Semi-Automated Semantic Annotation for Semantic Advertising Networks

Aseel A. S. Addawood
Department of Information Science
College of Computer & Information Sciences,
Imam Mohammad bin Saud University
Riyadh, Saudi Arabia
Addawood@ccis.imamu.edu.sa

Lilac A. E. Al-Safadi
Department of Information Technology
College of Computer & Information Sciences,
King Saud University
Riyadh, Saudi Arabia
lalsafadi@ksu.edu.sa

***Abstract*—In this work, we present a semi-automatic semantic annotation system which is designed to link publisher entries to an existing Ontology and instances. It assists in categorizing the content of Web sites and associating advertisements with publishers. Semantic annotation can enhance information retrieval and improve interoperability. Since manual annotation is inefficient, Automatic or semi-automatic annotation makes the process of annotation fast and easy. The suggested system is integrated in the publisher's registration platform of the semantic advertising network and serves to facilitate registration, semantic annotation and information utilization in Web sites. This system is part of a semantic advertising network prototype called Lexeme.**

Keywords-semantic web; semi-automatic annotation; Advertising Networks; RDF.

I. INTRODUCTION

Advertising networks or ad networks are companies that connect potential advertising Web sites with potential advertisers [5]. Ad networks have made advertising on the Web very easy. Advertisers pay Web site owners ("Publishers") to allow ads to be shown on their sites. The Publisher is an individual or corporation responsible for the distribution of digital publications who will be using the ad network to make a profit from facilitating easy and dynamic publishing of ads on his Web site.

The reliance of ad networks on the keywords (in the content) without an accurate interpretation of the context of the page results in a display of irrelevant and unappealing ads on a Web page [1]. The Semantic Web is a technology that can be utilized by publishers to analyze the meanings behind a word or words. It will help to place ads in prime web locations for the sole purpose of reaching their targeted consumers [2]. We envision an algorithm that contains not only information instructing machines as to what ads to display, but also structured data which make machines understand what ads have been displayed. Such structured information that can be read and understood by computers [3] is the key. It enables machine-to-machine exchange and automated processing in a way that computers can understand [4]. In a time of mass content creation, improving ad placement through more optimized, findable content ushers in a new era of Semantic technology . It delivers the right message to the right user. The Semantic advertising networks combine the desirable features of both advertising networks and the Semantic Web. Semantic documents, Web sites and ads are generally written in the Resource Description Framework (RDF) and Web Ontology Language (OWL) languages [4]. Currently, only a few semantic documents exist on the Internet.

The challenge addressed by this paper is related to automating the provision of semantic structure to publishing Web sites. The semantic structure is provided to individual pieces of information in the Web site and interlinks these pieces with semantic relations. This results in a meaningful organization of content.

The paper is organized as follows. In the next section, we present the semantic representation of Web site content, and in Section 3, we show the related work. In Section 4, we discuss the proposed semantic annotation system. In Section 5, we discuss what technology has been used in developing the prototype. In Section 6, we show the experiment results. The conclusion and future work are given in Section 7.

II. SEMANTIC REPRESENTATION OF WEB SITE CONTENT

In this section, we focus on annotating Web site content with semantic representations. Semantic annotation helps in effectively matching Web site content with relevant resources.

Currently, a document might be one page filled by text. The data itself are not structured in a way that can be interpreted by a computer. There is no complex logic or reasoning concerning the data; there are only simple keyword-matching algorithms. At this stage, it is necessary to establish a relation among the data so that it can be considered a semantic web.

However, the next generation of the Web, the Semantic Web, makes machines more intelligent as a result of better algorithms used to process data on the Web. Web 3.0 is about data that is connected and capable of being reassembled on demand. This reassembly of data and the reorganization of data pieces is a central factor of Web 3.0. [6].

