A Novel Method For Searching Metadata Using Classification Techniques

[1] Asaad Sabah Hadi , [2]Abbas M. Al-Bakry

[1],Babylon Uni.,Software Dep.,Iraq,Babylon,P.O.Box.4,asaadsabah@{yahoo.com,*uobabylon.edu.iq*}

[2],Babylon Uni.,Software Dep.,Iraq,Babylon,P.O.Box.4,abbasmoh67@{yahoo.com,*uobabylon.edu.iq*}

Abstract

In the world of Information Technology Revolution , the user search all the time for an automation techniques to complete his requirements in any field, therefore many researches appears from time to time that try to give the users a new techniques that facilitate their work. In the field of Web Search, the user need to search for the information in an easy way, but the search engines (like Google,Yahoo, ..,etc) give the user more results so that the user need to check these results to take his information and it is a time consuming task . The database for semantic web is RDF and SPARQL are used to search for the information inside RDF but it need from the user to know the information inside RDF before any search,and that is a difficult task for many users. We Suggest a method that take a data base in RDF format and translate it into a vector format then collect all the databases into one file then we use a machine learning algorithm (classification) to learn all the information inside that file. The user then search for the information inside these vectors in an easy way.

keywords : *Semantic Web, Linked Data, Linked Open Data (LOD), RDF, RDFS, SPARQL,Matrix, Vectorisation, Machine Learning ,and Classification.*

1. Introduction

Most search engines search for keywords to answer queries received from users[2][3]. Search engines are typically used to search web pages for the required information. However, they also filter pages from searching unnecessary ones[2][16][15]. They are useful for answering topic wise queries efficiently and effectively using state-of art algorithms. The challenge is to answer intelligent queries received from users based on information available inside the web pages. Therefore, the main focus of search engines is to answer user queries accurately[2][7][10].

Several approaches already exist that try to achieve this. For example keywords based searches are used to provide results from blogs (if available) , discussion boards and Web Pages but it lack a syntactic and semantic understanding of the content they are processing[2][4][6][11]. New approaches are needed that include domain knowledge and structured information to help search engines answer intelligent queries. The layered model of the Semantic Web provides one possible solution to this problem by providing tools and technologies to enable machine readable semantics in current web content[11][12].

The primary goal of the Linked Data is to make WWW useful for sharing and interlinking data at very details level[9][15].Instead of Web pages , the primary data model of the Linked Data is RDF (Resource Description Framework) which is a W3c standard for representing any information. The main formula for the RDF is : Subject - Predicate - Object. To retrieve information stored in RDF ,you need to use SPARQL (SPARQL Protocol And RDF Query Language)[5]. SPARQL is a W3C recommended query language for RDF[5]. To query linked data on the Web today , users must know exactly what the dataset that contain the information that he looking for and what data models describe these datasets before using this information to create structured queries[16][5]. Also the user must know the syntax of structured query languages such as SPARQL[13]. SPARQL enables values to be pulled from both structured and semi-structured data; it can explore data by querying unknown relationships; complex joins, of disparate databases, are able to be performed, in a single and simple query, and RDF data can be transformed from one vocabulary to another[8]. Many users do not have any experience in SPARQL, so that they can't access the information stored in RDF in a correct manner[13]. Therefore our contribution deals with converting the RDF into a MATRIX and using that matrix to access the information stored in RDF, and using Machine Learning Algorithm to learn that information.

The authors in [3] introduce the idea of mapping search engine to DBpedia, they propose and compare various methods for this idea by mixing machine learning techniques with information retrieval. they take a supervised machine learning methods to select the concept of the user query. This concept are obtained from an ontology and may be used to provide contextual information, related concept or navigational suggestion to the user submitting the query.

In [5] the Authors focuses on the role and usage of SPARQL. They also make a Comparison of SPARQL With SQL and present an execution analysis of SPARQL with some tools and they used two tools for execution Query (Jena & Twinkle).They show that querying SPARQL query using TWINKLE tool is more convenient than with Jena ARQ processor, and they give some important advantage of TWINKLE over Jena. They propose as a future work to design some methodologies which can give an optimization concept for efficient data retrieval from SPARQL.

In [13] the authors explain and investigate the main challenges in constructing a query for a linked data and make analysis of the existing approaches and trends.

