Lexical, Ontological & Conceptual Framework of Semantic Search Engine (LOC-SSE)

Gagandeep Singh Narula¹, Usha Yadav², Neelam Duhan³ and Vishal Jain⁴

Submitted in February, 2016; Accepted in July, 2016

Abstract – The paper addresses the problems of traditional keyword based search engines that process query syntactically rather than semantically. In order to increase degree of relevance and higher precision to recall ratio, it describes proposed architecture of Semantic Search Engine (SSE) which incorporates Google search results as input and processes them with the help of Semantic Web (SW) technologies. Modules to accomplish various tasks like query processing, importing existing ontologies and extraction of knowledge have been introduced in proposed framework. At last, the PROMPT algorithm is being applied to compare query graph and document graph which leads to improved results that are presented to user..

Index Terms- Semantic Web (SW), Ontology, PROMPT, Protégé 3.4.8, Jena, Resource Description Framework (RDF) and Knowledge Retrieval

1.0. INTRODUCTION

Traditional search engines are tools for retrieving information from massive sources on the web. The results are being produced by performing keyword based search most of the tme. The main drawback of search engines is lack of relevance. To illustrate the problem in a better way, consider a query "Mobile phones with red cover" submitted to a traditional search engine. It produces relevant as well as irrelevant results in relation with terms-mobile phones, red lotus, flower and cover. The search experience does not consider stopping words, auxiliary verbs that reflects the meaning of given statement. Likewise in above query, the term "with" has lost its significance due to which results are being produced in context of lotus and red flower. In order to reduce this ambiguity and perform intelligent search, concept of Semantic Web (SW) came into existence in 1996 as envisioned by Tim Berners Lee [1]. SW is defined as global mesh of information in machine interpretable format [2]. It is practically not feasible to annotate the entire web content into semantic tags so that current search engines could behave like Semantic Search Engines (SSE).

```
<sup>1</sup>Research Scholar, M.Tech (CSE), CDAC Noida
gagan.narula87@gmail.com
<sup>2</sup>Assistant Professor, CDAC-Noida
usha.yadav.912@gmail.com
<sup>3</sup>Assistant Professor (CE), YMCA University of Science &
Technology, Faridabad, India neelam.duhan@gmail.com
<sup>4</sup>Assistant Professor, Bharati Vidyapeeth's Institute of
Computer Applications and Management (BVICAM), New
Delhi
vishaljain83@ymail.com
```

So, there is need to develop search engine that analyses user query and produces meaningful results with higher precision and low recall.

The following paper is categorized into following sections. Section 2 describes objective and scope of research carried out in given paper. Section 3 presents brief survey of research conducted in context of evolution of SSE's and their methodologies. Section 4 provides bird's eye view of Semantic Web layered architecture and comparative analysis of studied literature survey. Section 5 describes proposed SSE framework along with its implementation. Section 6 validates higher precision to recall ratio in comparison to GOOGLE. Section 7 concludes the given paper followed by references.

2. 0. OBJECTIVE, SCOPE & FINAL OUTCOME Objective

"To enhance GOOGLE [2] search results with the help of Semantic Web technologies".

Scope

A user would be able to learn about semantic web technologies, semantic web tools, ontology development for knowledge representation and storing that knowledge using some open source framework.

Final Outcome

The intended final outcome of work carried out is precise and relevant search results produced by enhancing GOOGLE search results with the employment of SW technologies.

3.0. RELATED WORK

Several studies that have been conducted with an aim to build SSE and ranking of results as follows:

Debajyoti et.al [3] proposed semantic search framework that produces relevant results by performing mapping between classes and instances with the help of RDF codes. Fatima et.al [4] adds query optimizer, user interface and processor in its framework but it too has some limitations. Zhang et.al [5] performed keyword based search by finding RDF files and compares keywords with its contents. Swati et.al [6] proposed information retrieval system in context of university domain but it does not evaluate GOOGLE search results. Kumar et.al [7] made use of mapper and query processor for representation and scanning of keywords respectively.

For comparative analysis of these works, refer to Table 1.

