ISSN:0975-9646
Pushpa C N et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 8170-8177

EMET: Extracting Metadata using ElementTree to Recommend Tags for Web Contents

Pushpa C N[#1], Shankar R[#1], Thriveni J[#1], Venugopal K R[#1], L M Patnaik[*2]

[#1]*Department of Computer Science and Engineering,*
University Visvesvaraya College of Engineering,
Bangalore University, Bangalore, India.

[*2]*Indian Institute of Science,*
Bangalore, India

***Abstract*—Web search has become an important task for many individuals. As there is rapid growth of the internet, effective searches plays a vital role. Most of us, however, have tough frustration in making an attempt to search for something on the online. Metadata can be used to facilitate the searching of Web contents. We have proposed an algorithm to extract Metadata using ElementTree [EMET], new search methodology to provide keywords recommendation for Web user contents. The user is guided by an inventory of active keywords that is recommended dynamically throughout the search by a search engine. This active keyword list helps the user to select keywords that are more relevant to the search through recognition. The proposed EMET algorithm yields the average of 0.934 of Precision, 0.927 of Recall and the 0.93 of F-Measure.**

***Keywords*—** ElementTree, F-Measure, Keyword-based Search, Metadata Extraction, Precision, Query Response, Recall, Tags Recommendation, Search Engine.

I. INTRODUCTION

Web pages are typically full of free form text that is straightforward for humans to browse however difficult for computers to understand. Some sites have information with large structure that's straightforward to browse, similar to a page date embedded within the computer address or title of the page, or machine-readable fields embedded within the mark-up language code. Google extracts a spread of structured knowledge from sites.

Metadata is "data regarding data" or the knowledge about a video file, audio clip, or Web content. The two types of metadata are structural metadata and descriptive metadata. The Structural metadata is the design and specification of data structures or "data about the containers of data"; and descriptive metadata is the individual instances of application data or the data content.

Information will be used to facilitate the searching out of video content either by people or firms hosting their videos on their own Websites, moreover as for videos that are uploaded and hosted on video search engines and video sharing sites like Google, iTunes, YouTube, MySpace etc.,

Metadata could be a label on the video and it doesn't have an effect on the video content itself. However once data is formed or extracted from a video file, relevant keywords and descriptions will then be preferred to drive effective Search Engine Optimization (SEO). The metadata is one vital feature utilized by search engines to rank content in a search directory. Metadata is employed to obtain short descriptions of the online page(s) or videos within the actual search results that improves the quantity and quality of traffic to an internet Web site from the various search engines. For the past seven years, on-line enrolments [1] are growing rapidly than the overall pedagogy enrolments. The expectations of educational leaders have been that on-line enrolments would continue their substantial growth for a minimum of another year.

There are some key challenges encountered while submitting the files for mobile phones, podcasting and Web portals by the creators and the distributors of videos. Hence the significant metadata or keywords for video have to be created or extracted to find the video easily in a searchable environment. After the metadata has been created or extracted, they need to provide the video into the searchable repositories, in the right format, with the right metadata attached. This paper gives the information on the approach and solutions to address these key challenges by the video content creators.

Video resources ought to be represented precisely [2]. It is troublesome to use only one general description to accurately tell the entire story of a video as a result of one section of the video stream could have much information. However a number of them may not relate the main points of the video once it was created. Therefore, the conventional paragraph-based description method is not sensible for annotation videos exactly. Based on the timeline of the video stream a lot of correct description mechanism is needed.

There are several ways to extract and create metadata from a video:

1. Extract metadata from closed captions and other embedded video information from a source file. For example when an individual in a film is talking regarding flying airplanes-those words are regenerated to text. Operators will then search the extracted text manually, or through the use of automated software that will search for relevant keywords. The keywords extracted from this document are later used for searching.

2. Re-use existing metadata, when it is present. Previously some video formats may contain metadata such as descriptions of individual scenes mainly those which are used for web streaming or editing. It is possible to preserve and extract this information to facilitate searching in certain cases.

3. Create annotations or tagging – alternative way to create metadata from a video is to have an operator manually type information as the video is being viewed. This can be done in a separate database, MS Word, or notepad file, or with automated software and then link the created text to the video file.

