# A PROPOSED MULTI-DOMAIN APPROACH FOR AUTOMATIC CLASSIFICATION OF TEXT DOCUMENTS

Abdelrahman M. Arab[1], Ahmed M. Gadallah[2] and Akram Salah[3]

[1, 2] Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt
[3] Department of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt

***ABSTRACT***

*Classification is an important technique used in information retrieval. Supervised classification suffers from certain limitations concerning the collection and labeling of the training dataset. When facing Multi-Domain classification, multiple training datasets and classifiers are needed which is relatively difficult. In this paper an unsupervised classification system is proposed that can manage the Multi-Domain classification problem as well. It is a multi-domain system where each domain represented by an ontology. A document is mapped on each ontology based on the weights of the mutual tokens between them with the help of fuzzy sets, resulting in a mapping degree of the document with each domain. An experiment carried out showing satisfying classification results with an improvement in the evaluation results of the proposed system compared to Apache Lucene.*

***KEYWORDS***

*Information Retrieval, Ontology, Machine Learning, Document Classification, Fuzzy Sets*

## 1. INTRODUCTION

As well as classifying books in a library saves time and effort in searching for a specific book, so does classification of text documents in the IR system, especially with large corpus. It enhances the retrieval of text documents and increases the precision of the system. Classification is known to be a supervised machine learning (ML) technique [29], where a training dataset is labeled manually by an expert before the classifier's decision criterion is learnt automatically from it [23].

Notwithstanding the capability of supervised classification technique, it actually suffers from limitations. The first limitation is due to the fact that the precision of the classifier increases with the increased number of training documents. In order to reach higher degrees of precision, additional time, effort and money required for both collecting and labeling training documents, and with the increased number of documents to be labeled, the chance of mislabeling a document increases [9, 14]. The second limitation is the difficulty of handling multi-domain documents, where the document is being typically labeled as belonging to only one domain even if it actually belongs to other domain(s) too, that the system cannot recognize. For example, a document about drugs used in sport and doping analysis can be important to users searching for sport domain and users searching for chemistry domain as well, it may even be of interest to medicine domain users or others. Classifying multi-domain documents is not a trivial or an easy task, especially when the domains sharing the document are of equal importance.

This paper proposes an unsupervised classification approach capable of classifying text documents dealing at the same time with the multi-domain classification problem. The proposed system uses a number of readymade ontologies, each represents one domain, forming with the others a domain collection which is used to classify text documents without any training dataset. This paper assumes that a text document belongs to all domains with certain mapping degrees. Each document thus can have a list of domains with the first domain in the list (the one with the highest mapping degree) is considered as the main domain of the document. Fuzzy sets are being used to overcome the uncertainty associated with mapping the document onto each domain.

An experiment is carried out where the system succeeded in classifying 67% of the documents in the dataset. The evaluation results of the system showed an improvement when compared to the results obtained by Apache Lucene 5.5.0 using $F_1$ measure.

The rest of the paper is organized as follows. Section 2 gives a brief background of the main topics used in this research with a brief overview of the associated related work. Section 3 discusses the proposed system. Section 4 represents the experiment carried out to test the proposed system. Finally, the conclusion presented in section 5.

## 2. BACKGROUND AND RELATED WORK

This section is categorized into four sub sections presenting a brief background of the main topics used in this paper through shedding light on a number of related works in the area of text document classification.

### 2.1. Ontology

Ontology has been widely used in solving the semantic problem, it is considered as the choice of the W3C for handling the Semantic Web in general [20, 21]. In this research, an ontology classification performed on a number of papers based on the way they use ontology in the applied methods. **Figure 1** shows this classification:

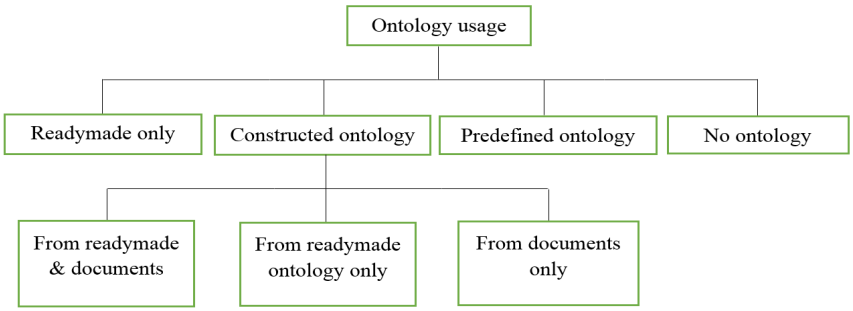


Figure 1. Graph representing the ontology usage classification.

