International Journal of Innovative Computing, Information and Control
Volume **12**, Number **2**, April **2016**

ICIC International ©2016 ISSN 1349-4198
pp. 519–531

ONTOLOGY-CONCEPTS WEIGHTING FOR ENHANCED SEMANTIC CLASSIFICATION OF DOCUMENTS

Salam Fraihat

Software Engineering Department
Faculty of Information Technology
Al-Ahliyya Amman University
P.O. Box 19328, Amman, Jordan
s.fraihat@ammanu.edu.jo

Received June 2015; revised December 2015

Abstract. *Automatic document classification has become increasingly important and difficult due to the large scale of the electronic documents used in the last years. Traditional information retrieval systems are based on the extraction of keywords from documents; these keywords serve as a basis for documents classification. This paper proposes a new semantic approach for documents classification. Specifically, our approach captures, in addition to the keywords frequency, the meaning of these keywords in documents using domain ontology. The main idea is to represent documents by concepts rather than keywords, and calculates weights for these concepts to reflect their importance in the documents where they appear. The presence of concepts in the same paragraph, section, document, or document set, provides important information to better extract and understand the semantic content of the document and therefore improves its classification. The experimental evaluation is carried out using the Reuters document collection RCV1-v2 and the GALEN medical ontology. The documents are classified using the SVM classifier. The experimental results demonstrate that the proposed approach yields higher accuracy, precision and recall compared to the traditional keyword-based information retrieval approaches.*

Keywords: Information retrieval, Documents classification, Domain ontology, Concept semantic weighting, Information extraction

1. Introduction.

Today, information is available in large quantity with varying quality. This complex information is irrelevant if there is no technology to access it effectively. For this, we need to develop systems allowing search, classify and analyze this information with minimum human involvement. One area that is trying to make improvements and reduce the human task is the information retrieval (IR) area. Traditional IR systems are based on a set of keywords extracted from document. These keywords constitute the set of features that are used to represent the documents. Traditional IR systems assign weights to keywords for each document to reflect the relative importance of keywords in the document. The performance of traditional IR systems is measured by their ability to classify relevant documents automatically using the extracted keywords. The evaluation programs of IR systems, such as TREC [1], confirm that IR systems show an interesting performance in documents classification when they are applied only on specific types of datasets. However, the main limitations of traditional IR systems are:

- All of the traditional IR systems are based on the hypothesis that the keywords are the best features to represent all the knowledge contained in the documents [2-6];

- The quality of the information retrieving process depends largely on the quality of keywords weighting approaches. Indeed, the keywords relevance weights are computed based only on the appearance or absence of the keywords in the document rather than the implicit semantic relations between the keywords [2].

In order to overcome these limitations, semantic IR systems based on ontology have arisen. In these systems, keywords denote concepts, optionally combined with representative concepts of the document semantic content. A concept can be defined as: *"Concepts, also known as classes, are used in a broad sense. They can be abstract or concrete, elementary or composite, real or fictious. In short, a concept can be anything about which something is said, and, therefore, could also be the description of a task, function, action, strategy, reasoning process, etc."* [7].

The ontology-based semantic IR systems use the ontology concepts and their semantic relationships, such as equivalence, synonymy, hyponymy and other types of relationships such as Is-a, Has-a, to represent the meanings that existed on documents [8].

This paper proposes a semantic IR system based on a new ontology-concept weighting approach. The purpose of concept weighting is to quantify the degree of importance of each concept in the document. The main idea of the proposed approach is based on the fact that the presence of concepts in the same paragraph, section, document, or document set, provides important information to better extract knowledge and understand the semantic content of the document, and therefore, improves its classification. Our ontology-concept weighting approach integrates the concept presence measure (named Intra-Concept Weight ACW) in the calculation process of the concepts connectivity measure (named Inter-Concept Weight ECW); this will enhance the representativity of concepts of a document.

