Dynamic Classification for Web Documents Using Semantic Knowledge (DBpedia)

Passent El-kafrawy[*1], Dina El-Demrdash[*2]

*Faculty of Science, Menoufia University, Egypt
[1]basant.elkafrawi@science.menofia.edu.eg
[2]dina.eldemrdash@science.menofia.edu.eg

Abstract—*we present a dynamic Web document Classification using semantic knowledge (DBpedia). We present a method for a dynamic Web document Classification and automatic classification. The proposed approach required only a domain ontology and a set of user predefined categories. Currently, most approaches to text classification represent document as (bag of words) and training the large set of documents to train the classifier. Our approach doesn't require a training set of documents. In our proposed method, we use DBpedia ontology as the main classifier, representing documents as (bag of concepts). We extract the terms from the document, extract their resources from DBpedia Spotlight, use Sparqle query to determine class ontology and map them to their concepts then we determine the best category.*

Key words*: Classification, ontology, DBpedia, DBpedia Spotlight.*

1 INTRODUCTION

The amount of information in the World Wide Web has been overloaded heavily and rapidly in current era. At the same time organizing and managing information for knowledge, extraction is a crucial problem. Web content consists of text documents and multimedia documents. Web Document Classification process plays an important role in organizing and managing data in the World Wide Web for better knowledge understanding.

Web document classification is the process of classifying documents into predefined categories. Classification is one of web mining techniques, to solve the problem of information overload. Moreover, web mining has been improved by utilizing semantic techniques. Semantic techniques provide deeper understanding of information by machine. Traditional web document classification methods represent documents as (bag of words). One of the advantages of Traditional Supervised Text Classification techniques is using machine learning in order to perform the desired tasks. These methods develop the classifier from a group of prepared documents, pre-classified into various specific classifications. Such strategies, including Support Vector Machines [1], Naïve Bayes [1], decision trees [1], and Latent Semantic Analysis [2] are viable, however they require a set of pre-classified documents to build the classifier. There are major problems when using context analysis of the words in the documents i.e. (bag of words, BOW)) First major drawback is to count word occurrences and not consider meaning. Second, in order to train classifier, large number of documents must be collected. Third, the meaning of web content is not machine accessible due to the lack of semantics. Fourth, it's simply difficult to distinguish the meaning between two sentences.

The DBpedia Ontology composes the knowledge on Wikipedia in 320 classes which shape a substitution hierarchy and is depicted by 1,650 various properties. It includes abstracts and labels of up to 3.64 million, up to about 97 various languages. Of these, 1.83 million are clearly classified in harmonious ontology including 416,000 people and about 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases. Also, there are 6,300,000 links to external web pages, 2,724,000 links to images, 740,000 Wikipedia classes and 690,000 geographic directions for places [3].There are more open source ontologies such as WordNet, open Cyc and SUMO. WordNet is a lexical ontology, is a large lexical database of English. Nouns, verbs, adjectives and adverbs grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The result is a network of meaningfully related words and concepts [4]. Open Cyc is open source version of the Cyc knowledge base. The Cyc knowledge base of general common-sense rules and assertions involving those ontological terms was largely created by hand axiom-writing; it grew to about 1 million in 1994, and as of 2017 is about 24.5 million and has taken well over 1,000 person-years of effort to construct[5]. SUMO (suggest upper merged ontology) is one of the largest and most comprehensive, free, formal ontologies SUMO is a large formal ontology (20.000 terms, 7.000 axioms). Axiomatization of general and domain specific concepts for applications requires basic "Common sense". Its disadvantages are work is done manual, slow process and low coverage, semantic wiki not accepted yet, and information extraction from corpora result of low accuracy[6].

In this paper, we proposed an ontology as a classifier. The advantage of the proposed method is that it does not require huge amount of training documents. Moreover, the inference mechanism is inherent in the model (i.e. the ontology), in other words there is no processing time consumed to reach a decision. In addition, it is not coupled with a

specific domain, as using general encyclopedic knowledge–based ontology such as DBpedia ontology. The paper is divided into introduction as explained in this section. Section 2 reveals the related work. Section 3 clarifies the background of semantic web document classification, while, the classification algorithm is presented in section 4. The results of the experimentation are listed in section 5, whereas, the conclusion and future work in section 6.

2 RELATED WORK

Bin Shi, Liying Fang, Jianzhuo Yan, Pu Wang, and ChenDong(2009) [7] proposed a uniform representation for the content, which includes concepts and relations, of semantic documents based on WordNet. They used WordNet (ontology) to map relations between concepts; then applied SVM to semantic classification of web document. This method only considered two semantic relations in WordNet. The method is to get only the path between two concepts in the WordNet.