Not using ontology at all can be substituted with ML techniques, where the classifier decision of how to classify new documents is learnt from a training dataset as in [5, 7, 11, 15, 17]. In other papers however, both techniques, ontology and ML, are being used together as the case in [1, 2, 4, 12], Where an ontology plays a crucial role in the learning process.

When the choice is made to use the ontology we have three types Figure 1. The first type involves the usage of a predefined (pre-built) ontology, an ontology that is being built for a specific task in the paper representing a specific domain decided by the author. This ontology is built with the

help of a domain expert as in [10, 13, 16, 18]. The ontology in [10, 13] used to expand the query into a list of keywords used to retrieve the documents. In [10] a Multi-Views Fuzzy Related Ontology, MVFRO, is built. Linguistic values and fuzzy number are used to express the relations between fuzzy ontology components, which are defined by a domain expert according to his own subjective view, and stored in a relational database together with the ontology components after stemming. In [13], two ontologies are used with the relationships between them expressed as fuzzy relations. The domain is defined as the agrometeorology domain in Brazil, and the ontologies used are a lightweight ontology referring to the geographical Brazilian territory and a lightweight ontology referring to the climate distribution over the Brazilian territory. The ontologies are manually constructed considering the Brazilian map. Although [16, 18] depend on the ontology to extract semantic concepts related to the terms in the document, they differ in the domain and the kind of documents subjected to the ontology. The domain in [16] is the economy domain and there is no training documents, while in [18] training documents in the philosophy domain are used.

The second type of ontologies is constructed ones, which are automatically built within the system. Building such an ontology depends on either the documents in the dataset (training or testing ones), a readymade ontology or both. We can find this type in [3, 6, 14], but they differ in the construction method. The dependence in [6] is on the training documents to construct the ontology. Each document constructs its own ontology then an Ontology for One Category (OOC) is built from the ontologies of the documents belonging to that category. This category ontology is then used to classify documents by subjecting the ontology of each new document to it. The construction of ontology in [3, 14] depends on a readymade ontology, which in case of [3] is the WordNet and in [14] is the Wikipedia. In [14] an RDF ontology constructed from a full English version of Wikipedia, on which the documents are being subjected in order to make a semantic graph of each document based on the hierarchy in the ontology. In [3] however the WordNet is not the main source for constructing the ontology. The WordNet is used to expand the query into a list of semantic instances which are then integrated with the keywords from documents, annotated by an expert, and the result is the construction of a hybrid ontology.

The last type is using only a readymade ontology, these are ontologies made for a specific domain(s) by different organizations or people, which can be used by researchers saving time and effort as in [1, 2, 4, 8, 9, 12, 19]. Training dataset is being subjected to a readymade ontology in [1, 2, 4, 8, 12], with the difference between them in the kind of ontology used. In [1] the MeSH medical domain ontology is used, [2, 12] use ACM CCS ontology for the computer science domain, in [4] General Finnish Ontology (YSO) is used. In [8] however, two ontologies being used which are WordNet and Wikipedia, the training documents are subjected to both of them forming two lexical chains, one for each ontology. The WordNet is being used differently by [9, 19]. While in [19] the query is expanded by being subjected to the WordNet, the category names are entered into the system in [9] then subjected to the WordNet to formulate proximity equations being transformed later on into proximity relations.

The proposed system belongs to the last type, it is a multi-domain system thus a number of readymade ontologies are being introduced into the system representing the different system domains. With each new domain a new readymade ontology is added to the system.

