Int. j. inf. tecnol. (March 2018) 10(1):99–110
https://doi.org/10.1007/s41870-017-0063-2

ORIGINAL RESEARCH

Query based approach for referrer field analysis of log data using web mining techniques for ontology improvement

Navjot Kaur[1] · **Himanshu Aggarwal**[1]

Received: 2 May 2017 / Accepted: 20 November 2017 / Published online: 29 November 2017

Abstract This work presents a new framework as to how web mining is helpful for information retrieval, using ontology and web log files. Ontology plays a major role in the retrieval of semantic data. The researcher has already constructed the string instrument ontology using protégé 5.0, which helps in refining the web search in music domain. The researcher has proposed a novel approach for ontology management in which the ontology is continuously updated using the knowledge extracted/discovered from the analysis of the log file (specifically the data related to the referrer field) in form of new concepts and new relationships between new and/or existing concepts. The goal of this study is to use data mining algorithms to analyse visitors and visited web pages of the website and somehow characterise or distinguish them in some way. During this the researcher has collected 'guitar' web access log from guitar selling website of 363 days of the year 2016. After pre-processing of this log file, two new feature sets have been extracted from 'guitar' log file and constructed two files namely 'File1' and 'File 2'. File 2 is also known as query log. Further clustering (EM), association rule finding (Apriori) and sequential patterns (n-gram) algorithms have been applied for suggestions of new concepts to continuously update and improve the existing ontology from time to time.

✉ Navjot Kaur
navjot@pbi.ac.in

Himanshu Aggarwal
himanshu.pup@gmail.com

[1] Department of Computer Engineering, Punjabi University, Patiala, India

Keywords Web mining · Web usage mining · Ontology · Log file · User sessions · Clustering · Knowledge discovery

1 Introduction

Web usage mining (WUM) also known as web log mining is the application of data mining techniques applied on web data to extract relevant data and discover useful patterns [1], with the aim of improving the usefulness of the various web based applications.

The process of web usage mining can be broadly divided into four phases—sourcing or collection of data, pre-processing or removal of 'noise', discovery of interesting patterns and the last analysis of the discovered patterns [2].

The first phase is simply sourcing of the data or information that is to be processed from various resources—which in our case will predominantly be the log files obtained from the Web servers, Web Proxy Servers and Client Browsers [3]. The quality of source log file is improved in the second phase by removing the extraneous, immaterial data termed as 'noise', to make the log file ready for further processing like segregating it into user and 'sessions'.

Identification [4–6]. In the third stage, various statistical techniques like 'association', 'classification' and 'clustering' are applied to the pre-processed data to discover interesting arrangements or patterns [7–9].

In the last stage the identified patterns are subjected to various analytical tools and mechanisms [1, 10] to finally extract the sublimate, the 'essence' or knowledge which has applications in a wider variety of fields like commerce, improvement of web based applications, identification of criminals and international security, attracting new customers and their retention, increasing website visits etc.

Present age, is rightly referred to as the age of information and knowledge, as having useful information/knowledge at 'right time' gives one a huge advantage over others leading to appropriate and efficient decision making and plan execution.

But in the last two decades, with the advent of internet technology and open source resources, there has been a humongous surge in the amount of information leading to 'information overload'. It becomes a challenging and time consuming task to sift through this huge volume of data to extract relevant information on a topic. To overcome this issue, certain techniques have been developed, which helps us to retrieve relevant results efficiently and accurately from the web. The plethora of these data mining techniques, methodologies, algorithms which are applied on the web data and web logs to extract relevant data and discover useful patterns, with the aim of improving the usefulness of the various web based applications is known as web usage mining.

Ontology is an explicit formal specification of the terms in the domain and relations among them [11]. Commonly it is defined to consist of abstract concepts and relationships only. In some rare cases, ontologies are defined to include instances of concepts and relationships [11].

In this paper it has been shown that how web log data can help to continuously update and improve the knowledge base of the existing ontology from time to time. Web mining techniques can be applied on web log files to find some suggestions to improve the existing ontology and some research has already been done in this field [11–16]. In this work the researcher has used protégé 5.0 (for ontology construction) and weka tool (for data mining algorithms). The thrust of this paper is facilitating information retrieval through a novel ontology management approach based using web log data.

2 Proposed methodology

The proposed methodology for novel ontology improvement approach is described in Fig. 1.

In Step: 1 of this methodology, the researcher has implemented each phase of preprocessing: data cleaning, user identification, and session identification. In session identification phase, the researcher has used the proposed Semantic-Time-Referrer based algorithm [17]. Side by side, the researcher has also constructed new string instrument ontology in the music domain using protégé 5.0 (the best and the most commonly used ontology Editor [18]) intended to enhance information retrieval and as illustrated in Step 1(a). In the next step (Step 2), the researcher has extracted two new features from the preprocessed log file and has implemented some web usage mining algorithms (Step 3) in order to extract some suggestions of the classes, concepts and relationship (Step 4) to update the knowledge base that has been build in Step 1(a). The algorithm corresponding to Fig. 1 is given below.

Algorithm 6.1: Proposed Methodology for Ontology Improvement.
Consider the following steps.
Step 0: Start
Step 1: In the first step the researcher has constructed an ontology for guitar (musical instrument) using Protégé 5.0 after thorough analysis of the content and/or structure of different websites [Step 1(a) from Figure 1].
***Step 2*:** In the Second step, the pre-processing of the guitar log file of an online guitar selling website has been done. This step further includes three sub steps [Step 1 from Figure 1]:

- Data Cleaning: Standard Method
- User Identification: Standard Method with time Constraint
- Session Identification: Semantically-Time-Referred Method

Step 3: Two new feature sets have been extracted according to the problem [Step 2 from Figure 1].