Bai Rujiang Shandong and Liao Junhua (2009) [8] proposed a framework which utilizes ontologies and Natural Language Processing (NLP) strategies in order to index texts. Bag of Concepts is an alternative Traditional BOW matrix. Automated classification of documents requires another ontology-based philosophy. For enhancing text classification, they improved documents with related ideas, and performed unequivocal disambiguation in order to decide the best possible importance of each polysemous concept communicated in documents. They used three ontology WordNet, open Cyc and SUMO to find concept of key words and compared concept in different ontology. They did not consider the meaning between all concepts in a document and relations between them.

Jun Fang, Lei Guo and YueNiu (2010) [9] proposed a novel ontology-based documents classification method by using ontology reasoning and similarity measure. They solved drawbacks of current ontology-based documents classification methods by training a classifier through dividing concepts to high concepts and low concepts. They implemented ontology reasoning and similarity measure but didn't consider relation between concepts.

Shikha Agarwal, Arachana Singhal and Punam Pedi (2012) [10] used weighted concept frequency–inverse document frequency (cf-idf) with background knowledge of domain ontology, for classification of RSS news feed items. They have shown that a rich and comprehensive ontology can be successfully used as text classifier. They considered the important concept only by (cf-idf), whereas, they used an ontology as a classifier, they did not benefit from it. News category are given by end user not by system according to keywords, resources and concepts in ontology.

Chaaminda Manjula Wijewickrema (2014) [11] improved the classification accuracy of an automatic text classification system by using ontology. He proposed a solution to reduce the number of misclassification due to vocabulary ambiguities of the language used. He utilized ontology to represent the relationship among the concepts. But he chose the first four highest frequency terms to decide the subject of the test document only. Although, the ontology increased the accuracy of automatic classification, the final decision still has to be made manually.

Henrihs Gorskis1 and ArkadyBorisov (2015) [12] inspected the possibility of utilizing principles and concepts found amid the characterization tree building process in the C4.5 calculations, in a totally mechanized manner, for the reasons for building a metaphysics from information. By building the philosophy straightforwardly from nonstop information, ideas and relations can be found without particular learning about the area. The primary oddity of this approach is the making of concepts that reflect extraordinary and essential information esteem interims or ranges for each quality. Ontology building schemes have a few disadvantages. The quantity of interims found by the C4.5 calculations can be extensive and not naturally justifiable to a human client. The sensibility of the discovered esteem interims must be estimated through a domain expert. Possibly the many-sided quality of the traverse chain of command is just an observed one, finally, perhaps to an expert the value traverses bode well and he will have the capacity to give them fitting names.

The DBpedia Ontology is a general encyclopedic knowledge–based ontology. The DBpedia data set uses a large multi-domain ontology which has been derived from Wikipedia. Hub on the web of data DBpedia ontology works automatically, speed process and high coverage, with semantic wiki accepted enough of high accuracy. DBpedia provides three different classification schemata for things.

1. Wikipedia Categories are represented using the SKOS vocabulary and DCMI terms.
2. The YAGO Classification is derived from the Wikipedia category system using Word Net (Yago: A Core of Semantic Knowledge – Unifying WordNet and Wikipedia (PDF)).
3. Word Net Synset Links were generated by manually relating Wikipedia info box templates and Word Net Synsets, and adding a corresponding link to each thing that uses a specific template. In theory, this classification should be more precise then the Wikipedia category system.

Using these classifications within SPARQL queries allows a user to select things of a certain type. Thus, the proposed approach in this paper concentrates on using semantic technology to improve web mining techniques (classification) using DBpedia ontology as a classifier. This method is distinguished because it does not involve a huge training set of documents. It has the ability to dynamically define classification categories; it considers all concepts not high only. It determines categories automatically not manual.

3 SEMANTIC CLASSIFICATION FOR WEB DOCUMENTS

This paper aims to present a semantic classification system that classifies web documents by a semantic approach based on the DBpedia Ontology. Hence, this section describes the necessary steps to build such system.

