A Survey on Entity mining and applications

Hemali J Damania[1], Khusboo Sawant[1], Kuntal Barua[1]

[1]Jaggadguru Dattatray College of Technology, Indore, M.P., India

Abstract

the data mining and it's approaches are helpful to identify the pattern of data. These techniques enable us to make and find essential relationship among the different kinds of data. In this generation where the data is frequently increasing in internet based applications. It is a need to analyze and compare the similar kinds of services offered and similar kind of product available in online service provider's products. In this context the method of entity comparison is a need to be developed. This paper provides an understanding for selection of entities among multiple data source objects, compare them and represent them in a common format. Therefore in this paper a review on traditional available techniques is provided first, in addition of that a theoretical model is represented which is used for future implementation and their strength analysis.

Keywords: *data mining, entity mining, entity mining applications, data representation, entity comparison*

I. INTRODUCTION

Data mining is the process of discovering knowledge, identify patterns and establish relationships from large amounts of data. The goal of data mining is to extract information from a data set and change it into understandable form to use it later. It is the automatic or semi-automatic analysis of large data to extract previously unknown and interesting patterns of data records. Today most organizations like retail, financial, communication and marketing primarily uses data mining to meet their needs.

Data mining enables companies to determine relationships among various factors like internal factors such as price, product positioning or staff skills and external factors such as economic indicators, competition and customer demographics. It enables the companies to determine the impact on sales, customer satisfaction and corporate profits. It also enables the companies to shortlist data into summary information to view detail transactional data [1].

For performing the data mining both kinds of data namely structured and unstructured can be used. In this work the structured data is targeted to be analyses. Structured data is a data that resides in a fixed field within a record or file which can be data contained in relational database or spreadsheets. Structured data has an advantage of being easily entered, stored, queried and analyzed. Structured data depends on how business data will be stored, processed and accessed which includes defining what fields of data will be stored and how that data will be stored, data type and any restrictions on the data input. Structured data thus needs a data model which describes the structure how data is stored and accessed [2].

An entity is something that exists as itself, as subject or as an object actually or potentially, concretely or abstractly, physically or not. Entity need not be of material existence. An entity can be real world object, either animate or inanimate that can be easily identifiable. Attributes are characteristics of the entities. An entity set is a collection of similar types of entities. There is a domain or a range of values that are assigned to attributes. Thus each attribute has a value. Entities can be represented by their attributes.Sometimes there is a need to combine more than one attributes. For example, in two different data sources where similar products description is available and both the data sources have their own product definitions. Here we need to find similar attributes by which the database queries results less duplicate values. A method is required which maps and aggregates different attributes in common format [3].

The proposed technique can be used for applications like Trivago which compares prices of various hotels from more than 200 booking sites. It provides comparison of hotel prices on larger scale. The proposed technique can provide more refined search results.

II. LITERATURE SURVEY

This section provides analysis of the recently made contributions and the research work for making entities comparable.

GuXuet al [4] addresses Named Entity Mining, in which knowledge is mined about named entities such as movies, games and books from a huge amount of data. It characterizes each named entity by its associated queries and URLs in the click through data. It uses the topic model to resolve ambiguities of named entity classes by representing the classes as topics.

Comparison of different things helps in human decision making process. In this paper, ***ShrutikaNarayane et al [5]*** review the background and state of the art of comparable entity data mining based on comparative questions. With each phase a related background is provided and discussed the technical challenges and review current research on the techniques used in that phase.

The integration of the classical web with the emerging web of data is a challenging vision. In this paper ***PavlosFafalios et al [6].***Focuses on an integration approach during searching which aims at enriching the response of non-semantic search system with semantic information i.e. Linked Open Data(LOD) and exploring the outcome for providing an overview of the related LOD. A Linked Analysis based method is used for ranking the more semantic information related to the search results and deriving and showing top-K semantic graphs.

Named Entity Extraction is the process of identifying entities in texts and linking them to related semantic resources. In this paper ***PavlosFafalios et al [7]*** show how we can exploit semantic information at real time for configuring a named entity extraction system. X-Link, a fully configurable named entity mining extraction tool is also presented. It allows the user to easily define the categories of entities that are interesting for the application at hand by exploiting one or more semantic knowledge base. The user is also able to update a category and specify how to semantically link and enrich the identified entities.

The "big data" era is characterized by an explosion of information in the form of digital data collections, ranging from scientific knowledge to social media, news and everyone's daily life. In this tutorial ***Jaiwan Han et al [8]*** summarizes the closely related literature in database system, data mining, web information extraction, information retrieval, natural language processing, overview a spectrum of data driven methods that extract and infer such latent structures from an interdisciplinary point of view and demonstrates how these structures support entity discovery and management, data understanding and some new database applications.

III. OBJECTIVES

The main goal of the proposed work is to find a method by which the different structures of the web databases are combined together for finding the user query relevance comparative study among the available attributes. Therefore the following key works are included as objectives of the work:

i. To investigate the techniques of entity mining

Different entity mining techniques and applications are investigated by which the entities are grouped and can be made comparable for different application areas.

ii. To design an approach for structured data processing and attributes comparison

There are different data formats available for evaluation and query making among which structured data formats are used for number web based applications. Thus the structured data formats are investigated and evaluated during the study.

iii. To implement the approach for real world case study

A new technique of data evaluation and entity comparison is developed and implemented. And additionally a case study is performed to find the effectiveness of entity mining approaches in real world applications.

iv. To perform comparative performance study among classical approach and proposed technique

The performance of the proposed approach is investigated using the parameters precision, recall, time complexity and space complexity. The comparative study of the performance parameters is also conducted for finding the improvement of the proposed technique over traditional technique.

