Annotation of web pages using semantic tagging and ranking model to effective information retrieval

Dr. Poonam Yadav
D.A.V College of Engineering. & Technology, India.
poonam.y2002@gmail.com

Abstract—Due to continuous growth of World Wide Web (WWW), huge number of web documents is published in web servers providing plentiful information. Because of the structure of web pages and huge amount of information, retrieving of relevant information among the web sites leads very challenging task. In order to extract more useful contents automatically from the web database, this paper presents a model for annotation of web pages using semantic tagging and ranking model to effective information retrieval. To accomplish this task, relevant words are extracted after performing tag removal process, stop word removal and stemming. Then, Semantic tagging is performed to tag all the important keywords to its attribute value through a set of constrain rule list. Finally, ranking model is developed to find the most suitable elements of the attributes to map into annotated database. For experimental evaluation, user query is submitted to google engine and the relevant web results are extracted. The extracted results are then given to the proposed annotation model to construct the annotated database and evaluation is done using precision and recall.

Keywords- World Wide Web, Annotation, web database, information retrieval, ranking model, semantic tagging.

I. INTRODUCTION

With the ever seen growth of World Wide Web (WWW) containing large number of web servers providing plentiful information, retrieving relevant information among the web sites can be a challenging task [2-5]. Because of the structure of the web pages which may be in variety of styles and format, it is very hard to extract the more useful contents automatically from the web pages. This lead to new research area, called annotation of web pages prescribed in the standard format to machine understandable [6-10]. Annotation of web pages refers to adding metadata to Web pages to give higher level machine understanding of the web pages and other web objects in the ways of including it as comments, notes, explanations, references, examples, advices, correction or any other type of external remarks that can be emotionally involved to the complete or element of web pages exclusive of modifications to the original pages. In olden days, semantic annotation was done through manually. Manual annotation is also a costly process as it would require direct intervention of human for annotation of web data and also, errors due to factors such as, annotator knowledge with the domain, amount of training, personal motivation and complex schemas. So, automatic annotation through different semantic and machine learning techniques is required for semantic web for existing and new documents on the Web.

In this paper, Annotation of web pages is presented using semantic tagging and ranking model to effective information retrieval. At first, input web pages are given to feature word identification phase where, the most import words are identified after removing stop words and stemming process. Then, these words are recognized using semantic tagging process. In semantic tagging process, semantic words are extracted and the matching of rule-based criteria is done to do tagging process. Once the tagging process is done, ranking measure is developed here to provide the rank measure for all the tag elements. Based on the ranking measure, tag elements are chosen and it is filled within the annotated database. The basic organization of the paper is given as follows. Section 2 discusses the existing system and section 3 presents the proposed system for web page annotation using semantic tagging and ranking model. Section 4 discusses the experimentation with results and finally, conclusion is given in section 5.

II. EXISTING SYSTEM

Data annotation problem was recently proposed and it was solved using different approaches. One of the approach recently developed is given in [1]. Here, automatic annotation approach was proposed for annotating the search result records retrieved from any given web database. In that approach, probabilistic method is combined with features for annotation. Also, a clustering-based shifting technique was utilized to align data units into different groups so that the data units inside the same group have the same semantic.

Disadvantages: The approach developed in [1] face critical challenge of achieving holistic and accurate annotation. Also, clustering-based method was utilized to automatic feature extraction. But, clustering of HTML

contents requires computation intensive algorithms for grouping the relevant contents. Also, semantic way of information handling is missed in that paper completely. So, to further improve the performance and accurate annotation, this paper brings the semantic way of information handling for annotation of information published in the web database.

III PROPOSED METHOD: ANNOTATION OF WEB PAGES USING SEMANTIC TAGGING AND RANKING MODEL TO EFFECTIVE INFORMATION RETRIEVAL

This section presents the proposed method of web page annotation model using semantic tagging and ranking model. The proposed method is developed towards achieving the objective of annotated database for the web pages related to book information. Suppose, user want to retrieve some information about the book titled as, "introduction of image processing". Lot of search results are obtained through any recent search engine. But, user's aim is to obtain only the author name, publisher name, price of the book and year of publication. That information is only enough for the user. With the perspective of obtaining that information, this work presents annotation model for the web pages. The block diagram of the proposed annotation model is given in figure 1.