- Feature 1: Similar Page Group Hits (File 1)
- Feature 2: All Keywords Searched in the Session (File 1)

Step 4: Third and an important step are to analyze the guitar Log file using web mining algorithms. The user has applied the following three algorithms to get relevant results [Step 3 from Figure 1]:

- Clustering Method: EM Algorithm
- Association Rule: Apriori Algorithm
- Sequential Pattern: n-gram Algorithm

Step 5: In the fourth and the last step, the user has updated the guitar ontology according to suggestions obtained after analysis in the previous step [Step 4 from Figure 1].
Step 6: *End*

3 Ontology construction

The ontology includes machine-interpretable definitions of the basic concepts of a specific domain and the relations among those concepts and entities. Various ontology applications are in the field of E-science, Medicine, Organizing complex and semi-structured information, Military/Government and the Semantic web. Ontology data models retrieve information semantically and the Protégé 5.0 is the best tool to create ontology easily, quickly and efficiently for every domain [19]. In this paper, the

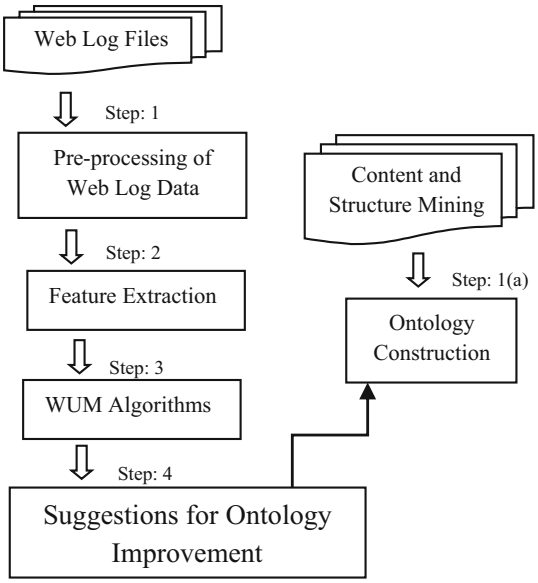

Fig. 1 Steps of proposed methodology for ontology improvement

ontology specific to the domain is constructed using the key concepts and words related to the string instruments in the domain of Music using Protege 5.0 [18]. The OWLviz visualization of string_instrument ontology is shown in Fig. 2.

For the maintenance of this ontology, it needs to be updated from time to time. Hence, the researcher has proposed a new approach according to which the log data of the various websites about the string instruments have been analyzed in order to retrieve new classes and concepts and relations between new and existing classes. During the analysis phase, the log file of 363 days of the year 2016 of a guitar selling website has been used. Now after ontology construction side by side the researcher has performed pre-processing on 'guitar' log file which has been discussed in next section.

4 Pre-processing

In this step, the focus has been to enhance the accuracy and quality of the log data to facilitate further analysis. The input 'guitar' log file is in the conditional log format collected from a website related to teaching, learning and selling of guitars and some other string instruments.

4.1 Data cleaning

The focus of data cleaning has been to improve the quality of the sourced log file by removing the irrelevant data termed as 'noise', to make the log file ready for further processing like segregating it into 'sessions' and finally, for user identification. For cleaning following steps have been followed:

- Failed and corrupted requests have been removed.
- Requests originated by web robots have been removed.
- Requests made by other than GET method has been removed.
- Requests, in which transfer bytes are nearly zero, have been removed.

The Table 1 shows the requests, in which transfer bytes were nearly zero, have been removed. The number of entries in original log file was 24,945 and number of entries in the cleaned log are 23,626, also removed 707 irrelevant URLs.

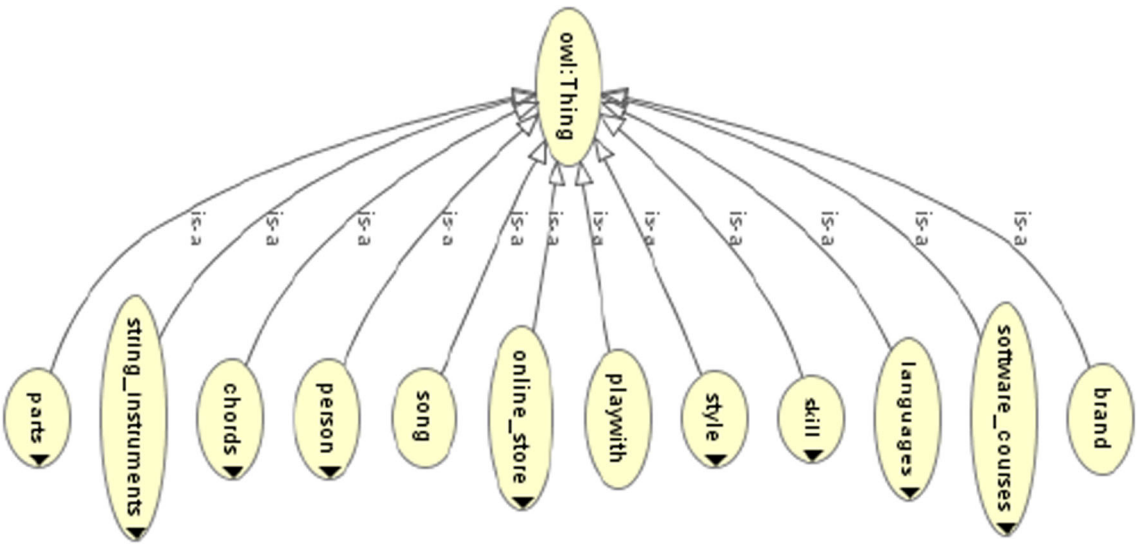

Fig. 2 OWLviz visualization of string_instrument ontology

Table 1 Details of log files before data cleaning

Features	Guitar web site
Size in KB	2110 KB
Time period	363 days
No of entries in log file	24,945
Format of log File	Conditional log format
Type of file	Excel file
Multimedia clicks	1227
Text file clicks	Nil
Robots.txt clicks	194
Error clicks	203
Other than GET method	215
Irrelevant URLS	707
Size of cleaned file in KB	1918 KB
Entries in cleaned file	23,626

