Ontology based Semantic Search Engine

[1]N.Vanjulavalli, [2]Dr.A.Kovalan

Department of Computer Science and Applications, PMU

[1]vanjulavallisn@gmail.com
[2]kovapmu@gmail.com

Abstract- **We live in a connected age when every entity is linked to each other through numerous relationships thus forming complex networks, information networks etc., In this meshy scenario, storing analyzing, managing extracting information is going to be highly challenging task. Information retrieval happened to be an activity that only a few people engaged in. due to the advent of search engines and communication facilities, millions of people are engaged in information retrieval and extraction process. The information retrieval deals with finding a set of documents relevant to the user query. Commercial search engines like Google deals with key word search which is based on Boolean logical queries. The major disadvantage of this kind of keyword search is that it returns a lot of irrelevant information to the users which results in low precision. Nowadays the field of information retrieval is moving towards semantic level from syntactic level. These semantic search engines are based on the concept of ontology which gives a Meta data representation of concepts. In this paper we focus on the retrieval mechanisms related to it. This includes various phases in which the ontology is designed first and then the indexing and retrieval phases work.**

Key words- **Ontology, semantic web, Information retrieval, Query builder**

I.INTRODUCTION

In this digital age each and every concept is linked to each other through various relationships forming the complex information web. By passing through the web we face some sort of challenges in storing and extracting the information from them. We therefore rely on search engines to overcome those difficulties. Any way these search engines facilitate only keyword search which does not returns accurate results to the external user. Aiming to solve the limitations of keyword based models, the idea of semantic search, understood as searching by meanings rather than by literals has been the focus of a wide body of research in the information retrieval and the semantic web communities

Semantic search has been present in the information retrieval since the early eighties. Some of these approaches are based on the statistical methods that study the co-occurrence of terms that capture and exploit tough and fuzzy conceptualizations. Other Information retrieval approaches apply linguistic algorithms modeled on human language processing structures and taxonomies, where the level of conceptualization is often shallow and sparse, especially the level of relations, which are commonly at the core of expressing user needs and finding the answers.

Ontologies are used to represent knowledge in a conceptual manner that can be distributed among various applications. The ontology has its application in information retrieval where it exploits the knowledge bases enhancing the semantic search, steering in one hand the use of fully fledged ontologies in the semantic based perspective, and on the other the consideration of unstructured content as target search space. In other words, this work explores the use of semantic information to support more expensive queries and more accurate results, while the retrieval problem is formulated in a way that is consistent with the Information retrieval field, thus drawing benefit from the state of art in this area and enabling more realistic and applicable approaches.

II. RELEATED WORK

A. Semantic web

Current World Wide Web (WWW) is a huge library of interlinked documents that are transferred by computers and presented to people. It has grown from hypertext systems, but the difference is that anyone can contribute to it. This also means that the quality of information or even the persistence of documents cannot be generally guaranteed. Current WWW contains a lot of information and knowledge, but machines usually serve only to deliver and present the content of documents describing the knowledge. People have to connect all the sources of relevant information and interpret them themselves

Semantic web is an effort to enhance current web so that computers can process the information presented on WWW, interpret and connect it, to help humans to find required knowledge. In the same way as WWW is a huge distributed hypertext system, semantic web is intended to form a huge distributed knowledge based system. The focus of semantic web is to share data instead of documents. In other words, it is a project that should provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by World Wide Web Consortium (W3C). The semantic web has got its peak in recent years by the use of explicit metadata representation of information in the web. The metadata representation is embedded in web page using RDF to enhance the visualization of results. Semantic Web can be seen as a huge engineering solution... but it is more than that. We will find that as it becomes easier to publish data in a repurposable form, so more people will want to publish data, and there will be a knock-on or domino effect.

We may find that a large number of Semantic Web applications can be used for a variety of different tasks, increasing the modularity of applications on the Web. But enough subjective reasoning. onto how this will be accomplished.

The Semantic Web is generally built on syntaxes which use URIs to represent data, usually in triples based structures: i.e. many triples of URI data that can be held in databases, or interchanged on the world Wide Web using a set of particular syntaxes developed especially for the task. These syntaxes are called "Resource Description Framework" syntaxes.