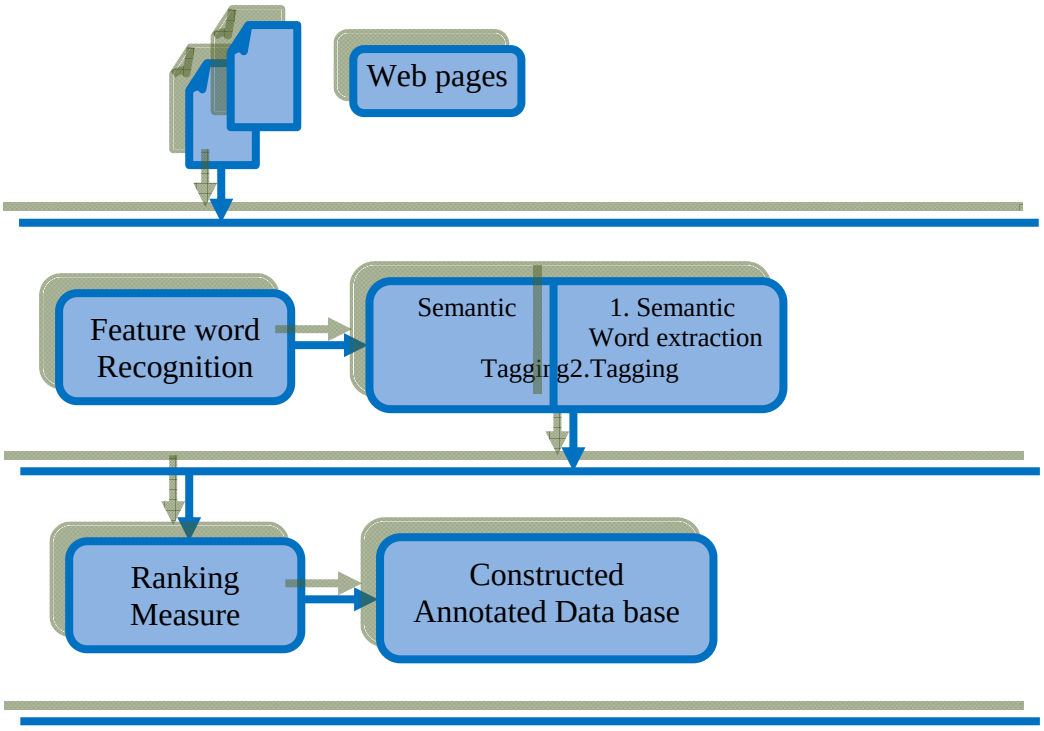

Figure 1. Block diagram of the proposed annotation model for web data extraction

A Feature word recognition

At first, user query is submitted to the search engine which provides m web pages as an output for the user query, Q. The aim is to identify only the most important keywords presented in every pages. To accomplish this task, HTML contents are extracted through tag removal process and then, stop words are removed to obtain only relevant words. The relevant words are then processed to obtain root words which are significantly important for the next process. The above given process provides a set of words which are analyzed to identify only the feature words. This process is known as feature word recognition. In feature word recognition phase, three steps are performed such as, Top frequent word identification, Numerical word identification, Non-dictionary word extraction. To identify the top N frequent words, the frequency of every word in the web page is computed by scanning the whole database. Then, words which are having high frequency is extracted from the top N list. Subsequently, numerical words presented in the web pages are also extracted. Non-dictionary words are also important which may be author name or publisher name so words are matched with dictionary to find whether it is a non-dictionary words. These three type of words extracted are then given for the next process, called semantic tagging.

B Semantic tagging of feature words

Semantic tagging is an important process to construct the annotated database, A_D which have author name, publisher name, price of the book and year of publication. This step aims to find four attributes from the web pages. The author and publisher name might be a non-dictionary word and also, both names should be

given in the web page before or after putting as author name and publisher name. Based on these two constraints, author name and publisher name is identified. The price of the book should be numerical word in most of the time. So, it can be recognized easily if, i) it is a numerical term, ii) neighbour words like price, cost and fee. The semantic word for 'price' is derived from wordnet ontology, and iii) neighbour words have currency notation. Year of publication can be recognized easily if the numerical term has the year starting from 1800 to 2015. But, the problem here is that if the price of the book is Rs. 2000, then there may be chance of confusion. But, this can be easily avoidable by matching the neighbour words with the above three constraints.

Based on the above process, the required four attributes such as, author name, publisher name, price of the book and year of publication can be multiple values. So, this step is to extract only one attribute value to put into the annotated database. This can be easily achievable through the ranking measure developed in this work. The ranking measure for every attributes is computed as follows,

$$R_{AN} = \sum_{i=1}^{n} AN_i$$

$$AN_i = \begin{cases} 1 & ;i \text{ if } rule\ i\ is\ matched \\ 0 & ; \text{ if } rule\ i\ is\ not\ matched \end{cases}$$

$$R_{PN} = \sum_{i=1}^{n} PN_i$$

$$PN_i = \begin{cases} 1 & ;i \text{ if } rule\ i\ is\ matched \\ 0 & ; \text{ if } rule\ i\ is\ not\ matched \end{cases}$$

$$R_P = \sum_{i=1}^{n} P_i$$

$$P_i = \begin{cases} 1 & ;i \text{ if } rule\ i\ is\ matched \\ 0 & ; \text{ if } rule\ i\ is\ not\ matched \end{cases}$$

$$R_Y = \sum_{i=1}^{n} P_i$$

$$Y_i = \begin{cases} 1 & ;i \text{ if } rule\ i\ is\ matched \\ 0 & ; \text{ if } rule\ i\ is\ not\ matched \end{cases}$$

Once the ranking measure for author name R_{AN} , publisher name R_{PN} , price of the book R_P , year R_Y is computed, an attribute value have the highest value is chosen as final element for the attributes.