Table 2 User identification details for guitar log file

Features	Log entries
Entries in cleaned log	23,626
Unique URLs	353
Unique IPs	11,712
Unique users	12,334
Number of sessions extracted	12,760

To identify the unique users, 'time constraints' have been used simultaneously with the IP address and the agent fields i.e. if the access time (crosses the threshold value) is very long from the same IP and agent filed, the proposed model automatically creates a new session. The determination of the threshold value is on the basis of the average access time of all unique users identified in the log file. Average access time has been taken as 2 h and 25 min. As shown in Table 2, the total entries in the cleaned log are 23,626. During this process, 23,626 entries have been grouped into 12,334 unique users.

4.2 User and session extraction

In this phase, the identity of the 'unique' visitor or user has been established and his navigation pattern is extracted (which pages of the website have been accessed) using this IP address. User's identification means to identify who accesses the website and more precisely which pages are accessed.

These 12,334 identified users have been used to extract the session. The appropriate session identification is a very important step in pre-processing of log data. Semantic-Time-Referrer method [17] is used to find out the number of sessions. From 23,626 entries and 12,334 users, 12,760 sessions have been extracted. Before applying data discovery techniques of web mining on these 12,760 sessions, in order to extract relevant information, the feature extraction is required to be done, this is discussed in next section of this paper.

5 Feature extraction

Before applying the data discovery techniques, first, the features have to be extracted according to our problem from the 12,760 identified sessions in pre-processing phase. Two new feature sets have been extracted as discussed here under.

5.1 File 1: similar page Group hits

This feature will group the web pages of website into some classes and sessions into similar types of visits. This feature will help us to identify that which type of pages has been accessed by which type of users. As a result of this feature extraction, the researcher has constructed a 'File 1' for further analysis which is discussed in next section (Sect. 6).

5.2 File 2: all keywords used to search in queries

In this feature, the user has retrieved only those sessions from main guitar log which contains a search engine in their referrer field and has created a new log namely "query" log. There are 563 transactions of search queries and out of these 417 and 423 are the unique users and sessions, respectively. From this data, the file for query analysis has been created. This file contains all queries of the query log. This feature will list the keywords which have been used for in the search queries related to the guitar or string instruments. The list is made session wise. The sample file is shown in Table 3. It shows the first four entries in the query log. An excel file has been made by splitting the query terms into cells assigning them values. If the keywords searched are less than five, then that attribute has no value. The resulted file (File 2) of this feature will include only those sessions which have at least one searched keyword.

In next section the web mining algorithms will be applied on these two file namely 'File' and 'File 2'.

6 Experiments and results of the log analysis

Two files have been created for two new feature sets extracted in previous section, namely 'File 1' and 'File 2' on which further web mining techniques have been

Table 3 A sample file of query log in each session

SID	K1	K2	K3	K4	K5
S10	Guitar	Online	Shopping		
S34	Guitar	course	In	Rewari	
S50	Online	Guitar			
S60	Online	Lessons	For	Electric	Guitar

Table 4 Linguistic distribution of the hits

Languages	Sessions		Page hits	
English	12,371	97.98%	22,397	94.79%
Other languages	389	3.04%	1229	5.20%
Total	12,760		23,626	

implemented to discover interesting and useful information.

6.1 Clustering results: on File 1

During the data cleaning phase, the guitar log file collected from the website with 24,945 entries has been cleaned. After removing the irrelevant data, the cleaned log file contains 23,626 entries shown in Table 1. The user and the session identification phase grouped the 23,626 entries into 12,334 users and 12,760 sessions, respectively. The session and page hit distribution on the different languages are shown in Table 4. The table shows that the English language of webpages of guitar website has been accessed more frequently than any other language. Hence, the web pages of English have been divided into more than one part and the webpages of other languages have been considered in a single.

The results that have been shown in the confusion Table 5 are obtained as a result of applying the crossed clustering method seven classes of the pages and the five classes of visits. The set of results obtained for the English language have been divided into 6 classes (E_HE, E1_M, E2_SC, E3_DD, E4_OS, and E5_SN). E_HE pages contain home page and the first page of the English language. E1_M contains the information like, sitemap, contact, mailing address etc. pages. E2_SC contain the webpages relating to the software and courses. E3_DD contains the audio, video, free trial and the downloaded pages. E4_OS contains the pages of an online store. E5_SN contain the pages related to the social networks. Olang contains all the web pages of other languages. This website is in four languages (Spain, Dutch, French, and English).

The percentage of clicks on the pages of the English language is 94.79% and just 5.20% on all the other language pages. From this statistics one thing is very clear, that people using this website prefer English language webpages for learning or buying a guitar or its accessories.

As mentioned earlier the visits are divided into 5 classes or continents. V_Asia contain all the visits or sessions from the Asian users. It has only 92 clicks out of 6991 clicks which accounts for 1.31% of the other language webpages out of 6991 clicks. Whereas 98.68% clicks are on the English webpages.