In [12] the authors describe the extraction of the DBpedia knowledge base and give an overview of application that facilitate the Web of Data around DBpedia. The DBpedia project leverages the gigantic source of knowledge by extracting structured information from Wikipedia and making this information accessible on the Web.

In [11] the authors give a survey that analyze the Semantic Web and the Web Mining and explain the Layers of Semantic Web. They have studied the combination of these two research area. They discussed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the Web. Also they show how to construction of the Semantic Web can use of Web Mining Techniques.

The authors in [6] describes the SIOC (Semantically Interlinked Online Communities) , which is a Semantic Web research project that aims to describe online communities on the Social Web, and also describe how SIOC and the Semantic Web can enable linking and reuse scenarios of data from Web 2.0 community sites.

In [10] the authors proposed a rule-based method for learning ontology instances from text that helps domain experts with the ontology population process. They define a lexico-semantic pattern language that make use o semantic information in addition to lexical and syntactical information that present in lexico-syntactic rules.

2.Linked Data

The classic World Wide Web is built upon the idea of setting hyperlinks between Web documents. These hyperlinks are the basis for navigating and crawling the Web; they integrate all Web documents into a single global information space. In classic search engines like Google, Yahoo, .. ,etc, they have started to provide access to their databases through Web APIs. As most Web APIs do not assign globally unique identifiers to data items, it is generally not possible to set hyperlinks between data items provided by different APIs. Web APIs therefore slice the Web into separate data silos, and developers must choose specific data sources for their application. They cannot implement applications against all the data available on the Web.

To overcome this fragmentation problem, Tim Berners-Lee outlined a set of best practices for publishing and connecting structured data on the Web: the Linked Data principles. In summary, the Linked Data principles provide guidelines on how to use standardized Web technologies to set data-level links between data from different sources to solve the fragmentation problem[2][4][8][12].

Tim Berners-Lee find the principles of Linked Data that overcome some of the problems of the previous web. the main idea of this work is focus on the data. The data is the main important thing in the web because all the users search about a data in several search engine and the governments and company put their information on the web, So the main idea of Linked data is to link these data into a general Graph that make these data more accessible to the users . By using Linked Data principles, sometimes we cannot find a person but we know one of his friends or know his car number , therefore we can access that person from these information which is linked to his profile[2].

The Linked Open data (LOD) extend the web by publishing various data sets and put a link between these data. The resulting data is termed the Linking Open Data cloud, and it gives the key for realizing the Semantic Web[3].

The Data may be Structured or Unstructured. We may find several problem when we deals with the unstructured data resources like textual document or queries that submitted to the search engines. we can Enrich this unstructured data resources by mapping its contents into structured knowledge repositories like LOD cloud[3].The DBpedia , which is the major linking hub of the LOD cloud, can help us by linking it with our unstructured data (search engine queries).This semantic mapping give the user more activity to acquire contextual information, suggesting another related concepts or terms that may be helpful for the search, and providing valuable navigational suggestions [3].

Today most knowledge base, which plays important role in enhancing the intelligence of web and enterprise search, covers a specific domain and are very expensive to keep it up to date. Wikipedia, at the same time, has grown into one of the central knowledge source of mankind that extracted from DBpedia[2][3][12].

There are several advantage of DBpedia over existing knowledge base:- it covers many domains, it represent real community agreement, it automatically evolve as Wikipedia changes , it truly multilingual, and it is accessible on the Web [4][5].

3.Machine Learning

Machine Learning is a branch of Computer Science that concerned with designing systems that can learn from the provided input. Usually the systems are designed to use this learned knowledge to better process similar input in the future. Machine learning can be considered as a subfield of Artificial Intelligence[14][1]. Over the past two decades Machine Learning has become one of the mainstays of information technology and with that, a rather central, albeit usually hidden, part of our life[14,15]. The Increasing amount of data that available in the web give the reasons to develop a smart data analysis Algorithm[15]. The use of computer algorithms and visualization techniques are considered fundamental to support the analysis of datasets, commonly referred to as Big Data[16]. Machine learning is actively being used today, perhaps in many times you'll encounter machine learning: You need to send a Card to your friend, and you search for a best card[1]. The search engine shows you the 20 relevant link, then you click the second one, the search engine learn from this[14]. Another example is the Spam Filtering: the spam filter catches unsolicited ads for pharmaceuticals and places them in the Spam folder[14].Machine learning lies at the intersection of computer science, engineering, and statistics[14]. Any field that needs to interpret and act on data can benefit from machine learning techniques[14,15]. It is useful to characterize learning problems according to the type of data they use.