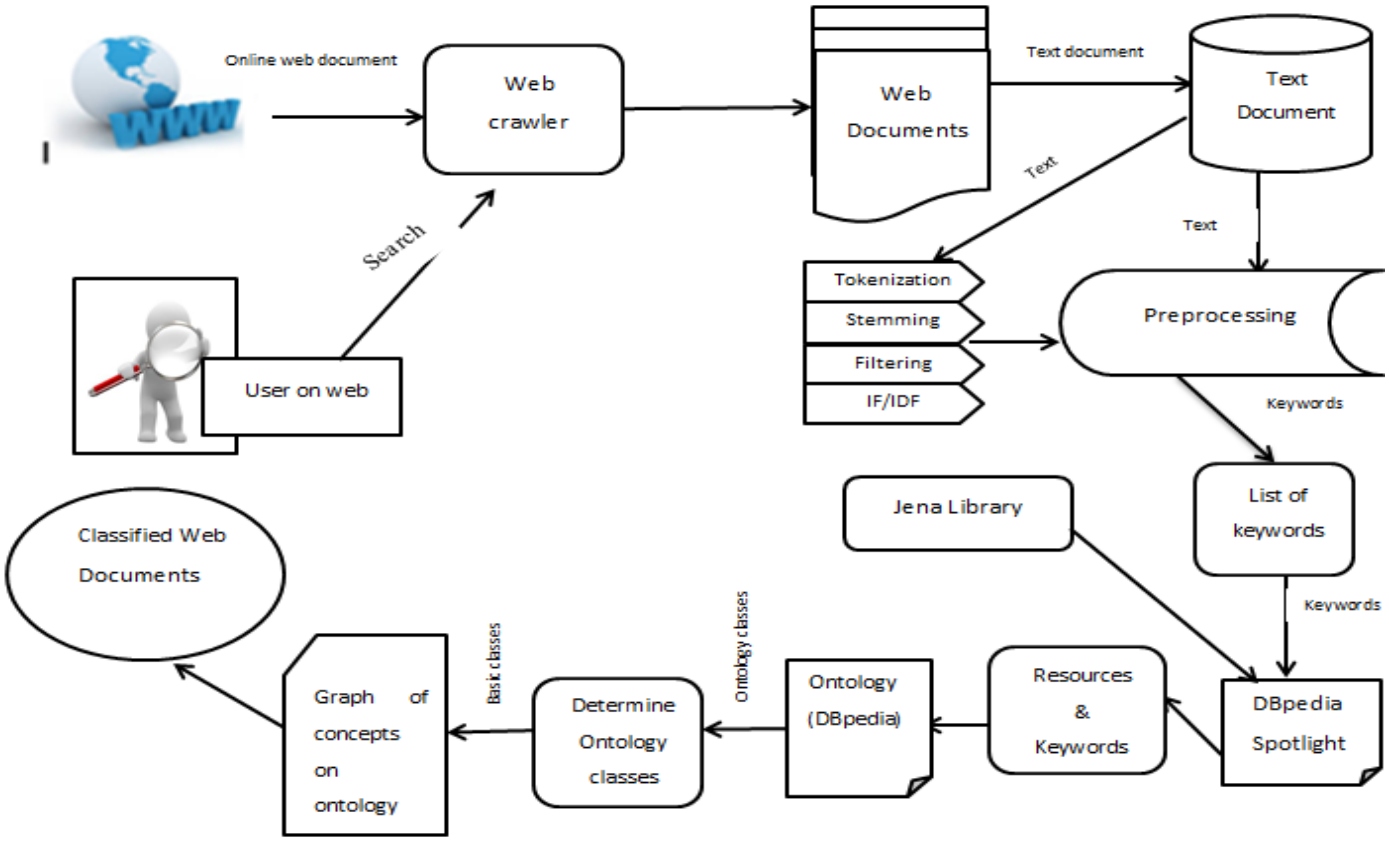

Figure 1: Classification Pipeline

A. First Step: Text document Preprocessing

Preprocessing phase plays a very important role in text mining techniques and applications. It is the first step in the text mining process. Preprocessing steps such as Tokenization, Stop-words removal, Stemming and TF/IDF algorithms shall be applied on the documents [13](figure 1)

- Tokenization: is the process of breaking a text into words, phrases, symbols.
- Stop-words removal: removes the unimportant words from document content by using a list of stop words. Stop-words are words that define non-linguistic view; they do not carry information such as (a, an, the, this, that, I, you, she, he, again, almost, before, after).
- Stemming: removes the affixes in the words and produces the root word known as the stem

- TF/IDF: Term Frequency–Inverse Document Frequency (TF/IDF) is a numerical statistic which reveals how important is a word to a document in a collection. The value of TF/IDF increases proportionally to the number of times a word appears in the document [14].

After Text document preprocessing we get the important words in document and the number of times a word appears in the document.