Regarding the statistics about the visitors, 16.87% of the visits are from Australia, South America, and Africa and 83.97% of the visits are from Europe, Asia, and North America. From the analysis result, it has been concluded also found that maximum visits are from Europe (30.83%) than from Asia (29.83%) and North America (22.70%) respectively shown in Fig. 3. Considering the statistics country wise, the maximum visits are from France. On further analysis, it is found that the website has been developed and maintained by an eminent French guitarist, so it can be safely concluded that this could be the strong reason for the highest number of clicks from France. The software and the courses available on this website are written by Amar Guerfi who has been a famous guitar Player since the late 1970s. Also, the maximum number of files has been downloaded from the European visitors. These visitors have been mostly visiting the home page, with very low clicks on the other pages, Asian visitors have been in the second place in terms of visits. The least number of visits are from Australia. Maximum clicks have been on the pages with software and courses are from Asia (47.48%). Clustering analysis of the EM algorithms

Table 5 Confusion table

Page Groups	E_HE	E1_M	E2_SC	E3_DD	E4_OS	E5_SN	OLang	Total
Visit Groups								
V_Asia	3466	675	1914	289	503	52	92	6991
V_Austr	146	27	44	5	14	0	2	238
V_SA	1635	58	248	163	32	2	559	2697
V_NA	2879	592	867	358	315	89	265	5365
V_Africa	533	25	113	271	46	12	51	1051
V_Europe	4786	414	845	599	335	45	260	7284
Total	13,445	1791	4031	1685	1245	200	1229	23,626

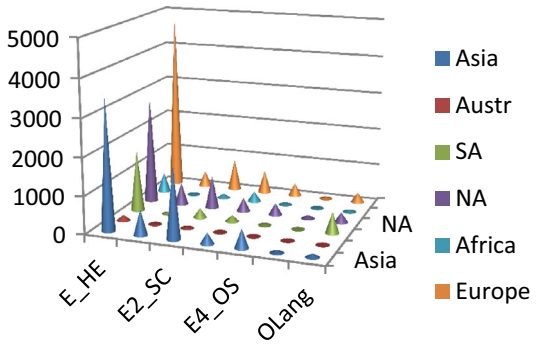

Fig. 3 Shows clustering analysis of page and visit classes

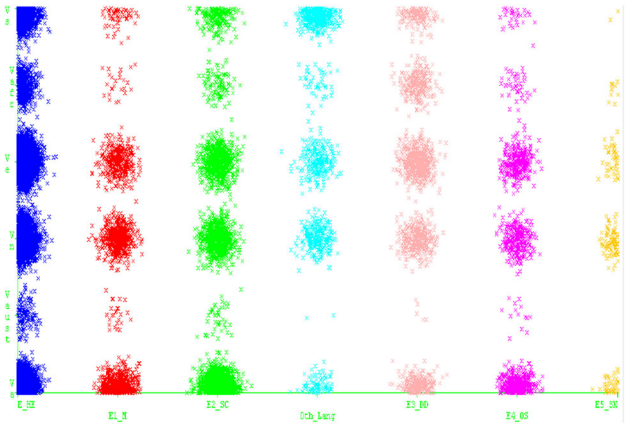

Fig. 4 Clustering analysis of EM algorithm between visits and visited pages

between the visits and the visited pages are also shown in Fig. 4.

Thus, it can be concluded that this guitar website receives most of the visitors on English web pages of the website and out of which, maximum are on the home page of the website. Also, the guitar log file contains visitors from 177 different countries out of which, the maximum visitors are from France, USA and India.

6.2 N-gram method: on File 2

In this section, the various attributes of the query data log files have been discussed along with the results obtained from the analysis which is presented by generating statistical description about the queries and user search sessions. The text analysis has also been performed on the queries submitted by the users to identify the most commonly used query and words specific to the domain and the possible relations between these words and the terms that have been used.

The cleaned 'guitar' log dataset consisting of 23,626 transactions has been shown in Table 6, out of which 786 transactions are the search queries and out of these 786, 222 are the empty queries, then left with only 563 queries for further analysis. Out of the 563 transactions of the search queries, there are 417 are unique users and 423 are the unique sessions. All the queries have been broadly classified into three groups—the unique, repeated and the blank queries. The string of terms/words used in a query which do not match the words of any other query, or which have been modifications of the previous queries, have been grouped under unique queries. Queries may also be repeated due to visiting/viewing of the result pages subsequently by the users. The blank or empty queries are those, which are without any terms. During this analysis, it has been found that 222 queries do not contain any keyword. Thus, these queries have been considered as empty queries. The Table 7 shows that out of the 786 queries, 145 (18.45%) are unique queries, 419 (53.30%) are repeat queries and 222 (28.24%) are the empty queries. The mean of the queries of the log is 1.66 (with a median of 1).

Further, the analysis of the distribution of numbers has been done in order to do the detailed study of the number of queries per session. As shown in Fig. 5, 74.46% (i.e. 315) of the sessions contain only one search query; 20.56% (i.e. 87) of the sessions contain two, 3.07% (i.e. 13) of the sessions contains three search queries, 1.18% (i.e. 5) of the sessions contain four search queries and 0.71% (i.e. 3) of the sessions contain 5 search queries. From the results, it

Table 6 Statistics on queries

Total number of search queries	786
Total number of unique queries	145
Total number of repeat queries	419
Total number of empty queries	222
Mean number of queries	1.66
Median number of queries	1

Table 7 Statistics on query terms

Total number of terms	1341
Total number of unique terms	162
Total Number of repeated terms	1179
Average number of terms per query	3.16
Median number of terms per query	3
SD of terms in query	1.9
Largest number of terms per query	10
Queries with one term	6.73%
Queries with two terms	20.39%
Queries with three terms	46.45%
Queries with four terms	11.87%
Queries with five terms	7.97%
Queries with six or more terms	6.38%