The rest of this paper is organized as follows. In Section 2, a brief overview of the research background and related work is presented. Section 3 shows our proposed ontology-concepts weight approach for documents classification. Section 4 presents the experimental environment of the proposed approach using the RCV1-v2 document collection and the GALEN ontology and the SVM machine learning. Section 5 presents the experimental evaluation and finally, conclusions and directions for future study are provided in Section 6.

2. Background and Literature Review.

The weighting process is a crucial problem in traditional IR systems, because the quality of document classification mostly depends on keywords or concepts weighting approaches. Different weighting approaches are reported in literature, and they can be classified into traditional and semantic approaches.

2.1. Traditional weighting approaches.

In traditional weighting approaches, each keyword in a document must be associated with a value (weight). There are a large number of approaches to calculate the keyword weight which can be classified based on different criteria such as: theory functions, statistic metrics, relevant probability, and supervised/unsupervised weighting approaches [2]. We invite readers to refer to [3-5] for more details of such approaches.

The *tf-idf* is the most common weighting method used to represent documents in IR system [4]. The *tf-idf* is the product of two statistics, the keyword frequency *tf* and the inverse document frequency *idf*. The *tf-idf* is a numerical statistic that reflects how a keyword is important in a document and in a collection of documents [4]. There are different methods used to calculate the *tf-idf* weight such as: Boolean, Logarithm, and Augmented frequency [4]. These methods belong to the unsupervised keywords weighting methods (i.e., which take into consideration only the frequency of existing keywords in the document) [6]. On the other hand, supervised keyword weighting methods take into consideration the keywords distribution in the document collection, when calculating the keywords weights [2]. For example, on the supervised keyword weighting methods [9], Debole and Sebastiani in [5], propose a supervised term weighting method that is based on replacing the *idf* weight by the values of three feature selections (i.e., *chi-square* metrics, information gain, and gain ratio). Erenel et al. [10] propose another supervised keywords weighting method that is based on keyword occurrence probabilities on the documents.

To summarize, all traditional weighting methods are based mainly on the presence or absence, and presence frequency of the keywords in the documents. These methods exploit only the syntactical and lexical level of keywords to retrieve it, without exploring the semantic level and meaning of the keywords.

2.2. **Semantic weighting approaches.** In semantic weighting approaches, concepts are considered through the senses they represent. Hence, a concept weighting aims at evaluating the importance of the corresponding senses in the document's content. This importance is estimated through the number of semantic relations between one concept and the other concepts in a document.

In semantic IR systems, several concepts weighting approaches have been proposed [11-18]. In these approaches, the concepts weighting is based on both of the concepts of the documents, and their associated synonym set which is defined in the WordNet ontology. These approaches require a disambiguation process because the resources are non-specialist. In [19] Doen et al. consider the concepts identified from ontology are weighted using classical weighting method such as the *tf-idf*. Tar and Nyunt [14] propose a mixed method to calculate the ontology concept weight. They calculate the concept weight based on the frequency, length, specific area and score of the keywords which appear in document and are associated to an ontology concept. Other methods calculate the concept weight based on the number of concepts contained in a collection of documents [15,16]. Although they are different one from another, they all follow the same principle which is the calculating of distances between all pairs of concepts. The distance calculations between concepts are largely utilized to resolve the concept similarity matching problem. Elavarasi et al. [17] propose the calculation of the distance between concepts based on the number of relationships existing between them. The number of relationships can be calculated via different measures like: path length, depth relative or mixed measures. In [18] the distance between concepts is calculated not only based on the relationships between concepts, but also based on the level of concepts in the ontology hierarchy.

The main limitations of the semantic weighting approaches are [12]:

- The concept weighting process depends largely on the quality of the ontology hierarchy and the calculation of the distance between the concepts which represent the documents (i.e., the more concepts are appropriately hierarchized; the better is the concepts weighting process which enhances the quality of documents classification);
- The concept weighting process depends largely on the efficiency of the replacement method used to replace the keywords extracted from documents with the concepts of the domain ontology. The main problem is to find the most closed keyword to the concept, in some cases; a keyword may be replaced by a concept that has different syntax of the keyword.

