International Journal of Computer Networks and Communications Security

VOL. 3, NO. 4, APRIL 2015, 138–150
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

A Decision Support Method for Finding Appropriate Information on the Web Documents

Reza Mohamadi Bahram Abadi[1] and Hassan Rashidi[2]

[1] Department of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

[2] Department of Mathematics and Computer Science, Allameh Tabataba'i University, Tehran, Iran

E-mail: [1]mohamadi_re@yahoo.com, [2]Hrashi@atu.ac.ir

ABSTRACT

Today, the Web has been expanded dramatically and hence, looking up desired information in a vast ocean of available data is a difficult task for users. So, we need methods which using a targeted search, help users in making decisions for choosing the appropriate documents according the desired content. In presented information retrieval technique, web documents are introduced to the user as search results. To resolve this problem can be used semantic extraction. That conclusion is valid for extraction if related subject pages identified initially. Semantic extraction ontology is one of these methods. This paper puts to evaluation the extent of relationship between a Semi structured HTML and ontology using some statistical techniques. Then with calculate the density of the document and compared with the expected density ontology in an acceptable limitation, documents related with ontology predicted. Then with calculate the density of the document and compared with the expected density ontology in an acceptable limitation, documents related with ontology predicted. If calculations for the two cases of expected value with density and view value are within the required range, then ontology would be related. According to experimental Results within a 99% reliable range, shows that the recommended method's ability to achieve value recall 100% and precision 83% is able.

Keywords: *Application Ontology, Web Documents, Information Filtering, Statistical Analysis.*

1 INTRODUCTION

World Wide Web consists of a large number of Web documents. Users to access the desired documents, work ahead are difficult [1]. For that users can find information about the interest, they need targeted search methods to find valid data is felt. The main problem is that, most of the information in web pages for humans is understandable if the machine cannot understand the meaning of them [2]. If the Web pages designed as a semantic then semantic information extraction from those pages is easy. But now all the pages of World Wide Web have been implemented, as a semantic [3]. We must use the technology uses the web pages of contemporary meaning of the simulation. Therefore, we need intelligence program that can read Web pages and data and communication between them to form into Structured [4]. Semantic extraction ontology is one of these methods. Extract information based on ontology is not affiliated web structure constants but also the detected documents described the content is dependent and in a specific field of knowledge is used [5].

For controlling could be among the vast and varied information on the web, before extraction of semantic documents about its relationship with the ontology to ensure [6]. In fact, filtering and separating documents, related or non-related from other documents, is related to search results for extraction of information will get better interest.

When we construct method to recognize which documents apply to a user's information needs, we

must be careful not to discard relevant documents and not to accept irrelevant documents [7].

In this paper, we offer an approach for recognizing whether a Web document is relevant for a chosen application of interest. specific domain ontology for web pages is defined. Then to determine the status of a sample document, its texts are matched with the ontology and then based on the results relation between document and ontology is decided [8].

The propose system is based on statistical techniques that has been implemented in three step. In step1, we used related ontology document to making a multiple linear equation. We can use this equation for two purposes:

- Angle Prediction between document vector and ontology vector
- Determine the weight and value of each of the independent variable (lexical object) in contrast to the dependent variable. In this paper we used second application.

In Step2, comparison between viewed value and expected value in a sample document and in Step3, calculation of document density, comparison of the expected density and viewed density in each record. This step is including four stages:

- stage1: density Calculation, number of characters and records of a document.
- stage2:Viewed density calculation for each lexical object in ontology in a document
- stage3:Expected density calculated for each lexical object of ontology in document
- Stage4: Comparison of two vectors, expected lexical object density and the viewed lexical object density.

After execute step2 if the calculations result is not within the acceptable range we can say that the document not related to ontology and don't need to execute step3.

2 RELATED WORKS

Before semantic extraction from web documents text, we need to be sure of its relatedness to the scope of ontology. In the past years, different methods have been applied for diagnosing the type of document relation to ontology. One of the methods was the use of heuristics, (H1) density, (H2) expected value and (H3) classification on a diagram [9]. H1 measures the density of constants and keywords defined in O that appear in D. H2 uses the Vector Space Model, a Common information-retrieval measure of document relevance, to compare the number of constants expected for each object set, as declared in O, to the number of constants found in D for each object set. H3 measures the occurrence of groups of lexical values found in D with respect to expected groupings of lexical values implicitly specified in O. In this method, machine learning is used to recognize the acceptable line for having a document within ontology. Upon calculation of the three heuristics on a sample document and evaluation of the results on the decision tree, one can state the idea on documents relation to ontology [9].

In 2001, Quan Wang posed the use of probabilistic retrieval model for distinguishing the type of documents relation to ontology. The three heuristics have been applied as used before, the difference lies in expected value heuristic, which is not calculated for the document in general [10]. Instead, the calculation is on expected value for lexical objects separately. In order to show the heuristics results on a document, vector is used with n+2 long. The two elements of vector including density value(y) and grouping (z) and other n variables including expected value for n number of lexical objects are in a sample document $\vec{D} = (X_1, X_2, ..., X_n, y, z)$.

For making decision on the type of document relation to ontology, we use logistic regression and probabilistic retrieval model. The degree of relation is shown using following formulas [11].

$$\text{sum} = \ln\left[\frac{p(R \mid X1, ..., Xn, y, Z)}{1 - p(R \mid X1, ..., Xn, y, Z)}\right] \quad (1)$$

$$p(R \mid X1, ..., Xn, y, Z) = \frac{1}{1 + e^{-\text{sum}}} \quad (2)$$

Considered the limit of probability for calculating ($0<p<1$) we can say the less difference in values out of above formulas, the more relation it has to ontology [11].

QuanWang used his experimental Results on the three types of different documents:

- Ten related web site to the ontology (Table 1)
- Ten nonrelated web site to the Ontology (Table2)
- Eight similar web site to the ontology (Table 3)

In the above tables extracted numbers of occurrences of the lexical objects of car ontology for any web document. These documents extracted from 10 different regions that cover 120 sites in the United States with 12 documents retrieved from each region [11]. In this article, we use these three sets in the proposed method implementation and experimental result.