ISBN: 978-1-61208-162-5

The resulting intelligence in the structure and format of the data yields a richer relationship and linking infrastructure of data on the Web. The Semantic Web specifications, in particular RDF and OWL, are the only technology specifications that were purpose-built for use as a metadata language entirely dedicated to describing and linking data of all sorts at Web scale [7]. Therefore, the Semantic Web introduces a logical language that human programmers can use to inform computers of the relationships among data. It pursues an important goal of creating a new form of Web content that is meaningful to computers.

Ontology is an explicit formal specification of the terms in the domain and relations among them [9]. Ontology can be used to define the underlying semantics of the Web site content through the semantic annotation system, as illustrated in Figure 1 below. Semantic annotation helps in linking Ontology with Web site content for an efficient and easy embedding of semantics. The annotation results are stored in an RDF metadata store.

III. RELATED WORK

This paper is motivated by the need for adding metadata to existing web pages in an efficient way and establishing relationships among the data. There have already been certain researches conducted which are connected with our mechanisms for semi-automatic semantic annotation, described in Section 4. Our attempt is to demonstrate the application of semantic annotation of publishing Web sites to serve advertising network applications. The suggested annotation system supports an integrated environment with the registration of publishing Web sites in a semantic advertising network system. In addition, it supports document-annotation consistency and separate annotation storage. It automatically links salient terms in a Web site to relevant ontological instances/classes and their properties. It uses simple lexicon-based parsing and linguistic rules to identify instances.

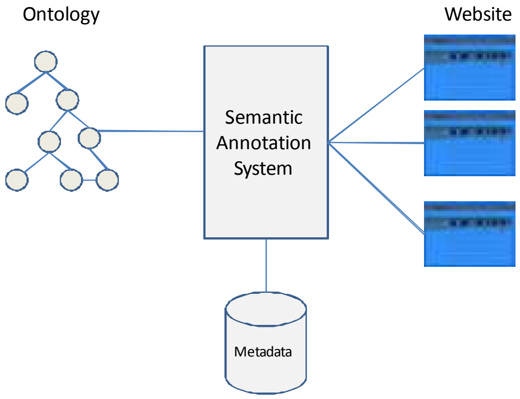


Figure1. Linking Web sites and Ontology by Semantic Annotation

Erdmann et al. [12] describe their approach to finite state technologies and support of lexical acquisition, and semantic tagging through them. The work coincides with our approach and uses the concepts that are stored on the Ontology. The source Websites have been manually annotated to explicitly represent the semantics of their contents. It introduces a proprietary extension to HTML that is compatible with common web browsers. Since there is a huge amount of relevant information for most communities, manual annotation is burdensome, and it is an impractical solution.

Blythe et al.'s [13] ACE system enables users to enter a free text into a parser. Then it compares the free text with the Ontology for term replacement. However, the ACE system cannot annotate the whole article. It only accepts user annotations as short statements of free text. The system is designed to be robust, allowing partial formalizations of the annotation and not relying on a successful parse of the user's input.

Steffen et al. [14] have been developing a tool, OntoAnnotate, that allows usage of domain-specific Ontologies for easy annotation of HTML documents and creation of meta data by (semi-automatic) annotating web pages. Starting from their Ontology-based annotation environment in OntoAnnotate, they have collected experiences in an actual evaluation study.

Liu et al. [15] propose a semi-automatic annotation system, which assists users in annotating textual web data and manages the terms defined by the user. The system uses OWL to describe semantic web data and annotate them. It happens in two ways: using a manual annotator with the help of the user, or an automatic annotator using the KMP algorithm. The work uses string matching to identify classes and neglect properties and relationships.

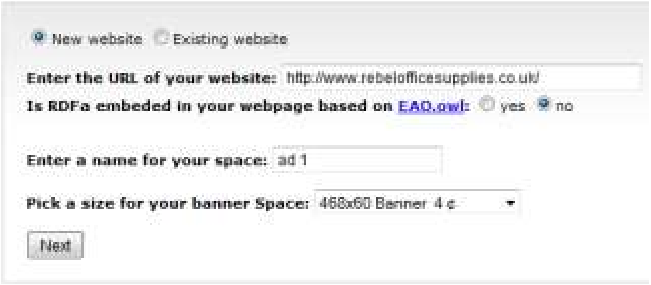
Most of the works on Ontology-based semantic annotation have been developed based on manual semantic annotation. With the huge amount of information on the Web and the upswing of the Semantic Web, there comes an urgent need for automating the process of adding semantic annotation to existing web pages.