Our proposal in this paper is to weight the concepts by taking into consideration the context where the concept is located. Specifically, we first exploit the domain ontology to calculate 1) the concept appearance frequency in the document, and 2) the correlation measure between the concept and the other concepts from their context in the document.

In the proposed weighting method, the concepts used to represent the documents are extracted from the GALEN ontology [20] which is used widely in the medical IR systems.

3. **Proposed Ontology-Concept Based Representation Approach for Documents Classification.** Taking into account that a concept is more representative of the document content than a keyword, the proposed weighting concept approach allows the exploiting of the semantic relationships between the existing concepts in the same document.

As shown in Figure 1, the calculation of weight $ECW_{m,i}$ of concept C_m in the document d_i, is based on three processes: 1) Keyword frequency weight $wk_{j,i}$; 2) Intra-concept weight in document $ACW_{m,i}$; and 3) Inter-concept weight in document $ECW_{m,i}$.

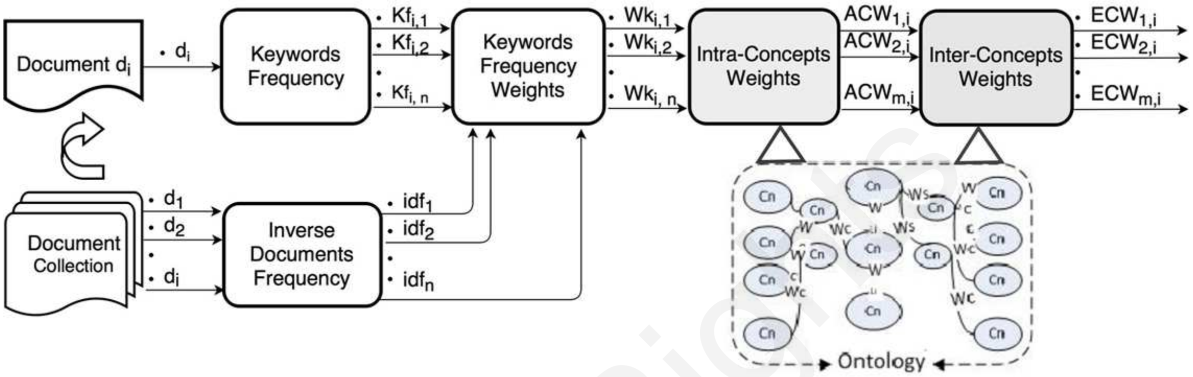

FIGURE 1. The weight calculation processes of concept in a document

3.1. **Keyword frequency weight (*wk*).** The document keywords weighting is a fundamental task in IR systems. It consists of measuring the importance of keyword k_j in a document d_i by assigning to it a weight $wk_{j,i}$ which expresses its degree of representativeness on the document.

As shown in Figure 1, the calculation of the keywords frequency weight $wk_{j,i}$ is based on the combination of two measures [4]:

- Keyword Frequency $Kf_{i,j}$: quantifying the importance of the keyword k_j in the document d_i;
- Inverse Document Frequency Idf_j: based on the idea that a keyword does not distinguish documents from each other if it is distributed in a uniform manner in all documents in the collection. In this case, a keyword has no discrimination power. Therefore, a keyword that appears in few documents are more discriminating and a weight is assigned to it. Idf_j quantifies the importance of a keyword k_j on a collection of documents.

Most weighting approaches used in IR are based on the combination of both $Kf_{i,j}$ and Idf_j where the keyword weight is defined by [4]:

$$wk_{j,i} = Kf_{i,j} \times Idf_j \tag{1}$$

In the RCV1-v2 documents collection, that is used in our experiments, the keyword weight is calculated using the Cornell keyword weighting, as defined by [21]:

$$wk_{j,i} = \underbrace{(1 + \log_e n(k_j, d_i))}_{Kf_{i,j}} \times \underbrace{\log_e(|D|/n(k_j))}_{Idf_j} \tag{2}$$

where,

$n(k_j, d_i)$ is the number of occurrences of keyword k_j in document d_i.