For doing the tests, first, he counted the number of lexical objects in each document. Then he showed the document vector. Then document vector was optimized. Also he calculated the density value and grouping values for the same document. Then upon using logistic regression, the probabilistic retrieval

model was applied for determining the type of document relation to ontology. Some other results showed, system suffers from some negative points.

Semantic extraction, most methods use such heuristics for distinguishing the document relation to ontology. In recommended method of this research, we use density heuristics and expected value for our purpose. Of course, the two heuristics we apply for lexical objects. Also we calculate the view value and expected value for the density. Decision-making on document relation to ontology is based on comparison of view values and expected values within an acceptable limit.

Table 1: List of related website to ontology [11]

URL	Occurrence numbers of the lexical objects of car ontology test set documents						
	Year	Make	Model	Mileage	Price	Feature	PhoneNr
http://www.delmarvaclassfieds.com	39	37	21	10	34	39	22
http://www.thetelegraph.com	30	51	63	7	58	26	25
http://www.vermontclassifieds.com	41	16	12	8	18	24	24
http://www.ndweb.com	12	6	6	4	8	18	12
http://www.adn.com	319	209	265	53	264	214	166
http://www.hawaiisnews.com/cars	128	74	44	21	73	116	51
http://www.brewtonstandard.com	9	5	9	3	9	11	5
http://www.aikenstandard.com/	72	47	48	19	54	216	74
http://adaeveningnews.com	37	13	15	9	8	25	35
http://www.tahoe.com	12	14	12	4	11	22	14

Table 2: List of non-related web site to ontology [11]

URL	Occurrence numbers of the lexical objects of car ontology test set documents						
	Year	Make	Model	Mileage	Price	Feature	PhoneNr
http://www.cs.byu.edu	13	4	1	0	0	0	2
http://www.dogpile.com	1	0	0	0	0	3	0
http://www.ecampus.com	10	0	2	0	0	0	0
http://www.cyberpages.com	25	3	3	0	0	29	7
http://www.ohio.com	295	15	63	17	9	10	72
http://www.crookstontimes.com	6	0	0	2	8	0	11
http://www.date-net.com	2	0	10	0	0	20	0
http://www.netbikes.yks.com	240	45	8	30	78	25	105
http://www.sunspot.net	61	1	7	5	51	12	76
http://www.internetclassiccars.com	5	11	6	0	10	15	0

Table 3: List of similar web site to ontology [11]

URL	Occurrence numbers of the lexical objects of car ontology test set documents						
	Year	Make	Model	Mileage	Price	Feature	PhoneNr
http://www.photoads.com	127	17	14	9	67	18	79
http://www.cyberus.ca/~obcweb/	54	2	7	1	4	9	17
http://www.prairietech.net	26	7	3	4	11	2	22
http://www.photoads.com	79	4	2	14	46	12	53
http://www.photoads.com	58	1	13	1	25	10	32
http://www.photoads.com/boats.htm	27	7	6	1	17	9	20
http://www.crookstontimes.com	6	2	2	1	3	5	25
http://www.photoads.com	49	20	6	6	24	26	36

3 PRELIMINARIES

3.1 Application Ontology

To Provide theoretical interest for this article, we define a sample application as an ontology as a cognitive model. In fact, this model shows a real environment in a limited space. This system uses the two methods, graphics and text. They are both equivalent. Application Ontology interested is in connection with the domain of car-ads [12].

In the Fig. 1 shows a portion of the textual representation of the car-ads ontology, which includes all object and relationship sets, cardinality constraints (lines 1-9), and a few lines of its data frames (lines 10-19). This figure shows only three set of the regular expression. For the representation of a complete ontology of car-ads, we need to 165 regular expressions. In a textual view, the symbol [→ object] shows the non-lexical object. In fact, the main title of ontology or ads is represented by this symbol. The min: max or min: ave: max, constraint specified next to the connection between an object set and a relationship set in a graphical representation is the participation constraint of the object set in the relationship set. min, ave and max denote the minimum, average, and maximum number of times an object in an object set can, or is expected to, participate in a relationship set, respectively, whereas * designates an unknown but finite maximum [13].

Number of times an object in an object set can participate in a relationship set. In the textual representation for the car-ads ontology, the participation constraints are listed from line 2 to line 9. Regular expressions consider some limits for lexical object. For example, lines 10 to 14 have constraints for the object make. Such that this object can be 10 characters maximum. The keywords in relation with considered object is defined in this section [14].

We can extract the related key words by using the data frame provided for ontology and by comparing the existing strings in the text and the regular expressions in the data frame.

3.2 Density Heuristic

A Web document D that is relevant to particular application ontology A should include many constants and keywords defined in the ontology. Based on this observation, we define a density heuristics. We compute the density of D with respect to O as follows [15]:

$$\text{Density}(D,O) = \frac{\text{total number of matched characters}}{\text{total number of characters}} \quad (3)$$

Where total number of matched characters is the number of characters of the constants and keywords recognized by O in D, and total number of characters is the total number of characters in D [15].

3.3 Expected-Values Heuristic

We apply the VSM model to measure whether a multiple-record Web document D has the number of values expected for each lexical object set of application ontology O. Based on the lexical object sets and the participation constraints in O; we construct an ontology vector OV. Based on the same lexical object sets and the number of constants recognized for these object sets by O in D, we construct a document vector DV. We measure the relevance of D to O with respect to our expected-values heuristic by observing the cosine of the angle between DV and OV.

To construct the ontology vector OV, we (1) identify the lexical object-set names these become the names of the coefficients of OV, and (2) determine the average participation for each lexical object set with respect to the object set of interest specified in O these become the values of the coefficients of OV.

Car ontology vector Based on lexical object defined in the ontology is as follows:

$$\vec{u} = (0.975, 0.925, 0.908, 0.45, 0.8, 2.1, 1.15)$$

The names of the coefficients of DV are the same as the names of the coefficients of OV. We obtain the value of each coefficient of DV by automatically counting the number of appearances of constant values in D that belong to each lexical object set. Observe that for document vectors we use the actual number of constants found in a document. To get the average (normalized for a single record), we would have to divide by the number of records—a number we do not know with certainty. Therefore, we do not normalize, but instead merely compare the cosine of the angles between the vectors to get a measure for our expected values heuristic.