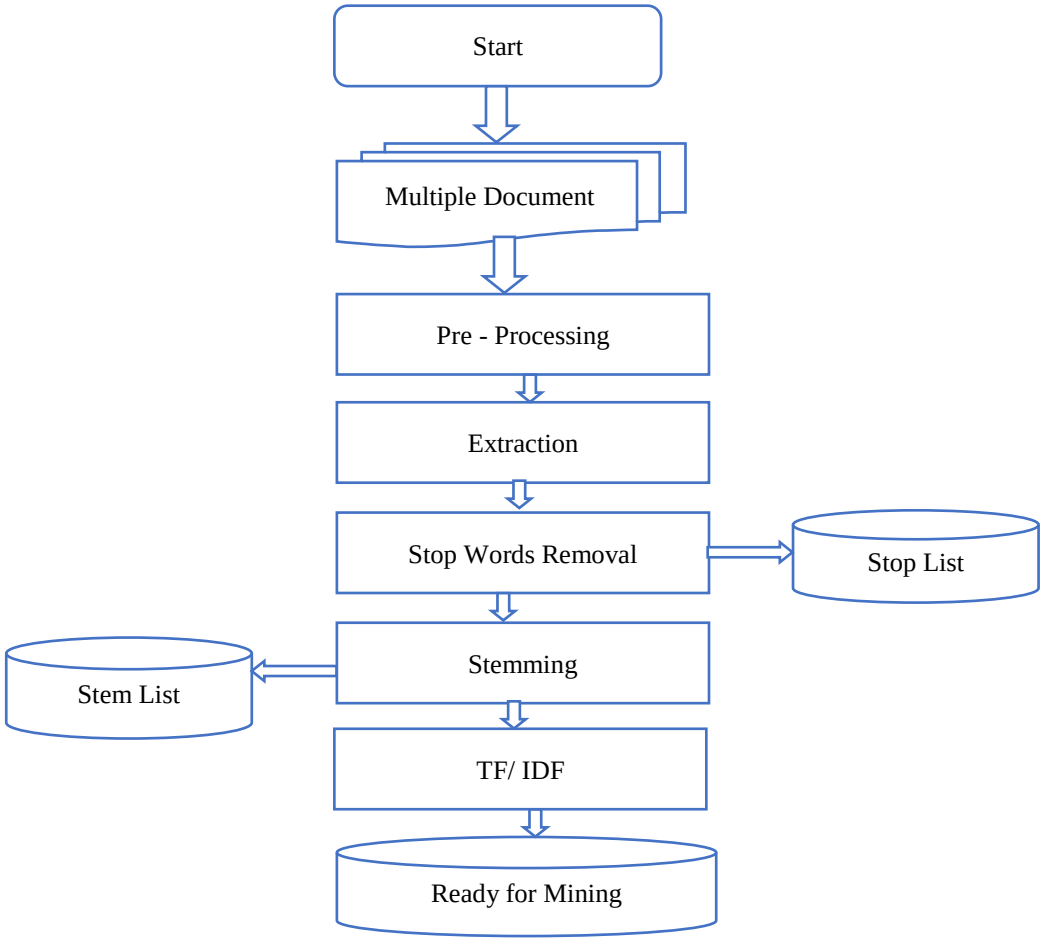

Figure 2:Text mining pre-processing steps

B. Second Step: Keywords and Resource Extraction

The DBpedia ontology has been formed in order to classify wiki data. It's a cross-domain ontology rely on info box templates in Wikipedia articles. In addition to, the ontology currently plaster about 359 classes that compose a utilization hierarchy and are depicted through 1,775 distinct properties.

The DBpedia 3.8 knowledge base describes 3.77 million things, out of that 2.35 million are categorized in a consistent Ontology, including 764,000 individuals, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases[15].

Resource Extraction: a DBpedia spotlight light is used to extract resources from the text document through our new system [16]. DBpedia spotlight is a tool designed to summarize system resource signals automatically in text [17].

Keyword extraction: The key words are drawn from the document so that the system can choose any word within the document that satisfies any of the following seven conditions [18]:

- The first condition: All non-stop words in quotations.

- The second condition: Words recognized as proper nouns by DBpedia Spotlight.
- The third condition: Complex nominal and its adjective modifiers.
- The fourth condition: All other complex nominal
- The fifth condition: All nouns and their adjectival modifiers
- The six condition: All other nouns
- The seven condition : All verbs

C. Third Step: Determine Ontology Class

After extracting resources and keywords, determine ontology class by using SPARQL query. Retrieve the DBpedia ontology classes and properties we have to build a SPARQL query with the resource itself. The result of the query (SPARQL query) is an RDF file, which holds ontology classes and properties and other information belonging to that resource.

D. Fourth Step: Determine Categories

After the last step, now we can collect the keywords (keywords extract from text document according to first and second step), resources (extract according to second step by DBpedia spotlight) and ontology classes (use Sparql query to determine the basic class of keywords and resources) are defining the system under consideration. The relationships between all can be shown to determine the best categories to these documents by using Protégé and Graphviz. Protégé is open source ontology editor and a knowledge management system and provides a graphic user interface to define ontologies. Graphviz is an open source graph visualization software and a way of representing structural information as diagrams of abstract graphs and networks. Graphviz is installed with Protégé to show relationships in ontology.

