The survey of co-occurrence analysis method in the structural depiction of scientific domain of ontology

Nooshin Hormozinejad[1] and Atousa Koochak[2]

This study is a part of a major design to be published in a scientific journal

[1]PhD Student in Information Science and Knowledge at University of Tehran, Iran
[2] PhD Student in Information Science and Knowledge at Shahid Chamran University of Ahvaz, Iran

Abstract

Appearance of semantic web and consequently, ontologies have led to a novel revolution in information retrieval field regarding their semantic structure. Web creation and emergence, on one side, has caused a faster information growth and data bank development, on the other side, in a way that the traditional methods of information retrieval and text processing are no longer adequate to meet the individual needs for necessary information. The present authors, due to the significance of ontology visualizations, analyzed the thematic content of this field in the Web of Science database by the words co-occurrence analysis. According to this method, the subjects of ontology visualization were extracted and their interrelationships were directly obtained from thematic content. Thus, the current study is to survey the thematic content of fields and subfields of ontology and the intra-relationships of these subfields. Then, the usage and applicability of this method is analyzed in depicting this scientific field structure. This study is practical in its type and various methods of taxonomy, co-words analysis and network analysis are used. So, some series of published essays on ontology in WOS database from 2000 to 2016 including 17,015 records were extracted. After probing the key terms of these essays, the terms of ontology and its subfields were identified; after acquiring the thematic pattern of this field by VOS viewer software, the tasks of analyzing the data collected from maps, formed clustering and classification and their interrelationships were done.

Keyword: Vos viewer, terms co-occurrence analysis, thematic map, network analysis, ontology.

1. Introduction

Ontology is a common subject considered and used in different fields such as philosophy, computer sciences, artificial intelligence, semantic web, linguistics, library and information science, etc. from different perspectives. Although, the origin of ontology is philosophy, after recognizing the various uses of ontology,

Received: 31.3.2017 Accepted: 21.5.2017 ISSN 2241-1925

it was initially found by the artificial intelligence experts in the 1990s. Then, it was used by the experts of other fields such as library and information science to describe and classify concepts.

Ontology creates the possibility of a common understanding of information, an obvious and clear description of assumptions, distinction of field knowledge from practical knowledge, a clear definition of the words and deduction of the knowledge related to the given concepts.

Nowadays, regarding the increasing application of ontologies in information systems, the stated issues among the scholars are the construction of ontologies, construction methodology, construction tools, automatic construction and learning ontologies (Shamsfard, Abdullahzadeh and Barforush.2002)[1]. There are various approaches working to construct, develop and update ontologies. Many approaches are identifiable in knowledge extracting and modelling in different texts. Regardless of the chosen architecture in construction and development of ontologies, studies should be made on the methods of extracting concepts and relationships.

Now, a huge volume of information are digitally produced and are put into production and reproduction cycle. The large body of existing textual information, especially in webbed, are a good opportunity to develop the studies on artificial intelligence field, construct and develop the tools for knowledge representation in order to organize information and knowledge. It shows a tendency toward automatic methods. Construction and development of ontologies as series of concepts in semantic relationships, are dependent on extraction of concepts and relationships.

The current data and resources in webs are increasingly growing and web users need a common understanding of them. Ontology plays the main role in information exchange and in developing lexical web into semantic web. Ontology is a conceptual model which obviously models the actual entities in a certain domain and their interrelationship.

In library and information science, ontology is used as a semantics tool which can exactly represent the concepts and their connections. Although glossaries have been known as instances for a long time, the limitations of glossaries in depicting the semantic relations and emergence of new information technologies which provide an accurate representation of concepts have led to the use of ontology structure and finally reengineering of glossaries into ontology.

Todays, role and application of ontology is noticeable in knowledge-based systems. Ontology, as a powerful tool for representation and expression of knowledge of one field, is stated in a formal and processable framework. The connections among heterogeneous systems could be improved using ontology. The above points are the advantages of ontology usage. However, it should be

notified that practically concentration is on ontology for different applications which are designed and produced by different people and developers in various fields and are used in multiple systems.

