Aggregating Similarity Measures based Ontology on Documents Retrieval

Ghaidaa A. Al-sultany **Samaher Hussein Ali**
College of IT University of Babylon, Iraq
Ghaidaa.balil@itnet.uobabylon.edu.net Samaher@itnet.uobabylon.edu.net

Abstract

This paper investigates a methodology for the ontology based semantic retrieval of annotated web documents with terms occurrence weighting. The semantic structural distance of document terms in terms of domain ontology is computed against new unknown queries to improve the documents ranking and retrieval. Furthermore, the role of aggregation methods to combine the weighting terms scheme based similarity measures with the similarity of semantic distance with respect to ontology have been carried out. Domain ontology is developed in which it defines a number of keywords and properties for semantic reasoning. The performance of the proposed method has been evaluated on health documents from a number of aspects, and experimental results are encouraging showing its effectiveness.

Keywords: Semantic distance, Resnik, Ontology, Aggregating

الخلاصة

مع التطور المتزايد في مجال تقنيات البحث الدلالي في شبكة المعلومات الدولية وازدياد الحاجة الى الدقة والسرعة في استرجاع البحوث والوثائق زادت الحاجة الى تبني طرق بحث وخوارزميات للاسترجاع وتقييم اداءها. يتناول البحث طريقة منهجية لبناء الأنطولوجيا لغرض الاسترجاع الدلالي للوثائق المذيلة بمجموعة من الكلمات الدلالية الموزونة والمعرفة نسبة الى الانطولوجيا. لاسترجاع ادق تم تجميع طريقتي البحث الدلالي المعتمد على احتساب المسافة الهيكلية الدلالية لكل كلمة مختارة من الوثيقة ومقارنتها مع مجموعة الاستعلامات الجديدة والمقدمة من قبل المستخدم اضافة الى احتساب وزن هذه الاستعلامات نسبة الى كلمات الوثيقة المسترجعة. وقد تم تقييم أداء الطريقة المقترحة باستخدام عدد من طرق التقييم المعروفة مثل الدقة والاسترجاع والتوزيع الهندسي، وقد اظهرت النتائج التجريبية ونسبة الاسترجاع فاعلية الطريقة المقترحة.

الكلمات المفتاحية : البحث الدلالي، طريقة ريزنك ، الانطولوجيا ، ارشفة الوثائق

1.INTRODUCTION

With the rapid growth of documents, web pages and other different textual content, great challenges of relatedness have been posed to the current content based systems. Semantic technology plays a vital role and offers a feasible approach to documents management in which it makes possible to retrieve the target information more accurate than keywords based methods 00 . Ontology is widely studied in many fields, such as artificial intelligence and information retrieval. It offers an easier access to the information of specific domain in addition to revealing the domain concepts and relations. Ontology can summarise the amount of information by encoding the structure of a domain components with entity classes, the relationships among them and distinguishing the features that describing the classes, subclasses and properties 00. Knowledge-based approach that derives similarity measures from ontologies is followed in this paper to support users in performing tasks semantically and adapt to context changes automatically for documents retrieval. The ontology based distance approach and annotation scheme significantly improves the retrieval performance especially for the top ranked document 0. In this paper, we discuss facilitating documents retrieval with semantic retrieval model based on domain ontology. The search system takes advantage of ontology based semantic annotation. Experimental results indicate that combining the ontology distance measure and semantic annotation weights improves the retrieval performance.

The rest of the paper is structured as follows. Section II presents the relevant research work.

Section III present the ontology based query reasoning, while in section IV the suggested similarity measures are explained. And section V discusses the retrieval evaluation. And finally, conclusion of the paper is presented in section VI, followed by the references.

2. RELATED WORK

Many works which employ the Semantic Web technologies for information and retrieval have been done recently 0. Ontology plays crucial role in reasoning and expressing contextual information in context aware research where the interoperability between them ensure extracting, interpreting, and sharing context information and present the status of things in the environment 000. Most of the ontologies and semantic similarity function have been emphasized recent investigation specifically in information retrieval research **خطأ! لم يتم العثور على مصدر المرجع.**0 0.