4 CLASSIFICATION ALGORITHM

The classification algorithm consists of:

(1) User enter URL if web document online or upload document if web document offline.

(2) The system works with the document to load Text from web document.

(3) The system works with text to make Text document preprocessing, user choose steps of preprocessing.

(4) The system connects to DBpedia spotlight to find resources and keywords according to the output of preprocessing.

(5) The system using Sparql query to determine classes of the resources and keywords by using online DBpedia ontology.

(6) Determine the best categories to these documents according keywords and concepts.

The algorithm below shows the steps of classification. Classification topics are defined dynamically.

Input:
O: DBpedia ontology
W: list of all words of web document
Output:
BC: a list of best categories for the web documents
Start:
// loop through all words in document
For i=1 to |w|
// document preprocessing and create list with all key words K
K = { };
// loop through the set of keywords k, connect to DBpedia spotlight to create list of resources
R= { };
// loop through the set of resources and create list of concepts
Con= { };
Find all entity belong to Con as E
E= { };
For i=1 to |w|
For i=1 to |cat|

```
//loop for all categories available
A (cat)={ };
// calculate the average for every category
BC = A;
Return BC
END
```

5 IMPLEMENTATION AND EXPERIMENTAL RESULTS

The Semantic Approach for Web document Classification System has been implemented using open source tools and software components. We used Java programming language, Jena, Sparql Query, DBpedia Spotlight and DBpedia ontology.

The designed system has the following steps:

1. User enters URL if web document online, or upload document if web document offline.
2. The system works with the document to load Text from web document.
3. The system applies preprocessing on document text, user choose steps of preprocessing.
4. The system connects to DBpedia spotlight to find resources and keywords according to the output of preprocessing figure (3).

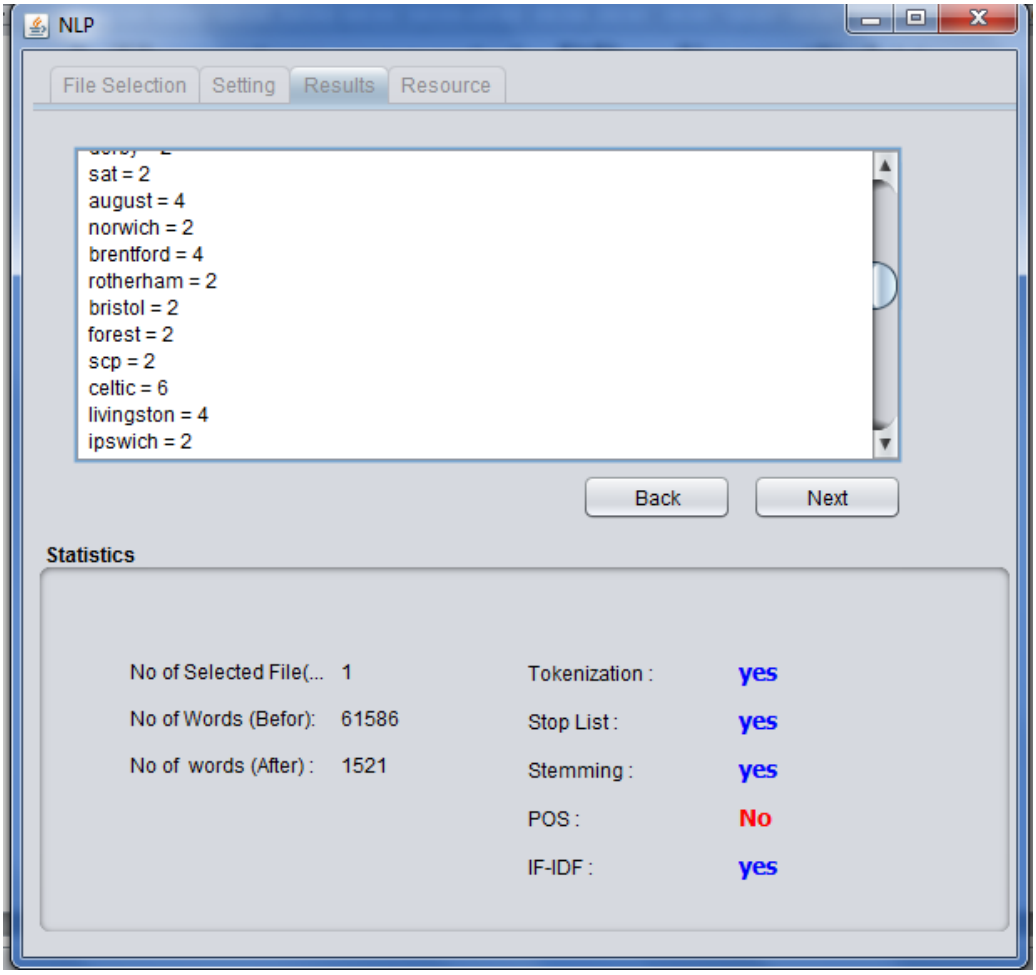

Figure (3) The output of preprocessing

5. Use DBpedia Spotlight to determine resources and keywords. Find classes by using online DBpedia ontology and determine category figure (4).