4.0. SW ARCHITECTURE

According to Kevin Kelly [8], it suffers from fax effect which means that development of semantic web is costly and its technologies have not been utilized fully. But, still most of researchers are trying their hands on this web technology to achieve machine- human interaction [8]

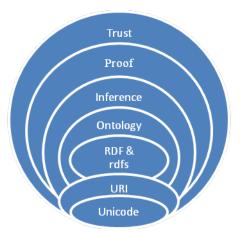


Figure1: SW architecture [9]

Table 1. Comparative Analysi	Table	1:	Comparative	Analys	sis
------------------------------	-------	----	-------------	--------	-----

	Table 1: Comparative	Allalysis
Research Work	Pros	Cons
	(a) Uses ontology to	(a) No ranking of
Debajyoti	maintain semantic	results is being
et.al [3]	relationships among	done.
[.]	classes and instances	
	rather than using NLP.	
	futifier than using (EF).	
	(b) Values of property can	
	be computed from RDF	
	codes and displayed to	
	user	
	(a) Query optimizer scans	(a) No updating of
Fatima	keywords and matches	ontology database.
et.al [4]	them with words stored in	ontology database.
et.ui [4]	ontology database.	(b) User interface is
	ontology database.	not connected to
		any semantic
		framework.
	(a) Combines Google	(a) Synonym
Zhang	search results with RDF	problem is not well
et.al [5]	and present them in	addressed in this
et.ai [5]	hierarchical fashion.	version of tool
	merarcincar fashion.	version of tool
	(b) OntoSearch acts as	
	visualization tool and can	
	be linked to other web	
	ontology editor tools	
	(a) Uses WordNet API for	(a) Does not
Swati et.al	generation of semantically	evaluate Google
[6]	similar words.	results
[0]	(b) Matches terms used	results.
	in user query with	
	designed ontology to	
	produce refined query.	
		(a) No comparison
Vumar	(a) Uses Mapper to	and evaluation of IR
Kumar	represent semantic results	
et.al [7]	into textual format.	performance.
	(b) Query processor scans	(h) Dana ang (
	keywords and matches	(b) Does not
	them with words stored in	evaluate Google
	ontology database.	results

5.0. COMPONENTS OF PROPOSED SSE FRAMEWORK

The proposed framework as outlined in Fig 2. consists of three phases:

- Generation of user query graph with the help of SW technologies.
- Generation of document relation graph by analyzing GOOGLE search results.
- Comparison of source and target ontologies that leads to improved results

First phase

(a) GUI: - The interface on which search is performed is treated as main component of any search engine. In context of traditional search engines, queries are written by developers and results are matched with pre-defined keywords stored in databases. But in proposed work, ontology is used as backend in interface.

In given framework, input query is being passed through user interface as well as GOOGLE search engine. It is passed to search engine in order to enhance search results with the help of SW technologies.

(b) Designing /Importing existing ontology: - The proposed framework uses PROTÉGÉ 3.4 beta [10] for importing existing ontology related to given domain. *Protégé is an open-source tool for editing and managing Ontologies. It is the most widely used domain-independent, freely available, platform-independent technology for developing and managing ontologies.*

(c) Extracting knowledge from given ontology: - Apache JENA framework can be used to represent relationship between classes, properties and instances from given ontology. It will lead to formation of knowledge base. *JENA is a java framework for building semantic web applications that provides programmatic environment for RDF, RDFS, and OWL and consists of rule based inference engine [11].*

Second Phase

Same user query is being entered in GOOGLE search engine and results are retrieved. These results are in form of HTML (Text) documents. So, relationship among those text documents is extracted by converting them into RDF documents. It is done with the help of Text2RDF application.

Third Phase

This phase requires comparison of target ontology graph and source ontology graph. In both graphs, concepts are represented by nodes while relations are represented by edges. It is done with the help of PROMPT [12] algorithm. Features of PROMPT are as:

- Besides merging ontology, it identifies locations for integration of ontologies, type of operations to be performed and resolves conflicts.
- Interactive merging process i.e. several choices are being performed by user and PROMPT selects them automatically on basis of user preferences.