Significance of ontology necessitates more serious survey in the investigations of this field. An analytic approach to the research interests of this field, its conceptual mapping, and categorizations of concepts can help finding out the scientific subjects of the considered field and their interrelationships directly from thematic content.

2. Research History

Ontology comes from a Greek term meaning universe recognition. The first part of the term refers to the universe and the other part stands for knowing a non-mythological micro-foundation which has been ideally chased by the ancient man. Thus, Greeks had been trying to recognize the universe by ontology and this the reason why it is referred to as universe recognition.

This is the same about web as a scientist named Berners-Lee stated development of world computers' internal connections in CERN(the European Organization for Nuclear Research) foundation in the early 1980s, and predicted availability of all data and files related to physics. The first practical hypertext was published in 1988. In March 1989, Berners-Lee and Robert Cailliau set up a project for a start. This project provided access to hypertext via a computer network and was called spider web which included a network of links. Web network was published as web in CERN after numerous researches in May 1991 and was publicly introduced in 1993.

Yoo et al. (2002)[2] found out that the used techniques in indexing the implication are based merely on mathematic calculations regardless of concept comprehension by the machine. cure (2003)[3] has emphasized on the traditional methods of information retrieval like the existing database in construction and development of ontologies. Raskin and Pan(2005)[4] focused on creating and using ontologies, especially on intranet environment. Casanovas et al.(2005)[5] express the stages and form of constructing thematic ontology of rights using natural language based on an ethnographic research. Currently, there are some projects on the full automatization of engineering ontologies such as weng et al. (2006)[6].

Zhang, Zu, Oh (2008)[7] extracted the semantic relationships by analyzing syntactic components. *Atefeh Sharif (2009)*[8] *did a research titled auto-engineering of ontology: feasibility study of lexical relationships extraction from Persian texts and concluded that lexical relationships can be extracted in Persian language through analysis of scientific texts. Using words networks partially solves the problem since they are networks of concepts along with their interrelationships which are structurally designed and work where only 49% of relationships are cited in text with partial and full compatibility.*

Lau (2009)[9] did a research titled a closed implementation of mono-core of wiki used in web services for efficiency improvement and analysis influence and medical ontology co-writings. He found that using Wiki being could develop standardization range of words, registering, exchanging and sharing. Also, Fenza[10] et al. in 2010, had a research titled knowledge structure to support the reality based on ontology projection which led to an effective ontology through conduction and representation of the key aspect of ontology construction throughout the extracted concepts. wang et al.(2012)[11] considered words co-occurrence analysis based on semantics and found out that, in China, four research methods have focused on it including "methods and theories of human intelligence network", "human intelligence network", " Competitive intelligence system(CIS)", and "mechanism and projection of human intelligence".

Ramakrishnan and Vijayan *(2012)[12], surveyed the development of future supportive features in the recent tools of otology projection and qualified the practical programs of the last task which used ontology to review their ideas.* Elahi et al., also in (2012)[13], had a research as the prevailing trends in innovation development in regions using words co-occurrence method. Findings of the quantitative analyses along with presented analyses based on the current knowledge study in this field have provided a good and noticeable classification of the prevailing trends of technology development in regions. Sanatju and Fathiyan[14] conducted a research in (2012) titled the methodology of design, construction and implementation of ontology: approaches, languages and tools. This essay explains the steps and processes of constructing the first ontology and its capabilities and facilities, as well.

Souri and khodamoradi (2013)[15], had a research called ontology application in lexical web and the history of ontology was considered.