As mentioned, we measure the similarity between an ontology vector OV and a document vector DV by measuring the cosine of the angle between them. In particular, use the Similarity Cosine Function defined in, which calculates the acute angle.

$$\cos\theta = P / N \quad (4)$$

P is the inner product of the two vectors, and N is the product of the lengths of the two vectors. When the distribution of values among the object sets in DV closely matches the expected distribution specified in OV, the angle θ will be close to zero, and Cosθ will be close to one.

For example, we run the Expected-Values Heuristic on the two documents D_a and D_b (Fig. 2) then we calculate the amount Cosine θ for each document. Initially, the number of values expected for each lexical object set of car ontology in the D_a and D_b counted and presented as $\vec{v}a = (16,10,12,6,11,29,15)$, $\vec{v}b = (4,0,0,2,8,0,11)$ vectors. In the first reviewed D_a document, with the help car ontology vector

$$\vec{u} = (0.975, 0.925, 0.908, 0.45, 0.8, 2.1, 1.15)$$

Calculated document optimized vector. We calculated the size of two vectors U and V.

$$\left|\vec{u}\right| = \sqrt{\begin{matrix}(0.975)^2+(0.925)^2+(0.908)^2+(0.45)^2 \\ +(0.8)^2+(2.1)^2+(1.15)^2\end{matrix}} = 3.03$$

$$\left|\vec{v}_a\right| = \sqrt{16^2+10^2+12^2+6^2+11^2+29^2+15^2} = 41.51$$

Document optimized Vector is equal:

$$\vec{v}\text{norm} = \frac{\text{Document Vectore}}{\text{Document Vector Size}/\text{Ontology Vector Size}} \quad (5)$$

$$\vec{v}a,\text{norm} = \frac{(16,10,12,6,11,29,15)}{41.51/3.03}$$

$$= (1.17, 0.73, 0.88, 0.44, 0.8, 2.12, 1.1)$$

The inner product of two vectors U and $V_{a,norm}$ calculated as book value.

$$P = \left\langle \vec{u}, \vec{v}_{a,norm} \right\rangle = (0.975)\times(1.17)+(0.925)\times(0.73)+$$
$$(0.908)\times(0.88)+(0.45)\times(0.44)+(0.8)\times(0.8)+$$
$$(2.1)\times(2.12)+(1.15)\times(1.1) = 9.17$$

Then we calculated the N value:

$$\left|\vec{V}_{a,norm}\right| = \sqrt{\begin{matrix}(1.17)^2+(0.73)^2+(0.88)^2+(0.44)^2 \\ +(0.8)^2+(2.12)^2+(1.1)^2\end{matrix}}$$

$$= 3.0355$$

$$N = \left|\vec{u}\right|.\left|\vec{v}_{norm}\right| = (3.03).(3.0355) = 9.197$$

$$\cos\theta = \frac{p}{N} = \frac{9.17}{9.197} = 0.997$$

Now we do calculations for the D_b document. With considering the V_b vector of this document, Cos θ is equal:

$$\cos\theta = \frac{p}{N} = \frac{1.02}{9.2} = 0.11$$

3.4 Regression Analysis

One of the main goals of many statistical researches is to create Dependencies that provide prediction of one or more variables according to others. One of the tools that we can achieve a good relationship is regression [16]. Regression analysis is a statistical tool to study the relationship between a dependent variable and a set of independent variables [17]. If more information that is associated with the subject could be considered, we can correct the predictions. The most common linear equation can be used on the regression relations between the two variables for implementation is as follows:

$$\mu_{y|x1,x2,...,xn} = \beta_0+\beta_1x1+\beta_2x2+...+\beta_kxk \quad (6)$$

In the equation above, y is a random variable that we want to predict their values according to known values x_1, x_2… x_k And $\beta_0,\beta_1,\beta_2,\ldots,\beta_k$ multiple regression coefficients, are constants, which must be determined on the viewed data. One of the main conditions of multiple linear regression independent variables is linear independency [17]. In this paper, we use in an application Ontology the lexical objects as independent Variables used in multiple linear regressions.

```
1. Car [-> object];
2. Car [0:0.908:1] has Model [1:*];
3. Car [0:0.925:1] has Make [1:*];
4. Car [0:0.975:1] has Year [1:*];
5. Car [0:0.8:1] has Price [1:*];
6. Car [0:0.45:1] has Mileage [1:*];
7. PhoneNr [1:*] is for Car [0:1];
8. PhoneNr [0:1] has Extension [1:*];
9. Car [0:2.1:*] has Feature [1:*];
10. Make matches [11] case insensitive
11. constant
12. { extract "\b chev \b"; }, { extract "\b chevy
    \b"; }, { extract "\b dodge\b"; },
13. …
14. end;
15. Model matches [16] case insensitive
16. constant
17. { extract "88"; context "\bolds\S*\s*88\b";
    },
18. …
19. end:
```

Fig. 1. Car-ads ontology – textual

Last Updated: Monday, January 24, 2000 12:19pm Cars for Sale
DEPENDABLE CAR, 1989 Subaru SW. Auto, AC, $1900 OBO. Call (336)835-8579. (61)
Factory Warranty, 1998 Elantra. Black 4 door W/tinted Windows. Auto, pb, Ps, cruise, am/fm cassette stereo. Excellent condition Pay off OBD. Call (336)526-5444 anytime & leave message
1994 HONDA ACCORD EX, Auto, power everything, jade green w/gold Package. Under 100 K miles. Call (336)526-1081.
1999 Grand AM 27,000 miles, silver, auto, still under warranty. $14,000. Call (336)366-499
'53 Chevy Bel Aire. All original, looks like new. Serious inquiries only. $8500. Call (336)468-8924 after 4 pm.(44)
Two GREAT CARS, 1973 MGB convertible. British racing green. Mags, New tires, 4-speed, 1 owner, excellent running condition. $4500. 1997 olds Cutlass Supreme. New white paint job W/ 1/2 red Landau top, original Mags & new tires.
95 FORD CONTOUR, 5-speed, great condition, one owner, $5300. Call (336)526-8853 & leave message if no answer. (92)
Seized Cars from $500, Sports, luxury& economy cars, trucks, 4x4's utility and more. For current listings, call 1-800-311-5048 ext. 10012 (118)
1996 VW JETTA GL, 26,000 miles. 4 door, 5-speed, AC, sunroof, 1 owner. $11,000. Call (336)874-7317 anytime. (90)
'85 Buick Park Avenue. $500. Head may be cracked. Will run. Body good condition. Call (336) 526-2768. (85)
'95 Ford Thunderbird. Loaded, V-8, 45K, $6995. Call S&J at (336)874-3403 (68)
'96 Mercury Tracer. 4 door, 5 speeds, 34K, $4995. Call S&J at (336)874-3403. (69)
'88 Firebird. V8,5.0, ful injected, T-tops, 109,000 miles, red, runs great. $1880. Call (336)526-1164 anytime. (96)