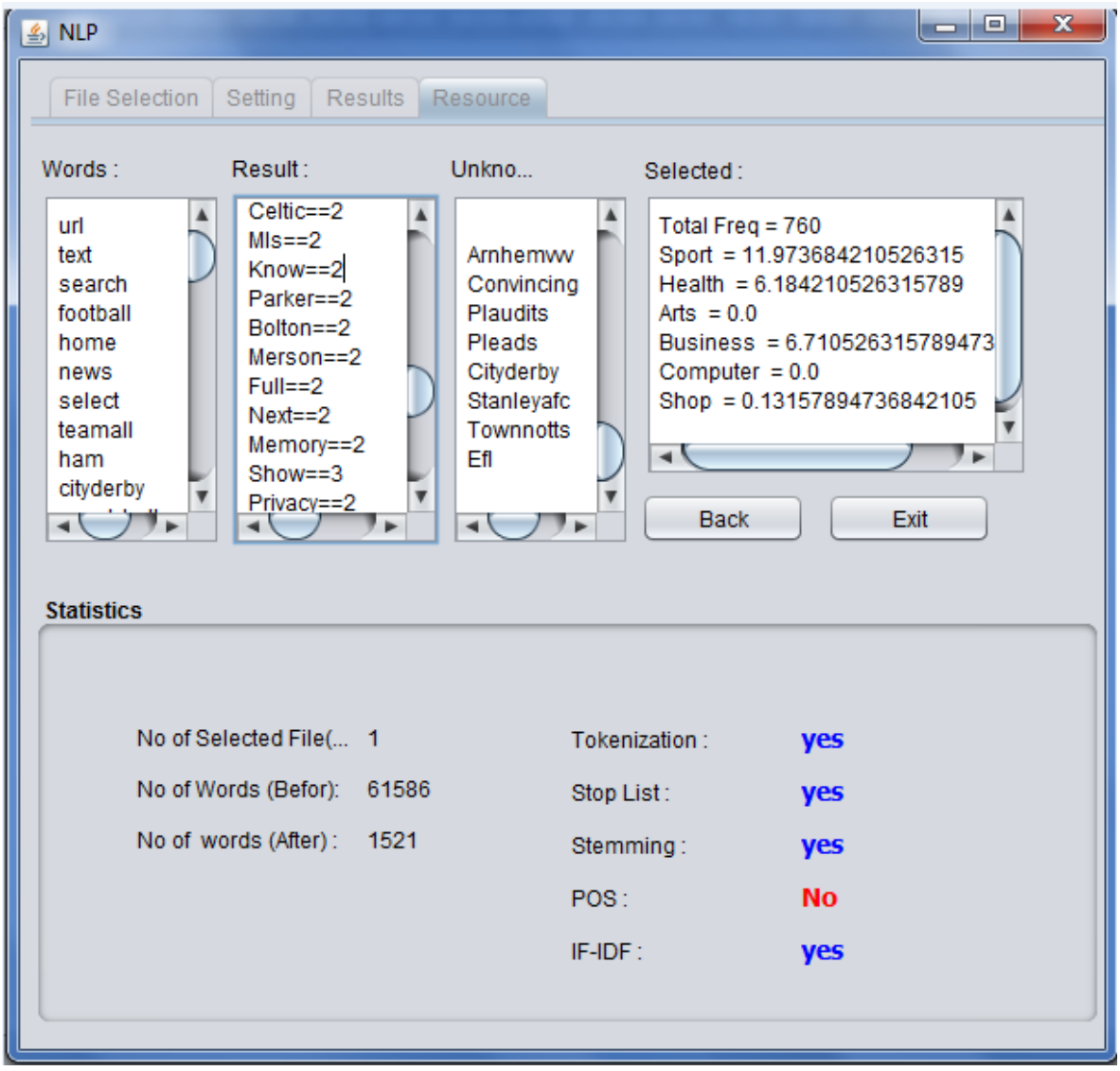

Figure (4) Determine category

There is rich history of research in classification web documents. These methods of classification can be classified in two ways: classification without using ontology, classification with using ontology. Research proved that the method classification using ontology is more efficient and gives more accurate results. Thus, we will compare one of those researches to our methodology by conducting two experiments.