Sediqi, (2014)[16], presented an essay as survey of words co-occurrence method usage in scientific domains depiction and found that, according to patterns extracted from the studied documents analysis, concepts such as information science, library, bibliometric analysis, innovation and text mining are of the most internationally used items in informing domain. Some words such as "bibliometric analysis" are present in all the studied years, whereas some other words disappear in time. New concepts appear as a recomposition of current words and in relations with new events and technologies.

Mustaphavi(2015)[17] dedicated his PhD thesis to identification and comparison of interdisciplinary relationships of information science and knowledge using scientific co-words and co-citation mappings of the scientific website based on Bordio theory. The results showed that the words of library and information make the base of studies prior to web appearance. The words like web, information, research, citation analysis, knowledge, library, magazines and technology have formed the study base of the new era. The studied concepts of

information science and knowledge, before web appearance, are categorized into 20 clusters and into 13 clusters in recent period.

As a final conclusion to the previous researches as this one, many works are done on ontology throughout the world. However, researchers did not confront similar items about co-occurrence analysis on ontology in the previous researches. On the other hand, findings in the other domains and those closed to ontology like information, scientometrics and knowledge management, etc. showed common events in words and concepts of these fields.

Then, current study tries to identify analytically the research cases of domain ontology using words co-occurrence method, and to depict the conceptual map of this domain in order to figure out the key concepts of ontology and their categorization and interrelationships. To get to such ends, the following questions should be answered:

1. How are research process and scientific productions in science ontology domain and its subdomains?
2. Which are the most productive institutes and universities in thematic domain of ontology on Web of Science from 2000 to 2016?
3. Who are the most productive authors in thematic domain of ontology on Web of Science from 2000 to 2016?
4. What are the base terms in domain of the science ontology?
5. Which are the active research domains in domain of the science ontology?
6. How are the relationships among the thematic subdomains of ontology science?

Research method is used to answer these questions as follow:
This study is practical in its type and various methods of taxonomy, co-words analysis and network analysis are used. So, some series of published essays on ontology in WOS database from 2000 to 2016 including 17,015 records were extracted.

First, 500 key words were selected as the base words. Then, due to the point that the relation of concepts number and the studied records number must be about 1 to 20, the words having less connection to the main concept are deleted and finally 324 words were selected and analyzed as the main concepts. The findings of co-words analysis in WOS Viewer software suggests 6 thematic branches in this domain. They are named by consulting with experts of the domain. Below are the answers to research and analytical questions about final results.

3. Data Analysis

To answer the questions of this research, the data of thematic ontology domain extracted from WOS database from 2000 to 2016 were surveyed and the analyses are as follow.

Scientific output growth trend in ontology domain

To answer the first questions, scientific output growth trend in ontology domain on WOS database from 2000 to 2016 was studied. This span is divided into four four-year periods (the scientific production quantity of the year 2000 was neglected due to its triviality and ability to divide the span into four periods). Totally, 17,015 documents were published in this span. A number of 1665(%10) of essays were published from 2001 to 2004. Other essays, in three periods (2005-2008, 2009-2012, 2-13-2016), had the same production rate (almost each 4-year period, %30 of the scientific productions of this domain). Scientific productions trend suggests a positive growth for the scientific outputs of ontology domain and shows a rather gradual and rising growth from 2005. Then, the most productive universities, institutes and scholars working in this period are elaborated.

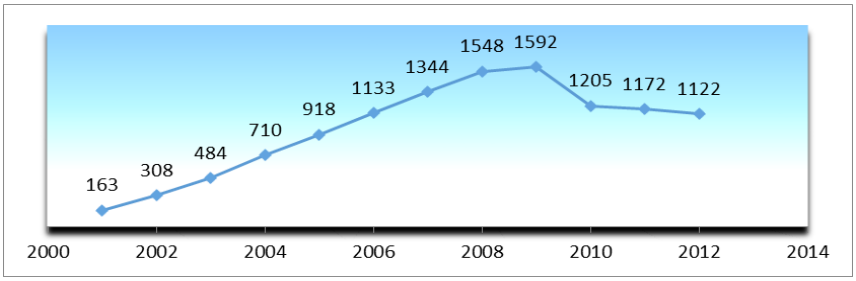

Figure 1. **The diagram of scientific outputs growth trend in ontology domain**

3.1 The universities and institutes with most contributions in creating scientific outputs in ontology domain are:

A number of 2,661 universities and institutes participate in creating scientific outputs in ontology domain. The high-productive universities and institutes with more than 100 outputs in this domain, are given in table1.