C Constructing of annotated database

The attribute values identified for every attributes like, author name, publisher name, price of the book and year is stored in the annotated database, A_D. The annotated database for the user those who want to obtain book related information is as follows.

URL of the web	author name	publisher name	price of the book	year

Table 1: Annotated database for getting book relevant information

The annotated database is stored in the web. So, when the user want to retrieve any information related with book, they can easily identify from this database.

IV RESULTS AND DISCUSSION

This section present the experimental outcome of the proposed annotation model developed using semantic tagging and ranking measure.

A Evaluation based on precision

For experimentation, user query "introduction to image processing" is submitted to google engine and the relevant web results are extracted. The extracted results are then given to the proposed system to build up the annotated database. The annotated database is evaluated using precision formulae,

$$\Pr ecision = \frac{\text{Re}levant \cap \text{Re}trieved}{\text{Re}trieved}$$

The evaluated outcome using Precision is given in figure 1. From the figure, the average performance of the proposed model is 65.8% and the existing system is 64.6%.

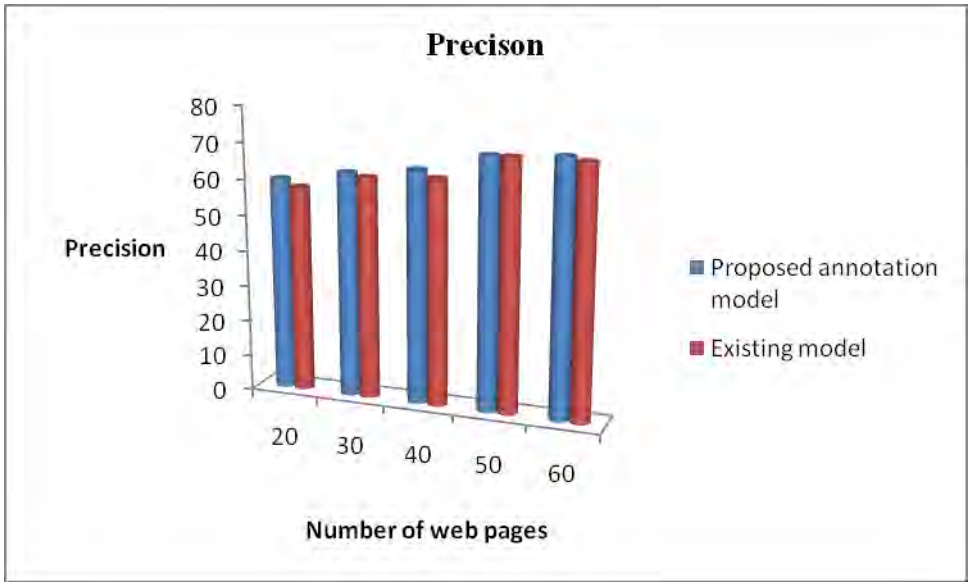

Figure 2. Precision plot for evaluation

B Evaluation based on recall

For experimental evaluation, user query "introduction to image processing" is submitted to google engine and the relevant web results are extracted. The extracted results are then given to the proposed annotation model to construct the annotated database. The annotated database is evaluated using recall formulae,

$$\text{Re}call = \frac{\text{Re}levant \cap \text{Re}trieved}{\text{Re}levant}$$

The evaluated outcome using Precision is given in figure 1. From the figure, the average performance of the proposed model is 69.2% and the existing system is 64.6%