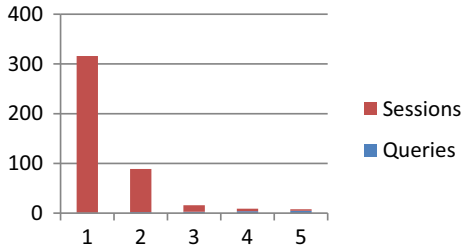

Fig. 5 Query distribution among sessions

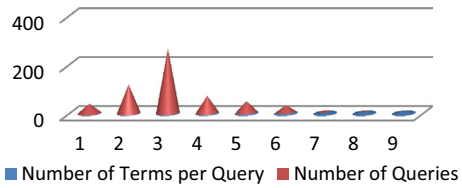

Fig. 6 Distribution of terms per query

can be safely concluded that, in terms of the number of queries submitted, the distribution is bent or skewed towards the lower end. As 95.03% of the users submitted only one or two queries, there is a possibility that the users might be very clear about what they have been looking for and hence have been able to better formulate their queries and thus have been obtained the relevant result from their first query, so they need not resubmit it.

The in-depth analysis of the terms used in the search query can reveal very useful information like 'HOW' the user formulates their queries. The statistics regarding these have been discussed below and also shown in Table 7: there are 1341 query terms out of which 162 are unique and 1179 are the repeated. The longest query contains 10 terms with a frequency of 1. The median of the query terms is 3 and the mean is 3.16. The distribution of the number of terms per query is shown in Fig. 6. When the queries are classified according to the number of terms used the results are as follows- Single terms queries are 6.73, 20.39% contain two, and 46.45% contain three terms and in 93.43% of cases queries having five or fewer terms so, from the results obtained it can be concluded that the users generally prefer short queries.

To study the terms and their relationship with each other, the n-gram algorithm has been applied on the query log. The researcher has created 5 files namely 1-gram, 2-gram, 3-gram, 4-gram, and 5-gram. From this analysis, suggestions to update existing string instrument ontology have been obtained, which have helped to create new or update existing classes and their relationships.

It will help us to update following:

- To add new leaf concepts in a hierarchy.
- To add a sub tree of concepts in the hierarchy.

Table 8 Top most frequently used 1 and 2 gram query terms with frequency

Term	Freq.
(a)	
Guitar	370
Online	247
Tuner	94
Multi	66
Instrument	66
Software	58
Learn	58
Shop	48
Metronome	48
Electric	45
Free	44
Play	41
(b)	
Guitar online	120
Multi instrument	66
Instrument tuner	66
Electric guitar	45
Learn electric	42
Online guitar	37
Online shop	33
For guitar	22
To play	17
Free	16

- To add a new relationship between existing or new concepts.

These files show the single terms, terms in a pair, three word terms, four word term and five word terms with their frequency. The sample files in this chapter show the 10 most used single terms, terms in a pair, three word terms, four word terms and five terms with their frequency.

In the end, the comparison of these individual terms has been done with data set of the guitar ontology. If there is any term used by most of the users, which is not present in the ontology then the researcher finds its frequency in 2-gram, 3-gram, 4-gram and 5-gram also check that how and in which context it has been used and try to find out the co-relation of that term with other terms.

The sample of 1-gram file is shown in Table 8a. The word "guitar" has been used 370 (27.48%) times in queries; "online" has been used 247 (18.35%) times and so on. Most frequently used 2-gram terms have been shown in Table 8b.

The results are almost similar to 1-gram. The top 2 terms of 1-gram have been used most frequently (i.e. 120 times in queries) in 2-gram. Similarly, the words "multi" and "instrument" used in 1-gram have been used together in bi-gram (66 times). Tri-gram and tetra-gram tables of 10 most

Table 9 Top 10 most frequently used 3 and 4 gram query terms with frequency

Terms	Freq.
(a)	
Multi instrument tuner	66
Learn electric guitar	42
Guitar online shop	27
Free metronome	16
Online metronome for	15
Metronome for guitar	15
For guitar free	13
How to play	13
How to play	12
To play the	12
(b)	
Online metronome for guitar	15
Metronome for guitar free	13
How to play the	12
To play the guitar	12
Play the guitar vol 1	9
How to play guitar	7
How to play the	5
To play the guitar	5
Play the guitar vol 1	5
Online tuner for guitar	4

Table 10 Top 10 most frequently used 5-gram of query terms with frequency

Terms	Freq.
Online metronome for guitar free	13
How to play the guitar	12
To play the guitar vol 1	9
How to play the guitar	5
To play the guitar vol 1	5
Learn to play guitar software	3
How to play guitar free	3
Publisher of the online guitar	2
Classical pieces for guitar vol 1	2
Play perfect music practice software learn	2

Table 11 Analysis of metronome concept

Terms used with metronome	Freq.
Metronome	
Free metronome	16
Online metronome for	15
Metronome for guitar	15
Online metronome for guitar	15
Metronome for guitar free	13
Online metronome for guitar free	13

Table 12 Analysis of guitar concept

Query prediction for term guitar	Freq.
Guitar	
Guitar online	120
Online guitar	37
Electric guitar	45
Learn electric guitar	42
Guitar online shop	27
Metronome for guitar	15
Online metronome for guitar	13
Metronome for guitar free	13
Online metronome for guitar free	13
Online tuner for guitar	4

frequently used terms have been shown in Table 9. The phrase "multi instrument tuner" has a maximum frequency (i.e. 66 times), "Learn electric guitar" has the second highest frequency i.e. 42 times (See Table 9a). In the tetra-gram table the "online metronome for guitar" has a frequency as 15; "metronome for guitar free" has a frequency as 13 shown in Fig. 9b. The 5-gram of terms is shown in Table 10, where "Online metronome for guitar free" is used 13 times in queries.