University	Number of Data	Percentage
UNIV MANCHESTER	187	1.080%
STANFORD UNIV	178	1.028 %
CHINESE ACAD SCI	164	0.947 %

WUHAN UNIV	120	0.693 %
UNIV OXFORD	114	0.658 %
UNIV POLITECN MADRID	112	0.647 %
UNIV KARLSRUHE	109	0.629 %
SHANGHAI JIAO TONG UNIV	100	0.577 %

Table 1. The high-productive universities and institutes in ontology domain.

3.2. The high-productive authors of ontology domain in a given period on Web of Science database:

A number of 8,969 authors published their academic productions on Web of Science database in the given time. The prominent figures with more than 50 essays in this span are introduced in table 2.

A survey of results of the scientific outputs of universities, institutes and scholars of this domain shows America and China as the most productive countries in thematic domain. The results were compatible with analytic data on WOS. The other active countries are England, Germany and French which were in third to fifth place respectively among 1,178 countries.

	Record Count	% of 17015	Bar Chart	
	SMITH B	76	0.439 %	
	CHANG E	62	0.358 %	
	MUSEN MA	56	0.323 %	
	GOMEZ-PEREZ A	54	0.312 %	
	STEVENS R	54	0.312 %	
	WANG Y	54	0.312 %	
	MIZOGUCHI R	53	0.306 %	
	STAAB S	51	0.294 %	

Table 2. The high-productive authors of ontology domain

3.3. Depiction of words co-occurrence map in ontology domain:

The three questions of the study end are answered in depiction of words co-occurrence map:

First, in order to survey the distribution and overlap of words in thematic subdomains, all the records extracted from the database Web of Science are restored using WOS Viewer software. So, co-lexical map of articles could be seen through the key words in all parts of the given records. Then, co-lexical analysis of co-occurrence threshold is possible for the key words.

In this study, the thematic clustering of scientific outputs of the domain ontology on Web of Science database during 2000 to 2016 were analyzed using WOS viewer software, which is especial for visualization and designed for creating and making scientometrics maps. First, co-occurrence threshold is considered 20 occurrences for each word, so that little domains do not come into the domain and will eliminate. Considering this threshold in the studied words, software has identified 324 words in the studied records and their co-lexical map would be depicted by that software (figure 1). Six clusters are seen in this picture, each one representing a certain thematic domain. Researchers, in this stage, try to name clusters consulting the thematic experts of the domain.

