w2, w3 \ldots . w_n\}.$$

Among, wj denotes a semantic vocabulary. For a definite domain, we collect many web sites. From some interface pages and result pages of each site, we collect enough many vocabularies representing data concept of this domain found on that domain knowledge, it needs to make vocabulary add-on and modification for consistency of semantic explanation in different sites.

B. Semantic annotation

Semantic annotation is to find an explicit semantic vocabulary for every data unit in result pages, for making these data understandable and process able for computers. It's characterized as follows:

Suppose that there is a vocabulary set, $W = \{w_1, w_2, \ldots. w_n\}$, a group of attribute values to be annotated, $V = \{v_1, v_2, \ldots\ldots, v_m\}$. Semantic annotation is that, for each $v_i \in V$, we can find an appropriate w_j, which can comparatively accurately describe the semantic of v_i. Finally it constructs a set of mapping pairs like $\{(vi, wj) \mid v_i \in V, w_j \in W, w_j$ is the semantic instruction of $v_i\}$.

It's noted that each vocabulary name in W should spontaneously and appropriately reflect the comfortable they describe. It can be a single vocabulary, word abbreviation, or words grouping.

C. Semantic Web and Its Agent

The main focus of this work is to present a concept to make existing web databases accessible and usable for software agents by external semantic annotations. The whole setup therefore includes a semantic agent for information extraction, the concept for the annotation of web pages and ontologies containing the vocabulary.

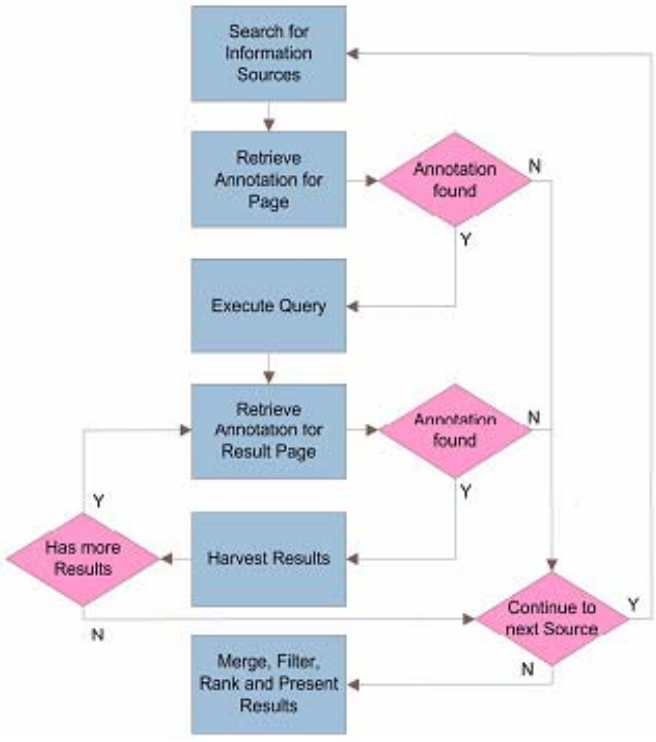

Figure. 3. Search Process Flow-Chart

Annotation of WebPages

One of the key principles in modern software engineering is the separation of concerns. This concept described in [3], states that by separating the basic algorithm from special purpose concerns makes each of the parts easier to write, maintain and test. For example, the separation of data and layout of a webpage into a style-sheet [4] and the html content page has become an engineering practice: It not only allows changing the layout and content independently, but it also allows different people with different skill-sets and training to work on the webpage independently, e.g. a developer and a designer. This paradigm is also at the core of other state of the art and emerging programming paradigms like Service-Oriented Architecture [5] or Aspect-Oriented programming [6].

In the context auf semantic annotations, we propose to follow this scheme and separate the page and the annotations. This would hold many aspired properties: The original page would not be cluttered with additional tags, a separate tool could be used for the annotation process, and a page could also be annotated by a third party.

III. Present Scenarios Of Annotation Research For The Semantic Web

This section introduces current web semantic annotation research. First, it surveys interactive annotation systems, and then it surveys automatic annotation systems.

Interactive Annotation: Interactive annotation lets humans interact through machine interfaces to annotate documents. In general, manual annotation incurs the problems of inconsistency, error-proneness, and scalability. Nevertheless, interactive annotation systems are still valuable for web semantic annotation. Compared to automatic systems, interactive annotation systems are easily implemented and can be used to accomplish small-scale annotation tasks and do experiments. Interactive annotation systems can also help people build sample annotated corpora to do performance evaluations for automated annotation systems.