(a) Car advertisements retrieved from http://www.elkintribune.com/

Last Updated: Wednesday, December 22, 1999 Select a category
Apartment For Rent For Sale or Rent Lost or Found For Sale House For Rent
Apartment For Rent, ONE EFFICIENCY, 2 & 3 bdrm, all utilities Paid. Call 281-2051-
For Rent, HOUSING Solutions – Free TV cable furn. $60/wk - $ 210/mo. 281-4060. -
For Sale. 1998 JD 455 mower, 60' deck. Call for price. Also, homemade GO-Cart. Call after 5:30 pm 218-281-1128.-
For Sale or Rent, 10,000 SQ.FT. Office building. Handicap accessible. Call 281-3631.
Help Wanted, NOW HIRING full time and part time customer service representatives. Advancement possible and weekly pay.
PART TIME AND Weekend help working with developmentally disabled adults. Call Melissa or Karen at 281-3872. -
REM-NORTHWEST Services, Inc. has a full time program Coordinator/Coordinator Position open in Crookston Working with four developmentally disabled adults. Must have a high school diploma or equivalent One year experience serve people with developmental disabilities preferred. Applicant must be 18 year of age or older. Must have a valid driver's license and driving record that meets REM's insurability requirements. Insurance and benefits available.
REM-NORTHWEST Services, Inc. has full and part time Coordinator Positions available in Crookston, Working with citizens. Excellent benefits are offered including health, 401K and profit sharing for full and Part time employees working 20 hours.
HOUSE For Rent, 3 BDFRM HOUSE $450/mo. 281-1970.22 STEEL BULDINGS, NEW, must sell. 40x60x14 was $17,500;
Lost or Found: Golden retriever about 4 months old Found 7miles south of Crookston

(b) Items for sale advertisements retrieved from http://w`ww.crookstontimes.com

Fig. 2. A car advertisement and a non-car advertisement Web document

4 IMPLEMENT RECOMMENDED SYSTEM

We use the statistical techniques to determine the type of Ontology relationship with the sample document. Acceptable error rate of α=0.01 is considered. This means that the calculations in 99% confidence intervals investigated. We used the Web documents in this project as semi-structured and HTML type also the ontology used is car-ads ontology. Recommended algorithm performed in three steps. In the present proposed method, related document D_a and non-related D_b reviews and

Documents relationship with the ontology will evaluate. In the fig. 2 shows two documents.

4.1 Step1: Using Related Ontology Document to Making Multiple Linear Regression Equation

To start this step we need number of related document to ontology. For this purpose used information table1 that it is a set of related documents including ontology website in 10 different sites. In this table for each document, its corresponding document vector specified. We calculate the vector optimized of document by using heuristic of expected values and we calculate the angle between the document vector and ontology vector for each document by using formula 4. Calculation results listed in Table 5.

To create a model of relationship between lexical object of car-ads ontology we use the information obtained and we formed a multiple linear regression equation. For this purpose, we consider the lexical object car-ads ontology as independent variables and the angle between optimized vector of document and ontology vector as the dependent variable .List of lexical object in the document I, and the regression variables used shown in Table 5 then we form this regression equation (Y_i is depended variable and X_{i1}-X_{i7} are independent variable).

The SPSS software use to implement regression. We used the date of table 4 to create regression equation. After implement regression specified coefficient β to each lexical object (table 6).After defining the early step of the formation and regression orders, β coefficients belonging to each lexical object specified in the regression. Using multiple linear regression formula $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ Coefficients obtained in the final model the desired formula shown below:

$$y = 0.017 + 0.104X_1 + 0.101X_2 + 0.098X_3 + 0.054X_4 + 0.082X_5 + 0.224X_6 + 0.12X_7 \quad (7)$$

The above formula can use for two purposes:

1. Angle Prediction between document vector and ontology vector using multiple linear regressions.
2. Determine weight and value of each lexical object

For example, we calculated angle between D_a vector from document of fig. 2 and ontology vector. This calculation implemented with Expected value heuristic in part 2.3 and equal is Cos θ=0.99. Now we implemented formula 7 for calculation Cos θ with using the optimized D_a document vector. This vector calculated in part 2.3 already.

$$\vec{v}_{a,norm} = (1.17, 0.73, 0.88, 0.44, 0.8, 2.12, 1.1)$$

The vector value $V_{a,norm}$ placed in the formula 7 and then y value is calculated.

$$y = 0.017 + (1.17 \times 0.975) + (0.73 \times 925) + (0.88 \times 0.908) + (0.44 \times 0.45) + (0.8 \times 0.8) + (2.12 \times 2.1) + (1.1 \times 1.15) = 0.99489$$

Result above show that Values obtained from both methods are almost equal.

$$(\text{Cos } \theta = 0.997) \approx (y = 0.99489)$$

Therefore we conclude that the angle between document vector and ontology vector can be calculated using the formula above. Then formula7 can be used instead Cos θ=P/N. But in this article, the second application desired. The set of the β coefficients belonging to lexical objects considered as vector B. In fact, this vector specifies weight and value that determine the lexical object.

$$\vec{B} = (0.104, 0.101, 0.098, 0.054, 0.082, 0.224, 0.12)$$

Table 4: List of optimized documents vector from documents vector of Table 1

Cos θ	Year	Make	Model	Mileage	Price	Feature	PhoneNr
0.9346	1.46	1.38	0.78	0.37	1.27	1.46	0.82
0.7953	0.82	1.40	1.73	0.19	1.59	0.71	0.69
0.8855	2.07	0.81	0.61	0.4	0.91	1.21	1.21
0.9837	1.32	0.66	0.66	0.44	0.88	1.98	1.32
0.8884	1.61	1.05	1.34	0.37	1.33	1.08	0.84
0.9309	1.82	1.05	0.63	0.3	1.04	1.65	0.72
0.9448	1.33	0.74	1.33	0.44	1.33	1.62	0.74
0.9659	0.86	0.56	0.57	0.23	0.64	2.57	0.88
0.8872	1.83	0.64	0.74	0.45	0.4	1.24	1.73
0.9908	1.01	1.18	1.01	0.34	0.93	1.85	1.18