Lexical, Ontological & Conceptual Framework of Semantic Search Engine (LOC-SSE)

restrictions. Handle conflicts like name conflicts, dangling references, redundant classes and slot value Querv GOOGLE User designed GUI Results Design/ Import existing (documents) ontology on given domain Document relation extraction Extract knowledge from given ontology Document graph Knowledge Base (Source ontology graph) Target ontology graph Comparison Improved results

Figure 2: Proposed LOC-SSE framework

5.1. Pros of Proposed Approach

•

- The given framework evaluates GOOGLE search results in addition to user query.
- User interface is connected to semantic framework called as JENA in order to retrieve knowledgeable results from ontology.
- Relationships among classes, properties and instances are represented in form of user query graph.
- On other hand, document graph is being created from GOOGLE search results.
- Thus, above methodology adds *lexical, conceptual and ontological* flavor to proposed framework.

5.2. Implementation

Above approach is being implemented as shown in steps below:

Consider user query as "List the faculties of CSE in IIIT Hyderabad"

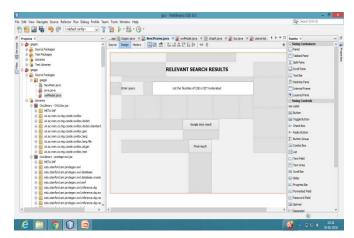


Figure 3: Home Screen

Step 1(a) User designed GUI: This form is drawn in NetBeans IDE 8.0

(b) Showing data connectivity among Protégé, NetBeans IDE and Jena

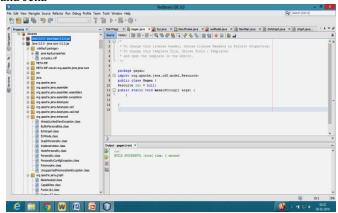


Figure 4: Importing libraries & its successful execution

Step 2: Designing of ontology on given domain (Educational_institute.owl)

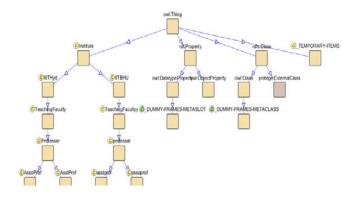
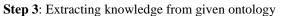


Figure 5: Educational_institute domain ontology



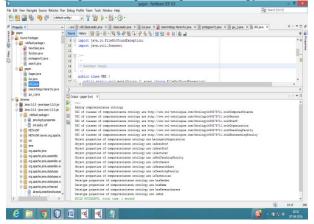


Figure 6: Displaying properties and URI's of Educational_institute ontology

From Fig 5, subsection of target ontology has been extracted further on basis of query "List the faculties of CSE in IIIT Hyderabad".

BELECT ?Institute
WHERE
{
<pre><http: ontology1460357725.owl#institute="" www.owl-ontologies.com=""> ?Institute</http:></pre>
}
PREFIX abc: <c:\program (x86)\protege="" 3.4.8="" files=""></c:\program>
SELECT ?name
WHERE
{
?abc:name?IIITHyd
}
}

Figure 7: Extracting target ontology portion using SPARQL query

Step 4: Creation of knowledge base involves: Generation of rules using Semantic Web Rule Language (SWRL) Four rules are being created that can lead to inferences related to given query. (i) Rule1_//_Hod_is_AssoProf_whose_Name_is Its expression in SWRL is CSE:isAssoProf(?A, ?S) CSE:isHod(?H, Λ ?A) CSE:hasName(?H, ?S) (ii) Rule2_//_AssoProf_is_senior_to_Lecturer_and_AsstProf Its expression in SWRL is CSE:isAsstProf(?A, ?S) ∧ CSE:isLecturer(?L, ?A)→ CSE:isAssoProf(?L, ?S) (iii) Rule3_//_AsstProf_for_TeachingFaculty Its expression in SWRL is CSE:isTeachingFaculty(?G, ?F) \rightarrow CSE:isAsstProf(?F, ?G)