IV. SEMANTIC ANNOTATION SYSTEM

This section illustrates how the semantic annotation system works. It gives an example, and then describes the suggested architecture of the system and the different components that compose it.

A. Example Scenario

The proposed system links Web site registration to an existing education advertising domain Ontology (EAO) [9] by semi-automatic semantic annotation. A Web site is described as a set of concepts followed by a set of properties. When the publisher registers his Web site, he has to enter the URL of the Web page that will host the



● New website ○ Existing website

Enter the URL of your website: http://www.rebelofficesupplies.co.uk/

Is RDFa embeded in your webpage based on EAO.owl: ○ yes ● no

Enter a name for your space: ad 1

Pick a size for your banner Space: 468x60 Banner 4 ¢

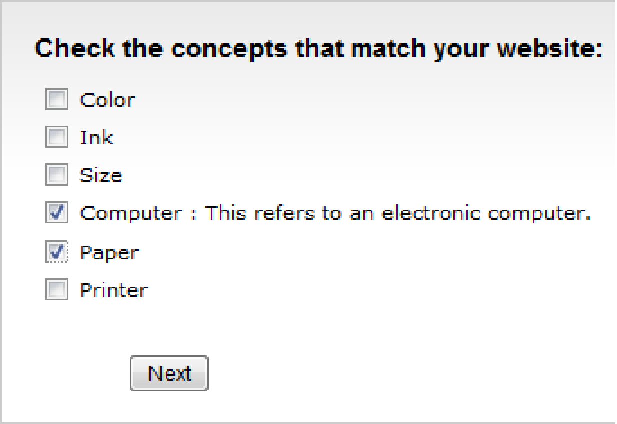
Figure 2. The registration page of the publisher

advertisements . For example, in our case, the Web site [17] is an educational Web site, and it was one of our samples. In addition, the publisher has the choice of entering a URL that has an RDFa embedded in his Web page (i.e., Semantic Web site), as shown in Figure 2.

Annotation starts with parsing the content of the registered Web site for extracting salient terms. This is automated by natural language processing of the Web site content.

Our system links extracted concepts to Ontology entries and suggests a number of terms that may describe the Web page, depending on the content of the publisher Web site and the Ontology. The publisher has to choose a number of terms that describe his Web page the best. In our example, the system will suggest the concepts "Color," "Ink," "Computer" " Paper," and "Printer" as shown in Figure 3. As long as we are suggesting a semi-automated annotation system, the system expects verification of the results suggested by the parser.

Then, the system gives a list of properties expressed in a natural language, which is suggested by the Ontology content. In our case it corresponds to the "Paper" and "Computer," namely: "is manufactured by," is dimensioned," "is colored," "is priced for," as shown in Figure 4. The publisher selects a property and then assigns to it a specific value. In our example, the price of the paper was set to "10$" and the computer manufactured by "Dell" and priced for "1500$". That represents the value of the selected property "is priced for" and the property" is manufactured by."



Check the concepts that match your website:

□ Color

□ Ink

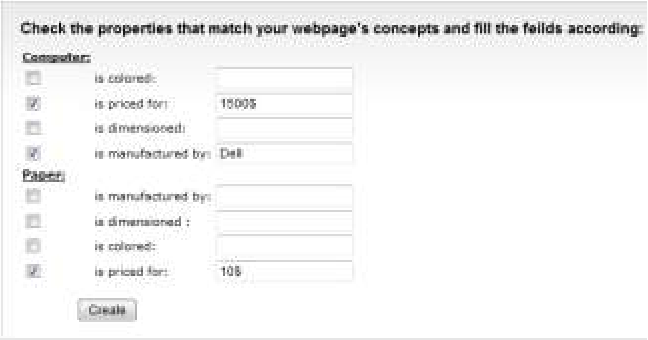
□ Size

☑ Computer : This refers to an electronic computer.