Annotations of Annotea are restricted to attribute instances. A user may decide to use complex RDF descriptions instead of simple strings for filling a template. However, Amaya provides no further help for filling in syntactically correct statements with proper references. Another problem with Annotea is that it does not support information extraction nor is it linked to an ontology server. Hence, it is difficult for machines to process Annotea annotations. Because the annotations must be done by humans, Annotea is not suitable for large-scale semantic annotation.

Annotea uses an RDF-based annotation schema for describing annotations as metadata and XPointer for locating the annotations in the annotated document. Annotea relies on an RDF schema as a kind of template that is filled by the annotator. For instance, Annotea users may use a schema from Dublin Core and fill the author-slot of a particular document with a name. Annotea stores the annotation metadata locally or in one or more annotation servers and is presented to the user by a client capable of understanding this metadata and capable of

interacting with an annotation server with the HTTP service protocol. When users retrieve documents, they can also load the annotations attached to them from a selected annotation server or several servers and see what annotations their peers have provided. Therefore, Annotea provides an open RDF infrastructure for shared web annotations.

IV. LITERATURE REVIEW

In this paper [9] they considers query planning and optimization problem .in the circumstance of a for deep web databases with dependencies system. The system design and implement a dynamic query planner to generate the top K query plans based on the user query and database dependencies. This approach provides different plans when the most efficient one is not feasible due to the non-availability of a database to support a very simple and easy to use query interface, where each query comprises a query key term and a set of query target terms that the user is concentration on them. The query key term is a name, and the query target terms capture the properties or the kind of information that is desired for this name. Here they do not need the user to provide us with a formal predicate-like query. In the circumstance of such a system, here they develop a dynamic query planner to generate an efficient query order based on the deep web database dependencies. They join together our query planner with a deep web mining tool SNPMiner [8] to build up a domain specific deep web mining system. Their query planner is able to select the top K query plans. This makes sure that when the most efficient query plan is not reasonable, for examples, because a database is not available, there are other plans possible. It shows that their algorithm can accomplish optimal results for a large amount of queries, and furthermore, their system has very good scalability.

Here they developed [11] a new method that can include as many possibly relevant objects as possible to facilitate approximate search through the Web database. This method, based on the Euclidean distance measurement, finds objects with similar or closely properties and then identifies objects that share closest properties. This paper presents the approximate search method based on the Euclidian distance concept. The purpose of their research is to find the most similar objects to the users‟ preferences in the Web-based applications. The users‟ queries recognize some attributes of the desired objects. This algorithm works out on Euclidian distances of the target object and the surrounding data. From the given distance threshold, the algorithm can produce the most appropriate objects to the user concentration. They implement the proposed method with the Erlang programming language and test the program with the telecommunication Web database. The experimental results reveal that the program can display similar objects within the short period of time. The proposed method can thus be applied to other Web applications.

In this paper [7] the proposed method is made for a Multi-domain record matching process which comprises an algorithm called N-Staged SVM that helps to divide the duplicate and non-duplicate records based on the classifiers. A latest move toward approach is called N-Staged SVM are made used which ensures for duplicates in multiple domain concurrently and make available of enormous results which are free from duplicates. The general SVM classifier classifies the duplicate and non-duplicate data from the web databases for single domain.

The N-Staged SVM checks for a part of two data sets and separates the duplicate and non-duplicate, the effected non-duplicate used as the input for other evaluation of data. This is how the N-Staged SVM is made used in the multi-domain Record matching for user query results. This process is replicated for multiple domains by constructing hyper-planes for each. Consequently the outcome produced will be proficient and more reliable outcomes are provided for the user query are extracted with non-duplicate data which are precise result. Hence by the above process, more number of results are gained by the exploring to multiple domains at a extend and those results are reliable to the user and in this way N-Staged SVM defeats the UDD algorithm.

Assigning semantic labels to the data units extracted from result pages is a difficult task. In this paper, they are analyzing the features of interface page and result page. As well, experimental results show that their method has a good annotation effect. After this, they need to think more concerning the uniformity of annotation results of different sites. Furthermore, our heuristic information-based method a short times cannot assurance the annotation comprehensiveness. So it needs us to investigate more useful heuristic information to improve the annotation completeness.