The Semantic Web could be a vision to solve this drawback. A new WWW architecture can support not only Web content, but also associated with formal semantics. The concept is that the Web content and additional semantics or metadata will be accessed by Web agents, allowing these agents to reason about the content and produce intelligent answers to user's queries. Multimedia plays a vital role in education, particularly for distance learning environments. With the rapid growth of the multimedia Web, huge numbers of educational resources are increasingly being created by several organizations. It is essential to explore, share, reuse and link these educational resources for enhanced e-learning experiences. Most of the video resources are presently associated in an isolated way, which implies that they lack semantic connections. Thus, providing the facilities for annotating these video resources is much demanded. These facilities produce the semantic connections among video resources and permit their metadata to be understood globally.

Motivation: In order to optimize the video file for semantic search as much as possible, the video file must contain as much as metadata about a file. Since the administrator has not provided the metadata, we are trying to populate tags for a video file by extracting the metadata from the Google Search Engine, the relevant keywords and descriptions can be selected to drive effective Search Engine Optimization (SEO).

Contribution: The objective is to create or extract the metadata for contents of Web automatically and recommending tags to the Web User and to find the Precision, Recall and the F-measure metrics of the system. We have proposed an Extracting Metadata using Element Tree Method (EMET) algorithm to extract metadata automatically for Web contents and recommend to the Web user and make it possible to preserve and extract this information to make web search easily.

Organization: The remainder of the paper is organized as follows: Section II reviews the related work of the existing works of the extraction of metadata, Section III explains the proposed system architecture, and Section IV gives the problem definition and the proposed algorithm. The implementation and descriptions of an Element Tree is given in Section V, the Performance Analysis and the results of the system are described in Section VI and Conclusions are presented in Section VII.

II. RELATED WORK

The education sector is facing dramatic changes with the role of universities quickly evolving from a place wherever information is created, delivered and licensed, to a (virtual) organization dedicated to the distribution and exchange of knowledge. Besides Open Universities, more and more institutions are dedicating a considerable a part of their activities within the production and distribution of (more or less) open educational material, moreover as in advertising their courses/qualifications on the far side their usual territory. It is in this context that the employment Linked Data [3] looks the most relevant. Indeed, whereas the content of open educational resources is by definition accessible and reusable, while the corresponding metadata as Linked Data will create the content of various repositories more discoverable, accessible and connectable.

Kovacevic et. al, [4] develops a system for automatic extraction of metadata from scientific papers in PDF format for the information system for monitoring the scientific research activity of the University of Novi Sad (CRIS UNS). The system is based on machine learning and performs automatic extraction and classification of metadata in eight pre-defined categories. The extraction task is realised as a classification process. Sergei Brin [5] has done interesting work on automatic extraction of bibliographic information from the Web. Brin relies on incrementally refining grammar definitions, similar to what we do when parsing text for reference anchors.

The Boston University College of Engineering Distance Learning Initiative (DLI) integrates computers, digital video and therefore the Internet to deliver graduate degree courses in engineering students in corporations distant from the Boston University campus. A key objective of the DLI is to support learning where it is most convenient, whether or not by groups in a classroom, at the workplace desk, or at home. The article describes the motivation, technology, and experiences in integrating a satellite-based digital video distance learning system combined with Web technology [6].

The use of common vocabularies to describe the topics treated is a way to make the content of these repositories addressable globally, and to retrieve resources first for their relevance, without having to consider their origin. An algorithm called Link Selection Algorithm [7] is an example used for vertical search engine proposes to collect the high topic relevance pages in the specific domain. There are two factors in this algorithm, one is expanded meta-data topic relevance score, which is calculated by combining analysis of link content with hyperlink structure characteristics; the other one is inherit score, which is the influence of father pages' topic relevance score calculated by authority value and hub value. Experiment results indicate that the spider which uses the proposed algorithm can get a high topic relevance search collection. In the past, many researchers have exploited ontologies to perform semantic annotation and retrieval from digital video libraries [8].

Ling Zheng et. al, [9] proposed a novel idea of creating a framework for multimedia content extraction and retrieval semantic Web. Semantic Web has been used for indexing of textual data. Swoogle [10] is a semantic Web based search engine for text documents. Han et. al, [11] describes a Support Vector Machine classification-based method for metadata extraction from header a part of research papers and show that it outperforms other machine learning methods on identical task. Kawtrakul and Yingsaeree [12] describe a framework for automatic metadata extraction from electronic documents that can be both text documents and images of paper documents to ease metadata creation process. The system consists of three main components: a text conversion module for converting electronic document into standard text file format, a task-oriented parser module for automatically extracting metadata from converted text using pre-defined grammar, and data verification module for identifying and correcting the errors in extracted metadata.

