Impact Factor (SJIF): 2.125

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 1, January-2016

Intelligent Information Retrieval in Semantic Way: Agents and Framework

[1]Sanjib kumar Sahu , [2]D. P. Mahapatra,, [3]R.C. Balabantaray

[1]Dept of Computer Science, Utkal University, Odisha, India
[2]Dept of Computer Science, NIT, Rourkela, Odisha
[3]Dept of Computer Science, IIIT Bhubaneswar, Odisha

ABSTRACT

The semantic web assumes the web as collection of data instead of set of documents. It aims not only to find keywords, but also understand intent of searcher and get contextual meaning of terms as they appear in the searchable dataspace. It can meaningfully correlate the keywords entered for search, thus, providing relevant and more accurate information and not just find all data related to those keywords.Instead of using ranking algorithm, in which pages are returned based on matching of keywords, semantic search takes into account " semantics" (the science of meaning in language) to produce exactly relevant search results. A multi-agent system(M.A.S.) is an electronics machine made up of multiple interacting intelligent agents within an environment. The problems those are impossible or difficult to solve for an individual agent can be completed with the help of Multi-agent Systems.

Keywords— *Search engine, semantic network, Ontology, information retrieval, Multi Agent System.*

I. INTRODUCTION

Semantic Web can be defined as linking of data between different entities that allows self-describing interrelations of variety of data available across the World Wide Web. Semantic Web structure can be described as: **Collection of data + Language for expressing that data=Semantic web.**

It standardizes the way of expressing the relationship that allows computers to easily understand the data and process it. There are three basic standards for Semantic Web, namely-OWL (*Ontology Web Language*), RDF (*Resource Description Framework*), and SPARQL (*SPARQL Protocol and RDF Query Language*). Certainly, there is huge amount of data that exists in form of documents and database. When a user wants to search anything, he types his query and wants the most accurate results to be shown. Currently, in tradition way to return results is to only march keywords. So, mostly blogs and discussion forums are retrieved as result which are not reliable sources. Instead, user must be looking for more reliable sources like journals, white papers etc. To retrieve such result, mapping of words in query is necessary.
The semantic search engine solves this problem. Each page stored in the database contains metadata with notes, meanings, list of words, definitions, vocabulary for the annotations etc., annotations are based on the classes of concepts and relations among them. Semantic search engines are aimed at having improved search strategies using the proposed architecture and relational algorithm to satisfy user's information needs despite increased complexity the search engine will be accurate with full efforts on performance time and scalability.

II. LOGIC AND NEED OF SEMANTIC SEARCH ENGINE

Certainly, there is huge amount of data that exists in form of documents and database. When a user wants to search anything, he types his query and wants the most accurate results to be shown. Currently, in tradition way to return results is to only march keywords. So, mostly blogs and discussion forums are retrieved as result which are not reliable sources. Instead, user must be looking for more reliable sources like journals, white papers etc. To retrieve such result, mapping of words in query is necessary.

There is nothing that can exist independently in the boundless universe. Everything is related to other things in various manners. When one tries to comprehend an entity, one comprehends this entity from the way it relates to other entities.

In fact, the semantics of an entity exist in the relationship between the entity and others. For communication between machines, the relationships between entities have to be defined before the machines can understand the semantics of each other.

The traditional search engine has no infrastructure or matching techniques to give correct or a related information for the query raised. As discussed above, what it relies on is just searching for the keywords

entered by the user in the pages available in the database and returning the result set without taking into consideration the fact that the result set may contain those words in context other than desired by the user or that their may be other pages with the information desired by the user but using some other words than those entered by the user. Moreover the pages are ranked according to the number of times the keywords occur in them, which leads to a result set with pages of least user interest ranked higher and the desired results ranked later thus troubling the user to browse a long list of pages that are not really useful and wastage of time. This post processing is a tedious job for the user. This is solved by semantic search

III. ARCHITECTURE OF SEMANTIC WEB

In semantic web, each page stored in the database contains metadata with notes, meanings, list of words, definitions, vocabulary for the annotations etc., annotations are based on the classes of concepts and relations among them. Semantic search engines are aimed at having improved search strategies using the proposed architecture and relational algorithm to satisfy user's information needs despite increased complexity the search engine will be accurate with full efforts on performance time and scalability. So, to do this, there should be a visionary model for the same.