$n(k_j)$ is the number of documents that contains the keyword k_j.

$|D|$ is the number of documents used to calculate the inverse document frequency weights (Idf_j weights).

The $wk_{j,i}$ measure is a good approximation of the keyword importance in a collection of documents, especially for collections composed of homogeneous size documents. However, for collections that contain varying size documents, the keywords in the longest documents appear very frequently with very high weight compared with short documents. Thus, long documents will have more chance of being selected. The cosine normalization method [21] is used in RCV1-v2 to normalize the keyword weight $nwk_{j,i}$. The cosine normalization formula is as follows:

$$nwk_{j,i} = \frac{wk_{j,i}}{\sqrt{\sum_l wk_{l,i} \times wk_{l,i}}} \tag{3}$$

The weight $nwk_{j,i}$ of each keyword k_j is then used afterwards to calculate the weight of the concept where the keyword belongs (called Intra-Concept Weight ACW). In the next two sections, we calculate the intra and inter concept weights in order to reflect their coherence within all other concepts of the document in the document collection.

3.2. **Intra-Concept Weight measure (ACW).** The ACW can be defined as a static coefficient to measure the importance of concept in the document. The concept importance is measured by the calculation of the concept frequency in the document through the frequency of keywords that exist in a document and which are linked to this concept.

Figure 2 shows the class diagram of the ontology entities (keywords or instance, concepts, relationships). Each concept can be associated to one or to a set of keywords extracted from the documents collection. As can be seen, each keyword can be associated to any concept or to a set of concepts.

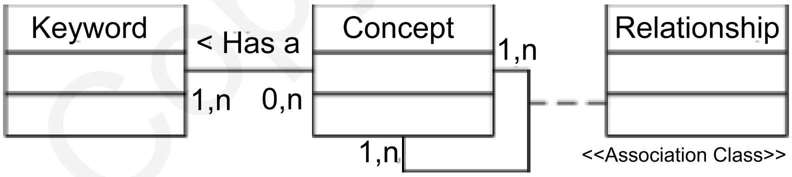

Figure 2. Ontology concepts relationships

However, to calculate the intra-concept weight, we use the ontology to extract the set of concepts which match with the set of keywords of the document (for more details about the concept/keyword matching method see Section 4.3), and then, for each concept C_m we assign an Intra-Concept Weight $ACW_{m,i}$, calculated based on all $nkw_{l,i}$ of the keywords k_j associated to the concept C_m (see Figure 3).

However, it is noted that a particular keyword can be polysemous, that means it can be associated with more than one concept in the ontology.

Algorithm 1 shows the $ACW_{m,i}$ calculation process. The output of this algorithm is a list of existing concepts C_m and their intra weight $ACW_{m,i}$ in each document d_i.

However, the $ACW_{m,i}$ is not enough to classify the documents. To improve the documents classification, we exploit the semantics of a document by the calculation of the inter-concept weight $ECW_{m,i}$ for each concept C_m in a document d_i. It is calculated using the relationships which exist between a concept C_m and other concepts that appear in the same document d_i (see Figure 3).