Ontology diversity is restricted to research needs for example, A holistic architecture of documents' semantic annotation and retrieval has been introduced by 0 in which an integrating information extraction using GATE was used to achieve fully automatic annotation and improve documents retrieval . Consequently, Lee in 0 was designed an ontology based retrieval model by suggesting ranking algorithm which uses semantic indexing based on annotation weighting techniques. In (Yu *et. al.* ,2006), the authors have adopted an approach to map text headings to ontology's entries. However, the mapping is based on exact matching between a specific ontology concept and the title of a text fragment using transformations methods such as N-grams and stemming algorithm to performance improvement. A new method is proposed in (Kaburlasos *et.al.*, 2007), in which data extraction ontologies for specific domain are utilised to annotate Web pages using automated semantic annotation. In spite of the notion that adopted to avoid the techniques of extracting information heuristics, in this research annotating candidate instances with concepts of a given domain ontology require an expert of that domain in order to import its formalised semantics. EgoIR is an ontology based information retrieval to manage, search, sharing and retrieve government documents in right time and right situation using ontology/search based server 0. Izumi et al have studied social context awareness ontology for elder people supervision 0. While GUO et al 0 built object ontology for smart indoor environment to detect hidden objects in physical artifacts. FLAME2008 platform was successfully developed to support mobile users with personalized context-aware services 0. Linguistic patterns that express semantic meaning of annotated text documents with named entities are implemented by (Li *et.al.*, 2007) where the proposed mechanism selects the best pattern that match to the annotated entity. Although the accuracy of this method is sufficiently high, its recall is limited as only named entities are annotated, which exist in specified documents in the Web pages. Similarly, Ontea system, in (Berners-Lee 2002) has adopted Web documents annotation based on lemmatization methods and regular expression patterns. The method limitation here is the need for predefined patterns for specific domain is required which affects the system performance.

3. ONTOLOGY BASED DOCUMENT ANNOTATION

Obviously, domain shares common concepts and knowledge of a specific field of documents such as fiction, non-fiction, entertainment, and medical. Domain ontology is modeled into set of ontological knowledge modules 0 with different levels that capture features of document terms $O_i \,|O_i \in O, i = 1,2, \dots, m$, where O is the structured domain ontology, and m is the number of modules in O.. These sub-ontological modules have set of concepts related with the document terms and may much more details of sub-concepts, properties and individuals. An OWL ontology reasoner is used to infer additional individuals, sub-concepts and extra statements from the main concepts to compose further connections from O_i and improve the information retrieval of the requested documents.

The annotation process of a document is done by selecting the words with high frequently occurrence and with high similarity relatedness. Therefore, the selection process of annotated terms for a document is calculated by combining its frequency (the term occurrence in the document) and the semantic distance between the selected term and the main document concept in the domain ontology.

The TF.IDF method as declared in the equation (1) is used to compute the weight of a term in a document in which, the term frequency (TF) is the number of times that a term t appears in a document. While the inverse document frequency (IDF) refers to the inverse of document frequency for a specified term.

$$tf\_idf = tf.\log({}^{N}/_{df}) \qquad (1)$$

Where, *tf* is the count of the word normalized by the total length of the document and *N* is the total number of documents.

For pre-processing purposes on the selected documents for semantic annotations and indexing, some common words are usually not considered in search engines in order to speed up the processing and the retrieving results. These filtered words are known as 'Stop Words' (such as is, are, the and so on) 0. Stemming is a technique to find morphological variants of search terms for improving information retrieval performance 0. It refers to the act of conflating or combining the variant, hence, this reduces all words with the same root to a single form and increase the recall accordingly. Stemming can be processed manually or programmatically using computer program called stemmers such as Porter 0.