The multi and instrument terms have been used 66(100%) times individually and together in 1-gram and 2-gram, respectively. That means whenever these words have been used they have been be used together. The term tuner has been used more than 70% of the times with multi and instruments in the tri-gram. The maximum chances are that tuner will be used with multi instruments. Therefore a new class 'multi instrument tuner' and tuner need to be added.

The word metronome has been used 48 times in 1-gram and in 2-gram, 3-gram, 4-gram, and 5-gram, it has been used for the maximum number of time either with the guitar or online or with both (see Table 11). The Table 12, below shows the frequency with which the word "guitar" has been used with other words in the search queries.

Complete queries have also been analyzed and their frequency has been calculated. The top ten queries used by the users have been shown in Table 13.

From the Table 13 it is evident that the query "guitar online" and "multi instrument tuner" have been requested the most number of times. Most of the people are interested in searching for Electric guitar than any other type of guitar. People are also interested in metronome instrument used for guitar. As a result, the researcher has retrieved the

Table 13 10 most frequently queries used for search

Query	Freq.
Guitar online	69
Multi instruments tuner	66
Learn electric guitar	42
Guitar online shop	26
Free metronome	16
Online guitar	16
How to play the guitar vol 1	14
Online metronome for guitar free	13
http://www.guitar-online.com	11
Guitar online	9

following suggestions. The term metronome, multi, instrument, the tuner does not exist in our ontology. After studying the suggestions, the researcher got to know that metronome is a practice tool that produces a steady pulse (or beat) to help musicians play rhythm accurately. The metronome can be used for piano, drums etc. The terms multi and instrument has been used together. The terms multi instrument has been also used for guitar accessories like multi-instrument gig-bags, multi instrument cases or multi instrument tuner etc.

The term metronome and multi instrument do not exist in the present ontology as shown in the Fig. 2. Therefore these terms can be added to the ontology by creating new concepts. As per the suggestions, the researcher has added 'metronome', 'multi-instrument-tuner' and tuner to the ontology. The updated ontology is shown in Fig. 8.

6.3 Apriori algorithm: on File 2

The association rules generated by the Apriori algorithm have been analyzed to find the interest of the visitors. Therefore after the n-gram method, the researcher has applied the Apriori algorithm on query log in the file 2 which contains all the queries made by users in each of the sessions. They have been obtained on the complete log file with minimum supports.

The results of Weka Apriori [W2] program using query log are almost similar those obtained with the n-gram method shown in Fig. 7. It also shows the similar type of rules that have been interpreted as the First rule is—if the term 'instrument' (66 times) has been used then, 'multi' (66 times) has also been used with it. The confidence of this rule is 100%. Rule 9 says that if the terms 'multi' and 'instrument' have been used together then the confidence of using the third term 'tuner' is 100%. Rule 14 says that

```
Apriori :
=======
Minimum support: 0.07 (39 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 19

Best rules found:
 1. K2=instruments 66 ==> K1=multi 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
 2. K1=multi 66 ==> K2=instruments 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
 3. K1=multi 66 ==> K3=tuner 66    <conf:(1)> lift:(7.51) lev:(0.1) [57] conv:(57.21)
 4. K2=instruments 66 ==> K3=tuner 66    <conf:(1)> lift:(7.51) lev:(0.1) [57] conv:(57.21)
 5. K2=instruments K3=tuner 66 ==> K1=multi 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
 6. K1=multi K3=tuner 66 ==> K2=instruments 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
 7. K1=multi K2=instruments 66 ==> K3=tuner 66    <conf:(1)> lift:(7.51) lev:(0.1) [57] conv:(57.21)
 8. K2=instruments 66 ==> K1=multi K3=tuner 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
 9. K1=multi 66 ==> K2=instruments K3=tuner 66    <conf:(1)> lift:(8.53) lev:(0.1) [58] conv:(58.26)
10. K2=electric 42 ==> K1=learn 42    <conf:(1)> lift:(11.04) lev:(0.07) [38] conv:(38.2)
11. K2=electric 42 ==> K3=guitar 42    <conf:(1)> lift:(9.71) lev:(0.07) [37] conv:(37.67)
12. K2=electric K3=guitar 42 ==> K1=learn 42    <conf:(1)> lift:(11.04) lev:(0.07) [38] conv:(38.2)
13. K1=learn K3=guitar 42 ==> K2=electric 42    <conf:(1)> lift:(13.4) lev:(0.07) [38] conv:(38.87)
14. K1=learn K2=electric 42 ==> K3=guitar 42    <conf:(1)> lift:(9.71) lev:(0.07) [37] conv:(37.67)
15. K2=electric 42 ==> K1=learn K3=guitar 42    <conf:(1)> lift:(13.4) lev:(0.07) [38] conv:(38.87)
16. K2=online 109 ==> K1=guitar 103    <conf:(0.94)> lift:(3.39) lev:(0.13) [72] conv:(11.23)
```

Fig. 7 Shows the output from WEKA Apriori program using the query log

the confidence of using "learn", "electric" and "guitar" together is 100%. Similarly, last rule is that the term 'online' is used with the 'guitar' term with a confidence level of 94%.

7 Ontology improvement

This work has given us the suggestion for three new classes 'metronome', 'multi-instrument-tuner' and 'tuner' to be incorporated in the ontology. The OWLviz visualization of the updated ontology has been shown in Fig. 8. After studying about these two suggested concepts, the researcher added a new class string_instrument_tuner to the ontology. Therefore muti_instrument_tuner has become the subclass of this class. The OWL visualization of the new class string_instrument_tuner is shown in Fig. 9.