According to Tim-berner-Lee, the model for semantic web is defined as in figure below. At lowest levels, we have URI and Unicode for modelling each element. Above that resides XML used for tags. Then RDF is used as data model. Encryption is used to provide high level of security.

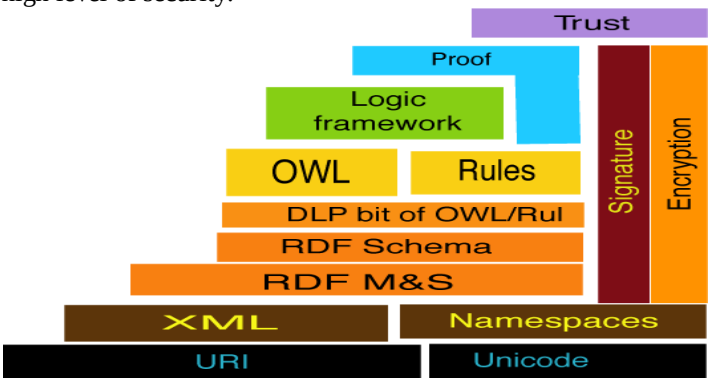

Fig -1 Architecture for Semantic Web [22]

As shown above, the major components that are needed to understand are: XML, RDF and OWL which makes semantic search meaningful. These are discussed in detail in upcoming sections

IV. FOUNDATION OF ONTOLOGY

Ontology can be defined as "formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain". Ontology categorises variables needed for some set of computations and establishes the relationships between them. Ontologies are the backbone of the Semantic Web by providing the vocabularies and formal conceptualization of a given domain (Gruber 1993) to facilitate information sharing and exchange. According to Thomas Gruber, Ontology is a formal, "explicit specification of a shared conceptualization". A conceptualization is an abstract model of a phenomenon, created by identification of the relevant concepts of the phenomenon. The concepts, the relations between them and the constraints on their use are explicitly defined. Ontology is said to be formal because it a machine readable and excludes the use of natural languages. Ontology is a "shared conceptualization" because ontologies aim at representing knowledge intended for use by a group with a common consensus for its meaning. Semantic web uses Ontology to represent knowledge share it and use it across various applications.

4.1 Structure of Ontology

Ontology is used in semantic search engine to relate documents to each other.In philosophical way, we use certain techniques to define structure of ontology. They are defined below:

4.2 Features of Ontology

4.2.1 Classes: Abstract group, sets, or collection of object is known as classes. Eg, Thing, Father, Human etc all are classes.

4.2.2 Individuals: These are ground level of classes, that is initialisation of classes. For example, Ram is a man.

4.2.3 Taxonomy: It is backbone of ontology that defines the hierarchy of the entities represented as keywords. It is generally shown in tree structure or lattice form. For example, every cow will be animal. So, in animal will belong to cow also.

4.2.4 Attributes: Parameters or properties defining the objects of ontology
4.2.5 Relationship: These define how the objects of ontology documents are related to each other and what they share
4.2.6 Restrictions: These are constraints we put on the certain objects. Like, if I search for age, it is restricted to be greater than zero.

Using above concepts, the documents available are related to each other and thus the modeled object are used to retrieve information for the query given.

V. ROLE OF XML,RDF,OWL IN ONTOLOGY CREATION

5.1 XML

XML is a markup language that gives serialization format that encodes information to pass between machines. It is used for user-defined tags. XML is just for syntax supported by RDF. RDF/XML is the one and only W3C standard syntax available for RDF.XML is similar to HTML with extensive capability to allow user to define its own tags. As an example, consider creating XML for "Sanjib is writer of this paper", then the XML for such would be:

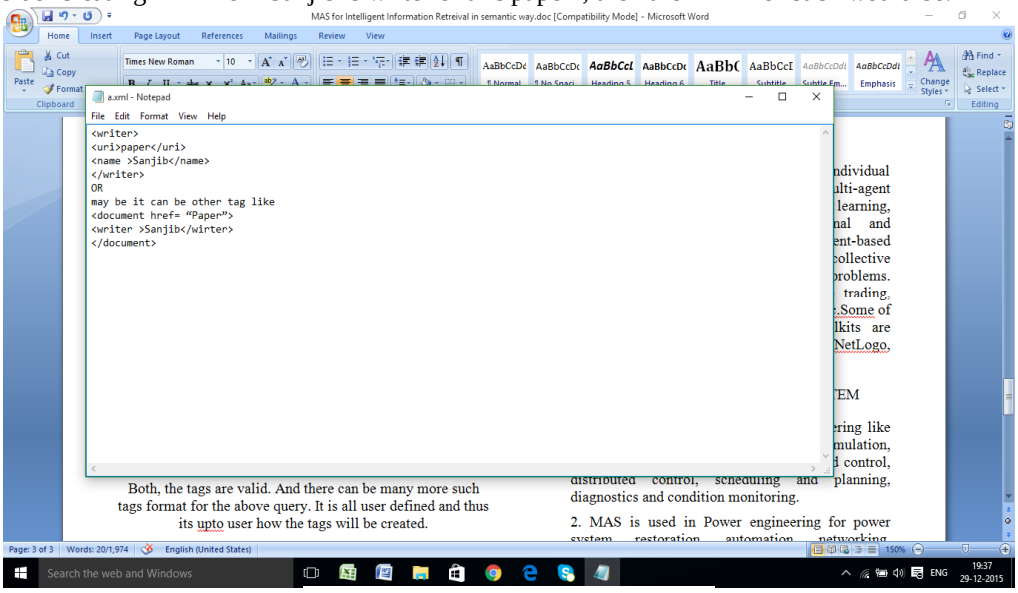

Fig-2 Sample XML Document Format

Both, the tags are valid. And there can be many more such tags format for the above query. It is all user defined and thus its upto user how the tags will be created.

5.2 RDF

RDF is a data model that stands for Resource Description Framework .Resources on web is described using this framework in graph form. XML is used to write RDF .It is designed to be read and understood by computers and not for being displayed to people. It is one of a W3C Recommendation. RDF supports triple scheme and provides data structure to it. All information is stored in triplet form. A triple represents a single edge, which is labeled with the predicate, and connecting two nodes (labeled with the subject and object); it describes a binary relationship between the subject and object via the predicate. The most vital thing defined by RDF is certainly a predicate called "rdf: type". This is used to express that things belong to certain types. The RDF model combines of 3 techniques:

5.2.1 **Subject** –It is resource that is described by the RDF statement
5.2.2 **Predicate** – It is property of the subject, uniquely identified by a URI
5.2.3 **Object** - value for the property, which can be any valid RDF data type (All XML data type are valid in RDF)

For example, in query, "Sanjib is writer of this paper", its RDFtriple structure would be:

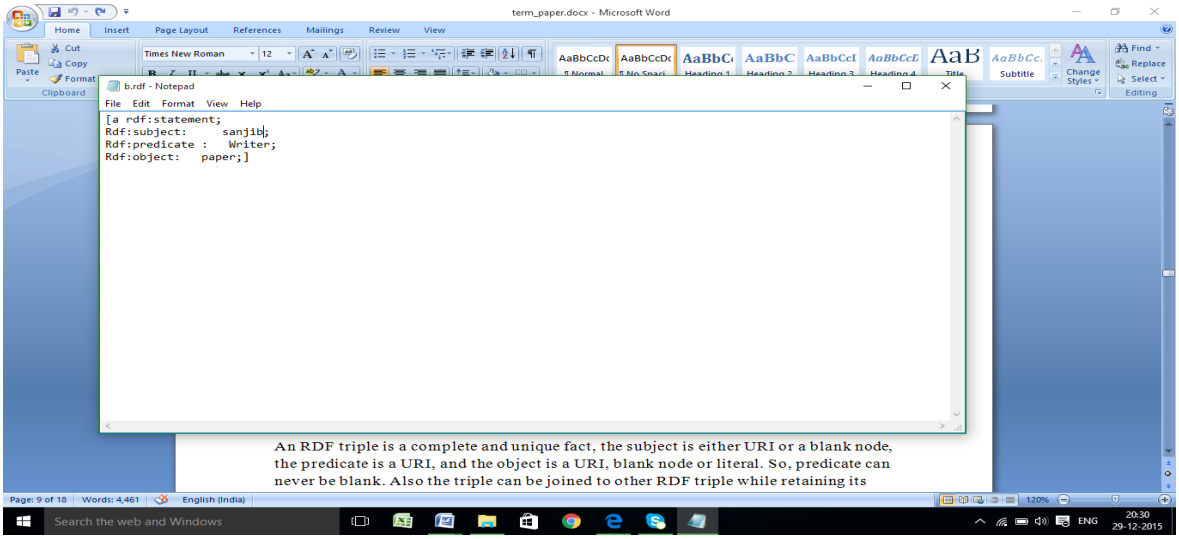

Fig-3 Sample RDF Document Format

An RDF triple is a complete and unique fact, the subject is either URI or a blank node, the predicate is a URI, and the object is a URI, blank node or literal. So, predicate can never be blank. Also the triple can be joined to other RDF triple while retaining its original meaning. The RDF element is the root of the RDF document.