4. SIMILARITY MEASURES

The relations between entities are discovered through the measure of their similarity. The ontology based similarity relatedness between sets of concepts helps in retrieving and filtering information in automatic way. Two similarity measures are discussed in this section, in which a mapping process between the terms of unknown query and set of predefined documents annotated terms are carried out. Each individual similarity matching measure is treated as a matcher while the result obtained from its process is considered as the similarity between two terms. Obviously, an isolated matching technique is not adequate enough to give an accurate match between two terms. Therefore, it has been proposed the combination of high occurrence and distance based matching techniques that provide a better accuracy about the overall similarity of the compared terms with respect to the ontology.

A. Resnik Based Similarity

Semantic similarity between ontology terms can be measured using different methods. .Resnik's measure 0 is a one of the measures in which a set of documents or terms within term lists are assigned a metric based on the similarity of their meaning / semantic content. This approach computes a probability of occurrence of an instance of the concept in a particular concept. Hence, it is based on the information content of a specific term to calculate its probability as defined in (2).

$$IC(C_i) = \log(p(C_i)) \qquad (2)$$

Where IC(C_i) is the entropy (information content) of the concept C_i . While p(C_i) refers to the *probability* of the concept occurrence that is computed by dividing the number of instances of Ci by the total number in the corpus. The semantic similarity using Resnik's measure can be

obtained per the frequency of terms appearance in the corpus by providing a systematic way to detect which entity classes are most similar to each other and, therefore, which entity classes are the best candidates for establishing the similarity between two terms with respect to the domain ontology.

B. Distance based Similarity

It is proved that the similarity based on the linguistic or string is insufficient to match between two terms. The structural similarity information plays vital role in information retrieval in which, the semantic distance between two entities are computed in terms of their structural features like, their relation with other entities and their direct properties. This means the structure similarity of the two terms $\boldsymbol{p}$ and $\boldsymbol{p_1}$ is computed by considering the similarities in terms of super-classes, sub-classes and properties. Hence, if those two terms have similar upper-classes or sub-classes in hierarchy, it is likely that they define the same concept. Based on the equation defined in (3) in 0, the similarity between query terms p_Q and document corpus p_A are calculated by assigning a numerical degree for each match to quantify the relationships between them

$$dom(P,P_1)= \begin{cases} exact\ match & \\ \frac{1}{2}+\frac{1}{e^{(||PQ,PA||-1)}} & plugin\text{match}, ||PQ,PA|| \geq 2, \\ \frac{1}{2\times e^{(||PQ,PA||-1)}} & subsume\text{match}, ||PQ,PA|| \geq 1, \\ 0.5 & uncertain\text{match}, \\ 0 & no\ match \end{cases} \quad (3)$$

Where, $||P_Q, P_A||$ be the semantic distance between p_Q and p_A in terms of the domain ontology O and $dom(p_Q,p_A)$ be the degree of similarity between p_Q and p_A. While *Exact match* means that p_Q and p_A are equivalent, *Plug-in match* denote that p_A subsumes p_Q, *Subsume match* indicates p_Q subsumes p_A and *Nomatch* refers to no subsumption between p_Q and p_A.

A degree of similarity is computed using the Equation (4), which is the mean value of the maximal match degrees of every property of a selected pattern when all the properties used in the query.

$$S(Q,s) = \sum_{i=1}^{M}\sum_{j=1}^{N} \max(dom(PQj,PAi)/M \quad (4)$$

Different techniques have been used to aggregate the results of different similarity matchers that address the different results obtained from multiple similarity measures for a specific term during the matching process.

In this research, the average of values returned by all similarity measures has been used to calculate the ultimate similarity value that determine the retrieved documents as well as represent the scale for document ranking.