IV. PROPOSED WORK

Now a day's number of products and service providers are available. To compare these services and products there is a need of evaluations of previous reviews and understanding of the products and services. There is also a need of some parameters for selection of appropriate service or product by which the comparison can be made between two objects. A system is required which can filter out the attributes on which comparison can be done [9][10].

To filter out these attributes from different databases is a tedious work. The same attribute can be named differently in different databases. And the attributes with similar mean are sometimes combined and sometimes defined separately. The attributes needs to be checked semantically and syntactically to find the matched attributes. And finally the matched attributes and their data is combined in the common structure of database [11][12].

The proposed work focuses on the solution for efficient and accurate data retrieval. The presented effort of named entity mining is dedicated to improve the information retrieval for data source aggregation and query time performance improvements.There exists number of service providers for the same service. They offer the similar product or services with different quality of service parameters. Therefore a technique is required by which the different quality of products and services are compared with each other. Traditional query processing is performed for finding the optimum results from the database. But to reduce the data redundancy and to improvise the quality of search results a new technique is to be developed.

This proposed technique involves the semantic and syntactic similarities for finding more effective outcomes from the databases. Additionally, this technique also deals with huge amount of data in same place, therefore data management cost and efforts are also effectively reduced.

In this presented work two key issues are targeted to stimulate the problem and obtain the appropriate solution.

- Semantic issues deals with entity or attribute which are to be combined for example data sources having similar name of attribute or different for same attribute and single data attribute defined with different datatypes.
- Syntactic issues deals with different attribute having similar mean when they are combined and checks if they are syntactically correct or not. For example full name can be defined as first name and last name attributes.

The proposed solution works in the following manner:

- First the system needs to prepare or manage multiple connections with different service provider's databases. Initially the system accepts the connection with all the available data sources. The data sources can be any kind of product information or any class of services offered through different kind of vendors
- Provision is needed for selection of data tables or attributes for further processing. Targeted data tables or schemas which are required to compare are extracted from the data sources.
- Once the data tables or schemas are extracted the attributes are analyzed semantically. A separate list of similar semantic data or synonym is prepared which is used to map the given attribute into possible semantic manner. For example the cost attribute can be addressed as price, amount or charges.
- In syntactic analysis of attribute different combinations of attributes are prepared and compared to the target attribute list. The most common attributes are separated.
- The system now compares the databases and attributes a single data source at a time to find the matched attributes. These attributes are now compared with other available attribute list. The matched attributes are separated and preserved at both the stages of evaluation.
- The outcomes of the semantic and syntactic analysis are used to identify the matched attributes. Using the obtained information from both the analysis the matched attributes and their data combined for finding the common structure of database.
- The search outcomes are generated on the basis of multiple data sources organized in a common format. After performing the search operation the search results are collected through the data structure and according to the query most relevant search results are produced.
- During the generation of outcomes for queried data the performance of the system is also computed in terms of their precision, recall and f-measures.

V. CONCLUSION

The proposed technique considers a number of data sources with different locations and different attributes. First both the database attributes are compared semantically and syntactically for finding similar attributes. Finally the data is transformed into a

common format for effective query satisfaction. Therefore the proposed technique generates the semantic outcomes in less redundant manner.

1. The current system is only able to distinguish the similar attributes from the similar categories of products. In future the technique can be modified for finding similar schema databases.
2. The system is demonstrated with the two party data sources. In future the technique can be extended for multiple data sources.

REFERENCES

[1] Hu, Ye, et al. "Resource provisioning for cloud computing", Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research, IBM Corp., 2009.
[2] Data Mining: What is Data Mining? Available online at: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[3] Structured data : Available online at: https://developers.google.com/search/docs/guides/intro-structured-data
[4] http://whatis.techtarget.com/definition/structured-data
[5] "ER Model : Basic Concepts", online available at: http://www.idc-online.com/technical_references/pdfs/information_technology/Er_Model.pdf
[6] GuXu, Shuang-Hong Yang, Hang Li, "Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation", KDD'09, June 28–July 1, 2009, Paris, France. 2009 ACM.
[7] ShrutikaNarayane, SudiptaGiri, "A Review on Comparable Entity Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 12, December 2014.
[8] PavlosFafalios and YannisTzitzikas, "Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time", 2014 IEEE International Conference on Semantic Computing (ICSC)
[9] PavlosFafalios, ManolisBaritakis and YannisTzitzikas, "Configuring Named Entity Extraction through Real-Time Exploitation of Linked Data", WIMS'14, June 02-04 2014, Thessaloniki, Greece Copyright 2014 ACM.
[10] JiaweiHan,Chi Wang, "Mining Latent Entity Structures from Massive Unstructured and Interconnected Data", SIGMOD'14, June 22–27, 2014, Snowbird, UT, 2014 ACM
[11] PavlosFafalios, and YannisTzitzikas, "Web Searching with Entity Mining at Query Time", 5th Information Retrieval Facility Conference, IRF 2012, Vienna, July 2012.
[12] ZhengXu, XiangfengLuo, Shunxiang Zhang, Xiao Wei, Lin Mei, Chuanping Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines", Future Generation Computer Systems,.
[13] PanikalaMadhavi, S. Vijayalaxmi, "Comparable Entity Mining from comparative Queries", Aurora's International Journal of Computing, 2015, Vol. 2. Issue 1, January-June 2015
[14] MaksimTkachenko, Hady W. Lauw, "Generative Modeling of Entity Comparisons in Text", CIKM'14, November 3–7, 2014, Shanghai, China, ACM.