☑ Paper

□ Printer

Figure 3. The concepts that match publisher Web site the best



Check the properties that match your webpage's concepts and fill the feilds according:

Computer:

□ is colored:

☑ is priced for: 1500$

□ is dimensioned:

☑ is manufactured by: Dell

Paper:

□ is manufactured by:

□ is dimensioned :

□ is colored:

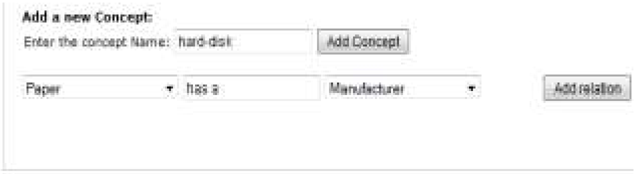
☑ is priced for: 10$

Figure 4. Properties of the selected concept

The system is self-learning. In cases in which some related instances or concepts are not defined in the Ontology, only the administrator may add a suitable instance or concept. The administrator may enter a new relation for an existing concept as shown in Figure 5 below.

Furthermore, if the concepts found on the publisher Web page cannot be matched with the Ontology concepts, the system suggests terms that have a similar meaning. For example, if it finds "Pen" or "Writing Instrument" in the Web page, then it will refer to the term "Pen." Figure 6 shows an XML file storing all the concepts that are in the Ontology and their synonyms. The XML file can be edited through administration access only. It is used to give each term a score that measures the relevancy to the concept that is stored in the Ontology. The relevancy score is a measuring function that is conducted by the system to help in matching the terms found in the Web pages to our Ontology concepts. The administrator can add and/or edit the concepts in the XML file and calculate the score of the relevancy.

ISBN: 978-1-61208-162-5



Add a new Concept:

Enter the concept Name: hard-disk [Add Concept]

Paper ▾ has a Manufacturer ▾ [Add relation]

Figure 5. Adding a new concept to the Ontology

```
<keyword name="Pen">
<term name="pen" score="100"/>
<term name="writing instrument" score="100"/>
</keyword>
```

Figure 6. Synonym of the term "Pen"

```
<keyword name="Notebook">
<term name="lined Notebook" score="90"/>
<term name="unlined Notebook" score="90"/>
<term name="notebook" score="100"/>
<term name="stationery" score="100"/>
</keyword>
```

Figure 7. Synonyms of the term "Notebook"

Also, if the system finds a type of concept, the system will suggest the concept to the publisher. For example, as shown in Figure 7, the system will suggest the term "Notebook" if it finds one of these terms: "lined Notebook", "unlined Notebook", "notebook" or "stationery".

B. Proposed Architecture

Figure 8 illustrates the architecture of our suggested system, which helps the publisher to add semantic description to his Web page in a semi-automated way. The publisher first adds his Web site metadata along with the URL through the Web site registration page. The Term Extractor is an automatic tool used to extract terms that best describe concepts in the Web page. We have used the Porter stemming algorithm (or 'Porter stemmer'). It is a process for removing the more common morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems [8].