5. Retrieval Evaluation

Computing the semantic similarity of the query against the documents corpora takes into account the semantic level of matched keywords in terms of the ontology. The modules of ontology comprise with different levels of information to form categorized clusters that share common properties. The accuracy of the proposed technique has been evaluated against the result set generated by utilizing web health documents. For example, let running the new unknown query keywords "cancer surgery operation". The new query keywords mapped with

respect to the domain ontology to retrieve the meaningful keywords. As a result, a set of documents annotated keywords are initiated to the new unknown query. Consequently, both the query keywords and document annotated keywords are matched with the domain ontology concepts to measure the semantic relatedness between them based on the equation 3 and 4. After that the average of distance similarity and resnik similarity are computed to calculate the final similarity degree.

Unlike The keyword based retrieval, the semantic information retrieval has been associated to the document corpora by using domain ontology's that describe set of relevant concepts linked to a specific domain.

Several measures such as precision, recall and geometric distribution are used to evaluate the performance of document retrieval.

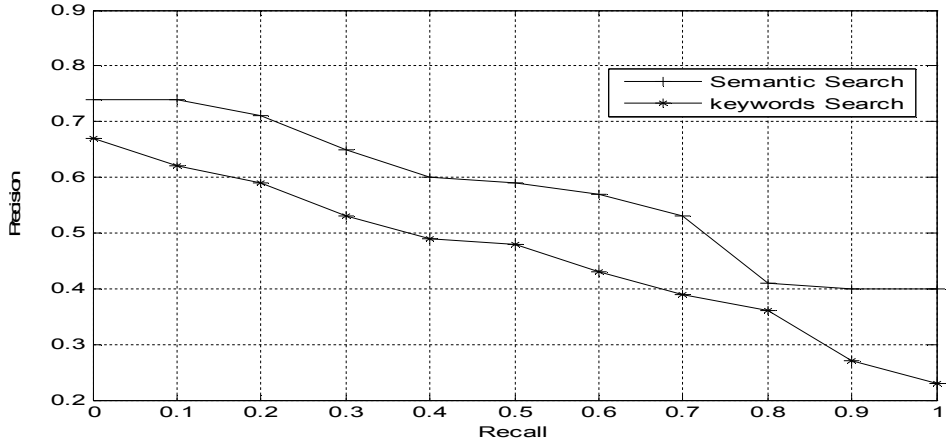
A. Precision and Recall

Recall (*R*) and precision (*P*) 0 can be calculated as follows:

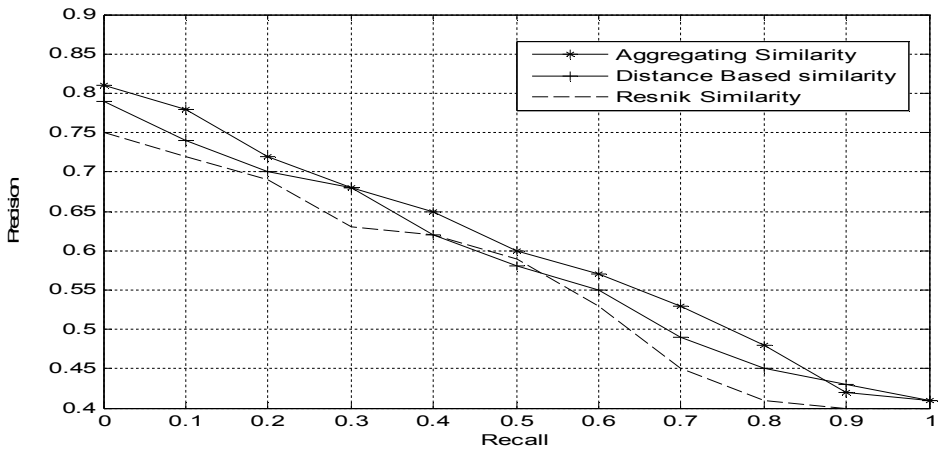
$$R = \frac{|Retrel|}{|Rel|}$$

$$P = \frac{|Retrel|}{|Ret|}$$

Where *Rel* refers to relevant documents, *Ret* be the number of returned documents and the returned relevant documents is denoted as *RetRel*. A set of 300 queries were chosen for testing the comparative performance measurement and to compute the match degree with compared to 1500 documents based on their clusters. Two test has been carried out, the first test was conducted by comparing between the proposed semantic - search against the traditional keywords based search. The input query keywords are selected randomly as shown in the Figure 1. Whereas, the second test was evaluated the precision and recall for the similarity measures combination in comapred with their perfomance seperatly as demonstrated in the Figure 2. The results in both tests show that a better precision and recall are attained using ontology based semantic distance search aggregated with the terms frequency.



Fiqure 1:Comparison between Ontology based and Keyword based search



Fiqure 2: Precision and recall for aggregated semantic search

B. Probabilistic Evaluation

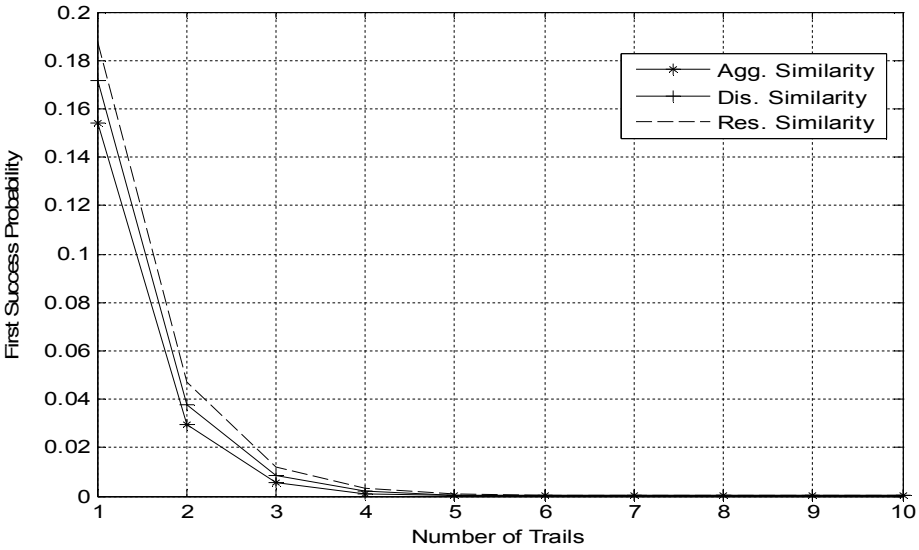
The Geometric distribution (C. M. Grinstead , Online) which refers to the probability of the number of independent trials that required to retrieve the first success value is applied to compute The probability of retrieving the first success documents using the Equation (5) .

$$G(x,p) = pq^{x-1} \qquad (5)$$

Where p and q the probability of success and failure respectively, G is a Geometric distributed variable, and $x=1,2,\ldots,n$ gives the number of n trials in which the first successes occurs. We have chosen 300 query keywords were chosen for testing to evaluate the match degree with compared to a set of 1500 documents in terms of the ontology. The test was carried out for 7 trials as shown in the Figure 3 and Table 1.

Table 1. Geometric distribution

Trail's No	1	2	3	4	5	6	7
$G_{Sa}(x,p)$	0.81	0.1539	0.0292	0.0056	0.0011	0.0002	0.0000
$G_{Sd}(x,p)$	0.78	0.1716	0.0378	0.0083	0.0018	0.0004	0.0001
$G_{Sr}(x,p)$	0.75	0.1875	0.0469	0.0117	0.0029	0.0007	0.0002