Once information is in RDF form, it becomes easy to process it, since RDF is a generic format, which already has many parsers. XML RDF is quite a verbose specification, and it can take some getting used to (for example, to learn XML RDF properly, you need to understand a little about XML and namespaces beforehand...), but let's take a quick look at an example of XML RDF right now:-

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:foaf="http://xmlns.com/0.1/foaf/" >
   <rdf:Description rdf:about="">
      <dc:creator rdf:parseType="Resource">
         <foaf:name>Sean B. Palmer</foaf:name>
      </dc:creator>
      <dc:title>The Semantic Web: An Introduction</dc:title>
   </rdf:Description>
</rdf:RDF>
```

This piece of RDF basically says that this article has the title "The Semantic Web: An Introduction", and was written by someone whose name is "Sean B. Palmer". Here are the triples that this RDF produces:-

```
<> <http://purl.org/dc/elements/1.1/creator> _:x0 .
this <http://purl.org/dc/elements/1.1/title> "The Semantic Web: An Introduction" .
_:x0 <http://xmlns.com/0.1/foaf/name> "Sean B. Palmer" .
```

This format is actually a plain text serialization of RDF called "Notation3.

B.Ontology.

The history of artificial intelligence shows that knowledge is critical for intelligent systems. In many cases, better knowledge can be more important for solving a task than better algorithms. To have truly intelligent systems, knowledge needs to be captured, processed, reused, and communicated. Ontologies support all these tasks.

The term "ontology" can be defined as an explicit specification of conceptualization. Ontologies capture the structure of the domain, i.e. conceptualization. This includes the model of the domain with possible restrictions. The conceptualization describes knowledge about the domain, not about the particular state of affairs in the domain. In other words, the conceptualization is not changing, or is changing very rarely. Ontology is then specification of this conceptualization - the conceptualization is specified by using particular modeling language and particular terms. Formal specification is required in order to be able to process ontologies and operate on ontologies automatically.

Ontology describes a domain, while a knowledge base (based on ontology) describes particular state of affairs. Each knowledge based system or agent has its own knowledge base, and only what can be expressed using ontology can be stored and used in the knowledge base. When an agent wants to communicate to another agent, he uses the constructs from some ontology. In order to understand in communication, ontologies must be shared between agents.

Semantic network (also called concept network) is a graph, where vertices represent concepts and where edges represent relations between concepts. Semantic network at the level of ontology expresses vocabulary that is helpful especially for human, but that still can be usable for machine processing. The relations between concepts that are used in semantic networks are as follows:

synonym - concept A expresses the same thing as concept B

antonym - concept A expresses the opposite of concept B

meronym, holonym - part-of and has-part relation between concepts.

hyponym, hypernym - inclusion of semantic range between concepts in both directions

Semantic nets were created as an attempt to express Interlingua, a common language that would be used for translation between various natural languages. A typical example is WordNet that describes relations between English words and defines the words using natural language. Parts of WordNet were translated to other languages and the links between various languages exist and can be used as the base for translation.

Topic Maps are (syntactically) standardized form of semantic networks. They allow using topics (concepts), associations (relations) between concepts (including specifying role of topic in the association), and occurrences (resources relevant to topic, in fact instances of topic). Topics, associations and occurrences are used to create ontology of a domain, and a particular topic map then uses them to expresses state of affairs in the domain

C.Kinds of Semantic Search approaches

The classification of semantic search approaches is complex not just because of their diversity in the sense how differently this has been approached in literature, but also because of the large no. of dimensions involved in the information search task. This section proposes a set of general criteria under which Semantic web & Information retrieval approaches can be classified and compared, identifying their key advantages and limitations.

Semantic Knowledge representation

Three main trends can be distinguished in the literature based on the type and use of semantic knowledge representation.

i)Statistical approaches use statistical models to identify groups of words that commonly appear together and therefore may jointly describe a particular reality;