Vectors are the most basic entity in our work. For instance, a life insurance company might be interesting in obtaining the vector of variables (blood pressure, heard rate, height , weight, cholesterol, level, smoker, gender) to infer the life expectancy of a potential customer. One of the challenges in dealing with vectors is that the scales and units of different coordinates may vary widely, so we need to normalize the data[14,15].

Matrices are a convenient means of representing pairwise relationships. In collaborative filtering applications the rows of the matrix may represent users whereas the columns correspond to products.

One task in Machine Learning is Classification(Statistical Classification), more precisely a supervised statistical classification[14]. A supervised learning system that performs classification is known as a learner or classifier[15]. The classifier is first fed training data in which each item is already labeled with the correct label or class. This data is used to train the learning algorithm, which creates models that can then be used to label/classify similar data. So that, The main data are divided into : training data and testing data .There are many types of Classification Algorithm like : kNN, Decision Tree, naive Bayes, Logistic regression, SVM (Support Vector Machine)[14,15].

4. Method Architecture

The suggested Algorithm has been designed to extract information from semantic metadata. We used is RDF(Resource Description Framework), and transform it into a Matrix representation to facilitate probabilistic search. Our suggested algorithm is work in any domain, because we work on RDF in its general form (rdf+xml, n-triples). This is the key feature of our work that alleviate the need to fully understand the domain before making a query for information. Our method treat the search of data as a classification problem, because we describe the features of the query and not the query itself.

Figure 1 below give a brief steps for our suggested algorithm:-

Step1 : Read new RDF file From any Semantic Website.

Step2 : Check the Syntax of RDF File , then Convert it into Triple Format (RDF/NTriples).

Step3 : Convert Triple into a Model.

Step4 : Convert a model into Binary Matrix.

Step5 : Convert a binary matrix into triple matrix.

Step 6 : Create a new predicate set which contain all the predicates (Features) in all files without repetition.

Step 7 : Remove Unwanted predicates (Features) from the new predicate set in step 6. We mean by unwanted features the feature that occur normally in any RDF File(identical for all RDF File).

Step 8 : Create a new object set which contain all the objects in all files without repetition

Step 9 : Repeat the Steps (1-8) until we complete reading all the RDF Files.

Step 10 : According to the new predicate set& new object set, that is updating after reading all the RDF File, We will Build a Vectors for every file according to the number of predicates and object in that file, but the features(predicates) are identical in all files.

Step 11: Merge all the Vectors, in all files, in one Huge File.

Step 12: Partition the Data in Huge File into : Training data set & Testing data set.

Step 13 : Use one of the Machine Learning Algorithm (Classification)to learn the training set.

Step 14 : Use the testing set to check the accuracy of the suggested method.

Figure 1. A brief Steps for Suggested Algorithm

5. Method Implementation

We develop a prototype system to implement the method in the previous section. We have develop a java application that utilizes the Jena Semantic Web API for processing RDF. In order to test our system we take Medicine Domain (Diseases)[16].

We take a medicine domain (100 different Diseases) as RDF File(s) taking from DBpedia Website(*http://live.dbpedia.org/ontology/Disease*). Figure 2 below provide an extract from RDF document that describe the Cancer Disease.

```
rdf:about="http://dbpedia.org/resource/Cancer">
        <rdf:type rdf:resource="http://umbel.org/umbel/rc/AilmentCondition"/>
        <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
<diseasesdb xml:lang="en">28843</diseasesdb>
        <icd10 xml:lang="en">—</icd10>
        <icd9 xml:lang="en">—</icd9>
```

Figure 2. RDF extract for Cancer

Each RDF File used is loaded to our method using the Jena API, and we check its validation and the loaded using *ModelFactory.createDefaultModel()* method. after that we convert RDF/xml into N-Triples .

The Model converted into Binary Matrix (0&1),Where '1' means existence of Triples & '0' means No Triple at that dimension ,and the binary matrix then converted into Two Dimension Matrix [N x 3] containing string element where N is the total number of Triple Statements in RDF File and 3 represent three possibilities: subject, predicate & object (matrix[][0] means subject of the triple, matrix[][1] means predicate of the triple & (matrix[][2] means object of the triple).