Li et al, [13] proposed CiteSeerx, is a scientific literature digital library and search engine which automatically crawls and indexes scientific documents in the field of computer and information science. Another framework for reference metadata extraction is using a hierarchical knowledge representation framework presented in [14]. Parse Citation (ParsCit) [15] performs Reference string parsing and logical structure parsing of scientific documents. ParsCit uses Conditional Random Fields as its learning mechanism. Mendeley [16] uses Support Vector Machines and Web based lookup for Extraction of embedded metadata and extraction of citation details from research publication. Mendeley provides windows based application that helps in organizing and collaboration of research publication. When research publication is added, Mendeley identifies the Author, Title, Journal, Year and keywords.

The paper [17] describes a method of speeding up the formalization and integration of new metadata. The method takes advantage of the fact that databases are often described in web pages containing natural language glossaries that define pertinent aspects of the data. Given a root URL, the method identifies likely glossaries, extracts and formalizes aspects of relevant concepts defined in them, and automatically integrates the new formalized metadata concepts into a large model of the domain and associated conceptualizations.

As the Web grows more complex and more numerous, there is a need for methods of metadata extraction to improve information retrieval from the Web.

III. System Architecture

The Metadata is one important aspect used by search engines to rank content in a search directory. Metadata is used to generate short descriptions of the Web page(s) or videos in the actual search results, which improves the volume and quality of traffic to a Web site from the various search engines. Fig. 1. shows the Architecture of Recommendation System for extracting the Metadata to populate the tags for any Web contents. Once the user inputs a keyword, then Metadata Parser will extract relevant keywords from the top Google web links via it's ajaxapi. Later, these extracted keywords will be provided to the user to upload his video files, any applications or to find any relevant keywords to his web contents.

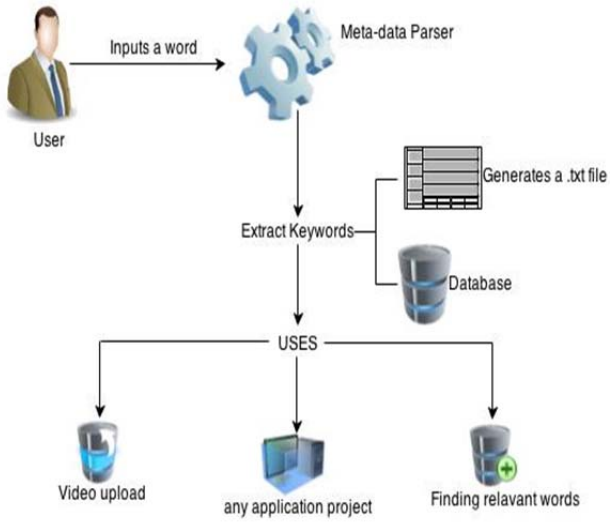

Fig. 1. System Architecture

The searching of the metadata for the video files gives more efficient results when the retrieved search results are stored in a tree pattern in file or a database using etree module provided by Element Tree API [18]. This structure is then searched for all the relevant information related to the file and the collected data is appended to the metadata of the video file. The search is performed online and appends huge chunks of collected metadata to the file. It uses lxml module which is XML toolkit in a Python [19] binding for the C libraries libxml2 and libxslt.

TABLE I : Extracting Metadata using Elementtree Algorithm (EMET)

```
Algorithm: EMET (query)
BEGIN
query← read("What do you want to search for ? ")
query = urllib.parse.urlencode( {'q' : query } )
response = req.urlopen (url + query ).read()
json_input=response
data = json.loads (json_input.decode('utf8') )

results = data [ 'responseData' ] [ 'results' ]
for result in results:
        url.append(result['url'])
for x in range(0, 10):
try:
f = req.urlopen( "%s" %url[x]).read()
tree = etree.HTML( f)
extracted=(tree.xpath("//meta[@name='keywords']")[0].get("content"))
meta.append(extracted.lower())
single=list(set(meta))
except IndexError:

try:
extracted2=tree.xpath("//meta[@name='Keywords']")[0].get("content")
meta.append(extracted2.lower())
single=list(set(meta))
except    IndexError:
pass
END
```

IV. Problem Definition and Algorithm

Extract the metadata information from the large number of keywords that are stored in the Google search engine. Our objectives are:

(i) To recommend tags automatically to the Web Users.
(ii) To increase the accessibility i.e., Effectiveness of searching can be significantly enhanced through the existence of rich and consistent metadata.
(iii) To improve the Precision, Recall and the F-measure metrics of the system.