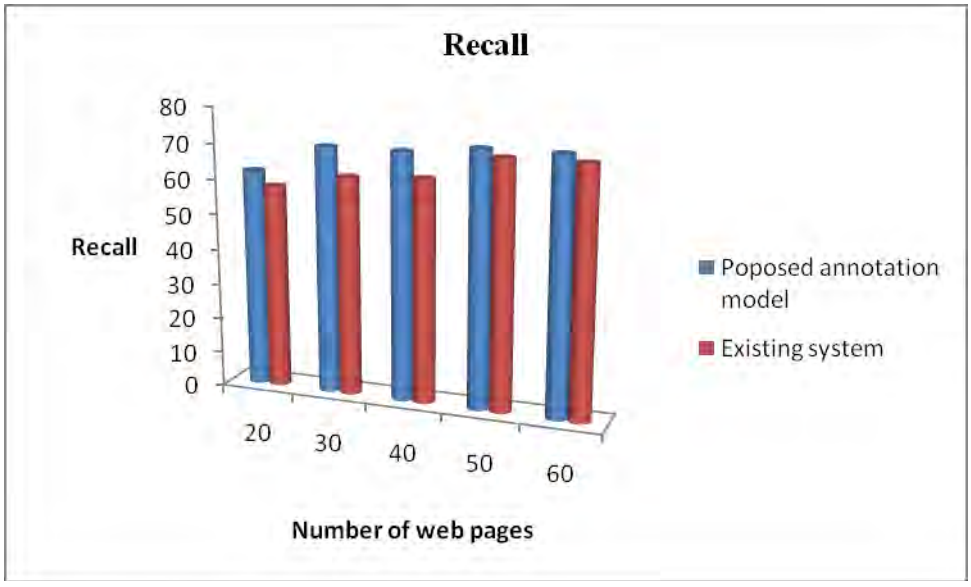

Figure 3. Recall plot for evaluation

V CONCLUSION

This paper presented a model for Annotation of web pages using semantic tagging and ranking model. Semantic tagging is performed by matching a set of constraint rules with text data and ranking model is developed to choose the most relevant attribute value for the attributes. The web pages are directly given to feature word recognition phase and the important word recognized are then given to semantic tagging which contains semantic word extraction and tagging. Finally, ranking of attribute elements is performed using ranking model and the identified elements are stored in the annotated database. So, when the user wants to retrieve any information, they can easily identify from this database. For experimental evaluation, user query "introduction

to image processing" is submitted to google engine and the relevant web results are extracted to construct the annotated database. The resultant outcome is evaluated using Precision and recall. The average performance of the precision for the proposed model is 65.8% and the existing system is 64.6%. Also, the average performance of the recall for proposed model is 69.2% and the existing system is 64.6%

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, and Clement Yu,"Annotating Search Results from Web Databases", IEEE transactions on knowledge and data engineering, Vol. 25, No. 3, pp. 514-527, March 2013.

[2] Junnila, V., Laihonen, T., "Codes for Information Retrieval With Small Uncertainty", IEEE Transactions on Information Theory, vol. 60, no. 2, pp. 976-985, 2014.

[3] Böhm, T, Klas, C.-P. ; Hemmje, M., "ezDL: Collaborative Information Seeking and Retrieval in a Heterogeneous Environment", computer, IEEE, vol. 47, no. 3, pp. 32-37, 2014.

[4] Sumiya, K., Kitayama, D. ; Chandrasiri, N.P., "Inferred Information Retrieval with User Operations on Digital Maps", IEEE Internet Computing, vol. 18, no. 4, pp. 70-73, 2014.

[5] Xiaogang Han, Wei Wei ; Chunyan Miao ; Jian-Ping Mei ; Hengjie Song, "Context-Aware Personal Information Retrieval From Multiple Social Networks", Computational Intelligence Magazine, IEEE, vol. 9, no. 2, 2014.

[6] Ciccarese, P., Soiland-Reyes, S. ; Clark, T., "Web Annotation as a First-Class Object", nternet Computing, IEEE, vol. 17, no. 6, pp. 71-75, 2013.

[7] Belhajjame, K.Embury, S.M. ; Paton, N.W., "Verification of Semantic Web Service Annotations Using Ontology-Based Partitioning", EEE Transactions onServices Computing, Vol. 7, no. 3, pp. 515-528, 2013.

[8] Zhigang Ma ;Feiping Nie ; Yi Yang ; Uijlings, J.R.R. ; Sebe, N., "Web Image Annotation Via Subspace-Sparsity Collaborated Feature Selection", IEEE Transactions on Multimedia, Vol. 14, no. 4, pp. 1021-1030, 2012.

[9] Stamou, G. van Ossenbruggen, J. ; Pan, J.Z. ; Schreiber, G. , "Multimedia annotations on the semantic Web", IEEE mulimedia, vol. 13, no. 1, pp. 86-90. 2006.

[10] Wilks, Y. "The Semantic Web: Apotheosis of Annotation, but What Are Its Semantics?", Intelligent Systems, IEEE , vol. 23, no. 3, pp. 41-49, 2008.

BIOGRAPHY

Dr. PoonamYadav obtained B.Tech in Computer Science &Engg. fromKurukshetra University Kurukshetra and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She had Awarded Ph.D inComputer Science& Engg. fromNIMS University, Jaipur. She is currently working as Principal in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Her research interests include Information Retrieval, Web based retrieval and Semantic Web etc. Dr. PoonamYadav is a life time member of Indian Society for Technical Education and her email id is poonam.ir@gmail.com.