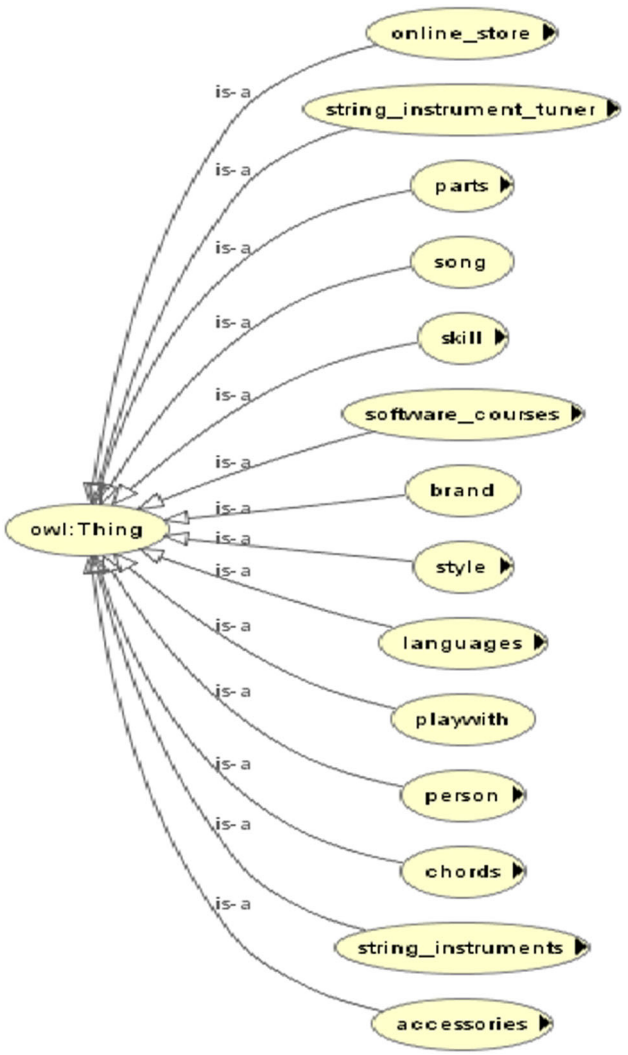

Fig. 8 Updated ontology

Object properties for this class are is_used_to_tune and tuned_by.

The usage of these new object properties has been shown in Fig. 10. The new term 'tuner' has been added.

8 Contributions

- A novel method to update the general ontology from log data.
- New feature selection for log analysis.
- Shows the relationship between Semantics web and Web usage mining.

Hence the proposed methodology is computationally simple and easy to deploy.

9 Conclusion and future work

Constructing ontology and its continuous improvement requires knowledge integration and updating it from varied sources, but specifically from web content belonging to a particular domain, in case of Semantic Web. During this study, the researcher has attempted to show the potential impact and use of web usage mining on updating the ontology. The researcher illustrated such an impact in the string instrument ontology in musical domain by considering the site of online guitar selling website maintained by Amar Grifu from France. The researcher has already constructed a new string instrument ontology from base using protégé 5.0 ontology editor and showed how the knowledge discovered from the analysis of specific type of log file data (referrer filed) of this domain can be immensely useful to update this ontology time to time. To prove this clustering (EM), association rule (Apriori) and sequential pattern (n-gram) mining algorithms in particular have been applied on 'guitar' log file of online guitar selling website. The original 'guitar' log file contain 24,965 transactions, after cleaning left with 23,626 transactions and 12,334 and 12,760 unique users and sessions, respectively. On this cleaned log file the researcher has applied clustering by grouping pages and visits into 7 and 6 classes, respectively and got some golden nuggets. (1) The percentage of clicks on the pages of English language is 94.79%. (2) The maximum visits are from Europe (30.83% and mostly from France). (3) Maximum downloads are from European visitors (35.54%). (4) Maximum clicks on software and courses pages are from Asia (47.48% and maximum from India). Reasons of these results are discussed earlier in clustering analysis phase.

In second experiment the researcher has extracted only those sessions from cleaned log file which contain query