ii) Linguistic conceptualization approaches are based on light conceptualization, usually considering few types of relations between concepts, and low information specifying levels;
iii) Ontology based proposals consider a much more detailed and densely populated conceptual space in the form of ontology based knowledge bases.

Scope: Semantic search has been applied in different environments such as the web, controlled repositories or even the desktop. Obtaining conceptualizations tom cover the meanings involved in all web content as well as the automatic annotation of these conceptualizations with some degree of completeness is still an open challenge. Restricting them to more reduced environments, many system have developed and tested over controlled repositories, where the available information is enclosed in one or few domains or knowledge. In a third degree of complexity, the desktop environment provides easier ways to attract the semantic information from semi structures contents such as e-mails, folders etc.,someworks do not explicitly state their potential or limitations in scale and scope, but the considerably computational complexity involved in their methods scalability as a non addressed issue.

Query: Another relevant aspect that characterized semantic search models is the way the user expresses his information needs. Four different approaches can be identified in the state of art, according to gradual increase of their level of formality and usage complexity.

Content-Retrieval: Semantic retrieval approaches can be characterized by whether they aim at data or information retrieval while the majority of IR approaches returns documents as response to user requests and therefore should be classified as IR models, a large amount of ontology based approaches return ontology instanced rather than document and therefore may be classified as data retrieval models.

Content-ranking; While IR approaches have traditionally addressed the ranking of documents, most ontology based approaches do not consider ranking query results in general or base their rank functionality or traditional keyword based approaches. A few approaches take advantage of semantic information to generate query result rankings.

D. Semantic Similarity

The Vector Space Model (VSM) is used to represent documents through the words that they contain. In generally, information retrieval employs the VSM and a set of document, traditional information retrieval model usually measures the similarity of a query and different documents, and then returns the documents with top-ranked similarities as the results. As with the vector space model, a query and documents are both regarded as a certain point on the vector space. The similarity between two documents is perceived by the cosine similarly. The authors [15] pointed out that the smaller the angle cosine value, the greater the similarity. Figure 1 show the cosine value applied as sin (d_j, q) as the traditional IR model when a user issue a query, it's recently stored context is classified to create the user's contextual profile. Once the list of document if formed, the search engine computes a semantic similarity value between the query and each document results.

III. METHODOLOGY

In our research work we propose a methodology as the first step we design the ontology and it is followed by the mapping of ontologies and the data is extracted and the searching phased is executed. We design a central ontology, which is utilized by every aspect of the system, especially in information extraction, inference and retrieval phases. Thus the overall performance of the system is highly dependent on its quality. We follow an iterative development process in the ontology engineering phase. First we started with a core ontology including the basic concepts and a single hierarchy. Then we experimented with this ontology and fix the issues in the reasoning and extraction.

A. Ontology Population

Ontology population is the process of knowledge acquisition by transforming or mapping unstructured, semi-structured and structured data into ontology instances. Information extractor module already does most of the labor by extracting structured information from unstructured text narrations.. Having the output of the IE module, the ontology population process now becomes creating an OWL individual for each object extracted during IE. If the IE module cannot extract some attribute of an event, we still create an instance with empty properties. Thus, the recall performance for simple queries will not be affected even IE fails to extract some details of the event. Moreover, if no event is detected in a narration, an instance with the type Unknown Event is created. Unknown events are not discarded because of the reasons. Ontology population is not restricted with the events extracted from the IE module. As mentioned earlier, the crawled information also contains some basic information about the match including players, teams, referees, stadium etc. This information is also added to the ontology by creating an OWL individual for each of them if they do not already exist in the knowledgebase.