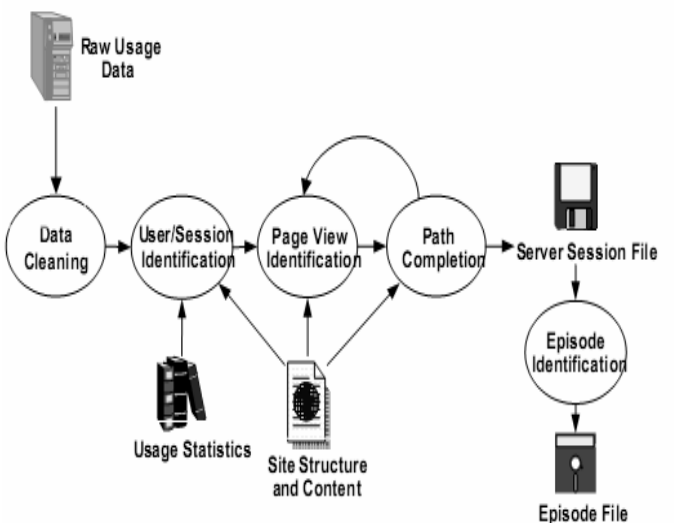

Figure 4 Preprocessing of Web Usage Data

In this paper, [10] they proposed that establishing and exploiting relationships between web search results and structured entity databases significantly enhances the effectiveness of search on these structured databases. A new integrated search architecture which effectively influences existing search engine components in order to efficiently implement the integrated entity search functionality. They establish and make use of the relationships between web search results and the items in structured databases to identify the appropriate structured

data items for a much wider range of queries. Here the given architecture influences existing search engine components to implement this functionality at very low operating cost. Specifically, they demonstrate their techniques the quality and efficiency of our techniques through an extensive and add very little space and time overheads to current search engines while returning high quality results from the given structured databases experimental study.

This paper proposes a heuristics-based semantic annotation method. According to the distinctive analysis of interface page and result page, this paper reviews some heuristic information. This technique uses this heuristic information in revolve to analyze the data to be annotated, for recognizing a semantic vocabulary for each data unit. At last, it performs a semantic annotation experiment on the Deep Web data of various regions in the UIUC standard dataset. The experimental result specifies that their approach is extremely efficient. Compared with Ontology-based annotation (OBA) method, their method has a better performance.

This paper proposes [12] a heuristic information-based annotation method. First for a certain domain, we collect many semantic words on behalf of data concepts of this domain as semantic annotation word set. All the way through analyzing the distinctiveness of interface page and result page data, we sum up quite a lot of pieces of heuristic information. One by one by means of them to analyze data to be striking, to end with finds a semantic vocabulary for each data unit. In actual fact, their method is similar to literature [13]. But compared with basic annotators in [13], their proposed method new heuristic information that is, the neighboring position feature of attributes and the word count of attribute value. This latest position will contribute to the annotation entirety, thus they remember of the annotation will be enhanced.

CONCLUSION

The various techniques implemented for the annotation search based on semantic similarity and their various advantages and applications are analyzed and discussed here. Hence on the basis of their various issues or shortcomings in the existing technique a new and efficient technique is implemented which is more efficient as compared to the other techniques.

REFERENCES

[1] Bin He, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang. Accessing the deep web: A survey. Communications of ACM, 50:94–101, 2007.

[2] Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. The Very Large Data Bases-VLDB Journal, 10(4):334–350.

[3] W.L. Hürsch and C.V. Lopes, "Separation of Concerns". Technical Report NU-CCS-95-03, Northeastern University, 1995

[4] Cascading style sheets: www.w3.org/Style/CSS.

[5] A. Arsanjani et al, "S3: A Service-Oriented Reference Architecture". IT Professional, Vol. 9, Issue 3, pp. 10-17, 2007

[6] T. Elrad, R.E. Filman and A. Bader, "Aspect-oriented programming: Introduction", Communications of the ACM, Vol. 44, Issue 10, pp. 28-32, 2001

[7]P.Kowsiga, T.Mohanraj, "Multi-Domain Record Matching over Query Results from Multiple Web Databases", International Journal of Scientific & Engineering Research, Volume 3, Issue 5, May-2012, ISSN 2229-5518.

[8] Fan Wang, Gagan Agrawal, Ruoming Jin, and Helen Piontkivska. Snpminer: A domain-specific deep web mining tool. In Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, pages 192–199, 2007.

[9] Fan Wang, Gagan Agrawal, and Ruoming Jin, "Query Planning for Searching Inter-Dependent Deep-web Databases"2008.

[11] Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti Arnd Christian König, Dong Xin, "Exploiting Web Search Engines to Search Structured Databases" ACM 978-1-60558-487-4/09/04, 2009.

[12] Sarawuth Sonnum, Somtida Thaithieng, Sittichai Ano, Krerkchai Kusolchu, and Nittaya Kerdprasop, "Approximate Web Database Search Based on Euclidean Distance Measurement," Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol-I, IMECS 2011,March16-18,2011.

[13] Yong FENG,Wei LU, "Heuristics-based Semantic Annotation for Deep Web Query Results", Journal of Computational Information Systems 9: 14 (2013) 5685-5692.Available at http://www.Jofcis.com