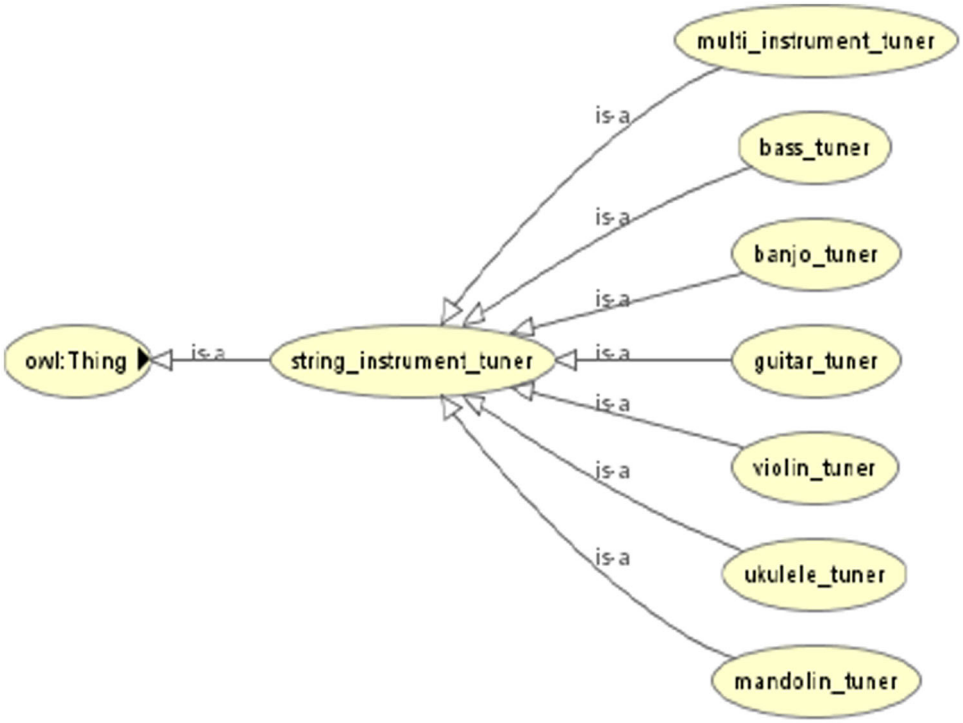

Fig. 9 OWLviz visualization of string_instrument_tuner class

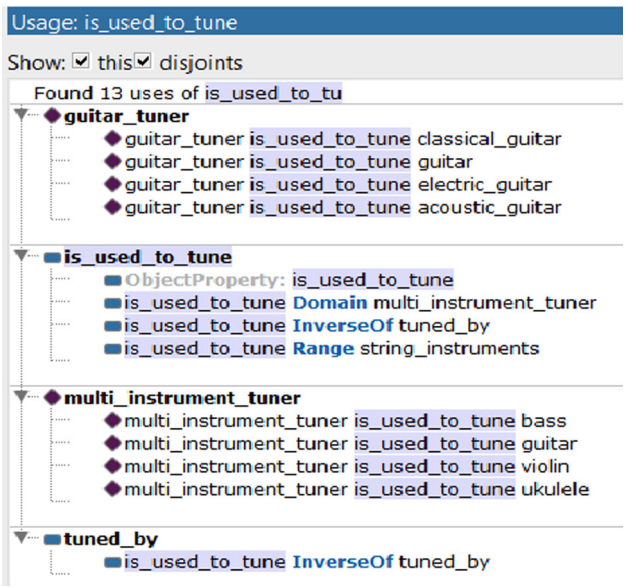

Fig. 10 Usage of is_used_to_tune object property

from any search engine in their referrer field and has come up with 786 transactions out of which 222 have been the empty queries. So after removing those transactions the researcher is left with 563 queries and 417 unique users and 423 are the unique sessions. N-gram and Apriori algorithm have been applied on this data set to get some suggestions for ontology improvement. As a result the researcher has concluded that the term metronome, multi instrument and tuner does not exist in the string_instrument ontology in Fig. 2. After adding these concepts in ontology the updated ontology is shown in Fig. 8.

The objective of the research has been to accelerate and to improve the ontology development process by semi-automatically generating a hierarchal ontology. This work has been expanded to build the semantic web from the generated string ontology, which has helped in refining the web search on generic search engines in music domain.

References

1. Kaur M, Gurm RK (2016) Survey paper on frequent pattern mining on web server logs. Int J Technol Comput (IJTC) 2(4)
2. Joshila Grace LK, Maheswari V, Nagamalai D (2011) Analysis of web logs and web user in web mining. Int J Netw Secur Appl 3(1)
3. Kaviarasa S, Hemapriya K, Gopinath K (2015) Semantic web usage mining techniques for predicting users' navigation requests. Int J Innov Res Comput Commun Eng 3(5):4261–4270
4. Ciesielski V, Lalani A (2003) Data mining of web access logs from an academic web site. In: International conference on hybrid intelligent systems, pp 1034–1043
5. Berendt B, Hotho A, Stumme G (2005) Semantic web mining and the representation, analysis, and evolution of Web space. In: Proceedings of RAWS' 2005—Workshop on the representation and analysis of web space, September 2005
6. Jiang Y, Li Y, Yang C, Armstrong EM, Huang T, Moroni D (2016) Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. ISPRS Int J Geo Inf 54(5):1–14
7. Kannan N, Shanthi E (2010) Classification and clustering of web log data to analyze user navigation patterns. J Glob Res Comput Sci 1(1)

8. Facca FM, Lanzi PL (2005) Mining interesting knowledge from weblogs: a survey. Data Knowl Eng 53(3):225–241
9. Patra R, Sunani SM (2016) A review on different computing method for breast cancer diagnosis using artificial neural network and datamining techniques. Int J Adv Res (IJAR) 4(11):598–610
10. Om Prakash PG, Jaya A (2016) Analyzing and predicting user behavior pattern from weblogs. Int J Appl Eng Res 11(9):6278–6283
11. Kumari M, Garg K (2013) Ontology approach to web mining. Int J Sci Eng Comput Technol 3(1):60–63
12. Al-Hegami AS, Kaity MS (2014) An ontology framework based on web usage mining. Int J Appl Inf Syst (IJAIS) 6(9):28–35
13. Agirre E, Ansa O, Hovy E, Martinez D (2000) Enriching very large ontologies using the WWW. In: Proceedings of ECAI workshop on ontology learning
14. Deitel AC, Faron C, Dieng R (2001) Learning ontologies from RDF annotations. In: Proceedings of the IJCAI '01 workshop on ontology learning, Seattle, WA
15. Chau M, Fang X, Liu Sheng OR (2005) Analysis of the query logs of a web site search engine. J Am Soc Inform Sci Technol 56(13):1363–1376
16. Trousse B, Aufaure MA, Grand B (2008) Web usage mining for ontology management. In: Nigro HO, González Císaro SE, Xodo DH (ed) Data mining with ontologies: implementations, findings, and frameworks, 1st edn. Information Science Reference, pp 37–64
17. Kaur N, Aggarwal H (2017) A novel semantically-time-referrer based approach of web usage mining for improved sessionization in pre-processing of web log. Int J Adv Comput Sci Appl (IJACSA) 8(1)
18. Kaur N, Aggarwal H (2017) Evaluation of information retrieval based ontology development editors for semantic web. Int J Modern Educ Comput Sci (IJMECS) 9(7):63–73
19. Protége Ontology Editor, Retrieved on Dec, 2016 from Stanford University School of Medicine. http://protege.stanford.edu/