## 2.2. Machine Learning

ML techniques are of great help in information retrieval and have been widely used in classifying documents. Mainly the presence or absence of a training dataset determines the type of technique used, whether supervised, unsupervised or semi-supervised. For a supervised technique the classifier's decision criterion is learnt from the training documents that have been already

classified. Training documents are absent in the unsupervised technique, the classifier has to depend on the unlabeled documents that need to be classified, where some features from the document are extracted to measure the similarity between the document and the category to be classified within. Semi-supervised technique is an intermediate one between supervised and unsupervised techniques. It depends on both labeled (training) and unlabeled documents.

Although many researches use ML, they may differ in the way the techniques are applied. [1, 2, 4, 8, 12] use supervised learning with an ontology used to construct a semantic representation of the training documents. [7, 11, 15, 17] on the other hand, do not use ontology at all during the supervised learning. [6, 18] use supervised learning and use ontology somehow during their approach. While in [6] ontology is constructed from the training documents, in [18] ontology is used as a second step after the similarity between labeled and unlabeled documents has been calculated. Semi-supervised technique is applied in [5], where both labeled and unlabeled documents exist, and the classifier's decision criterion is learnt based on expectation maximization algorithm. The unsupervised technique used in [3, 9, 10, 13, 14, 16, 19], where ontology is used, to compensated for the absence of the training dataset, with the help of the unlabeled documents, the query or both.

This research uses an unsupervised ML technique, no training dataset is being used. The proposed system depends on the words of the document itself to perform classification with the help of readymade ontologies.

## 2.3. Transparency

Transparency as described in [22] considers the interaction between the user and the system, it can be either transparent, interaction or hybrid. In transparent systems the interaction between the user and the system is minimum, while in the interaction systems this interaction is crucial as in [10]. In [4, 9, 19] the hybrid system is used where the system looks like the transparent one but with a little interference from the user in a specific task.

The proposed system is a hybrid one, where the user is asked through an interface to select a domain from a list of the available system domains. This would be the only interaction of the user with the system besides formulating his query.

## 2.4. Fuzzy Theory

Fuzzy theory is used to face the ambiguity and uncertainty that usually accompanies the semantic problem. Having more than one meaning of the same word, and the possibility that a word can belong to more than one domain at the same time creates the need for the fuzzy sets to overcome this fuzziness. In [10] fuzziness is used within the ontology itself where fuzzy ontology is used to expand the query with semantic instances, the fuzzy values between ontology components determined by an expert. In [13], fuzzy relations formed between two ontologies also used in query expansion. A fuzzy matching technique used in [3] to construct a hybrid ontology from the integrated list between the document and the query which is then used in classifying documents. The similarity between document keywords and the ontology is measured in [1, 9] using also fuzzy theory. In [7, 8] fuzzy rules have been generated from the documents themselves to classify them based on the document features.

In this research, the document is subjected to many domains. With the fuzziness resulting from such case, fuzzy sets are used in order to help overcoming this fuzziness and classifying the document among the system domains.

# 3. THE PROPOSED SYSTEM

## 3.1. Architecture

The proposed IR system consists of 1) document collection 2) domain collection 3) classification module 4) internal storage 5) retrieval module. It has three inputs (document, ontology and query), and one output which is the retrieved documents. **Figure 2** shows a diagram representing the system components.

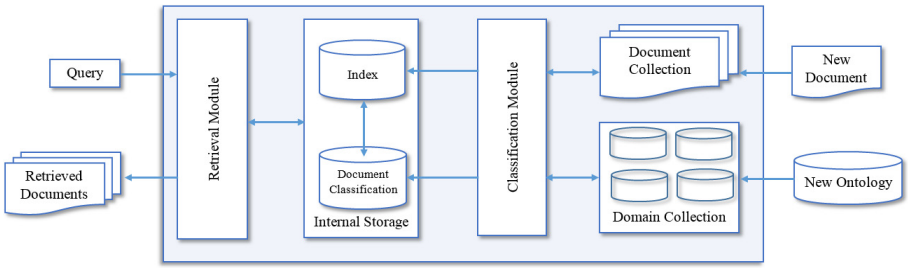


Figure 2. A diagram representing the proposed system.

## 3.2. Methodology

The general idea behind the method used in this research is represented in **Figure 3**. The methodology is discussed in light of the Classification and Retrieval modules in the system diagram **Figure 2**. It has therefore two stages (Classification, Retrieval).