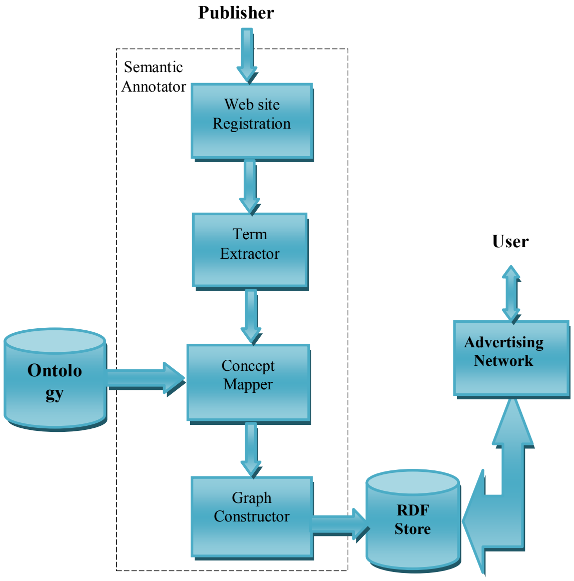


Figure 8. The proposed architecture for the semi-automated semantic annotation system

```
<rdf:RDF
 xmlns:eao=http://www.lexeme-ads.com/EAO.owl/
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >
<rdf:Description rdf:about="http://www.lexeme-
ads.com/#Computer">
<eao:hasManufacturer
xml:lang="en">Dell</eao:hasManufacturer>
<eao:hasPrice xml:lang="en">1500</eao:hasPrice>
<rdf:type rdf:resource="http://www.lexeme-
ads.com/EAO.owl/Computer"/> </rdf:Description>
<rdf:Description rdf:about="http://www.lexeme-
ads.com/#Paper">
<eao:hasPrice xml:lang="en">10$</eao:hasPrice>
<rdf:type rdf:resource="http://www.lexeme-
ads.com/EAO.owl/Paper"/>
</rdf:Description>
</rdf:RDF>
```

Figure 9. A fragment of the RDF Representation of the www.rebelofficesupplies.co.uk content

The Concept Mapper links extracted concepts to Ontology entries. The component automatically suggests related instances and saves verified annotations in the metadata store. Further, the system suggests relationships defined between instances by finding related instances from the Ontology. Finally, the publisher is provided with a list of properties associated with each instance and asks for supplying values. The set of concepts, relationships and properties describing the Web site semantic content is referred to as the RDF model.

The RDF model will be converted into an RDF graph through the Graph constructor. An RDF graph is set of RDF triples (subject, verb, and object). Triples are the basis of information representation. Figure 9 above shows part of the RDF code that has been generated for the Web site [17].

V. IMPLEMENTATION

To develop the semi-automated system, we used Jena's APIs [18] as our semantic framework. We also developed

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-162-5

an inference engine that links both advertisers to publishers and vice versa. In addition, we asked a few experts to gather some samples that they considered to be related to the educational domain. We then made sure that they were 100% strictly XHTML pages.

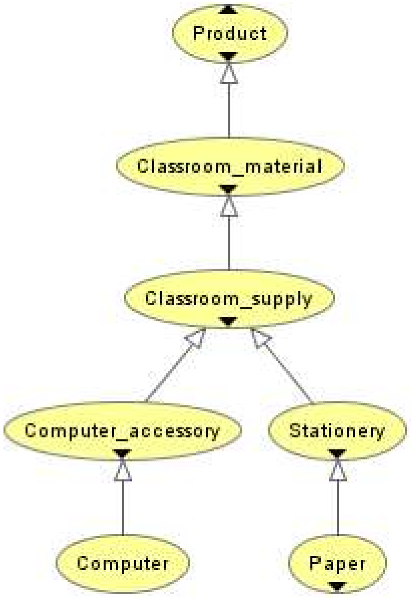


Figure 10. The relationship among the entities

Failure to find a proper Ontology for putting relevant ads on WebPages stimulated us to create our own product. By developing a custom-made Ontology, we were able to define countless education specialties, such as publications, school supplies, and writing instruments, that had not been present in any educational Ontology at the time. To build the Ontology, we used the Protégé editor [19]. Figure 10 shows a snapshot of the structure of the Ontology of the above scenario.

VI. EXPERIMENT RESULTS

The system provides indexed RDF matches and metadata matches. In this section, we test the system using two evaluation methods. First, we assess the RDF files that have been generated by the system for each website using the W3C validation service [20] to validate the RDF files. Experiments have been carried out using ten educational Web sites, the advertisement collection is approximately 50 ads, and all generated RDF files have been validated successfully through this tool.