Jun Fang, Lei Guo and YueNiu [9] presented a method for classification web document by using ontology (experiment A). They used ontology to classified web document by reasoning the ontology based on similarity, moreover, they divided the concepts to high concept and low concept according to TFD/ID factor. In order to assess the basic classification methodology, we shall experiment the same technique and compare the results to our proposed methodology. The second experiment, presents a method for classifying web document by using ontology (experiment B). We are using DBpedia ontology to classify web document generally without being limited to a certain domain. Experiment (B), our proposed methodology, uses all terms in the document without categorization as in Fang et. al. [9] as conducted in experiment (A). Experiments (A) and (B) evaluated the system through the Reuters (21578) data set [19] as input corpus.

The documents in the Reuters-21578 collection appeared as Reuters newswire stories. All the documents are indexed manually and they were made available for information retrieval research purposes in 1990. Currently the most widely used test collection for text categorization research, though likely to be super ceded over the next few years. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the

CONSTRUE text categorization system. Further details, including discussion of previous versions of the collection (e.g. Reuters-22173), various researchers have prepared data files useful for work with Reuters-21578 [19].

Each of the experiments were evaluating by calculating the precision, recall, accuracy and f-measure according to the following equations:

$$\text{Accuracy}[\%] = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100[\%]$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F}_1 = 2 \cdot \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \qquad \text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}$$

Where true positives (TP) denote the correct matches of positive examples, true negatives (TN) are the correct matches of negative examples, false positives (FP) represent the incorrect matches of negative examples into the positive class, and false negatives (FN) are the positive examples incorrectly classified into the negative cases.

TABLE (1) EXPERIMENTAL RESULT

	Fang et. al. **Method (A)**				**Proposed Method (B)**			
Category	**Accuracy**	**Recall**	**Precision**	**f-measure**	**Accuracy**	**Recall**	**Precision**	**f-measure**
Sport	84.3	96.1	82.1	88.5	**94.3**	93.2	92.1	**92.64**
Arts	87	96.1	85.4	90.4	**96.1**	95.9	96.6	**96.2**
Health	85.6	95.0	81.5	87.7	**97.3**	92.2	97	**94.5**
Business	87	92	85.3	88.5	**98**	97	98.5	**97.7**
Computer	85	93.1	85.1	88.9	**94.2**	92.2	92.4	**92.2**
Shop	**96**	95.3	85.5	90.1	92	92	92	**92**

From the above table, we can observe that experiment (A) gives higher similarity values but not accurate as the results of experiment (B). Experiment (B) provides higher accuracy and achieved high precision and low recall than experiment (A). Which reveals that classifying web document using ontology considering all concepts provides higher accurate results and reduces or prevents irrelevant knowledge in classification than the categorizing concepts methods of Fang et. al.[9].

6 CONCLUSIONS AND FUTURE WORK

We presented a tractable approach to web document classification, that depends on DBpedia ontology. The use of an open source ontology provides having a classifier without the need for extensive training. In addition the ontology is a general knowledge and continuously updated with new concepts, providing general knowledge that is not domain specific. We extract terms from document and extract resources to determine classes and map them to their concepts then determine the best category with the highest similarity. Our method depends on the semantics of terms and sentence. In future work, terms that do not exist in the used ontology need to be added and a merge methodology need to be developed. Moreover, other open source ontologies can be validated and used and results shall be compared.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, pp. 1-47, 2002.
[2] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis". Discourse processes, vol. 25, pp. 259-284, 1998.
[3] Morsey, M. Lehmann, J. Auer, Sö. Stadler, C. & Hellmann, S. (2012) DBpedia and the Live Extraction of Structured Data from Wikipedia. Program: electronic library and information systems, 46, 27.
[4] https://wordnet.princeton.edu/ [online last visited 5/9/2018].
[5]www.cyc.com/opencyc/ [Online last visited 5/9/2018].
[6]http://www.adampease.org/OP/ [online last visited 5/9/2018].

[7] Bin Shi, Liying Fang, Jianzhuo Yan, Pu Wang, and Chen Dong. "Classification of Semantic Documents based on WordNet". International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, pp. 173-176,Vegas, Nevada, USA2009.

[8] Bai Rujiang Shandong and Liao Junhua, "Improving Documents Classification with Semantic Features". Second International Symposium on Electronic Commerce and Security. Guangzhou, China 2009.