	computerscience P	totégé 3.4.8 (file:\C\Program%20Files%20(x86)\Protege_3.4.8\computerscience.ppr). OWL / RDF File	ei) - 🗆 🎑
e Eat Briject DWL	Beasoning Code Incis EkoPortal Mind	ow Prompt Colaboration gep	
	22 0 0 0 2 0 2 0	Þ	protég
Metadwin(Drittlogy145827)	1711 owli C OVA Dasses Paper	tes 1 individuals II Forms @ Janitalays* Jans . Strill, Rules Prompt	No. 1 State State State
Alla, Hulaw			0.0000000
Enabled	Name	Expression	
	coppes.comRule1_I_Hod_k_AssaProf_whose	GE:sAssiProt(7A, 15) ∧ CSEstHod(7H, 1A) → CSE/hashlane(7H, 75)	
 Mp.Peww.swiatt Mp.Peww.swiatt 	stopes conflue2_I_AssoProf_is_server_to_L.	CSEInAastProf(7A, 75) ∧ CSEInLochare(7L, 7A) → CSEInAastProf(7L, 75) CSEInTeaching*south(70, 17) → CSEInActFrof(77, 70)	
Htp://www.ewi-eni	ologies com/Rule4_IF Technician in Lecturer's	GSE:sResearchAstd(7R, 7S) ∧ CSE:isLecturer(7L, 7R) → CSE:isTechnician(7L, 7S)	
Mailwww.awian	otopes.com/Rulet_iffacuity_Teacher_Oude_c.	CSE Infaculty in Infeactor in InGuide in InAnelProty Assorboty Lectured	
			😽 - 12-4 - Bal

Figure 8: Rules generated using SWRL Step 5: Target ontology graph

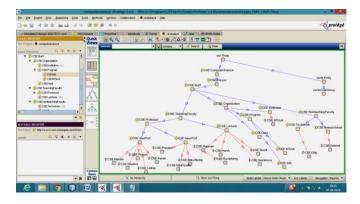


Figure 9: Target ontology graph

Step 6: Now, query is entered on Google and it produces links of other faculties of IIT BHU, IIT Hyd in addition to IIIT Hyd



Figure 10: GOOGLE search results page

Step 7: Results (documents)

It involves conversion of HTML links to semantic web resources like RDF so that ontology can be created which can be said as "GoogleCSE.pprj" or "GoogleCSE.owl"

	803	-	the second s	10 C C C C C C C C C C C C C C C C C C C		
			TEXT2RDF Applic	1000		- 8 .
Assistant Ph.D. (In Coding Th Madhava H Associate Ph.D. (II systems a	Abr Ser Ser Professor diam Taetinute of mory, Information Iziehom K. (IIIT H) Professor T Eargur) much and methanism desig Dame	Ditanow, Bangalorw) Theory, Signal Proces (0)	alog, Distributed Durage Jys vision, outdoor ribbolics, Ho	Luna	Addhand (shrmatan hpe of Cinten) Const Bandh (steet	
 deven thead deven thead deven thead deven thead deven thead deven and the 	> CARELPESTS > Manufil Devices CET Ha > Assets (Professor - Americ > The CET Has > Assets (Professor - Americ > The CET Has > Commonly - Camera - America) Common - Camera - America > Common - Camera - America > Common - Camera - America > Common - Camera - Camera - Camera > Camera Camera - Camera - Camera - Camera > Camera - Camera - Camera - Camera > Camera - Camera - Camera - Camera - Camera > Camera - Camer	end > alt> > ret spe="Resource"> dereca idea o > doerophrase >ALP for Inda La o > doerophrase >ALP for Inda La	ural Language Processing , Hachres Lawring , J 2016 - Martin J. Lawrence Processing / American and 2016 - J. Language Processing / American days, 147 for Sold Languages - American		ec right for -sharees a classes on the statement of the s	
To diedk and	I vasaket fos RDF, var W3C	vidator service <u>tata/inco.at</u> .			a Marcine da	

Figure 11: Conversion of HTML links of IIIT Hyd into RDF

Similarly, conversion of IIT Hyd and IIT BHU can be done into RDF.

Step 8: Document Relation Extraction

It involves designing of ontology from above RDF results.

De Lit Drant Litt. Breatring Cola	Des privat in	Huntyd A.A. Hantal. Hoganitha	De .	SE WAS OWN / HOF Files	- 6
		4 Þ			protég
Contractions of the second sec				46 T	
		Compress They Compress Theorem	Ing I	en e	Marine Starter
	Virses	W increasedy	5. Non parting	Non-Labora Salara Sum (Salar)	· An Labor C Receiptor Tapoly
A B B B		4		0	🔐 - 4 0 4

Figure 12: GoogleCSE ontology graph (source ontology)

Step 10: Comparison

It is done by comparing both ontologies using PROMPT algorithm where source ontology is "GoogleCSE.owl" and target ontology is "computerscience.owl"