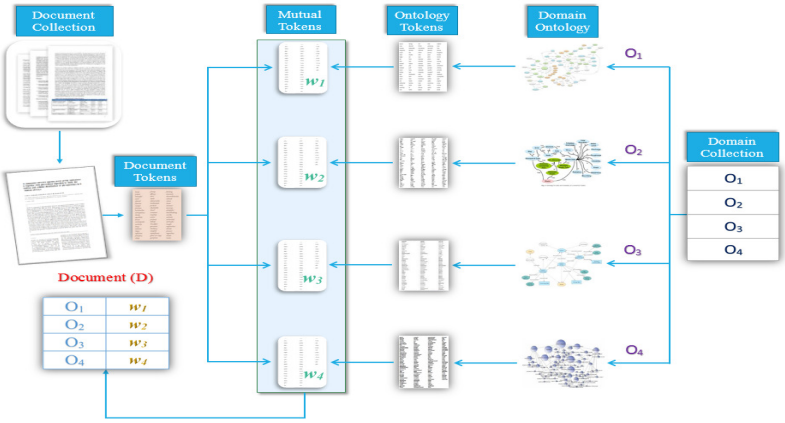


Figure 3. A diagram representing the general idea behind methodology.

### 3.2.1. Classification

Classification has three phases: 1) Preprocessing 2) Weight Calculations 3) Mapping. These phases are briefly discussed in the following subsections.

#### 3.2.1.1. Preprocessing

The preprocessing phase contains four steps (Tokenization, Lowercase, Stop Words and Stemming). A document is first tokenized converted thus into a list of tokens. All tokens are converted to lowercase, and the stop words (union of some stop word lists namely: English,

Gerard and Chris, Snowball, Terrier and MySql) are removed. Finally all tokens are stemmed using Porter Stemmer. An ontology is preprocessed the same way as the document except of the stop word step. The result of this phase is two lists of preprocessed tokens, one for a document and one for an ontology.

### 3.2.1.2. Weight Calculations

The duplications in the two lists, formed as a result of the preprocessing phase, are removed giving rise to two lists of distinct tokens, with each token has its frequency calculated. The total weight ($w_d$) of a token in the document list is given by:

$$w_d = wt + wc + wf_d$$

Where ($wt$) is the weight of the token due to the presence in the document's title, ($wc$) is the weight of the token due to uppercase and ($wf_d$) is the frequency weight of the token in the document. The total weight ($W_o$) of a token in the ontology list is given by:

$$w_o = wl + wf_o$$

Where ($wl$) is the weight of the token due to its level in the ontology and ($wf_o$) is the frequency weight of the token in the ontology. A list of mutual tokens is created from the document and the ontology lists. Each token in the mutual list will have a weight ($w$) given by:

$$w = w_d + w_o$$

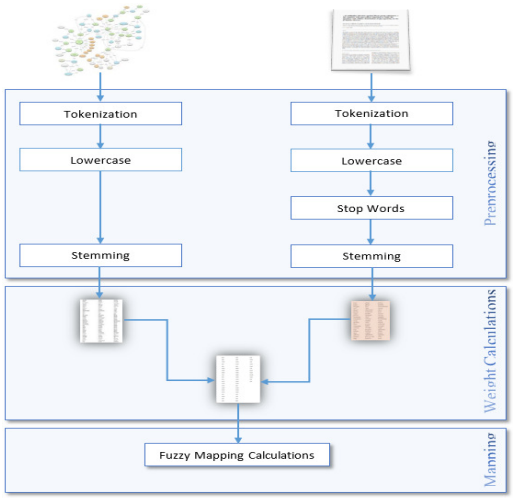


Figure 4. Classification stage of the methodology.