Table 5: List of lexical objects and variables used in regression

Cos θ	Year	Make	Model	Mileage	Price	Feature	Phone Nr
Y_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}	X_{i7}

Table 6: Output SPSS software after executes the regression over the data Table 4

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Constant	Year	Make	Model	Mileage	Price	Feature	Phone Nr
0.017	0.104	0.101	0.098	0.054	0.082	0.224	0.120

4.2 Step2:Comparison Between Viewed Value And Expected Value In a Sample Document

After receiving a document to determine its relationship with ontology, in the first step, we form the document vector then we calculate document optimized vector using the formula 5. For evaluate the model, focus on two documents D_a and D_b (Fig.2). The number of lexical object in the D_a and D_b counted and presented as vectors $\vec{v}a=(16,10,12,6,11,29,15)$, $\vec{v}b=(4,0,0,2,8,0,11)$. In the first reviewed D_a document. We use to calculate document optimized vector in the part 2.3 that equal is:

$$\vec{v}a,\text{norm}=(1.17,0.73,0.88,0.44,0.8,2.12,1.1)$$

Then vector values $V_{a,norm}$ and U multiplied in the vector B and provide them with vectors $V_{a,view}$ and U_{Expect} Thus each element of vectors $V_{a,norm}$ and U takes the value and weight.

$$\vec{V}_{a,view}=\vec{V}_{a,norm}\times\vec{B}=(1.17\times0.104,\ 0.73\times0.101,\ 0.88\times 0.098,0.44\times0.054,\ 0.8\times0.082,2.12\times0.224,\ 1.1\times0.12)$$
$$=(0.122,0.074,0.086,0.024,0.066,0.475,0.132)$$

$$\vec{U}_{Expect}=\vec{U}\times\vec{B}=(0.975\times0.104,0.925\times0.101,0.908\times 0.098,0.45\times0.054,0.8\times0.082,2.1\times0.224,1.15\times0.12)$$
$$=(0.101,0.093,0.089,0.024,0.066,0.47,0.138)$$

By the following formula and test χ^2, can be defined an area sure to accept or not accept related document with ontology.

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i} \quad (8)$$

(e_i: Expected Frequency, *fi:* View frequency, K: number of lexical objects in the ontology defined)

If the relationship $\chi^2\le\chi^2_{1-\alpha,k-1}$ is true, then our calculations within the confidence 1-αwould be accepted. Desired Values for $\chi^2\le\chi^2_{1-\alpha,k-1}$ are extracted from the Chi-square distribution table. To perform calculations in the above formula, we use the e_i value of the $\vec{U}_{Expect}$ vector and we use the f_i value of the $\vec{v}_{norm}$ vector Extraction.

The χ^2 For D_a document is equal:

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}=\frac{(0.122-0.101)^2}{0.101}+\frac{(0.074-0.093)^2}{0.093}+\frac{(0.086-0.089)^2}{0.089}+\frac{(0.024-0.024)^2}{0.024}+\frac{(0.066-0.066)^2}{0.066}+\frac{(0.475-0.47)^2}{0.047}+\frac{(0.132-0.138)^2}{0.138}=0.008608$$

Considering the amount $\chi^2_{1-\alpha,k-1}=\chi^2_{0.99,6}=0.872$, if the condition $\chi^2\le\chi^2_{1-\alpha,k-1}$ is established then can be said that, Document optimized vector would be acceptable in the range of the ontology valid vector. For D_a, Condition $\chi^2\le\chi^2_{1-\alpha,k-1}$ is true (0.0086608<0.872). With establishing the conditions for definitive diagnosis in the document, we will go to the third step Otherwise the document not related to ontology.

Now we do the calculations for V_b Vector the same as V_a.

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}=0.882957$$
$$0.882957\succ0.872$$

Given that 0.882957>0.872, then we can declare that D_b document definitely is not related with ontology and don't go to the third step.

4.3 Step3:Calculation Document Density And Comparison Between Expected Density And Viewed Density In Each Record

This step based the computing density records. Calculation expected density in each record is according ontology. Also we can calculation viewed density in each record based is according document text and ontology. Comparison of these values helps us to do a correct statement about the relationship ontology and document. Calculations performed in four stages:

4.3.1 Firs Stage: Calculation Density, Number Of Characters And Records In The Sample Document

Web documents used in this project as semi-structured and HTML type. We can count the total

records of documents structurally. To continue working we have to count the number of document records and the total characters in the sample (D_a). Records number 15 and characters sum 1992 counted in the D_a

4.3.2 Second Stage: Calculation Viewed Density for Each Lexical Object in The Sample Document

In ontology, the maximum acceptable number of each lexical object in one record specified. Using these limitations and considering number of record in the sample document, about number of viewed lexical object in each document we decide. This observation presented in the document vector. The limitations desired car ontology as $\vec{W} = (1,1,1,1,1,5,2)$ vector shown. Other than the Feature and PhoneNr, the maximum number of lexical objects in the ontology is equal to one but ontology does not have limitation for these two particular cases. In this project, the calculations in an acceptable range considered. For each record usually not used more than five cases for Features and two PhoneNr. With consider these conditions and the counting records number of the document, the document vector modified and then will be display as the V_{opt} vector.