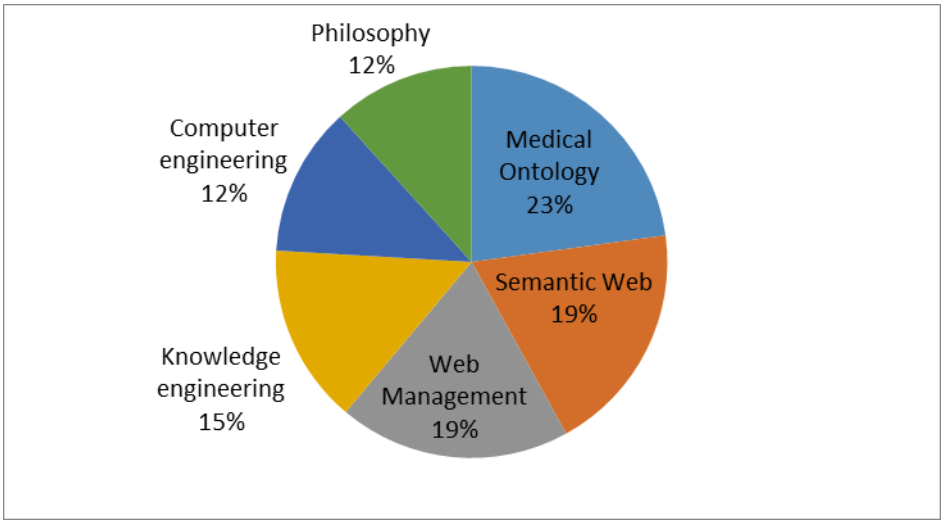

Figure 2. Contribution of each cluster in formation of co-lexical network of the domain ontology

First, for thematic division of each cluster, each cluster was named with the key word confirmed by the thematic expert of the domain. Then, the contribution each cluster has in formation of co-lexical network is displayed. The most contribution is for medical ontology with %23 and the least is for clusters of computer engineering and philosophy with %12. Following, the thematic clusters formed about ontology domain on WOS and the given span are discussed.

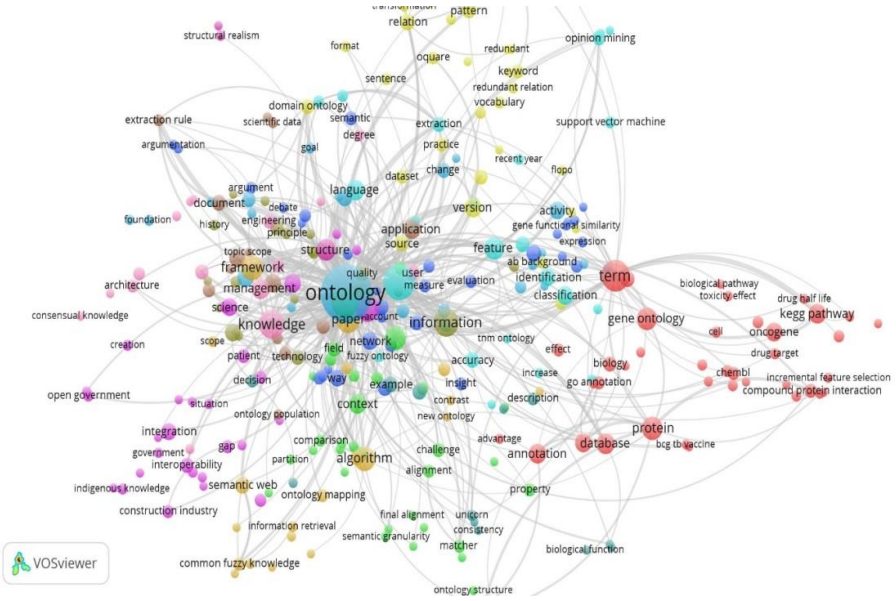

Figure 3. The co-occurrence map of words of the texts studied in ontology domain

3.4. First cluster: Thematic domain of medical ontology

This red cluster is related to medic with 74 words. The most significant word of this cluster "gene ontology" with occurrence rate of 352 and co-occurrence of 1,281 is presented on the co-occurrence map of domain ontology in figure 2. The other important words of this domain are "semantic similarity", "biology", "databases", "evolution", "bio-ontology", etc.

3.5. Second cluster: Thematic domain of semantic web

This cluster, green in figure 1, has 62 words. The most significant word of this cluster with 1,214 occurrences and co-occurrence rate is 3,059 is "semantic web" (occurrence and co-occurrence rate difference is to be discussed). The most important words of this cluster include "big data", "e-learning", "business intelligence", "ontology investment", "information management", "conceptual models", etc.

3.6. Third cluster: The thematic domain of web management

It covers the items such as technical perspectives of web, ontology of web and knowledge management. This thematic domain, red in figure 2, has 62 words. The important words of this cluster include web (433 occurrences and 1,227 co-occurrences as the most important word), information retrieval, ontology domain, algorithm, knowledge discovery, etc.