Second, we evaluate the performance of the system using Precision and Recall. We do not intend to suggest a sophisticated semantic annotation system. Instead, we provide a simple demonstration of semantic annotation in advertising networks.

Semantic match in the system retrieves publishing web sites relevant to a submitted ad of interest. A simple algorithm is outlined below.

1. Get and conceptualize the ad.
2. Find relevant publisher's web sites by matching the ad RDF against the RDF repository of publisher's web sites.
3. Retrieve the metadata of the relevant Web sites.
4. Place the ad in the ad place defined for ad displays in the publisher's Web site.
5. Capture the number of clicks on the ad and add to web site metadata.
6. Direct the visitor to the corresponding ad home page.

Jena's matching method is used to match the semantic content of both the ad and the publishing web sites.

The following experiment aims at testing our generated RDFs for the publishers' Web sites against the stored RDFs of the ads. The matching process for both approaches is the same.

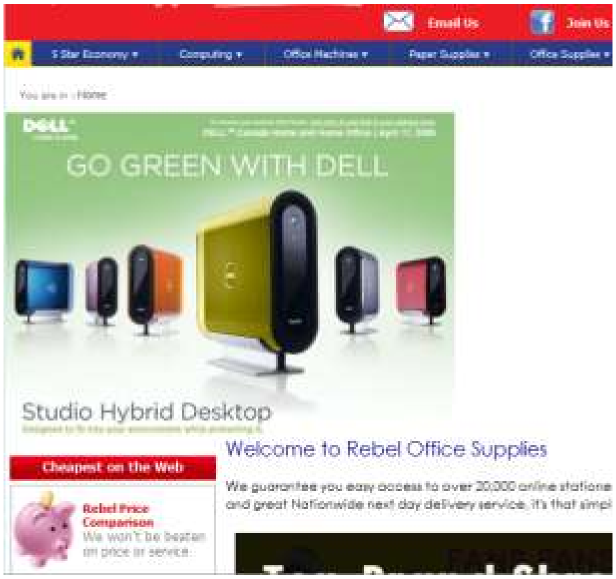
The system will match the RDF that has just been created with the stored RDFs in the database and try to find a match. The matching process will depend on the number of triples in the RDF graph that matches with the Web site graph. If it finds a matching triple or a partial matching, it will put the matching advertisement in the Web site space. Figure 11 shows part of the code that is used for matching. The system uses the getSubject() and getPredicate() methods that are impeded in Jena to match the two graphs depending on their subject and predicate.

```
public static int Match(Graph g1, Graph g2)
{
   Triple T1,T2;
   int count;
   ExtendedIterator iter1, iter2;
   count = 0;

    /* if the two graphs are isomorphic; stop the loop and the ad will be
placed in this website. the count will be = -1 */

    if(g1.isIsomorphicWith(g2))
     return -1;
     iter1 = g1.find(Triple.ANY); //returns an Iterator for all the triples in
the graph
  while(iter1.hasNext())
    {
     iter2 = g2.find(Triple.ANY);
      T1 = (Triple) iter1.next(); // T is a triple from the ad graph
      while(iter2.hasNext())
      {
         T2 = (Triple) iter2.next();
         if (T1.getSubject().equals(T2.getSubject()) &&
T1.getPredicate().equals(T2.getPredicate()))
          {
             count++;
          } } }
return count;
}
```

Figure 11. Part of the matching code

```
<rdf:RDF
 xmlns:eao=http://www.lexeme-ads.com/EAO.owl/
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" >

  <rdf:Description rdf:about=http://lexeme-ads.com/#Computer>
<eao:hasDimension xml:lang="en"> 11 inches</eao:hasDimension>
<rdf:type rdf:resource="http://lexeme-ads.com/EAO.owl/Computer"/>
  </rdf:Description>
</rdf:RDF>
```

Figure 12. A fragment of the RDF Representation of the ad



We guarantee you easy access to over 20,000 online statione and great Nationwide next day delivery service. It's that simpl

Figure 13. Displaying the result of the match

The RDF file that is shown for the matching ad is shown in Figure 12.