### 3.2.1.3. Mapping

The three lists from the previous phase (document list, ontology list and mutual list) are represented here as the finite sets (D, O and L) respectively as follows:

D = {Set of the distinct tokens in the document list}
O = {Set of the distinct tokens in the ontology list}
L = D ∩ O

The finite set (D) represents the domain of the fuzzy set (A) which is defined as:

$$A = \{(\mu_A(x_i) \mid x_i) \mid x_i \in D, j = 1,2,\dots,n\}$$

The degree of membership of each element ($x$) to the fuzzy set ($A$) is determined by the membership function $\mu_A(x)$ which is defined as follows:

$$\mu_A(x) = \begin{cases} 0 & if\ x \notin L \\ w & if\ x \in L, w < h \\ 1 & if\ x \in L, w \geq h \end{cases}$$

Where ($w$) is the weight of the token in the mutual list corresponding to element ($x$), and ($h$) is a constant. The cardinality of the fuzzy set ($A$) denotes the mapping degree of the document to the domain and is given by:

$$Card\,(A) = |A| = \sum \mu_A(x_j), x_j \in L, j = 1,2,\dots,n\}$$

The three phases of the Classification stage are repeated for every domain (ontology) in the domain collection. Then if the number of domains in the system equal ($i$), an equivalent number of fuzzy sets ($A_i$) is produced, one for each domain (ontology). The cardinality values obtained for these fuzzy sets denote the mapping degree of the document to each domain. The cardinality general equation is given by:

$$Card\,(A_i) = |A_i| = \sum \mu_{A_i}(x_j), x_j \in L_i, i = 1,2,\dots,n \text{ and } j = 1,2,\dots,n\}$$

### 3.2.2. Retrieval

Retrieval stage involves the interaction with the user and has two phases 1) Preprocessing 2) Ranking.

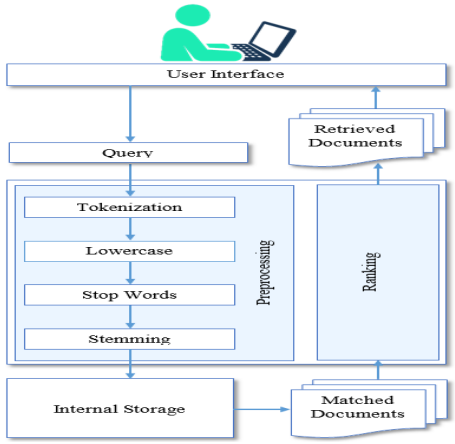


Figure 5. Retrieval stage of the methodology.

#### 3.2.2.1. Preprocessing

The query is preprocessed the same way the document does in the classification stage passing through the four preprocessing steps (tokenization, lowercase, stop words and stemming). The list of query tokens is then passed to the internal storage where matched documents being retrieved using Boolean retrieval.

#### 3.2.2.2. Ranking

Matched documents are ranked, based on the document classification and the weight of query tokens within the document, giving rise to the retrieved documents introduced to the user through the user interface.

# 4. EXPERIMENT

## 4.1. Domain and Document Collections

The proposed system was tested using four domains each represented by one ontology **Table 1**.

Table 1. The ontologies used to represent the specified domains.

| No. | Domain | Ontology | URL |
|---|---|---|---|
| **1** | chemistry | Chemistry-cplx | http://dumontierlab.stanford.edu/ontologies.php |
| **2** | medicine | MeSH | https://www.nlm.nih.gov/mesh/ |
| **3** | philosophy | PhiloSURFical | http://philosurfical.open.ac.uk/onto.html |
| **4** | computer science | ACM CCS | http://www.acm.org/about/class |

Only sections A, B, C and F have been used from the MeSH ontology. The dataset contains 40 randomly collected documents **Table 2**.

Table 2. The sources used to collect documents in the specified domains.

| No. | Domain | source | No. of documents | Total |
|---|---|---|---|---|
| **1** | chemistry | Chemistry Central Journal | 2 | 10 |
| | | Journal of Analytical Chemistry | 8 | |
| **2** | medicine | American Journal of Clinical Medicine Research | 6 | 10 |
| | | The American Journal of Medicine | 4 | |
| **3** | philosophy | Test dataset used in paper [18] | 10 | 10 |
| **4** | computer science | Journal of Computer Sciences | 7 | 10 |
| | | International Journal of Computer Science & Engineering Survey (IJCSES) | 3 | |

## 4.2. Evaluation

The system evaluated based on $F_1$ measure and the results compared with those of Apache Lucene 5.5.0 using the English Analyzer. The evaluation contained 25 queries, and is done under two assumptions:

1- The queries are chosen arbitrarily from the most frequent tokens in the index, and the domain for each query is chosen randomly.
2- The retrieved document is considered relevant if it contains the query tokens, and its source domain (the domain of the source from which the document is collected) is the same as the query domain.