3.7. Fourth cluster: the thematic domain of knowledge engineering

This cluster, yellow in figure 2, includes 48 words. It's most significant word is ontology which includes 4,148 occurrences and 7,379 co-occurrences links. The number of occurrences and co-occurrences of this word suggests the importance and noticeability considered by the experts of this domain in the given period. The other important words of this cluster include modelling, OWL, semantic web services, stimulation, privatization, etc. Totally, from the connections between the words of this domain and its general concept, the importance of construction of ontologies in sematic restoration of information considered by researchers could be referred.

3.8. Fifth cluster: Thematic domain of computer engineering

This domain is purple in figure 2 and its main word is OWL with 491 occurrences and 1,372 co-occurrences links. It includes 40 words. The current cluster whose subject refers to computer and engineering and is made up words such as knowledge modelling, go terms, fuzzy logic, project designing samples, etc. Its words would have close relationships with the words of cluster 4 regarding engineering in designing ontologies.

3.9. Sixth cluster: The thematic domain of philosophy and human sciences

This domain is cyan in figure 2. It deals with cognitive and behavioral aspects of ontology. The key word of this cluster is knowledge with 433 occurrences and 1,580 co-occurrence links. The other concepts of this domain with 38 words include cooperation, concepts, epistemology, history, metaphysics, philosophy, etc.

It should be notified that the interrelationships of words in the final map are based on distance, i.e. the distance of circles shows the relationship between concepts. Magnitude and smallness of each circle reveals the existing knowledge of each concept. Then, considering figure 2, concentration of the studied texts would mostly be upon the subjects as" ontology", "semantic web", "knowledge", "gene ontology", "OWL", " Web" . The concepts reflect a rather good distribution on map. The distribution of one class shows that scholars have focused on different subjects in this domain.

It's worth mentioning that the distance between concepts is determined according to the approximate distance of one concept from other concepts. Vicinity or distance of concepts on this map suggests that how the current texts have talked about the relations between these two concepts and have measured their influence upon each other. In figure 2, for example, where "ontology" is near "semantic web", shows that the effect of these two words upon each other are more regarded. In fact, the scholars of this domain working on their researches in the given period have shown more concentration on ontology and its application in semantic web. However, the longer distance of the word "ontology" from the concepts "information", "categorization", "language" and

"systems" represent that scholars have shown less concentration on the relations among these concepts in their researches in a given time.

4. Conclusion

The purpose of the study was to determine the conceptual structure of this domain and forms of relationships among the thematic sub-domains using co-lexical analysis and co-occurrence study of words in domain ontology. In this study, the findings of the subject ontology show that the structure of thematic domains of this field has gradually changed and continuously developed. The permanent trend of scientific productions leads to permanent structural change in this domain as well as in the other scientific domains.

Considering the span of this scientific domain, there would be increasing interactions with other sciences. Depending on people's needs for web and receiving information on it, some scientific domains of ontology are more highlighted. The results, as well, showed that configuration of the texts of this domain is rich in using the resources of different fields. In other words, it has broad interdisciplinary relations.

The drawn maps give an obvious picture of research topics on ontology and their connections with different topics. This map represents changes and stabilities in the concepts and words related to ontology in different periods. Some words are present in all the studied years while some disappear in time. Also, some words are present in some or all clusters and some are special for one cluster. New concepts are created as recomposition of current words and in relation with novel evolutions and technologies. The other finding could be discovery of many similar words from all clusters of ontology using information concept.

A survey of clusters and the vicinity and distance of words on the map of this domain recognized the close relationship of ontology with semantic web and knowledge engineering with information retrieval among the findings of this domain in the given time.