- Description -element which describes a resource.
- About-an attribute which identifies the resource.
- Property - used to describes the resource within Description.

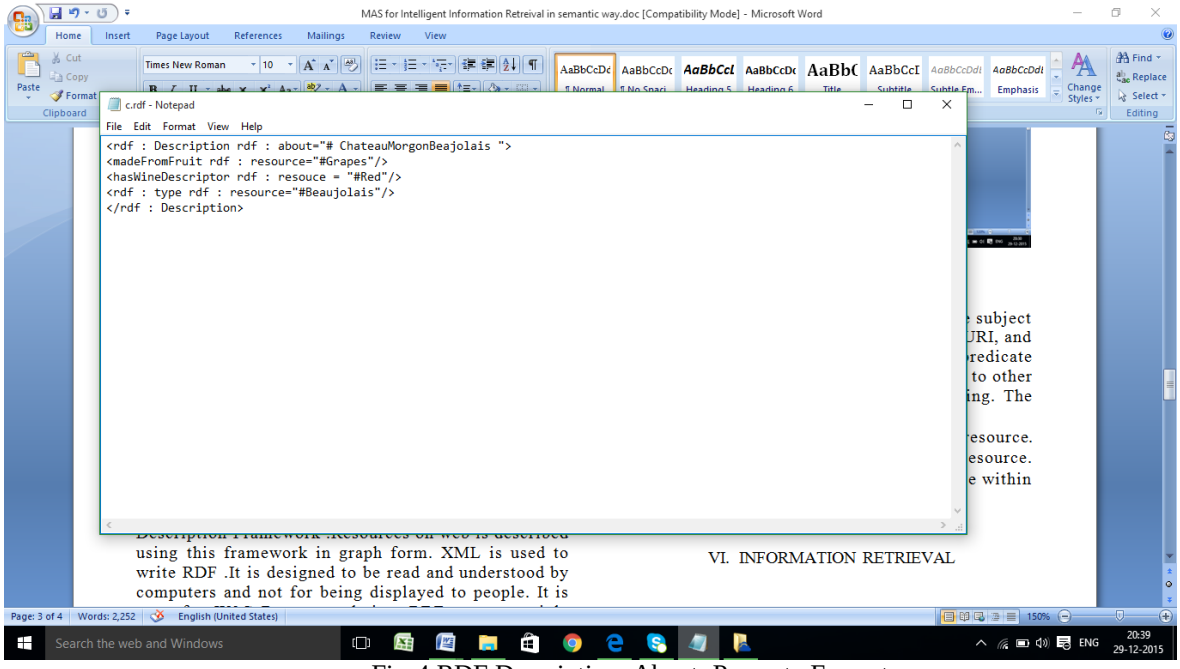

Fig-4 RDF Description, About, Property Format

RDF thus defines data model that uses the subject, predicate and object to classify the words of query and then map them to the documents available for search engine.

5.3 OWL

OWL is a W3C Web Ontology Language. It is computational logic-based language for semantic web. It has really complex as well as rich knowledge about various things. It also covers knowledge for groups of things and relations between them. Computer programs can exploit all the knowledge expressed in OWL. It is used by those applications which need processing the information content and not simply presenting it. It supports much greater machine

interpretability in comparison to XML, RDF etc. Semantic search engine uses XML capabilities to define tagging scheme. It also used RDF to exploit the representation of data. But then, there is a need for a language above RDF, which can formally describe all the meanings of the terminologies that are described in all Web documents. We needed a language that can go beyond RDF schema and thus need for OWL came into picture.