### 4.3. Results

In **Table 3** the domain with the biggest mapping degree is considered as the main domain.

Table 3. Classification results of the document collection.

| No. | Domain | No. of documents in the document collection | No. of successfully classified documents | Ratio of successful classification |
|---|---|---|---|---|
| **1** | chemistry | 10 | 6 | 60% |
| **2** | medicine | 10 | 8 | 80% |
| **3** | philosophy | 10 | 7 | 70% |
| **4** | computer science | 10 | 6 | 60% |
| **Total No. of successfully classified documents** | | | **27** | **67.5%** |

**Table 4** shows part of the evaluation results of both the proposed system and Apache Lucene. For each query precision (P), recall (R) and $F_1$ values are calculated, then the difference between $F_1$ values between the two systems is calculated.

Table 4. Part of the evaluation results table.

| query | domain | Apache Lucene 5.5.0 | | | proposed system | | | difference in $F_1$ |
|---|---|---|---|---|---|---|---|---|
| | | P | R | $F_1$ | P | R | $F_1$ | |
| method | computer science | 0.36 | 1.00 | 0.53 | 0.75 | 0.67 | 0.71 | 0.18 |
| organs | medicine | 0.39 | 1.00 | 0.56 | 0.88 | 0.78 | 0.82 | 0.26 |
| measure | medicine | 0.26 | 1.00 | 0.41 | 0.83 | 0.83 | 0.83 | 0.42 |
| layer | chemistry | 0.55 | 1.00 | 0.71 | 1.00 | 0.50 | 0.67 | -0.04 |
| procedure | computer science | 0.33 | 1.00 | 0.50 | 0.67 | 0.80 | 0.73 | 0.23 |
| Procedure | medicine | 0.40 | 1.00 | 0.57 | 1.00 | 0.67 | 0.80 | 0.23 |

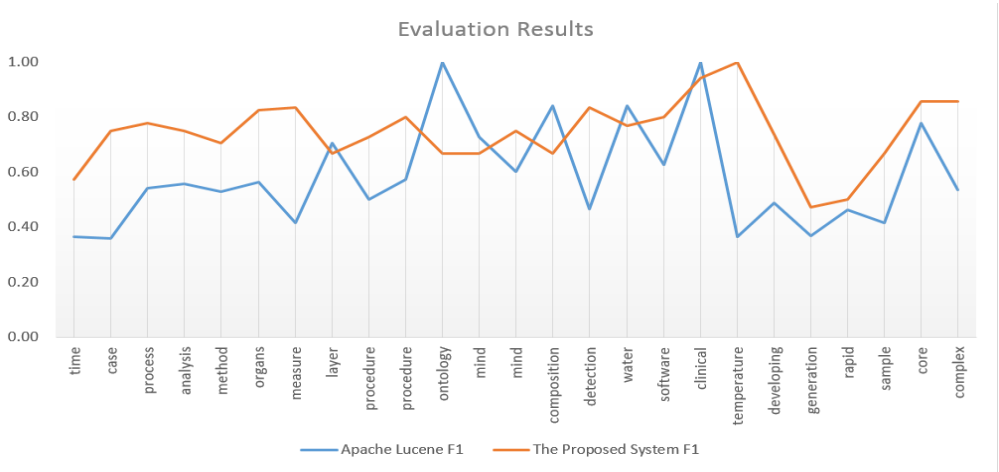


Figure 6. Comparing $F_1$ values between the proposed system and Apache Lucene.

### 4.4. Discussion

- Although the system succeeded in classifying only 67.5% of the documents, it did better by succeeding in classifying 80% of the medicine documents, while chemistry and computer science domains had the worst results.

- By testing more than one ontology for the same domain, the results changed correspondingly.