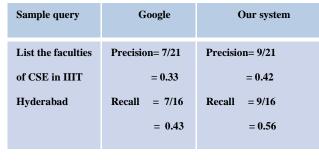
	file Team Tools Window Help	Q - Search (Ch1+L)
cefault cefault cefault cefault centres	V 7 19 D . B. Q.	
iecta X	- and in forderer faring fand toos X all exceptions toos X all exceptibilitest	java 🗴 🗟 MappingDonainPDM.java 🗴 🗟 MappingDonapeRugin.java 🛪 🗟 P 🗘
gigan	* Source Hatery 10 10 - 10 - 10 - 10 - 10 - 10 - 10 -	
Source Packages	1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 1 1 1 1 1 1 1	
default package>	7 package gagan;	
NewClass. Java	8 [import java.util.krrayList;	
- ill function take	9 import java.util.Collection:	
protegesvrl1.java	10 import jeve.util.Iterator:	
in search lava	11 import org.apache.jens.rdf.model.impl.Util;	
E E atom	12	
R FinalresultsafterPrompt.java	13 public class MappingDomainFSM implements Mapping	StorapePlugin (
- B Gegen.jeve	14 private static final String MAPPING DOMAN	IN PSM DIRECTORY = "edu.stanford.emi.proteges.prompts.mappin
- its GoogleCSE.jeve	15 private static final String JAR NAME = "	mappingComainP3H, 5ar*;
- K Googleresults Java	16	
Gui java	17 private static final String INSTANCE MAPS	PING CLASS - "instance-menuing";
HashHapForCollections. Java	18 private static final String TARGET CLASS	
Knowledgeflase.java	19 private static final String SOURCE CLASS	
- B NeppingDomenPDM.Serve	20	and the second se
MappingStoragePlugin.java	21 private static final String TARGET CLASS	DESCRIPTION CLASS = "computerscience.ovl";
- PlagnUtities.java	22 private static final String SOURCE CLASS	DESCRIPTION CLASS = "GoogleCSE.owl";
- Regect java	23 private KnowledgeBase mappingKb = null;	
- BURI Java	24 private Project _mappingProjects	
Userontology/hierarchy.java	25	
gui_Ljeva	26 private KnowledgeBase _sourceKb = null;	
🗄 🛖 Test Packages	27 private KnowledgeBase targetRb = nulls	
🖯 🍙 Libraries	28	
🗉 🖀 Jena-3.0.0 - jena-base-3.0.0.jar	29 private HashMapForCollections mappingsHa	ap; // <frame, collection="" mappings="" of=""></frame,>
🕀 🖀 Jena-3.0.0 - jena-core-3.0.0.jar	** <	
🕀 🔣 <default package=""></default>		
🖯 🗟 jena-log4j.properties	Output-gagan (run) ×	
ant-policy.rdf	💭 zuzi	
- E META-DAP	BUILD SUCCESSFUL (total time: 1 second)	
In the META-DM maven.org.apache.jena.jena		
🖗 🔝 etc	80	
🐵 🔛 jena	v 94	
		(1) 22.14

Figure 13: Execution of PROMPT algorithm

Step 11: Improved results

On Children and Anti-	State No. H. Ch.	
Event August and	<pre>unit Table Mark Mark</pre>	
Income Income Income and a comparison of the comparis		
in an and a second seco		

Figure 14: Enhanced Results



6.0. EVALUATION MEASURES

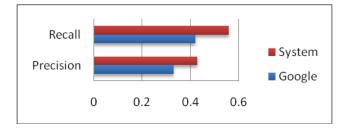


Fig 15: Higher P to R ratio of our system than GOOGLE

7.0. CONCLUSION AND FUTURE SCOPE

The given paper presents a Lexical, Ontological & Conceptual framework of Semantic Search Engine (termed LOC-SSE) with the help of semantic web technologies. The proposed system is implemented and evaluated on basis of Precision- Recall Ratio. From implementation & analysis of the proposed framework, it is concluded that given system produces more accurate results as compared to Google.

As a future work, it can be extended by developing agent based middleware search engine with the help of JADE (Java Agent Development Environment).