There were many features that were missing if only RDF and XML were used. They don't possess many qualities. So, OWL brought into picture many features like:

5.3.1 **Disjointness of classes**: In RDF, we can only state superclass-subclass relation of classes. But, we cannot specify whether 2 classes share any object in common or not. For example, a girl and a boy are 2 disjoint sets. This can be defined only with help of OWL

5.3.2 **Boolean operation on classes**: RDF schema lacks the capability to define classes that are result of Boolean operations on other classes. For example, there is no provision to define a set that is intersection, union or combination of 2 classes. This feature is brought by OWL.

5.3.3 **Cardinality of classes**: RDF schema again lacks ability to define classes. It cannot define that 1 male will marry only 1 female. This is also done with help of OWL then.

5.3.4 **Range for properties**: If OWL would not have been used, then we could not possess any restriction on range of properties. As an example, we could not say that goat can only eat grass, but other animals can digest meat also.

5.3.5 **Special characteristics of properties**: RDF schema does not define special characteristics like a feature saying that a property is transitive (like "is taller than") or unique (like "is mother of"). So, OWL is useful for defining such characteristics.

This above features, hence, [illegible] an ontology language that is richer than RDF Schema. We need a language which offers the above features [illegible] language we [illegible] 1.0 148.3

efficient reasoning mentioned before. RDF Schema has some very powerful modeling primitives, such as the rdfs: Class (the class of all classes) and rdf: Property (the class of all properties).

VI. INFORMATION RETRIEVAL

The process of information retrieval starts when a user enters a query into the system. Queries can be called as formal statements of data that requires. A single query may match to collection of objects with different degrees of relevancy. An object can be considered entity that is used to represent the information in database. Generally Information Retrieval systems identified by numeric score depending on how often or seldom the object matched the query in the database. In 1940s, information retrieval (IR) systems were used to manage the scientific literature but now almost all university, libraries, journals, encyclopaedia databases are using IR systems[1]. The amount of data on WWW is increasing day by day and it is not necessary that we always retrieve the best quality of data every time. To retrieve best and useful data form huge website that contain bulky data various types of links are available on WWW that result effective data and information with searching and filtering. Information retrieval is used to retrieve information from different documents available that are related to a user query. For every query Q, set of D documents is retrieved.

There are different mechanisms in which we can retrieve information from the documents available. They have changed over time to time. Below discussed are models for information retrieval, which have been used over different time frames:

6.1 Term weighting model

This is one of the simplest representations of documents for information retrieval, where documents are considered as set of different words. Weights indicate relevance. More the frequency of a word in documents, more is weight assigned to that document. Constant rank-frequency law of Zipf indicates the occurrence characteristics of words.

Frequency. Rank = constant

Simplest approach for term weighting is term frequency, where each term t(i) is weighted according to number of occurrences of word in term of document d(j).

The documents are then selected by defining the threshold, by rejecting the most frequent and the least frequent words. The reason for rejecting such words is that none of them a good candidate for search.

6.2 Boolean Model

This was most dominating search model in mid-nineties. It is simple and intuitive. It is based on Boolean algebra and set theory. Queries are specified as Boolean expression, in terms of and, or etc. According to Wikipedia, given a Boolean expression - in a normal form - Q called a query as follows:

Q = (Wi OR Wk OR ...) AND ... AND (Wj OR Ws OR ...),

with -- Wi=ti, Wk=tk, Wj=tj, Ws=ts, or Wi=NON ti, Wk=NON tk, Wj=NON tj, Ws=NON ts
where, ti means that the term ti is present in document Di, whereas NON ti means that it is not.

Example, for query of a “car vacations”, we will form the query as below:

Q= (“car” V “automobile” V “auto”) ^ (“holiday” V “vacation”).

So, the query is modelled and then it is mapped to all documents.Each term is either present or not in the given document, On basis of given query, each term is mapped

Alternatively, Query Q, can also be represented in a disjunctive normal form (DNF). A retrieval operation consists of 2 steps:

A) The sets S(i) of all documents are obtained that contain term t(i) or not(depending on whether W(i)=t(i) or W(i)=NON t(i)) :
   Sj = {Di | Wj element of Di} means documents having the term
B) Those documents are retrieved in response to Q which are the result of the corresponding sets operations, i.e. the answer to Q is as follows:
   UNION (INTERSECTION Sj)

The use of AND and OR makes the retrieval simple. But, the problem is that the document is either marked as relevant or not. It is on binary decision. There is no level of relevance of any document.