B.Inference and Rules

Type subsumption defines a relationship between a subtype and a super type. That relationship contributes to the semantics of both types and therefore should enable some semantics of one type to be transferred to the other, such as in the case of relation and meta-relation types where the extra arguments of one type can be transferred to the other. This is one of the main ideas behind the process of completing missing arguments in a relation or meta-relation type, which is essential for Inferencing in some cases. Inferencing is also

most useful when applied to real world objects and events, that is, when it is performed on the set of individuals in the ontology. Semantic axioms can improve reasoning over the ontology. In the case of the proposed ontology formalism, Inferencing axioms (also called rules).

C.Semantic Indexing and Retrieval

For the retrieval part, we adapt a semantic indexing approach based on Lucene6 indices. The idea is extending traditional full-text index with the extracted and inferred knowledge and modifying the ranking so that documents containing ontological information gets higher rates.

3.3.1*Index Structure:* The structure of semantic index has utmost importance in the retrieval performance. We constructed a Lucene index such that each entry represents a soccer event. As we have mentioned in the previous sections, each event has its own properties associated with it, such as subjects and objects. That information is also included with each event. We also include full-text narrations associated with events to the index. This is especially important if the event type is unknown (an event which is not recognized by the information extractor). Adding full-text narrations to the index tolerates the incomplete event information, thus ensures at least the recall values of traditional full-text search.

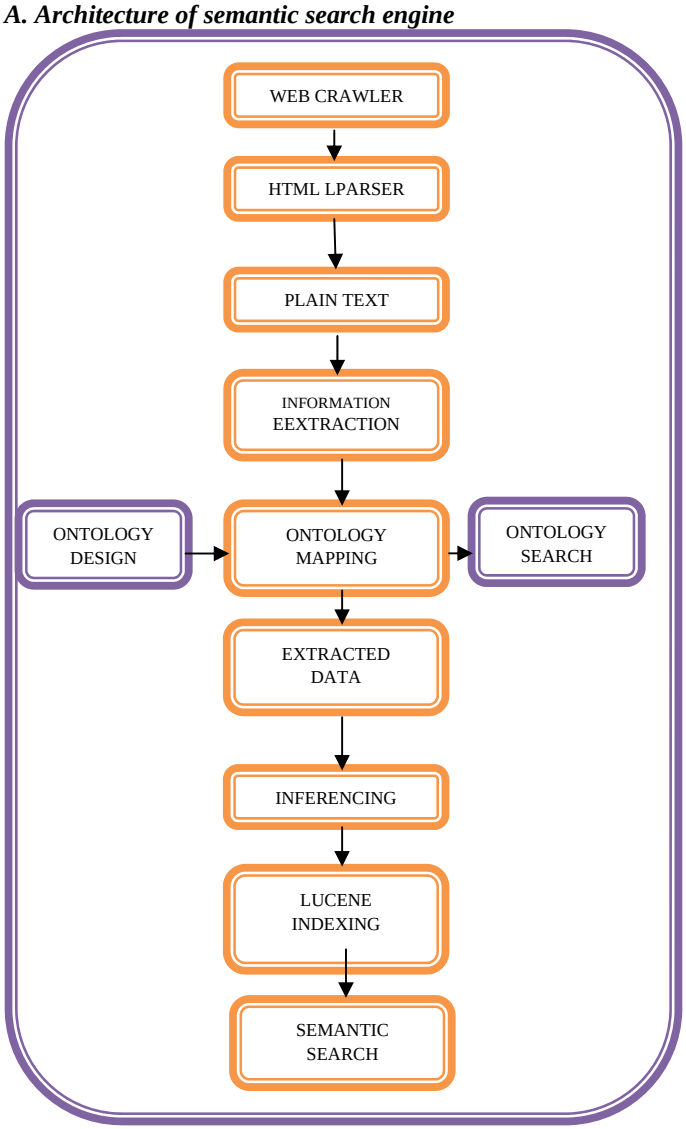
3.3.2*Searching and Ranking:* In traditional keyword search, indexed documents usually contain nothing but raw text associated with that document. Lucene can easily handle such indices and its default ranking gives usually good results. However, complex indices should be handled carefully. In order to take the advantages of our ontology-aided index structure, we slightly modified default querying and ranking mechanism of Lucene. First of all, we boosted the ranking of fields containing extracted and inferred information to stress the importance of them. Secondly, these fields are re-ranked according to their importance. For example, the “event” field is given the highest ranking. This approach prevents misleading stemming from ambiguous words in full-text. For example, let’s say a narration contains “Ronaldo misses a goal”: Searching for a “goal” in a traditional search may return this document in the first place, which is a false positive. However, in ontology aided index, the events whose type is Goal will have higher ranks. Since the type of the event above is a Miss, it will have a lower rank.

We have presented a novel semantic retrieval framework and its application to soccer domain, which includes all the aspects of Semantic Web, namely, ontology development, information extraction, ontology population, inference, semantic rules, semantic indexing and retrieval.

IV.COMPONENTS

Web Crawler
HTML Parser
Information extraction
Ontology Design
Ontology Mapping
Inferencing
Semantic Indexing
Searching

A. Architecture of semantic search engine



4.1 Web crawler

The focused crawler takes at least as input ontology with its lexicon. If according to the ontology existing metadata is already it serves also as input. An important aspect is that for the metadata also the metadata lexicon serves as input to the crawler. This includes the definition of ontology and metadata constraints, threshold, and the selection of a relevancy measure. The output of the focused crawler is a set of documents (for each document a set of most relevant concepts is assigned), discovered metadata according to the selected ontology, and suggestions for the evolvement of the ontology.

Based on the user input as described above, the crawling process is started, resulting in a first set of retrieved documents. The retrieved documents are preprocessed using the preprocessing module. Preprocessing is splitted into several steps, roughly distinguishable in RDF metadata extraction and validation against the ontology, text processing

and normalization and hyperlink extraction. The preprocessed segments of a document serve as input to the relevance computation process that extends the URL list for further processing, the document list and the RDF metadata container. The user may now inspect the results of the crawling process, add RDF metadata to the local system and refine the evolving ontology based on analysis of the documents contained in the document list.