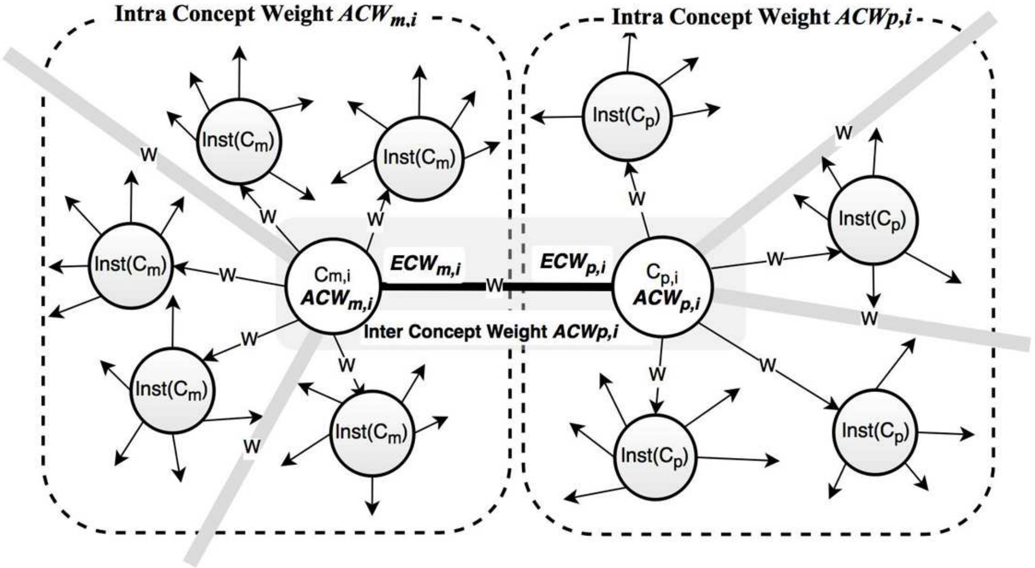
Intra Concept Weight $ACW_{m,i}$
Intra Concept Weight $ACWp,i$
Inst(C_m)
Inst(C_p)
W
$C_{m,i}$
$ACW_{m,i}$
$ECW_{m,i}$
$ECW_{p,i}$
$C_{p,i}$
$ACW_{p,i}$
Inter Concept Weight $ACWp,i$

FIGURE 3. Inter and Intra concept weighting

Algorithm 1: Intra-Concept weight $ACW_{m,i}$ calculation

```
$SetK_i\{\}$ = Set of keywords existing in the document $d_i$.
$OntoC\{\}$ = Set of Concepts and their instances on the Ontology.
$SetKC_{m,i}\{\}$ = Set of keywords associated to the Concept $C_m$ and existing in the document $d_i$.
For each concept $C_m \in OntoC\{\}$
    For each keyword $k_j \in SetK_i\{\}$
        If $k_j = C_m$ or $k_j \in Inst(C_m)$ then
                Add $k_j$ to $SetKC_{m,i}\{\}$
        end if
    end For
    $ACM_{m,i} = \frac{1}{N_m} \sum_{N_m}^{l=1} nkw_{l,i}$
end For
where: $N_m$: Number of keywords in the $SetKC_{m,i}\{\}$
       $Inst(C_m)$ = Instances set of concept $C_m$.
       $nwk_{l,i}$ = normalized Keyword $K_l$ Weight in the document $d_i$.
```

3.3. **Inter-Concept Weight measure (*ECW*).** The $ECW_{m,i}$ is defined by the distance between the concept C_m and the other concepts that appear in the same document d_i. This allows to:

1) The exploitation of the semantic information that is illustrated by the fact that these concepts appear in the same document;

2) The adjustment of the concept weight for better representation of the document.

Algorithm 2 shows the $ECW_{m,i}$ calculation process. The output of this algorithm is a list of concepts with their inter weight with other concepts in each document d_i.

However, it is noted that, in case there are different paths between two concepts, the algorithm retains the shortest one to ensure that the relationships between the concepts are well considered.