6.3 Vector Model

In vector model, the queries and the documents are represented in form of vectors.

$$\vec{d_k} = (w_{k,1}, w_{k,2}, \ldots, w_{k,t})$$
$$\vec{q} = (w_{q,1}, w_{q,2}, \ldots, w_{q,t})$$

Here the documents are represented by vector d(k) and queries by vector q(k). Unlike Boolean model, where only 0 and 1 are assigned to documents, the vector model covers this limitation. Weight w(i,k) is assigned to each document where w is greater than 0 but can be any number. Then, the documents are sorted by

descending order of their weights with respect to given query. The more the weight, the relevance of document increases.

$$RSV(\vec{q}, \vec{d_k}) = \sum_{i=1}^{t} w_{q,i} \cdot w_{k,i}$$

Ranking function is used to rank the documents. RSV (retrieval status value) is thus the value of similarity expressed by ranking function. Advantages of vector based scheme are: Partial matching, Weight assigning and ranking of documents, which were missing in binary model.

VII.MULTI AGENT SYSTEM AND PROPOSED FRAMEWORK

According to Wooldridge definition [2], An agent is "a software (or hardware) entity that is situated in some environment and is able to autonomously react to changes in that environment."A multi-agent system (M.A.S.) is an electronics machine made up of multiple interacting intelligent agents within an environment. The problems those are impossible or difficult to solve for an individual agent can be completed with the help of Multi-agent Systems. Intelligence may include reinforcement learning, method, algorithmic searching, functional and procedural approach. The principle being agent-based model is to focus on explanatory insight into collective behaviour of agent following rules rather than problems. Agent-based model typically used in online trading, modelling social structures, disaster response etc. Some of commonly used agent based modeling toolkits are Brahms , GAMA, FLAME, Mesa, NetLogo, SeSam,Visual Bots etc.

7.1 Advantages of Multi Agent System

1. MAS is very helpful in control engineering like automation, system restoration, market simulation, network control, congestion control, hybrid control, distributed control, scheduling and planning, diagnostics and condition monitoring.
2. MAS is used in Power engineering for power system restoration, automation, networking, condition monitoring etc.
3. MAS also find application in Distributed Artificial Intelligence to understand the concept of reactive and cognitive agents, Rational Agents, Reactive Agents, Finin, Estraillier etc.
4. MAS are useful for simple programming. Here a complicated program can be split into small modules and assign control of those modules to different agents.

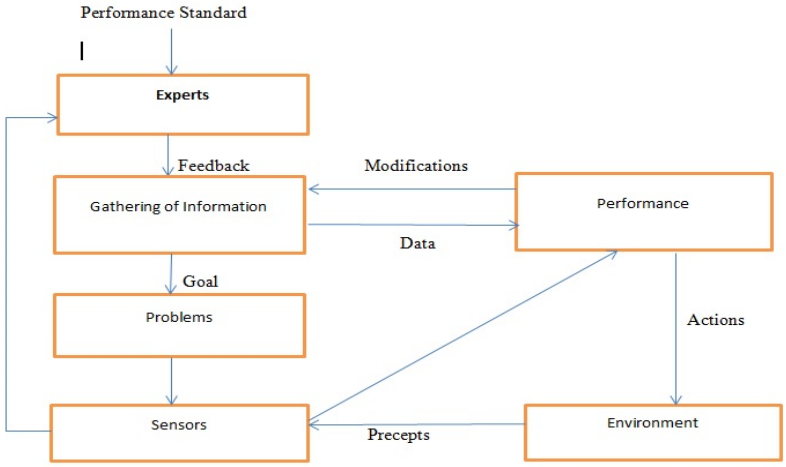

Fig-6 Proposed framework of MAS for Information Retrieval

VIII.CONCLUSION

The evaluation of different Components of Semantic Web model and information retrieval systems in this paper has shown that most available systems are not efficient at all aspects of the knowledge creation cycle. Some systems are more efficient at the internalization of knowledge, while others are better suited for the externalization. The proposed MAS model tries to overcome some of the shortcomings of current information retrieval and classification systems.

In general we can say that MAS is effective information retrieval technique in the field of distributed systems, computer supported cooperative work, organizational theory, cognitive science, knowledge representation, software engineering, distributed artificial intelligence, sociology and organizational theory.