After reading all the required RDF Files, then the suggested method extract the predicate values for all RDF Files except unwanted features. for our 100 diseases there is 58 distinctive predicates. A particular disease may have numerous object instances because specific features may have multiple values. For example , for the Cancer disease the final matrix containing several object instances as illustrated in Figure.3

```
[cancer 546 537 555 -1 556 544 530 551 -1 24 531 -1 -1 534 538 531 527    535 541 527 527 535 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 543 -1 -1 -1 -1 -1 -1 -1 531 536 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1]

[cancer 546 537 555 -1 556 544 530 551 -1 24 531 -1 -1 534 538 531 527 535 541 527 527 535 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 543 -1 -1 -1 -1 -1 -1 -1 531 539 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1]
```

Figure 3. An excerpt Matrix for RDF information on a Cancer

The Matrix provide one-to-one mapping with the information contained in the RDF documents, it means we can return the matrix back to its original RDF representation.

6. Evaluation

As we mentioned before we take 100 RDF File as a sample for 100 Different Diseases and after we convert each RDF into a Vectors, according to the number of objects in each one ,we combine them into one Big CSV File, we use JAVA Eclipse 2012 for this stage.

According to our Method we find there is a 58 features between all 100 Diseases, and it is very important knowledge and very helpful for scientist to know the behavior of multiple diseases according to identical features.

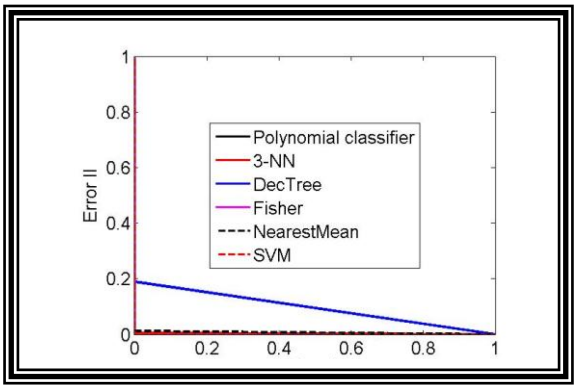
After that we take this Big CSV File and imported it into Matlab and random permutations of the vector objects are performed to shuffle the Vectors. We know that the Vector contain the label (In our sample : ,from Figure.3 , Label =Cancer) and the object values , therefore we Build a feature Matrix from the original matrix and also we build a label matrix . By using these two matrices we create a dataset by using PRTools. In order to Validate the Classification Algorithm we split this dataset into training set and testing set. Table-1 below show multiple runs on different value for training & testing set and give the accuracy of classification using multiple Classifiers like : Polynomial, kNN , Decision Tree, Fisher, Nmc & Support Vector .

No.	Training Data	Testing Data	Polynomial	kNN, k=3	Decision Tree	Fisher	Nearest Mean	Support Vector
1.	5%	95%	82%	83%	98%	82%	99.4%	99%
2.	10%	90%	84%	93.6%	98%	84%	99.7%	99.7%
3.	15%	85%	83.5%	96%	99.6%	83%	99.7%	99.8%
4.	20%	80%	83%	97.2%	99.6%	83%	99.7%	99.9%
5.	25%	75%	83%	97%	99.7%	83%	100%	100%
6.	30%	70%	84%	99%	100%	84%	100%	99.8%
7.	35%	65%	83%	99%	100%	83%	99.6%	99.8%

Table 1. Result for Classification using multiple classifier

The first row in Table-1 gives the accuracy for classification by multiple classifier using just 5% of the data as training data and it gives us a very good result. We see that some classifier gives us a 100% classification for testing data , and other make some misclassification, and our method can give the accuracy of classification for user query.

In order to determine the performance of a classifier, we used Receiver Operating Characteristics (ROC). By using PRTools we plot the false negative and the false positive . Figure.4& Figure.5 respectively Show the confusion matrix for the row No.1 & row No. 7 in Table-1, where Error I = False Negative (FNs) and Error II =False Positive (FPs).