The Table I give the algorithm to extract the metadata keywords using ElementTree approach for the given query words. It first takes the user's query and pings the url - http://ajax.googleapis.com/ajax/services/search/Web?v=1.0 & to get json response.

This response data has to be padded with other top Google Web links after searching the relevant metadata keywords. The whole process is achieved by parsing the HTML content and appending the DOM's (Document Object Model) elements to etree data structure. The first try block searches for the metadata with keywords as its content while the second try block searches for keywords just to make sure to cover the alphabetical cases.

V. IMPLEMENTATION

In receiving the desired result, a simple ajaxurl http://ajax.googleapis.com/ajax/services/search/Web?v=1.0 & gets the Google search services via the ajaxapi's. We are using Python, an Open source interpreted language to extract the metadata for a given keyword from the user. In the code all the retrieved metadata keywords gets appended to the etree. This unit is displayed as an output in the case of above program which can be extended to store the metadata to the file and thereby increasing the information of the file. The entire operation gets stored in a database as a **sqlite** application file with a **.db** extension. Also it gives us the separate text file containing metadata.

The search is performed iteratively in order to improve the time taken to search and append huge chunks of collected metadata. The data collected can be stored in a heap structure for even greater efficiency. It uses lxml module which is XML [20] toolkit is a Python binding for the C libraries libxml2 and libxslt. It is unique in that and it combines the speed and XML feature completeness of these libraries with the simplicity of a native Python API, mostly compatible but superior to the well-known ElementTree API.

The Element Tree is a flexible container object, designed to store hierarchical data structures in memory. The type can be described as a cross between a list and a dictionary. Each element has a number of properties associated with it:

- A tag. This is a string identifying what kind of data this element represents.
- A number of attributes, stored in a Python dictionary.
- A text string.
- An optional tail string.
- A number of child elements, stored in a Python sequence

lxml.etree tries to follow established APIs wherever possible. Sometimes, however, the need to expose a feature in an easy way led to the invention of a new API. This page describes the major differences and a few additions to the main ElementTree API. lxml.etree follows the ElementTree API as much as possible, building it on top of the native libxml2 tree.

A. Advantages of Element Tree

- ElementTree is much easier to use, because it represents an XML tree (basically) as a structure of lists, and attributes are represented as dictionaries.
- ElementTree needs much less memory for XML trees than DOM and thus is faster.
- ElementTree has only a small feature set compared to full-blown XML libraries, but it is enough for many applications.

Compared to the original ElementTree API, lxml.etree has an extended tree model. It knows about parents and siblings of elements.

For example,

```
>>>root = etree.Element("root")
>>> a = etree.SubElement(root, "a")
>>> b = etree.SubElement(root, "b")
>>> c = etree.SubElement(root, "c")
>>> d = etree.SubElement(root, "d")
>>> e = etree.SubElement(d, "e")
>>>b.getparent() == root
True
>>>print(b.getnext().tag)
c
>>>print(c.getprevious().tag)
b
```

Elements always live within a document context in lxml. This implies that there is a notion of an absolute document root. You can retrieve an ElementTree for the root node of a document from any of its elements.

B. Other Features

Note that this is different from wrapping an Element in an ElementTree. You can use ElementTrees to create XML trees with an explicit root node:

```
>>>tree = etree.ElementTree(d)
>>>print(tree.getroot().tag)
d
```

ElementTree objects are serialised as complete documents, including preceding or trailing processing instructions and comments. All operations that you run on ElementTree (like XPath, XSLT, etc.) will understand the explicitly chosen root as root node of a document. They will not see any elements outside the ElementTree. However, ElementTrees do not modify their Elements. The rule is that all operations that are applied to Elements use either the Element itself as reference point, or the absolute root of the document that contains this Element.

C. Locating Required Tags

After the etree formation using lxml module is done, the required html/xml tags have to be located in the tree. This is done using xpath. Now what exactly is XPATH?

- XPath is syntax for defining parts of an XML document
- XPath uses path expression
- XPath contains to navigate in XML documents and library of standard functions
- XPath is a major element in XSLT
- XPath is a W3C recommendation

XPath uses path expressions to select nodes or node sets in an XML document. These path expressions look very much like the expressions you see when you work with a traditional computer file system. In our case the XPath which we are looking for is the one that contains all the keywords of the Metadata.