Among the set of advertisements returned by the system to be displayed on the publisher Web site, we selected the result shown in Figure 13, from an advertisement for the Dell company, which was placed in the matched website.

We have used the Precision and the Recall to compute the matching Web sites with relevant ad rates. The following table shows the Precision and Recall rates for each website. Also, it shows the F-score, which is defined as the harmonic mean of Precision and Recall to reflect the actual performance of the system.

Precision is a measure of correctness. As Table 1 shows, the Precision rates are mostly 100% because the system only retrieves the ads that have matching RDF graphs with the website graphs. If the graph does not match, the system will discard the advertisement as a choice and test another advertisement graph.

TABLE 1. PRECISION, RECALL AND F-MEASURE FOR A SEMANTIC ADVERTISING NETWORK

website No.	Precision(%)	Recall(%)	F-measure(%)
1	75	30.8	44
2	100	100	100
3	100	100	100
4	100	75	85.7
5	100	75	85.7
6	100	55	70.9
7	100	88	93.6
8	100	55	70.9
9	100	33.3	49.6
10	100	37.5	54.5

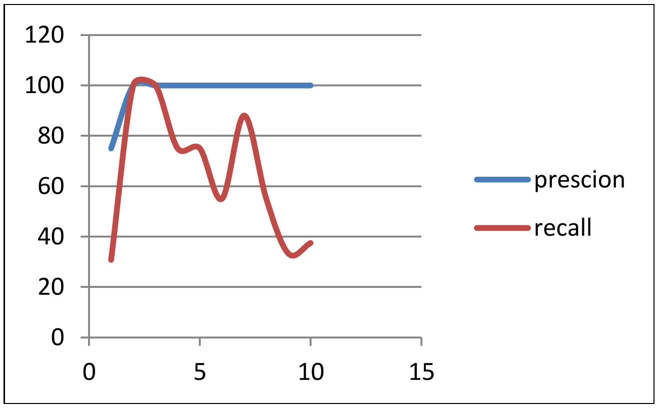


Figure 14. the relation between the Precision and the Recall

On the other hand, Recall is a measure of completeness. It shows the rate of retrieving all the relevant ads, as shown in the table below, where it ranges from 30% to 100%. The reason for that is that when the system chooses one ad and tries to find another ad that matches, it will encounter a relevant ad; however, sometimes that relevant ad will have fewer matching triples then the original ad. The system will not consider it as retrieved.

As a result of a greater Precision rate, the Recall rate is decreased, as shown in Figure 14, and as the Precision rate becomes higher, the number of relevant ads that are retrieved is lower.

If we combine the two metrics, Precision and Recall, we get their harmonic mean, known as the F-measure. This is a measure of a test's accuracy. As shown in Table 1, the F-measure ranges from 44% to 100%. The F-measure average is 75.5%. This average is considered high and it points to the high accuracy of the system. Considering the experiment results, we believe semantic advertising networks have considerable potential.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a prototype of a semi-automated semantic annotation system for semantic advertising networks. This helps the publisher in describing his Web page using semantic technology and Ontology. We have demonstrated and proven how current advertising methods can be improved. The introduction of Semantic Technology is essential for reaping larger financial gains. It is important for analyzing the meanings between the lines in order to place ads in prime web-locations for the sole purpose of reaching their targeted consumers. With this system, there will no longer be the Web 2.0 emphasis on factors such as keyword matching. The emergence of Web 3.0 means tapping into more important elements such as understanding the context of Web pages, and latent or hidden connections amid these contexts. It serves to discover relationships among concepts and ideas.

In the future, a significant amount of work should be done. We will try to provide the publisher with the privilege of adding/editing concepts to the Ontology by an administration approval that matches his preference. We will support the RDF standard for representing metadata on the web, representing both Ontologies and generated annotated facts in RDFs. This standard will make annotated facts reusable and machine processable on the web [10].