Algorithm 2: Inter-Concept weight $ECW_{m,i}$ calculation

```
For each $C_{m,i} \in SetC_d\{\}$
    For each $C_{p,i} \in SetC_d\{\}$
        If $C_{m,i} = C_{p,i}$ then
            CDist$(C_{m,i}, C_{p,i}) = 0$
        Else
            CDist$(C_{m,i}, C_{p,i}) =$ ShortestPath$(C_{m,i}, C_{p,i})$
        end If
    end For
$ECW_{m,i} = ACW_{m,j} * \left(1 + \frac{1}{\sum_{Nc}^{i=1} CDist\left(C_{i,d}, C_{j,d}\right)}\right)$
end For
where, CDist(): matrix of distance between each pair of concepts.
    ShortestPath$(C_{m,i}, C_{p,i})$: function that returns the number of edges in the
    shortest path that connects the two concepts $C_{m,i}$ and $C_{p,i}$.
    $ECW_{m,i}$: Inter-concept $C_{m,i}$ weight in the document $d_i$.
    $Nc$: Number of Concepts in the document.
    $SetC_{d,i}\{\}$ = Set of Concepts existing in the document $d_i$.
```

The proposed weighting approach allows further weighting of concepts, the weighting of isolated concepts that have not any relationships with the other concepts.

4. **Experimental Environment.** The experimental environment includes the domain ontology used to extract the concepts that are representing the documents, a description and characteristics of the documents collection used for performing our experiments, and a description of the machine learning method used to generate our classification model.

4.1. **The domain ontology.** We use ontology to extract the concepts that represent the documents. The construction of a new ontology is expensive in terms of time and requires heavy design and construction efforts. Therefore, we decided to use an already well-established ontology. A thousand of domain and application ontologies are reported in literature, but none of them can represent the topics covered by the RCV1-v2 data collection documents (see Section 4.2). Thus, we used the GALEN medical ontology [20] for many reasons summarized as follows:

- Open: it is an open source medical ontology. It can be downloaded with sources and documentation for free [20].
- Language: it is written in several formal languages as GRAIL (GALEN Representation and Integration Language) and also distributed in OWL (Web Ontology Language).
- Usability: it is reusable for a wider range of applications and semantic medical information systems.

GALEN is a large ontology for human anatomy, pathophysiology, function, surgical procedures, diseases and drugs, contains about 23141 concepts organized in multiple hierarchies relationship "*is-a*" and other 25 relationships [22]. Figure 4 shows a part of GALEN ontology hierarchy.