This is obtained by using python command
tree.xpath("//meta[@name='keywords']")[0].get("content")

This command basically goes to the tag
<meta content="keywords" name="keywords"/> under<head></head>tags.

VI. Performance and Analysis

We tested our system with 1000 queries and analysed through parameters like Precision, Recall, Accuracy and F-measure.

In the field of information retrieval, **precision** is the fraction of Retrieved keywords that are Relevant to the find:

$$Precision,\ P = \frac{|\{\text{Relevant Keywords}\} \cap \{\text{Retrived Keywords}\}|}{|\{\text{Retrived Keywords}\}|} \quad (1)$$

Recall in information retrieval is the fraction of the keywords that are relevant to the query that are successfully retrieved.

$$Recall,\ R = \frac{|\{\text{Relevant Keywords}\} \cap \{\text{Retrived Keywords}\}|}{|\{\text{Relevant Keywords}\}} \quad (2)$$

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F - Measure,\ F = 2\frac{PxR}{P+R} \quad (3)$$

In simple terms, high **precision** means that an algorithm returned substantially more relevant results than irrelevant, while high **recall** means that an algorithm returned most of the relevant results. The Table III shows the number of extracted metadata keywords for the given set of queries.

TABLE III: Number of Extracted Metadata Keywords of given Queries

Sl. No.	Query	Number of Metadata Keywords Extracted
1	India	117
2	XML	84
3	SQL	29
4	MIT600	41
5	JSP	12
6	JAVA	20
7	HTML	51
8	Goa	12
9	English	20
10	Amazon	41
11	C Language	37
12	.NET	9
13	ROR	32
14	Lenovo	17
15	Python	4
16	LUA	10
17	Federer	15
18	IBM	5
19	Microsoft	43
20	php	43

The Table IV compares the results of proposed system with the existing system named 'asianfoxdevelopments'. The results are comprehensive and apparent.

Sl. No	Query	Keywords from Asianfoxdevelopments	Keywords from our EMET	Remarks
1	lua	Lua, Luann, Luann De Lesseps, Luau, Luanda, Luang Prabang, Luau Food, Lua Lyrics, Luan, Luau Party (0/10)	lua, language, extension, embedding, configuration, scripting, rapid prototyping, free, source, portable (10/10)	Lua is a recent language. Our tool recognized it whereas the other didn't.
2	goa	Goat Simulator, Goarmyed, Goal, Goanimate, Goat, Goat Simulator Free, Goat Locker, Go Ape, Goalsarena, Goapele 0/10	gun owners of America, wikitravel, tourism, travel guide, hotels, restaurants, nightlife, things to do, goa, goa tse, beaches in goa 10/11	Our result showed atleast what goa meant for and its attractions whereas the other didn't meet the essential meaning.
3	Excel	Excelsior College Excel Excel Formulas Excelsior Excel lookup Excel Pivot Table Excel Drop Down List Excellence Punta Cana Excel Counting Excellent Sheep 6/10	microsoft excel online, free excel, spreadsheets, excel web app, excel templates 5/5	Excel's basic meaning is a spreadsheet. We recognized it.