- **Table 5** shows that the mapping degrees of the domains are close that sometimes only less than 1% can favor one domain over the others, which suggests that further Adjustment of weights is needed.

- The evaluation results in **Figure 6** shows an improvement in the $F_1$ value for most of the queries using the proposed system when compared to Apache Lucene. This improvement is due to the classification capability of the system and the fact that the proposed system is a domain specific one, on the contrary to Apache Lucene.

Table 5. Part of the document mapping results table.

| Document | Main domain | Chemistry | Medicine | Philosophy | Computer science |
|---|---|---|---|---|---|
| **001** | chemistry | 25.85 | 25.50 | 25.50 | 23.15 |
| **002** | chemistry | 26.93 | 22.25 | 25.15 | 25.67 |
| **003** | philosophy | 26.42 | 23.44 | 27.24 | 22.90 |
| **004** | chemistry | 28.58 | 25.56 | 24.34 | 21.52 |
| **005** | chemistry | 27.92 | 24.71 | 24.18 | 23.19 |
| **006** | computer science | 25.04 | 24.03 | 25.29 | 25.64 |
| **007** | Philosophy | 27.72 | 21.65 | 27.98 | 22.64 |
| **008** | Chemistry | 27.85 | 26.29 | 23.92 | 21.94 |
| **009** | Philosophy | 25.71 | 23.05 | 26.32 | 24.91 |
| **010** | Chemistry | 28.95 | 21.57 | 26.02 | 23.46 |

## 5. CONCLUSION

The proposed system proved a considerable capability in classifying text documents among multiple domains. The results encourages further work on the light of two conclusions; first, the mapping degree between the document and the domain increases proportionally with the quality of the ontology regarding its view, size, domain description and concepts formulation. Second, enriching the weighting features (both in document and ontology) of the word increases its mapping power.

## REFERENCES

[1] François-Élie Calvier, M. P., Gérard Dray, Sylvie Ranwez (2013). "Ontology Based Machine Learning for Semantic Multiclass Classification." TOTH : Terminologie & Ontologie : Théories et Applications, Jun 2013, Chambéry, France <hal-00838262>: 100.

[2] Periakaruppan, R. and R. Nadarajan (2015). Automatic Clustering of Research Articles Using Domain Ontology and Fuzzy Logic. Advances in Web-Based Learning – ICWL 2013 Workshops: USL 2013, IWSLL 2013, KMEL 2013, IWCWL 2013, WIL 2013, and IWEEC 2013, Kenting, Taiwan, October 6-9, 2013, Revised Selected Papers. D. K. W. Chiu, M. Wang, E. Popescu et al. Berlin, Heidelberg, Springer Berlin Heidelberg: 42-51.

[3] Uthayan, K. R. and G. S. A. Mala (2015). "Hybrid Ontology for Semantic Information Retrieval Model Using Keyword Matching Indexing System." The Scientific World Journal 2015 (2015): 9

[4] Nyberg, K. (2011). Document Classification Using Machine Learning and Ontologies. SCHOOL OF SCIENCE. Norway, AALTO UNIVERSITY. Degree Programme of Information Networks: 71.

[5] Nigam, B., et al. (2011). "Document Classification Using Expectation Maximization with Semi Supervised Learning." International Journal on Soft Computing ( IJSC ) 2, No.4.

[6] Asta Bevainyte, L. B. (2010). "DOCUMENT CLASSIFICATION USING WEIGHTED ONTOLOGY " Materials Physics and Mechanics 9 (2010) 246-250.

[7] Nogueira, T. M., et al. (2011). "Fuzzy rules for document classification to improve information retrieval." International Journal of Computer Information Systems and Industrial Management Applications 3: 210-217.

[8] Pandey, U., et al. (2011). "Semantic Document Classification using Lexical Chaining & Fuzzy Approach." International Journal of Soft Computing and Engineering (IJSCE) 1(5).

[9] Romero, F. P., et al. (2013). "Classifying unlabeled short texts using a fuzzy declarative approach." Language Resources and Evaluation 47(1): 151-178.

[10] Zeinab E. Attia , A. M. G., Hesham A. Hefny (2014). "Semantic Information Retrieval Model: Fuzzy Ontology Approach." International Journal of Computer Applications 91(13): 9-14.