REFERENCES

- [1]. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web.", *Scientific american* 284.5 (2001): 28-37.
- [2]. Tim Berners-Lee, The Semantic Web Revisited, IEEE Intelligent Systems, 2006
- [3]. Debajyoti Mukhoupadhyay, Aritra Banik, Jhalik Bhattacharya, "A Domain Specific Ontology Based Semantic Web Search Engine" 2011 IEEE 6TH International Conference on Intelligent systems and Artificial Intelligence, Jaypee University, Shimla, 11-14 February 2011.
- [4]. Arooj Fatima, Cristina Luca, George Wilson "New Framework for Semantic Search Engine" 2014 IEEE UKSim-AMSS 16th International Conference on Computer Modeling and Simulation, 26-28 March 2014, Cambridge, pages 446-451.
- [5]. Yi Zhang, Wamberto Vasconcelos, Derek Sleeman "OntoSearch: An Ontology Search Engine" In Proceedings of AI-2011, the twenty-fourth SGAI

International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, pages 58-69

- [6]. Swati Rajasurya, S.Swamynathan, "Semantic Information Retrieval using Ontology in University Domain" 2012 IEEE 6th International Conference on Intelligent systems and Artificial Intelligence, July 2012, Jaypee University, Shimla.
- [7]. S.S. Kamath, Garima Meena, K.Kumar "A Semantic Search Engine for Answering Domain Specific Userqueries" 2014 IEEE International Conference on Communications and Signal Processing (ICCSP), 3-5 April 2014, pages 1097-1101
- [8]. Gagandeep Singh Narula, Dr. S.V.A.V. Prasad and Dr. Vishal Jain, "Use of ontology to secure the cloud: A Case Study", International Journal of Innovative Research and Advanced Studies (IJIRAS), Vol 3 Issue 8 July 2016, ISSN 2394-4404
- [9]. Gagandeep Singh, Vishal Jain, "Information Retrieval (IR) through Semantic Web (SW): An Overview", In proceedings of CONFLUENCE-The Next Generation Information Technology Summit, 27-28 September 2012, pp 23-27
- [10]. Gagandeep Singh, Vishal Jain , Dr.Mayank Singh, "Ontology Development Using Hozo and Semantic Analysis for Information Retrieval in Semantic Web" in 'ICIIP-2013 IEEE Second International Conference on Image Information Processing ', Jaypee Univ. Shimla, 9-11 Dec 2013
- [11]. https://jena.apache.org/
- [12]. http://protegewiki.stanford.edu/wiki/PROMPT
- [13]. Gagandeep Singh Narula, Dr. Subhan Khan, Yogesh. "DST's Mission Mode on Program Natural Resources Data Management System (NRDMS)", BIJIT-BVICAM's International Journal of Information Technology, Jan-June 2016 Vol.8 No.1 pages 973-978 having ISSN No. 0973-5658 with impact factor 0.605, indexed with IET (UK), INSPEC
- [14]. Meenu Dave, Mikku Dave, Y S Shisodia, "Cloud computing and knowledge management as a service", BIJIT-BVICAM's International Journal of Information Technology, July-Dec 2013 Issue 10 Vol.5 No.2 pages 619-622
- [15]. Usha Yadav, Gagandeep Singh Narula, Neelam Duhan, Vishal Jain and BK Murthy, "Development and Visualization of Domain Specific Ontology using Protégé", Indian Journal of Science & Technology (INDJST), Vol. 9 No. 16, April 2016 having ISSN No. 0974-5645 and indexed with Thomson Reuters (Web of Science), Scopus (Elsevier), Index Copernicus, SJR=1.3
- [16]. Anil Kumar, Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area", BIJIT-BVICAM Journal of Information Technology, July-Dec 2009, Vol 1 No 2, Issue 2 pages 137-140.
- [17]. S. Ajitha, T.V, Suresh Kumar & K. Rajnikanth, "Framework for multi-agent systems performance

prediction", BIJIT-BVICAM Journal of Information Technology, Issue 12, July-Dec 2014, Vol. 6 No.2 pages 774-778

[18]. Parul Gupta, AK Sharma, "A Framework for hierarchical clustering based indexing in search engines", BIJIT-BVICAM Journal of Information Technology, Issue 6, July-Dec 2011 Vol. 3 No.2 pages-329-334