ACKNOWLEDGMENT

Our thanks to Arwa Al-Tameem, Ghada Abuguyan, May Abu Melah, Nora Al-Zaid, Nouf Al-Najran, and Nadeen Al-Abdullatif for developing the prototype.

We are thankful to Dr. Muhammad Abulaish working at the Center of Excellence in Information Assurance, King Saud University for his review and guidance.

REFERENCES

[1] W. D. Wells, S. Moriarty, and J. Burnett, "Advertising: principles and practice", 7th ed. New York: Prentice Hall, 2005.

[2] M. Strckland, "Guide to Semantic Web: Creating more relvent ads",[Online]. Available: https//www.threeminds.organic.com/ [Accessed: Sept. 9, 2011].

[3] T. Wilson, "How semantic web works", [Online]. Available: http//computer.howstuffworks.com. [Accessed Sept. 9, 2011].

[4] D. Allemang, and J. Hendler, "Semantic web for the working ontologist: effective modeling in RDFS and OWL". Burlington, Massachusetts: Morgan Kaufmann, 2008, pp. 1-40.

[5] L. Al-Safadi, A. Al-Dawood, and N. Abdulateef, "Lexeme: An Ontology-based semantic advertising networks". Journal of Computing, 2(9), 2010, pp.1-5.

[6] M. Marshall, "The semantic web & its implications on search marketing",[Online].Available:http//www.searchenginejournal .com/. [Accessed Sept. 9, 2011].

[7] J. Davies and R. Struder, "Semantic web technology: trends and research in ontology system-based systems". Chichester, Wiley, 2006, pp.29-50.

[8] "The Porter Stemming Algorithm",[Online]. Available: http://tartarus.org/~martin/PorterStemmer [Accessed Sept. 9, 2011].

[9] L. Al-Safadi, N. Abdulateef, "Educational advertising ontology: a domain-dependent ontology for semantic advertising networks". Journal of Computer Sciences 6(9), 2010, pp.1-8.

[10] R. Benjamins, D. Fensel, and S. Decker. 1999. "KA2: building ontologies for the internet: a midterm report". International Journal of Human Computer Studies, 51(3):687.

[11] Lassila, O. and Swick, R. (1999). Resource description framework (RDF) model and syntax specification. Technical report, W3C. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax.

[12] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab, "From manual to semi-automatic semantic annotation: about ontology-based text annotation tools". In P. Buitelaar & K. Hasida (eds). Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, pp.3-7.

[13] Blythe J. and Gil Y. "Incremental formalization of document annotations through ontology-based paraphrasing". In Proceedings of the 13th International World Wide Web Conference (New York, New York, May 2004), pp. 455-461.

[14] Steffen Staab, Alexander Maedche and Siegfried Handschuh. "An annotation framework for the semantic web". In P. Buitelaar & K. Hasida (eds). in proceedings of the first workshop on multimedia annotation, pp. 1-4.

[15] C.-H. Liu, H.-C. Chen, J.-L. Jain, and J.-Y. Chen. "Semi-automatic annotation system for owl-based semantic search". International Conference on Complex, Intelligent and Software Intensive Systems (2009 IEEE), pp.1-5.

[16] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, “Semantic annotation for knowledge management: requirements and a survey of the state of the art,” Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol.4(1), 2006, pp. 14–28.

[17] "Rebel Office Supplies", [Online].Available: www.rebelofficesupplies.co.uk. [Accessed Sept. 18, 2011].

[18] "Jena – A Semantic Web Framework for Java", [Online].Available: jena.sourceforge.net. [Accessed Sept. 18, 2011].

[19] "Protégé", [Online]. Available: protege.stanford.edu. [Accessed Sept. 18, 2011].

[20] "W3C Validation Service", [Online]. Available: www.w3.org/RDF/Validator/. [Accessed Sept. 18, 2011].