For example In the document, number of year lexical object is 16 but this document has only 15 records, so only 15 Occurrence of Year lexical object is accepted. Other lexical objects V vector is accepted. Edited V vector is:

$$\vec{V}_{Opt} = (15,10,12,6,11,29,15)$$

The maximum number of characters that each lexical object can have, defined in the destination ontology limitations (For example line 10 of the car ontology, fig. 1). This limitation about car ontology displayed in the document y vector.

$$\vec{y} = (4,10,12,8,10,20,10)$$

Using the following formula, we can calculate maximum number of viewed character and acceptable for each lexical Object.

$$\overrightarrow{Ch}_{view} = \vec{V}_{Opt} \times \vec{y} \quad (9)$$

Maximum number of viewed character for the initial vector $\bar{v}$ from the D_a document is:

$$\overrightarrow{Ch}_{view} = (60,100,144,48,110,580,150)$$

Also by dividing the number of accepted Characters for each lexical object by the number of Viewed characters of document, we calculate the density value for each lexical object.

$$\overrightarrow{Density}_{View} = \frac{\overrightarrow{Ch}_{View}}{\text{Number Of Characters in the document}} \quad (10)$$

Viewed Density for each lexical object in instance document, we represent as follows:

$$\overrightarrow{Density}_{View} = \frac{(60,100,144,48,110,580,150)}{1992}$$
$$= (0.03,0.05,0.07,0.02,0.05,0.29,0.07)$$

All Steps for calculate the viewed density of lexical objects in the D_a document presented in the table7.

4.3.3 Step3: Calculate Expected Density For Each Lexical Object Based The Ontology in The Sample Document

Using the ontology vector and the records number of sample document for each lexical object, we can calculate approximate number of expected value, this calculation done based on the following formula

$$\vec{D}_{expect} = \vec{U} \times \text{Total Records Document} \quad (11)$$

Thus for the V vector of D_a document can say:

$$\vec{D}_{expect} = (0.975,0.925,0.908,0.45,0.8,2.1,1.15) \times 15$$
$$= (14.63,13.88,13.62,6.75,12,31.5,17.25)$$

Then, using the following formula for the expected number of characters for each object is calculated. Then by the use of the following formula, the number of expected characters would calculate for each object.

$$\overrightarrow{Ch}_{expect} = \vec{D}_{expect} \times \vec{y} \quad (12)$$

We use the formula above and we calculate the expected number of characters for each lexical object in the D_a document.

$$\overrightarrow{Ch}_{expect} = (14.63,13.88,13.62,6.75,12,31.5,17.25)$$
$$\times (4,10,12,8,10,20,10)$$
$$= (58.5,138.8,163.4,54,120,630,172.5)$$

After calculating the Expected characters number for each object and considering the characters number in document, using the following formula done calculated the expected density of each lexical object in the sample document.

$$\overrightarrow{Density}_{expect} = \frac{\overrightarrow{Ch}_{expect}}{Number\ Of\ Document\ Characters} \quad (13)$$

The expected density of each lexical object in the D_a document shown as follows.

$$\overrightarrow{Density}_{expect} = \frac{(58.5,138.8,163.4,54,120,630,172.5)}{1992}$$
$$= (0.029,0.069,0.082,0.027,0.06,0.316,0.086)$$

Because we do not know about the method of occurrence of expected lexical values in future, we consider the maximum number of characters allowed for each lexical object defined in the ontology. Therefore, for uniformity of comparisons, we ignore

the actual number of characters for viewed lexical values and let in the maximum quality value for each lexical object, but there is a little difference between this number and the real size. Steps Summary for calculate the expected density of lexical objects in the D_a document in table 8 presented Steps.

Table 7: All steps for calculate the viewed density of lexical objects in the Da document

Character sum = 1992, Record number = 15							
Vectors	Year	Make	Model	Mileage	Price	Feature	PhoneNr
V	16	10	12	6	11	29	15
W	1	1	1	1	1	5	2
V_{opt}	15	10	12	6	11	29	15
Y	4	10	12	8	10	20	10
Ch_{view}	60	100	144	48	110	580	150
$Density_{View}$	0.03	0.05	0.07	0.02	0.05	0.29	0.07

Table 8: All steps calculated the expected density of lexical objects in the Da document

Character sum = 1992, Record number = 15							
Vectors	Year	Make	Model	Mileage	Price	Feature	PhoneNr
V	16	10	12	6	11	29	15
U	0.975	0.925	0.908	0.45	0.8	2.1	1.15
D_{Expect}	14.63	13.88	13.62	6.75	12	31.5	17.25
y	4	10	12	8	10	20	10
Ch_{Expect}	58.5	138.75	163.44	54	120	630	172.5
$Density_{Expect}$	0.029	0.069	0.082	0.027	0.06	0.316	0.086

4.3.4 *Step4: Comparison of Two Vectors Expected Lexical Object Density and The Viewed Lexical Object Density*

In this step, we compare the two vectors $Density_{Expect}$ and $Density_{view}$ of the sample document, and then we will put to test the credit calculations within a reliable range. Before doing this, all the values of the two vectors would divide by the acceptable density value of the sample document. By doing so, it's acceptable density in the computing is more effective. If density value is more, result divide two vectors $Density_{Expect}$ and $Density_{view}$ by density is closer to zero. Resulting, to the reliable area of %100(1-α) for test χ^2 is closer. To calculate of acceptable density from document, the following formula is used:

$$\overrightarrow{\text{density(D)}}_{\text{accept}} = \frac{\text{Total haracters accept}}{\text{Number Of document Characters}} \quad (14)$$

The total accepted characters of total elements would obtain from ch_{view} vector. With the division of vectors $Density_{Expect}$ and $Density_{view}$ by value $Density(D)_{accept}$ vector, will be formed two vectors optimized $Density(D)_{EcceptNormal}$, $Density(D)_{viewNormal}$. We use the following formula for measuring the two vectors.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

The values f_i of $Density_{viewNormal}$ vector and the values e_i of $Density_{ExpectNormal}$ vector are used. Considering the amount $\chi^2_{1-\alpha,k-1} = \chi^2_{0.99,6} = 0.872$, if condition $\chi^2 \le \chi^2_{1-\alpha,k-1}$ is established, it can say that values Acceptable viewed density vector to the values expected density vector within the acceptable range.