B. Ontology mapping

Ontology mapping is the process whereby semantic relations are defined between two ontologies at conceptual level which in turn are applied at data level transforming source ontology instances into target ontology instances. Ontology mapping faces new challenges in the context of Semantic Web, especially concerning heterogeneity, dynamics, distribution and limitations on representation technology. The MAFRA – Mapping Framework is a conceptual description of the ontology mapping process, in which its phases are identified and described.

V. CONCLUSION

In this paper, we present the developing procedure of ontology based semantic search engine. Ontology based semantic annotation are needed when building the system. In addition, key techniques are also discussed in our paper. Compared to previous research, our works concentrate on the semantic similarity and the whole process including query submission and information annotation. We believe that our information retrieval can give more precise answer to user than the work without ranking document.

REFERENCES

[1] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American May 2001

[2]D.Fensel. W.Wahlster, H. Lieberman. J. Hendler. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press 2003

[3]Mehrnoush Shamsfard, Azadeh Nematzadeh and Sarah Motiee, ORank: An Ontology Based System for RankingDocuments, International Journal of Computer Science, vol .1, pp.225- 231, 2006.

[4]Wang Wei, Payam M.Barjaghi and Andrzej Bargiela, "Semantic enhanced information search and retrieval", Sixth International Conference on Advance Language and WebI formation Technology, pp.218-223, 2007.

[5]S. Lu, M. Dong and F. Fotouhi, "The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications". International Journal of Information Research, 7(4), 2002

[6]T. R. Gruber. A translation approach to portable ontologies. Knowledge Acquisition, 5(2): 199-220, 1993

[7]Geroimenko, V., Chen, C. (eds) (2002). Ontology-based Information Visualization. Visualising the Semantic Web. Springer Verlag, London, ISBN 1-85233-576-9

[8]van Harmelen, F.; Broekstra, J.; Fluit, C.; ter Horst, H.; Kampman, A.; van der Meer, J.; Sabou, M. Ontology-based Information Visualization.Proceedings of Fifth International Conference on Information Visualisation Conference, 546–554 2001

[9]R.Guha, R.McCool and E.Miller, Semantic Search. International Conference on World Wide Web, pp.700-709. 2003.

[10]Pablo Castells, Mriam Fernandez and David Vallet. An Adaption of the Vector Space Model for Ontology based Information Retrieval", IEEE Transaction on Knowledge and Data Engineering, vol. 2, pp.261-22, 2007

[11]Dik L. Lee, Huei Chauang and Kent Seamons, " Document Ranking and the Vector Space Model" , IEEE Software, pp.66-75, 1997.

[12]Mayfield, J., and Finin, T.: Information retrieval on the Semantic Web: Integrating inference and retrieval. In: Workshop on the Semantic Web at the 26th Intl. ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada (2003)

[13] S. Luke, L. Spector, D. Rager, Ontology-based knowledge discovery on the World-Wide Web, in: Internet-based Information Systems: Papers from the AAAI Workshop, AAAI, Menlo Park, CA, USA, 1996, pp. 96–102.

[14] J. Contreras, V.R. Benjamins, M. Blázquez, S. Losada, R. Salla, J. Sevilla, et al., A semantic portal for the international affairs sector, in: Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004), Whittlebury Hall, UK, 2004, pp. 203–215.

[15] A. Bernstein, E. Kaufmann, Gino— guided input natural language ontology, in: Proceedings of the 5th International Semantic Web Conference (ISWC 2006), Athens, GA, USA, 2006, pp. 144–157.

[16] P. Cimiano, P. Haase, J. Heizmann, Porting natural language interfaces between domains—an experimental user study with the ORAKEL system, in: Proceedings of the International Conference on Intelligent User Interfaces (IUI 2007), 2007, pp. 180–189.

[17] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, A. Kirilov, KIM—a semantic platform for information extraction and retrieval, Journal of Natural Language Engineering 10 (3–4) (2004) 375–392.

[18] F. Giunchiglia, U. Kharkevich, I. Zaihrayeu, Concept search, in: proceedings of the 6th European Semantic Web Conference (ESWC 2009), Heraklion, Greece, 2009, pp. 429–444.

[19] T. Finin, J. Mayfield, C. Fink, A. Joshi, R.S. Cost, Information retrieval and the Semantic Web, in: Proceedings of the 38th Annual Hawaii international Conference on System Sciences (HICSS 2005), Big Island, HI, USA, 2005, pp. 4–14.

[20] M. Fernández, V. López, M. Sabou, V. Uren, D. Vallet, E. Motta, P. Castells, Semantic search meets the Web, in: Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC 2008), Santa Clara, CA, USA, 2008, pp. 253–260.

[21] P.A. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl, R. Paiu, Activity based metadata for semantic desktop search, in: Proceedings of the 2nd European Semantic Web Conference (ESWC 2005), Heraklion, Greece, 2005, pp. 439–454.