[9] Jun Fang, LeiGuo and Yue Niu. "Documents Classification by Using Ontology Reasoning and Similarity Measure". Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010) , Yantai, Shandong, China. IEEE2010,

[10] Shikha Agarwal, Archana Singhal and Poonam Bedi, "Classification of RSS feed news items using ontology," *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*,pp. 491-496.Kochi, 2012IEEE.

[11] Chaaminda Manjula Wijewickrema "Impact of an ontology for automatic text classification". Annals of Library and Information Studies. Vol.61, pp.263-272, December 2014.

[12] Henrihs Gorskis1 and Arkady Borisov2 "Ontology Building Using Classification Rules and Discovered Concepts", Information Technology and Management. Information Technology and Management Science Journal of Riga Technical University, Volume 18: Issue 1, pp. 37-41, 2015.

[13] Vishal Gupta , Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging technologies in web intelligence, vol. 1, no. 1, pp.60-76, August 2009.

[14] Jacob Perkins "Python Text Processing with NLTK 2.0 Cookbook". 2010.

[15]http://blog.dbpedia.org/category/dataset-releases/ [online last seen 9/8/2018].

[16] Tartir, Samir; McKnight, Bobby and Arpinar, I. Budak (2009) "Semantic QA: web-based ontology driven question answering". In Proceedings of the 2009 ACM symposium on Applied Computing (SAC '09). Honolulu, Hawaii, USA 2009.

[17] Pablo N. Mendes1, Max Jakob1, Andrés García-Silva2, Christian Bizer1 (2011). "DBpedia Spotlight: Shedding Light on the Web of Documents". In 7th International Conference on Semantic Systems , SEMANTICS 2011, Graz, Austria, 2011.

[18] Guo, Qinglin and Zhang, Ming "Question answering system based on ontology and semantic web". In Proceedings of the 3rd international conference on Rough sets and knowledge technology (RSKT'08), 2008, *Tianjin, China*.

[19] David D. Lewis. Distribution 1.0 readme file (v1.2) for reuters-21578. AT&T Labs - Research, 1997

BIOGRAPHY

PassentEl Kafrawy, Professor, Faculty Science, Menoufia University

Dr. Passent M. ElKafrawy is a Professor since 2018, she got her PhD from the University of Connecticut in United states on 2006 in Computer Science and Engineering. In the field of computational geometry as a branch of Artificial Intelligence. Then she taught in Eastern State University of Connecticut for one year. In 2007 she worked as a Teacher in Faculty of Science, Menoufia University, Mathematics and computer science department, then as an Assistant Professor in 2013.

Dina ElDemrdash, Computer Instructor, Information center, Menoufia University

Graduate from Faculty of Science, Menoufia University, Mathematics and Computer Science Department2009. Certified computer trainer from UNESCO and ICDL ARABIA since 2011 Then, in 2012 she works Specialist Systems Analysis and Design, Menoufia University until now.

ARABIC ABSTRACT

المنهج الدلالي لتصنيف متغير وغير ثابت للوثائق على شبكة الإنترنت

بسنت محمد الكفراوي، دينا الدمرداش
كلية العلوم جامعة المنوفية
[1]basant.elkafrawi@science.menofia.edu.eg
[2]dina.eldemrdash@science.menofia.edu.eg

الملخص:
نحن نقدم وسيلة للنهج الدلالي لتصنيف متغير وغير ثابت للوثائق على شبكة الإنترنت .
النهج المقدم يتطلب فقط انطولوجيل نطاق محدد ومجموعة من فئات المستخدمين معرفة مسبقا .في الوقت الحالي ، تعتمد معظم المناهج لتصنيف الوثائق النصيةعلي ان تمثل الوثيقة باعتبارها (حقيبة من الكلمات) ويتطلب تدريب مجموعة كبيرة من الوثائق لتدريب المصنف علي اثارها.ولكن لا يتطلب النهج الدلالي الذي نستخدمة مجموعة من الوثائق للتدريب، ويقوم النهج بتصنيف الوثائق تصنيف متغير وغير ثابت على حسب محتوها من كلمات ومفاهيم .
في طريقة استخد امنا الأنطولوجيا نستخدم**DBpedia** كمصنف، وتمثل الوثيقة باعتبارها (حقيبة من الكلمات و المفاهيم).فنحن بمعالجة الوثائق النصية واستخرج الكلمات من الوثيقة، ونستخرج مواردها من **DBpedia spotlight** ، ونستخدم الاستعلام **SPARQL**و**DBpedia ontology**لتحديد الفئة الأنطولوجيا ثم نحدد الفئة الأفضل.

الكلمات المفتاحية:
التصنيف، النهج الدلالي ، الانطولوجي، DBpedia spotlight ، DBpediaانطولوجي.