With Calculating density of acceptable for the D_a document optimized values for the two vectors is:

$$\overrightarrow{\text{density(D)}}_{\text{accept}} = \frac{\Sigma(\overrightarrow{\text{ch}}_{\text{view}})}{\text{Number Of document Characters}} = \frac{1192}{1992} = 0.598$$

$$\overrightarrow{\text{Density}}_{\text{expectNormal}} = \frac{\overrightarrow{\text{Density}}_{\text{expect}}}{\overrightarrow{\text{Density(D)}}_{\text{accept}}} = (0.0485, 0.1154, 0.137, 0.0451, 0.1003, 0.528, 0.144)$$

$$\overrightarrow{\text{Density}}_{\text{ViewNormal}} = \frac{\overrightarrow{\text{Density}}_{\text{View}}}{\overrightarrow{\text{Density(D)}}_{\text{accept}}} = (0.05, 0.084, 0.117, 0.0334, 0.084, 0.485, 0.117)$$

As follows, two vectors compared:

$$x^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}=\frac{(0.05-0.0485)^2}{0.0485}+\frac{(0.084-0.1154)^2}{0.1154}+\frac{(0.117-0.137)^2}{0.137}+\frac{(0.0334-0.0451)^2}{0.0451}+\frac{(0.084-0.1003)^2}{0.1003}+\frac{(0.485-0.528)^2}{0.528}+\frac{(0.117-0.144)^2}{0.144}=0.0223$$

Considering the amount $\chi^2_{1-\alpha,k-1}=\chi^2_{0.99,6}=0.872$,If the condition $\chi^2 \pounds \chi^2_{1-\alpha,k-1}$ is established, it can be said that View density vector values is acceptable to the values expected density vector within the acceptable range and we said the document with instance application ontology is related.

5 EXPERIMENTAL RESULTS

In this section, is evaluated the proposed model. For this purpose, we tested three different documents. The first group consisted of ten relative document (Table 1), the second group include ten non-relative document (Table2) and third group include eight non-relative document but similar to the car ontology (Table 3).

The model performance evaluated by computing the recall and precision ratios on each test document set.

$$\text{recall}=\frac{TP}{TP+FN} \quad (15)$$

$$\text{precision}=\frac{TP}{TP+FP} \quad (16)$$

- TP: Number of documents that related with the ontology and method they detected related too.
- FN: Number of documents that related with the ontology, but method they detected unrelated.
- FP: Number of documents that unrelated with the ontology, but method they detected related.

The experimental result of propose method on different documents in the table 9 summarized. Reference first 10 rows is table1, second 10 rows is table2 and third 8 rows is table3. Two heuristics mentioned in the tables as follows:

- The first heuristic is the second step of the proposed method as Comparison between viewed value and expected value in a sample document.
- The second heuristic is the third step of the proposed method as calculation document density and comparison between expected density and viewed density in each record.

For both of heuristic, two columns with the title of acceptance and rejection is considered. If calculation results, is in the acceptable range selection acceptance otherwise selection rejection.

In the First evaluating, the propose method in an acceptable range α-1=99%with error α=1%also evaluate opportunities. Whereas the object number of ontology vector is equal to seven then k=7. If the condition $\chi^2 \le \chi^2_{1-\alpha,k-1}$ or $\chi^2 \le \chi^2_{0.99,6}=0.872$ is established, it can say that values Acceptable viewed density vector to the values expected density vector within the acceptable range then Experimental result in the range α-1=99% shown in the table 9. In this table, number of Record represents the number of records counted in a sample document also detection type represents type of document relation to ontology that by the propose method specified.

According to the information in Table9, values TP=10, FN = 0 and FP =2. Thus, the values of recall and precision follow:

$$\text{recall}=\frac{10}{10+0}=100\%$$

$$\text{precision}=\frac{10}{10+2}=83\%$$

Also for further evaluation, the propose method in an acceptable range α-1=95% with error α=5% evaluate again. That results of this evaluation shown in the Table 9. According to the information in this table, values TP=10, FN = 0 and FP =10. Thus, the values of recall and precision follow:

$$\text{recall}=\frac{10}{10+0}=100\%$$

$$\text{precision}=\frac{10}{10+10}=50\%$$

If we increase the accepted error rate then the system accuracy would be reduced. The results of the tests in a 95% confidence interval was evaluated .The recall rate reduced of 100% and the precision decreased to 50%.

In normal conditions, the rate system performance is much more. Because a set of irrelevant documents selected for the system evaluation is very similar to that of the car ontology, which was already considered. However, Normal, usually non-related documents that determine the assignment given to the system, is not similar to the related documents.

Table 9: Result of proposed method tested on a set of documents

Row	Number Of Records	First heuristic			Second heuristic			1-α = 99% $\chi^2 \leq (\chi^2_{0.99,6}=0.872)$		1-α = 95% $\chi^2 \leq (\chi^2_{0.95,6}=1.635)$	
		χ^2	Accept	Reject	χ^2	Accept	Reject	Detection type	Correct diagnosis	Detection type	Correct diagnosis
1	13	0.1281	√	-	0.1071	√	-	Related	√	Related	√
2	12	0.4001	√	-	0.0703	√	-	Related	√	Related	√
3	11	0.2249	√	-	0.0795	√	-	Related	√	Related	√
4	6	0.0321	√	-	0.1278	√	-	Related	√	Related	√
5	160	0.2189	√	-	0.0868	√	-	Related	√	Related	√
6	33	0.1354	√	-	0.1712	√	-	Related	√	Related	√
7	15	0.1067	√	-	0.7997	√	-	Related	√	Related	√
8	31	0.0680	√	-	0.5440	√	-	Related	√	Related	√
9	19	0.2213	√	-	0.1731	√	-	Related	√	Related	√
10	8	0.0181	√	-	0.0700	√	-	Related	√	Related	√
11	16	1.044	-	√	9.963	-	√	UnRelated	√	UnRelated	√
12	15	0.475	√	-	18.987	-	√	UnRelated	√	UnRelated	√
13	11	1.229	-	√	13.980	-	√	UnRelated	√	UnRelated	√
14	41	0.331	√	-	2.814	-	√	UnRelated	√	UnRelated	√
15	85	0.972	-	√	1.840	-	√	UnRelated	√	UnRelated	√
16	38	0.848	√	-	12.965	-	√	UnRelated	√	UnRelated	√
17	27	0.434	√	-	2.920	-	√	UnRelated	√	UnRelated	√
18	91	0.734	√	-	1.122	-	√	UnRelated	√	Related	*
19	82	0.667	√	-	2.521	-	√	UnRelated	√	UnRelated	√
20	11	0.239	√	-	0.4261	√	-	Related	*	Related	*
21	53	0.649	√	-	0.963	-	√	UnRelated	√	Related	*
22	19	0.816	√	-	1.447	-	√	UnRelated	√	Related	*
23	12	0.653	√	-	1.016	-	√	UnRelated	√	Related	*
24	26	0.687	√	-	0.961	-	√	UnRelated	√	Related	*
25	31	0.654	√	-	1.309	-	√	UnRelated	√	Related	*
26	20	0.434	√	-	0.960	-	√	UnRelated	√	Related	*
27	11	0.70	√	-	1.331	-	√	UnRelated	√	Related	*
28	14	0.298	√	-	0.1165	√	-	Related	*	Related	*

6 REFERENCES

[1] Al-Kamha, Reema, and David W. Embley. "Grouping search-engine returned citations for person-name queries." In Proceedings of the 6th annual ACM international workshop on Web information and data management, pp. 96-103. ACM, 2004.