References

Shamsfard, mehrnoosh and Abdollahzadeh Barforoush, Ahmad (2002). " Conceptual knowledge extraction from core with use the Linguistic and semantic ". Tazehaye oluom shenakhti. Vol.13,No.19, 48-66.

Yu, C., et al (2002). "Patterns in Unstructured Data: Discovery, Aggregation, and Visualization". Retrieved Oct ,13, 2004, From: http://javelina.cet.middlebury.edu/lsa/out/cover_page.htm

Cure, O (2003). "Mapping Databases to ontologies to design and maintain data in a semantic web environment". Retrieved Des ,25, 2006, From http://www.iiisci.org/journal/cvs/sci/pdfs/p704935.pdf

Raskin, R. G. ; Pan. M. J (2005). "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)". *Computers & Geosciences,* 31 : 1119–1125. Retrieved May ,5, 2007, From Elsevier Database.

Casanovas, P. et al (2005). " Iuriservice II Ontology Development". Retrieved sep, 5, 2007, from http://www.aifb.uni-karlsruhe.de/WBS/dvr/publications/ivrcasanovas.pdf

Weng, S. et al (2006). " Ontology construction for information classification". *Expert Systems with Applications*, 31 : 1–12. Retrieved May ,5, 2007, , From Elsevier Database.

. Zhang, M. ; Zhou, G. ; Aw, A (2008). "Exploring syntactic structured features over parse trees for relation extraction using kernel methods". *Information Processing and Management* , 44 :687–701 .Retrieved Jan ,20, 2008, From ScienceDirect Database.

Sharif, Atefeh, (2009). Ontology Auto-engineering: Feasibility of Extracting Semantic Relationships in Persian Texts and Their Visibility Rate. Library and Information science,Vol.46, No.21, 243-263.

LAU, ADELA S. M. (2009). Implementation of an onto-wiki toolkit using web services to improve the efficiency and effectiveness of medical ontology co-authoring and and analysis. Information for health & social care. Vol.34, No.1, 73-80.

De Maio, Carmen, Fenza, Giuseppe, Loia, Vincenzo & Senatore, Sabrina (2010). Knowledge Structuring to Support Facet – Based Ontology Visualization. International Journal of Intelligent System. 25. 1249-1264.

Wang, Z. Y., Li, G. Li, C. Y., & Li, A. (2012). Research on the sementic-based co- word analysis. Social Studies of Science, Vol.19,No. 3, 473-496.

Ramakrishnan, Sivakumar, Vijayan, Arivoli (2014). A Study on development of Cognitive Support Features in Recent Ontology Visualization Tools. Artif Intell Rev. 41. 595-623.

Elahi, Shaaban, Naqizadeh, Reza, Qazinoor, Sepehr and Manteqi, Manochehr(2012). *Identifying Dominant Trends on Developing Innovations in Regions using Words Co-occurrence Analysis Method*,. Vol.6,No.30, 136-158.

Sanatjo, Aazam, Fathian, Akram(2012). "Methodology of Designation, Construction and Implementation of Ontology: Approaches, Languages and Tools (the Case Study of Ontology Design ASFAOnt in Library and Information Science)." *Library and Information Science Quarterly*,Vol. 1,No.15, 113-142.

Khodamoradi,, Mohammad and Suri, Danial(2013)."Ontology Usage in Semantic Web". The First National Conference in Computer and Information Technology Engineering.

Sediqi, Mehri(2014). "Application of Words Co-occurrence Analysis Method in Depiction of Scientific Structure: the case study of information domain." *Information management and processing*, Vol.30, No.2, 373- 396.

Mustaphavi, Ismael(2015). Identification and Comparison of Interdisciplinary Relationships of Information and Scientometrics Using Scientific Mappings of Co-words and Co-citation in Science Website Based on Bordio's Theory. Master of Art Thesis. Shahid Chamran University of Ahvaz. Faculty of Training Science and Psychology. *Information Science and Knowledge Group*.