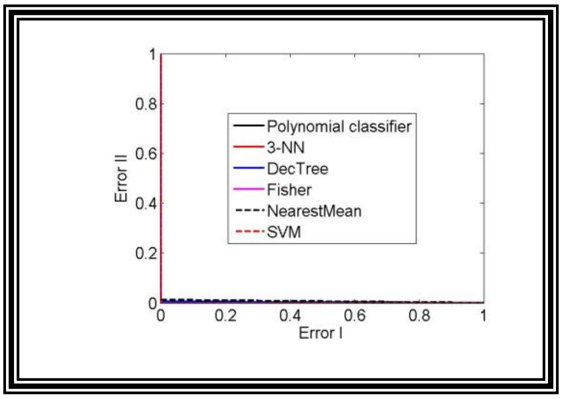


Figure 4. Confusion Matrix for training set= 5%

Figure 5. Confusion Matrix for training set= 35%

7. Conclusion

Current approaches are choosing to structured data in a form that is more understandable to both human and computer and that give the data more meaning . By Using RDF, the data is described and linked in a much informative way that allow it to be processed in a more formal way. The language that take information from RDF is SPARQL , and it is similar to SQL, but the SPARQL is based on a full understanding on the structure of the information inside the RDF so that the user must be familiar with that information in order to make a question?, and that is a weak point in SPARQL, and the query either match the information or not matched it . So, it is better to check for a new approach that is give a probability (percentage) of how much the query is closed to our information.

In this Paper we take this weak point and check to find a solution for it, and we propose a probabilistic approach that treat the querying for information as a classification problem. The approach take the RDF and convert it into a matrix format and after that convert all the information inside the RDF into a vectors, and these vectors make the feature space which is divided into a training set and a testing set . We check our algorithm by using multiple classifiers and the result is very good for the sampling data(Disease).

8. References

[1] G.Nagarajan, K.K. Thyagharajan, " A Machine Learning Technique for Semantic Search Engine", Proccedia Engineering 38,ELSEVIER, 2012.

[2] Christian Bizer," The Emerging Web of Linked Data " , IEEE,2009.

[3] Edgar Meij , Marc Bron, Laura Hollink , Bouke Huurnink , Maarten de Rijke, " Mapping queries to the Linking Open Data cloud: A case study using DBpedia", 2011 , ELSEVIER.

[4] Rada Mihalcea, Andras Csomai, " Wikify! Linking Documents to Encyclopedic Knowledge ", 2007 .

[5] Rupal Gupta , Sanjay Kumar Malik, " SPARQL Semantics And Execution Analysis In Semantic Web Using Various Tools " , 2011 .

[6] U. Boj¯ars, J.G. Breslin, A. Finn, S. Decker, " Using the Semantic Web for linking and reusing data across Web 2.0 communities ", 2007 , ELSIVIER .

[7] Bettina Fazzinga , Giorgio Gianforme , Georg Gottlob , Thomas Lukasiewicz , "Semantic Web search based on ontological conjunctive queries ",2011,ELSEVIER

[8] Dennis Pfisterer, Kay Römer, Daniel Bimschas, Oliver Kleine, Richard Mietz, and Cuong Truong, et all, " SPITFIRE: Toward a Semantic Web of Things " , 2011, IEEE.

[9] Aidan Hogan, jurgen Umbrich,et al. , "An empirical survey of Linked Data Performance", 2012,ELSEVIER.

[10] Wouter IJntema, Jordy Sangers, Frederik Hogenboom, Flavius Frasincar, "A lexico-semantic pattern language for learning ontology instances from text" , 2012 , ELSEVIER.

[11] Gerd Stumme, Andreas Hotho, Bettina Berendt, " Semantic Web Mining State of the art and future directions " ,2006, ELSEVIER.

[12] Christian Bizer, Jens Lehmann, Georgi Kobilarova, Soren Auer,Christian Becker, Richard Cyganiak, Sebastian Hellmann , " DBpedia - A crystallization point for the Web of Data ", 2009 , ELSEVIER.

[13] André Freitas,Edward Curry,João Gabriel Oliveira,and Seán O'Riain, "Querying Heterogeneous Datasets on the Linked Data Web Challenges, Approaches, and Trends", 2012, IEEE.

[14] Harrington, peter, " Machine Learning In Action", Book , Manning publishing co., 2012.

[15] Alex Smola, S.V.N. Vishwanathan, " Introduction to Machine Learning", Cambridge University Press, 2008.

[16]Matthias Samwald, et al, " Linked open drug data for pharmaceutical research and development", journal of Cheminformatics,2011.