[11] Han, E.-H. and G. Karypis (2000). Centroid-Based Document Classification: Analysis and Experimental Results. Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Springer-Verlag: 424-431.

[12] Lee, Y.-H., et al. (2009). Use of Ontology to Support Concept-Based Text Categorization. Designing E-Business Systems. Markets, Services, and Networks: 7th Workshop on E-Business, WEB 2008, Paris, France, December 13, 2008, Revised Selected Papers. C. Weinhardt, S. Luckner and J. Stößer. Berlin, Heidelberg, Springer Berlin Heidelberg: 201-213.

[13] Leite, M. A. A. and I. L. M. Ricarte (2008). A Framework for Information Retrieval Based on Fuzzy Relations and Multiple Ontologies. Advances in Artificial Intelligence – IBERAMIA 2008: 11th Ibero-American Conference on AI, Lisbon, Portugal, October 14-17, 2008. Proceedings. H. Geffner, R. Prada, I. Machado Alexandre and N. David. Berlin, Heidelberg, Springer Berlin Heidelberg: 292-301.

[14] Maciej Janik, K. K. (2008). Training-less Ontology-based Text Categorization. In Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008) at the 30th European Conference on Information Retrieval (ECIR'08). Glasgow, Scotland.

[15] Muhammad Faheem Khan, A. K., Shahid Khan,Aziz Ullah Khan (2014). "CONTENT BASED AUTOMATIC CLASSIFICATION OF RESEARCH ARTICLES." Sci.Int.(Lahore) 26(5): 2495-2499.

[16] Song, M.-H., et al. (2006). Ontology-Based Automatic Classification of Web Pages. Applied Soft Computing Technologies: The Challenge of Complexity. A. Abraham, B. de Baets, M. Köppen and B. Nickolay. Berlin, Heidelberg, Springer Berlin Heidelberg: 483-493.

[17] Oakes, M., et al. (2001). A method based on the chi-square test for document classification. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. New Orleans, Louisiana, USA, ACM: 440-441.

[18] Wijewickrema, C. M. and R. Gamage (2013). An Ontology Based Fully Automatic Document Classification System Using an Existing Semi-Automatic System. (IFLA WLIC 2013) IFLA World Library and Information Congress. At Suntec City, Singapore, 79th IFLA General Conference and Assembly.

[19] De Luca, E. W., et al. (2004). Ontology-based semantic online classification of documents: Supporting users in searching the web. Proc. of the European Symposium on Intelligent Technologies (EUNITE 2004), Aachen.

[20] Kück, G. (2004). "Tim Berners-Lee's Semantic Web." South African Journal of information management 6(1).

[21] Shadbolt, N., et al. (2006). "The Semantic Web Revisited." IEEE Intelligent Systems 21(3): 96-101.

[22] Mangold, C. (2007). "A survey and classification of semantic search approaches." Int. J. Metadata Semant. Ontologies 2, No. 1(1): 23-34.

[23] Christopher D. Manning, P. R., Hinrich Schütze (2008). An Introduction to Information Retrieval. Cambridge, England, Cambridge University Press.

[24] INGWERSEN, P. (1992). INFORMATION RETRIEVAL INTERACTION. 500 Chesham House 150 Regent Street LONDON W1R 5FA United Kingdom, Taylor Graham Publishing.

[25] Blair, D. C. (2003). "Information retrieval and the philosophy of language." Annual Review of Information Science and Technology 37(1): 3-50.
[26] Griffin, E. A. (2012). A first look at communication theory. New York, USA, McGraw-Hill.
[27] BLAIR, D. C. (1992). "Information Retrieval and the Philosophy of Language." THE COMPUTER JOURNAL 35, NO. 3(3): 8.
[28] Engelbrecht, A. P. (2007). Computational Intelligence: An Introduction. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Wiley Publishing.
[29] Stefano Ceri, A. B., Marco Brambilla, Emanuele Della Valle, Piero Fraternali and Silvia Quarteroni (2013). WEB INFORMATION RETRIEVAL. Springer Heidelberg New York Dordrecht London, Springer.