REFERENCES

[1] Shah, U., Finin, T., Joshi, A., Cost, R. S. and Mayfield, J.'Information Retrieval on the Semantic Web.' *10th International Conference on Information and Knowledge Management*, November 2002.
[2] M. Wooldridge, "Intelligent agents," in Multi-Agent Systems, M. Wooldridge and G. Weiss, Eds., pp. 3–51, MIT Press, Cambridge, Mass, USA, 1999.
[3] H. Rui, L. Fen, S. Zhongzhi. (2008). Focused Crawling with Heterogeneous Semantic Information. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-3496-1/08, IEEE.
[4] J. LUO, X. XUE. (2010). Research on Information Retrieval System Based on Semantic Web and Multi-Agent. 2010 International Conference on Intelligent Computing and Cognitive Informatics. 978-07695-4014-6/10, IEEE.
[5] Jan Paralic and Ivan Kostial, "Ontology-based Information Retrieval," In Proc. of the 14th International Conference on Information and Intelligent systems, 2003, p. 23-28
[6] Pablo Castells, Miriam Ferna ndez, and David Vallet, "An Adaptation of the Vector-Space Model for OntologyBased Information Retrieval," IEEE Transactions on Knowledge and Data Engineering, vol. 19, 2007, p. 261-272.
[7] Rachid Ahmed-Ouamer and Arezki Hammache, "Ontology-Based Information Retrieval for e-Learning of Computer Science," International Conference on Machine and Web Intelligence, 2010, p. 250-257.
[8] Nenad Stojanovic, Rudi Studer and Ljiljana Stojanovic, "An Approach for Step-By-Step Query Refinement in the Ontology-based Information Retrieval," International Conference on Web Intelligence, 2004, p. 36-43.
[9] Nenad Stojanovic and Ljiljana Stojanovic, "A Logic-based Approach for Query Refinement in Ontology-based Information Retrieval Systems," Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, 2004, p. 450-457.
[10] Jing-Yan Wang and Zhen Zhu, "Framework Of Multi-Agent Information Retrieval System Based On Ontology And Its Application," Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, 2008,p. 1615-1620.
[11] Jibran Mustafa, Sharifullah Khan and Khalid Latif, "Ontology Based Semantic Information Retrieval," 4th International IEEE Conference Intelligent Systems, vol. 3, 2008, p. 14-19.
[12] Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, and Eero Hyv onen, "Ontology-Based Query Expansion Widget for Information Retrieval," Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web, 2009.
[13] Huiying Gao, Jinghua Zhao, Qiuju Yin and Jingxia Wang, "Ontology-based Enterprise Information Retrieval Model," Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, 2009, p. 13261330.
[14] R.Suganyakala and Dr.R.R.Rajalaxmi, "Movie Related Information Retrieval Using Ontology Based Semantic Search," Information Communication and Embedded Systems, 2013, p. 421-424
[15] Huiyong Xiao and Isabel F. Cruz, "A Multi-Ontology Approach for Personal Information Management," In Proceedings of the 1st Workshop on Semantic Desktop, 2005, p. 19-33.
[16] Jian Guan, Xiang Zhang, Jianming Deng and Yuzhong Qu, "An Ontology-Driven Information Retrieval Mechanism for Semantic Information Portals," Proceedings of the First International Conference on Semantics, Knowledge, and Grid, 2005, p. 63-66.
[17] Antonio Jimeno-Yepes, Rafael Berlanga-Llavori and Dietrich Rebholz-Schuhmann, "Ontology refinement for improved information retrieval," Semantic Annotations in Information Retrieval, vol. 46-4, 2010, p. 426–435.
[18] Poonam Yadav and R.P. Singh, "An Ontology-Based Intelligent Information Retrieval Method For Document Retrieval," International Journal of Engineering Science and Technology, vol. 4-9, 2012, p. 3970-3974.Label,"
[19] Gábor Nagypál, "Improving information retrieval effectiveness by using domain knowledge stored in ontologies," Springer Berlin Heidelberg, On the Move to Meaningful Internet Systems, vol. 3762, 2005, p. 780-789.
[20] HU Jun, LI Zhi-lu and GUAN Chun, "A Method of Rough Ontology-based Information Retrieval," IEEE International Conference on Granular Computing, 2008, p. 296-299.
[21] David Vallet , Miriam Fernandez , and Pablo Castells, "An Ontology-Based Information Retrieval Model," Springer-Verlag Berlin Heidelberg 2005, The Semantic Web: Research and Applications, vol. 3532, 2005, p. 455470.
[22] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web," Scientific American, 2001.