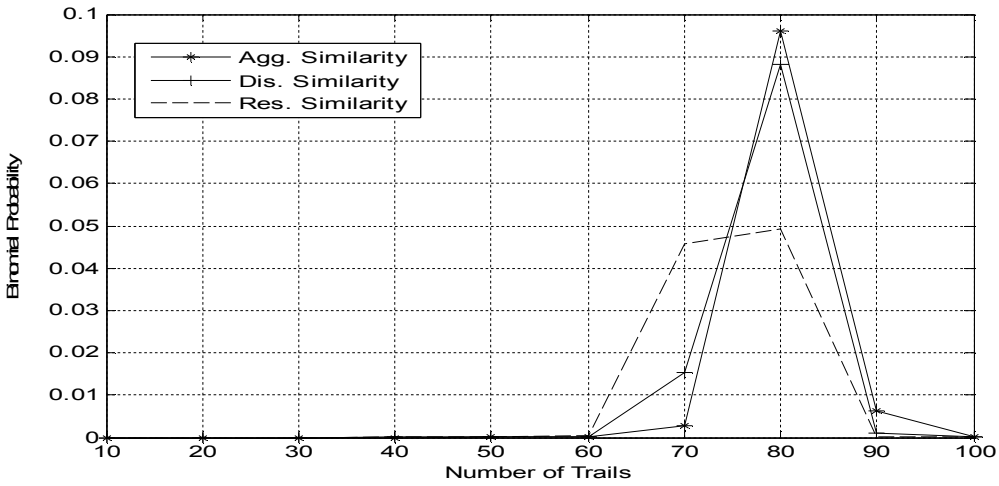
Fiqure 3: The first success probability of aggregated similarity

The performance of the ontology-based method was discussed with respect to the efficiency of the document retrieval against the keyword-based method. Probabilistic approach (C. M. Grinstead , Online) was used to validate the framework efficiency, which gives a general idea from the user's perspective about the success probability of retrieving required documents.
Based on the outcomes of the search, in which two possible values can have, p (probability of success) and q (probability of success), the efficiency of a system can be assessed by computing the probability of successes. The Binomial Distribution b (x; n, p), as defined by the Equation (6), is calculated to estimate the probability of successes for the proposed method based on number of trials.

$$b(x, n, p) = \binom{n}{x} p^x \; q^{n-x} \qquad (6)$$

Where are x is the number of successes, n is the number of trials and p is the success probability respectively.

A comparison is carried out by computing the probability of getting x successes for a number of trials, in which, 1300 documents were carried out based on first success probability of 0.81 for Aggregated similarity, 0.78 for Distance based similarity and 0.75 for Resnik similarity. The Figure 4 shows the result of the Binomial distribution.



Fiqure 4: The Binomail distribution

6.CONCLUSION

In this paper we have presented an aggregating occurrence and semantic distance based similarity in documents retrieval based on domain ontology concepts. The results of aggregating show an enhancement in retrieval accuracy against applying each measurement seperatly in terms of several evaluation experiements. The similarity findings are evaluated with respect to precision and recall and probapbilistics measures. Aggregated similarity approach has shown better performance in all test with compared with Resnik similarity. Howeever, it has shown good performance in most test against similarity based distance because of the latter followed the semantic searching in its retrieval process. cases. Currently, we are working to deal with uncertainties in similarity values during the mapping process between the query keywords and retrieved documents using naïve Bayes algorithm.

REFERENCES

Al-Sultany, G. M. Li, S. Jan, and H. Al-raweshidy, 2010 ***"Facilitating Mobile Communication with Annotated Messages,"*** 10th IEEE International Conference on Computer and Information Technology, pp. 755-760.

Asadi N. and J. Lin 2010 ***"Fast Candidate Generation for Two-Phase Document Ranking□: Postings List Intersection with Bloom Filters",*** 2419–2422.

Berners-Lee, T. J. Hendler And O. Lassila, 2002 ***"The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities"***. Scientific American Special Online Issue.

Gómez-Pérez, F. A. Ortiz-Rodriguez, and B. Villazón-Terrazas, 2006 ***"Ontology-based legal information retrieval to improve the information access in e-government,"*** Proceedings of the conference on World Wide Web - WWW '06, pp. 1007- 1008.