[2] Wessman, Alan. "A Framework for Extraction Plans and Heuristics in an Ontology-Based Data-Extraction System." PhD diss., Brigham Young University, 2005.

[3] Embley, David W., Douglas M. Campbell, Randy D. Smith, and Stephen W. Liddle. "Ontology-based extraction and structuring of information from data-rich unstructured documents." In Proceedings of the seventh international conference on Information and knowledge management, pp. 52-59. ACM, 1998.

[4] Zhou, Yuanqiu. "Generating data-extraction ontologies by example." PhD diss., Brigham Young University, 2005.

[5] Vickers, Mark S. "Ontology-based free-form query processing for the semantic web." (2006).

[6] Reema A: Conceptual xml for system analysis, August (2007).

[7] Studer. Semantic Issues in Multimedia Systems: Ontology based Access to Distributed and Semi Structured Information Database Semantics, (1998).

[8] Woodbury C. Brigham Young University: Retrieving Danish Genealogical Records on the Semantic Web, December (2004).

[9] Embley, David W., Yiu-Kai Ng, and Li Xu. "Recognizing ontology-applicable multiple-record Web documents." In Conceptual

Modeling—ER 2001, pp. 555-570. Springer Berlin Heidelberg, 2001.

[10] Olson, Lars E. "Querying Disjunctive Databases in Polynomial Time." PhD diss., Brigham Young University, 2003

[11] Wang, Quan. "A Binary-categorization Approach for Classifying Multiple-record Web Documents Using a Probabilistic Retrieval Model." PhD diss., Brigham Young University, 2001.

[12] HASSARD, TH. "APPLIED LINEAR-REGRESSION-WEISBERG, S." (1981): 158-158.

[13] Ding, Yihong. "Study of Design Issues on an Automated Semantic Annotation System." AIS SIGSEMIS Bulletin 2 (2005): 45-51.

[14] Al-Muhammed, Muhammed. "Dynamic matchmaking between messages and services in multi-agent systems." PhD diss., Brigham Young University. Department of Computer Science, 2004.

[15] Embley, David W., Norbert Fuhr, Claus-Peter Klas, and Thomas Rölleke. "Ontology suitability for uncertain extraction of information from multi-record web documents." Datenbank Rundbrief 24 (1999): 48-53.

[16] Boes, Duane C., F. A. Graybill, and A. M. Mood. "Introduction to the Theory of Statistics." Series in probabili (1974).

[17] Wonnacott, Thomas H., and Ronald J. Wonnacott. Regression: a second course in statistics. New York: Wiley, 1981.

[18] Lonsdale, Deryle, David W. Embley, Yihong Ding, Li Xu, and Martin Hepp. "Reusing ontologies and language components for ontology generation." Data & Knowledge Engineering 69, no. 4 (2010): 318-330.

[19] Lynn, Stephen G. "Automating mini-ontology generation from canonical tables." PhD diss., Brigham Young University, 2008.

[20] Lynn, Stephen, and David W. Embley. "Automatic Generation of Ontologies from Canonicalized Web Tables." submitted manuscript (2008).

[21] Muhammed, Al, and Muhammed Jassem. "Ontology aware software service agents: Meeting ordinary user needs on the semantic web." (2007).

[22] D'Avanzo, Ernesto, Antonio Lieto, and Tsvi Kuflik. "Manually vs semiautomatic domain specific ontology building."

[23] Tao, Cui, Yihong Ding, and Deryle Lonsdale. "Automatic creation of web services from extraction ontologies." In Advances in Conceptual Modeling-Theory and Practice, pp. 415-424. Springer Berlin Heidelberg, 2006.

[24] Walker, Troy. "AUTOMATING THE EXTRACTION OF DOMAIN-SPECIFIC INFORMATION FROM THE WEB—A CASE STUDY FOR THE GENEALOGICAL DOMAIN." PhD diss., Brigham Young University, 2004.

[25] Chen, Xueqi Helen, David W. Embley, and Stephen W. Liddle. Query rewriting for extracting data behind HTML forms. Springer Berlin Heidelberg, 2004.

[26] Ding, Yihong. "Semiautomatic generation of resilient data-extraction ontologies." PhD diss., Brigham Young University, 2003.

[27] Zhang, Ning, Hong Chen, Yu Wang, Shi-Jun Cheng, and Ming-Feng Xiong. "Odaies: Ontology-driven adaptive web information extraction system." In Intelligent Agent Technology, 2003. IAT 2003. IEEE/WIC International Conference on, pp. 454-460. IEEE, 2003.

[28] Embley, David W., D. M. Campbell, Y. S. Jiang, Y. K. Ng, R. D. Smith, Li Xu, S. W. Liddle, and D. W. Lonsdale. "Extracting and Structuring Web Data." Brigham Young University (2002).

[29] Yau, Sai Ho. "Automating the extraction of data behind web forms." PhD diss., Brigham Young University, 2000.

[30] Arocena, Gustavo O., and Alberto O. Mendelzon. "WebOQL: Restructuring documents, databases and Webs." In Data Engineering, 1998. Proceedings., 14th International Conference on, pp. 24-33. IEEE, 1998.

[31] Jiang, Yuan. "Record-Boundary Discovery in Web Documents." PhD diss., Brigham Young University, 1998.

[32] Kaufmann M, Mateo S: Programs for Machine Learning, Quinlan. C4.5, California, (1993).