Sl. No	Query	Keywords from Asianfoxdevelopments	Keywords from our EMET	Remarks
4	Theater	Theater J Theater Washington Dc Theater Church Theater Theaters In Dc Theater Seating Theater Lab Theatermania Theaters Near Me Theater Vs Theatre	silovsky, joseph, robots and robotics, theater, vanzetti, bartolomeo, sacco, Nicola, theater, parks, suzan-lori,trent, adam, theater, magic and magicians, eisenhauer, peggy, krieger, henry,russell, bill (1949-),theater,tantleff, jack zimmer, fisher, jules,rockwell, david, paradigm talent agency,tazewell, paul,condon, bill,longbottom, robert,kymal, rohan,corren, donald,new brunswick (nj),mastro, michael,walton, jim,theater,dunn, wally (1960-),george street playhouse,dolce, richard t,eisenberg, ethan,shepherd, jean,milo, gina,reese, griffin,engeman, john w, theater,theater,luker, steve,flannery, evan,northport (ny),gifts,holidays and special occasions,christmas,hanukkah,theater,austen, jane,bedlam (theater co),theater,tucker, eric,hamill, kate a (1981-),chekhov, anton,yen, stacey,epstein, brett (actor),langdon, brent (actor),knight, adam harrison,smart, mat,greenhill, susan (actress),dellapina, matthew,theater,slant theater project,st ann's warehouse (brooklyn, ny),kneehigh theater,theater,rice, emma,behar, joy,theater,mobile shakespeare unit,melrose, rob,shakespeare, william,theater,public theater,movies,art,income inequality,books and literature,theater,mutu, tam,hewitt, tom,barrett, kelli (1984-),theater,nolan, paul,chamberlain, richard,theater,actors equity assn,advertising and marketing,actors and actresses,theater,chicago (ill),stroman, susan,dancing,peck, tiler,flaherty, stephen,kennedy, john f, center for the performing arts,degas, edgar,luker, rebecca,ahrens, lynn,gaines, boyd,theater,books and literature,france,theater,libraries and librarians,shakespeare, william,europe,art,race and ethnicity,theater,bailey, brett (1967-),paris (france),silovsky, joseph,theater,here arts center,culture (arts)	This is awesome! We recognized the theater artists as well thus giving the perfect epitome of it.
5	Movie	Moviefone Movies Moviestarplanet Movie Theaters Movies 2014 Movie Trailers Movietickets Movie Reviews Movie Showtimes Movie 25	movies, films, movie database, actors, actresses, directors, hollywood, stars, quotesmovie trailers, movie times, movie reviews, movie news, celebrity interviews, new movies, red carpet, movies, theaters	Ours looked into the overall structure of the movie than focusing on any individual concept like tickets and trailers.
6	Flight	Flight Tracker Flight Club Flights Flight Status Flight Trampoline Park Flights Google Flightstats Flight Of The Conchords, Flight Tickets, Flight 370	cheap flights, airline tickets, cheap tickets, discount airline tickets, airfare, flight deals, plane tickets, airlinesdiscount airline tickets, discount airfare, discount flights, discount tickets, cheap flights, cheap flight, cheapflights, cheap tickets, cheap ticket, cheaptickets, cheap airlines, airline flights, flight deals, airfare deals, airfares, airfare, airlines, flights, flight	The results are apparent. A user would normally want the information about tickets, fares, cheap flights, airdeals than the one showed by our counter tool.
7	English	English To Spanish Translation English Premier League English To Spanish English Bulldog, English To French, English Dictionary English Banana, English Bill Of Rights, English Mastiff English To German 2/10	english lessons, free english lessons, online english lessons, english grammar, learn english, english online, online english, esl lessons, esl quizzes, english quizzes, grammar, exercises, language learning, esl, efl, toefl 15/16	Just “English ” means a user would normally like to learn it or play around with it than converting it to other languages.
8	MIT600	Mit 6006 Mit 600 Mit 6002 Mit 6004 Mit 6002x Mit 600x Mit 6005 Mit 6001 Mit 6003 Mit 600sc	computer science, computation, problem solving, python programming, recursion, binary search, classes, inheritance, libraries, algorithms, optimization problems, modules, simulation, big o notation, control flow, exceptions, building computational models, software engineering, computer science, programming languages python programming, algorithms, dynamic programming, object-oriented programming, debugging, problem solving, recursion, iteration ,search algorithms, program efficiency, order of growth, memorization, hashing, object classes, inheritance, monte carlo simulation, curve fitting, optimization, clustering, queuing networks, data sampling, computer science	Well, mit 600 essentially means a course in mit college. Our result listed out various paradigms available in that course while the other tool didn’t even realize what it is all about.

Sl. No	Query	Keywords from Asianfoxdevelopments	Keywords from our EMET	Remarks
9	mit	Mitchell And Ness Mitsubishi, Mit Mitre, Mitch Mcconnell Mitochondria, Mitt Romney Mitbbs, Mitchell Gold Mitch Hedberg	massachusetts institute of technology, mitopencourseware, mit ocw, courseware, mit opencourseware, free courses, class notes, class syllabus, class materials, tutorials, online courses, mit courses	"Mit" is a world renowned college. We apparently resulted out the information about it.
10	django	Django Unchained Django, Django Reinhardt Django Tutorial, Django Unchained Soundtrack Django Unchained Cast Django Python, Django Forms, Django Unchained Full Movie, Django 1966	python, django, framework, open-source reviews, showtimes, dvds, photos, message boards, user ratings, synopsis, trailers, credits	Just a word "django" is an open source python framework. Then comes the movie django unchained.