Grinstead C. M. and J. L. Snell. **" Introduction to Probability"**, Available at: http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book.

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

Guo, B. S. Satake, and M. Imai, 2008 ***"Home-explorer: ontology-based physical artifact search and hidden object detection system,"*** Mobile Information System, vol. 4, no. 2, pp. 81–103.

Horridge, M. J. Simon, M. Georgina, R. Alan, S. Robert, and W. Chris, 2007 ***"A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools,"*** Edition 1.1, The University of Manchester.

Izumi, S. K. Yamanaka, Y. Tokairin, H. Takahashi, T. Suganuma, and N. Shiratori, 2009 ***"Ubiquitous Supervisory System based on Social Contexts using Ontology,"*** Mobile Information Systems MIS , vol. 5, no. 2, pp. 141–163.

Jearome, E. P. Shvaiko, 2007 ***"Ontology Matching"***, Springer-Verlag, Berlin HeidelbergDE , ,isbn:3-540- 49611-4.

Jiang J. and D. Conrath, 2008 ***"Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,"*** Proc. Int'l.

Kaburlasos, V. G. I. N. Athanasiadis and P. A. Mitkas 2007 ***"Fuzzy Lattice Reasoning Classifier and its Application for ambient ozone estimation"***, International Journal of Approximate Reasoning, 451 pp. 152-188.

KARA, S. 2012 ***"An Ontology-Based Retrieval System Using Semantic Indexing"***. *Information Systems*.

Lee, M. Kim, J. and Y. Lee, 1993 ***"Information Retrieval Based on Conceptual Distance in IS-A Hierarchies,"*** J. Documentation, vol. 49, pp. 188-207.

Li, Y. Y. Wang, and X. Huang, 2007 ***"A Relation- Based Search Engine in Semantic Web"***, IEEE Transaction on Knowledge and Data Engineering, vol.19, pp.273-282.

M. Li, B. Yu, Rana O., Wang Z. and Member S. 2008 ***"Grid Service Discovery with Rough Sets"***, *Knowledge Creation Diffusion Utilization,* 206 pp. 851-862.

McGuinness, D.L. and F. Harmelen,2004 ***"OWL Web Ontology Language Overview,"*** World Wide Web Consortium W3C recommendation. *www.w3. org/TR/owl-features.*

Porter, M. F. 2012 **"An Algorithm For Suffix Stripping, Computer Laboratory, Cambridge"**. www.emeraldinsight.com/0033-0337.htm

Resink. P. 1999 ***" Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language",*** Journal of Artificial Intelligence Research, Vol 11, pp. 95-130.

Shamsfard, M. A. Nematzadeh and S. Motiee, 2006 ***" ORank: An Ontology Based System for Ranking Documents"***, International Journal of Computer Science, vol .1, pp.225- 231.

Su L. T. 1994 ***"The relevance of recall and precision in user evaluation",*** *Journal of the American Society for Information Science*, April 1994, 453 pp. 207-217.

Wang, X. H. D. Q. Zhang, T. Gu and H. K. Pung , *2004* ***"Ontology Based Context Modeling and Reasoning using OWL"***, *Pervasive Computing and Communications Workshops,. Proceedings of the Second IEEE Annual Conference*, 14-17 March 2004, pp. 18-22.

Weißenberg, N. A. Voisard, and R. Gartmann. 2004 ***"Using ontologies in personalized mobile applications,"*** in D. Pfoser and I. Cruz Eds. , Proceeding of the Intl. ACM GIS Symposium, ACM Press: New York,

Wilbur, W. J. K. Sirotkin, 1992 ***"The Automatic Identification Of Stop Words",*** *Journal of Information Science,* 181 pp. 45-55.

Yu and M. Li, B. 2006 ***"RSSM: A Rough Sets based Service Matchmaking Algorithm"***, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.7384&rep=rep1&type=pdf.