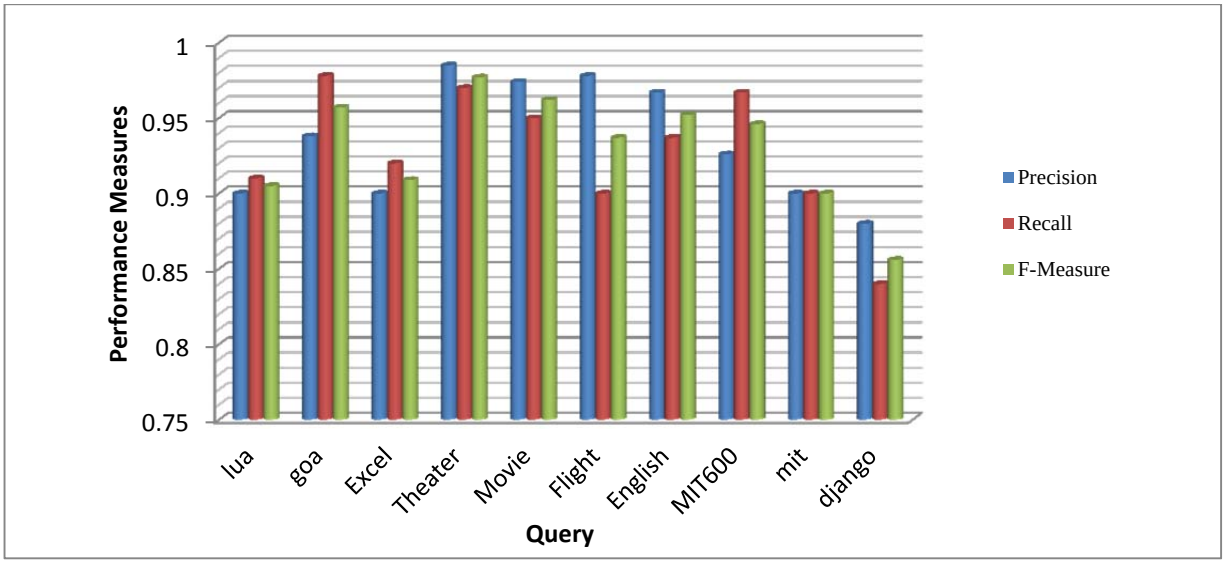

Fig. 2. Precision , Recall and F-Measure analysis of EMET Algoritm.

Table V depicts the performance analysis of the proposed system for a sample input set.

TABLE V: PERFORMANCE ANALYSIS OF EMET SYSTEM

Query	Precision	Recall	F-Measure
lua	0.9	0.91	0.905
goa	0.938	0.978	0.957
Excel	0.9	0.92	0.909
Theater	0.985	0.97	0.977
Movie	0.974	0.95	0.962
Flight	0.978	0.9	0.937
English	0.967	0.937	0.952
MIT600	0.926	0.967	0.946
mit	0.9	0.9	0.9
django	0.88	0.84	0.856
Average	**0.934**	**0.927**	**0.93**

The Fig. 2 shows the Precision, Recall and F- Measure of the EMET algorithm for 10 Sample Query words.The algorithm yields the average Precision is 93.4%, Recall is 92.7% and the F–Measure is approximately 93%

VII. CONCLUSIONS

As the Web grows more complex and more numerous, there is a need for methods of metadata extraction to improve information retrieval from the Web. The proposed Extract Metadata using ElementTree [EMET] algorithm is used to provide keywords recommendation for Web user contents. The proposed EMET algorithm can be implemented in a wide variety of areas where the availability of information about a file is very less or if the information about a file is not enough to be considered it for Semantic Web technology. The proposed EMET algorithm yields the average of 0.934 of Precision, 0.927 of Recall and the 0.93 of F-Measure.

The future work of this is to enhancing EMET algorithm with parsing the title to extract metadata along with the meta tag and the redundancy of the metadata can be improved by preprocessing the data, so that further efficiency can be improved.

REFERENCES

[1] E. Allen and J. Seaman, "Class Differences Online Education in the United States, " http://sloanconsortium.org/publications, 2010.

[2] Hong Qing Yu, Carlos Pedrinaci, Stefan Dietze, and John Domingue, "Using Linked Data to Annotate and Search Educational Video Resources for Supporting Distance Learning", *IEEE Transactions on Learning Technologies*, vol. 5, no. 2, April-June 2012.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data-The Story So Far," *International Journal on Semantic Web and Information Systems*, vol. 5, pp. 1- 22, 2009.
[4] Kovacevic, Aleksandar, Ivanovic Dragan, Milosavljevic Branko, Konjovic Zora, Surla Dusan, "Automatic Extraction of Metadata from Scientific Publications for CRIS Systems", *Electronic Library and Information Systems*, vol. 45, no. 4 pp. 376-396, 2011.
[5] Sergey Brin. Extracting Patterns and Relations from the World Wide Web. In *WebDB Workshop at EDBT '98*, 1998.
[6] J.W. Brackett, "Satellite-Based Distance Learning Using Digital Video and the Internet," *IEEE Multimedia*, vol. 5, no. 3, pp. 72-76, 1998.
[7] Ling Zheng, Yang Bo, Ning Zhang, "An Improved Link Selection Algorithm for Vertical Search Engine", *IEEE International Conference on Information Science and Engineering (ICISE)*, pp. 778-781, 2009.
[8] Marco Bertini, Alberto Del Bimbo and Giuseppe Serra, "Learning Ontology Rules for Semantic Video Annotation, " *Proceedings of the 2nd ACM workshop*, 2008.
[9] S.C. Sebastine, B.M. Thuraisingham and B. Prabhakaran, "Semantic Web for Content Based Video Retrieval," *Proceedings of IEEE International Conference on Semantic Computing (ICSC '09)*, pp. 103-108, 2009.
[10] Swoogle Search Engine: URL: http://swoogle.umbc.edu/
[11] H. Han, C. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. Fox, "Automatic Document Metadata Extraction using Support Vector Machines," *Proceedings of Joint Conference on Digital Libraries*, pp. 37-48, 2003.
[12] A. Kawtrakul and C. Yingsaeree, "A Unified Framework for Automatic Metadata Extraction from Electronic Document", Proceedings *of the International Advanced Digital Library Conference*, pp. 71-77, 2005.
[13] H. Li, I. Councill, W. Lee, and C. L. Giles, "CiteSeerx: An Architecture and Web Service Design for an Academic Document Search Engine, " Proceedings of the 15th international conference on World Wide Web, p. 883-884, 2006.
[14] Isaac G. Councill, C. Lee Giles and Min-Yen Kan, "ParsCit: An Open Source CRF Reference String Parsing Package".
[15] http://www.mendeley.com/
[16] Eduard Hovy, Andrew Philpot, Judith Klavans, Ulrich Germann , Peter Davis, Samuel Popper "Extending Metadata Definitions by Automatically Extracting and Organizing Glossary Definitions", *National Conference on Digital Government Research*, *Boston*, MA. 2003.
[17] M Y. Day, R. T H. Tsai, C L. Sung, C C. Hsieh, C W. Lee, S H. Wu, K P. Wu, C S. Ong, and W L. Hsu, "Reference Metadata Extraction using a Hierarchical Knowledge Representation Framework", Decision *Support Systems*, vol .41, no.1, pp. 152-167.
[18] Element Tree: http://lxml.de/api/xml.etree.ElementTree-module
[19] Python tutorial: https://www.python.org/
[20] XML Tutorial: http://www.w3.org/XML/

AUTHORS PROFILE

Pushpa C N has completed Bachelor of Engineering in Computer Science and Engineering from Bangalore University, Master of Technology in VLSI Design and Embedded Systems from Visvesvaraya Technological University. She has 14 years of teaching experience. Presently she is working as Assistant Professor in Department of Computer Science and Engineering at UVCE, Bangalore and pursuing her Ph. D in Web Mining.

Shankar R has completed Bachelor of Engineering in Computer Science and Engineering from Visvesvaraya Technological Univesity, Master of Engineering in Web Technologies from Bangalore University. He has 4 years of teaching experience and 1 year industry experience. His intent is to persue R&D in the field of Web mining and semantic web services.

Venugopal K R is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, and Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Master's degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D.in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored more than 50 books on Computer Science and Economics. He has more than 600 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

Thriveni J has completed Bachelor of Engineering, Masters of Engineering and Doctoral Degree in Computer Science and Engineering. She has 4 years of industrial experience and 20 years of teaching experience. Currently she is an Associate Professor in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore. Her research interests include Networks, Data Mining, Distributed Systems, Cloud Computing and Biometrics. She has authored one book and has more than 50 research papers to her credit.

L M Patnaik is a Honorary Professor in Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and refereed International Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.