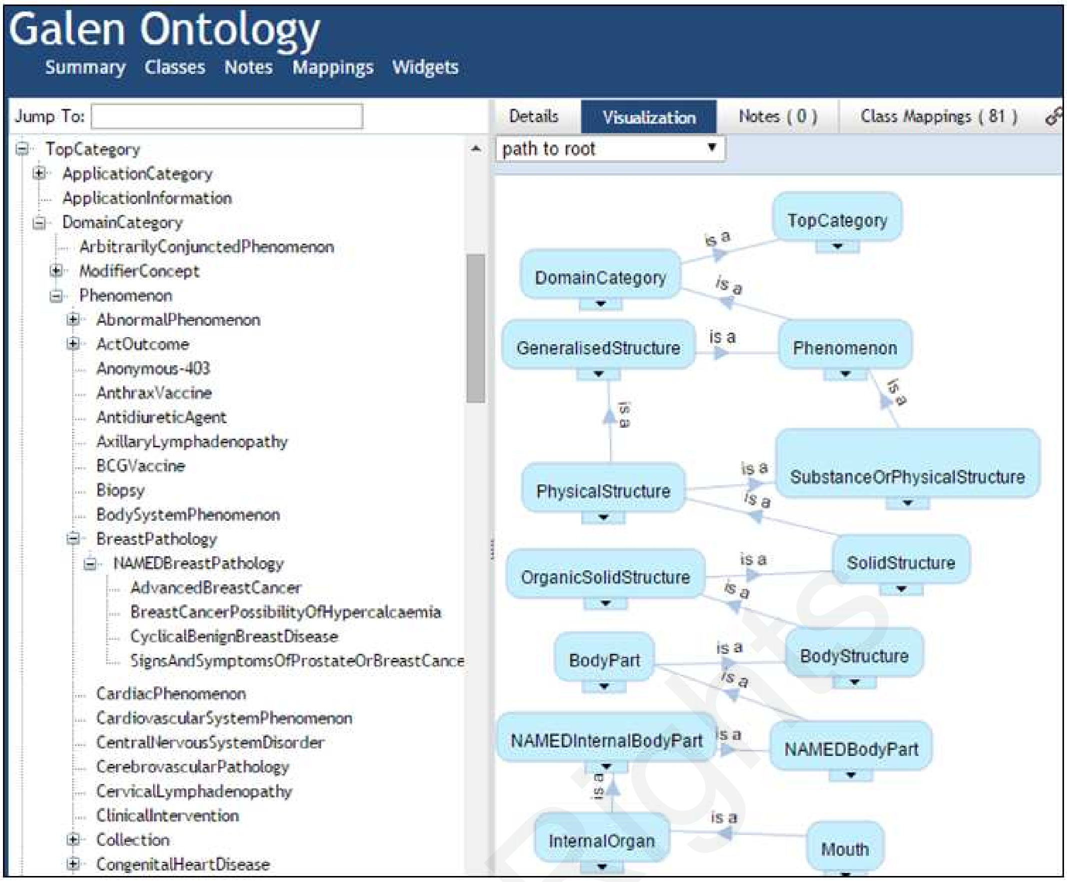

FIGURE 4. A part of the concepts hierarchy (on the left), and example of the path from the concept mouth to the concept TopCategory via is-a relationship [23] (on the right)

4.2. **Document collection.** To perform our experiments, we use the well-known benchmark RCV1-v2 Reuters Corpus [21] as it is currently the most widely used collection for text classification research. This collection has a set of documents represented as vectors. The LYRL2004 partition with 23149 training, and 781265 testing vectors, was used. In our experiment we use the following.

- The keywords list extracted from the RCV1-v2 document collection. The keywords list was extracted after preprocessing the documents texts, in which lower case characters reduction, tokenization, punctuation removal and stemming, stop word removal, keyword weighting and length normalization were applied. The list contains 47,236 keywords. This list contains, for each keyword, an inverse document frequency weight Idf_i (see Section 3.1) calculated on the whole documents collection.
- The hierarchy list of topics (named here classes), which contains the 103 RCV1 classes organized by categories using the relationship parent/child. For example, class-parent: Government/Social, and class-child: Defense, Health, Art, etc.
- The list of classes and the document which belongs to. However, it is noted that one document can belong to more than one class.

In our experiments, 3000 documents have been selected by randomly extracting 300 documents from each class of ten selected classes distributed on four categories of the RCV1-v2 collection.

Table 1 shows the ten selected classes from the categories GCAT, ECAT, CCAT and MCAT.

TABLE 1. Classes and categories of documents using in the Train and Test

Class Label	Class Description	Category Description
GHEA	Health	GCAT: GOVERNMENT/SOCIAL
GWEA	Weather	
GSPO	Sports	
GPOL	Domestic politics	
E12	Monetary/economic	ECAT: ECONOMICS
E14	Consumer Finance	
C12	Legal/Judicial	CCAT: CORPORATE/INDUSTRIAL
C17	Funding/Capital	
M11	Equity Markets	MCAT: MARKETS
M14	Commodity Markets	

The choice of classes is made to demonstrate the classification of documents compared to the different classes and also compared to the different categories. One tier of these documents is used as a training dataset and two tiers as a testing dataset.

The GALEN medical ontology is used to improve the classification of documents in the GHEA class. Firstly, several experiments were carried out to classify the documents that belong to the GHEA class compared to other classes using the keywords. Secondly, only for the documents of the GHEA class, we replace the keywords by their related concepts which have been extracted from the GALEN ontology. To resolve the syntax matching problem between the keywords of documents of the GHEA class and the GALEN concepts ontology, we applied a stamping and normalization method on the keywords [24].

4.3. **Keyword and concept matching.** To validate our approach, each keyword on the documents was replaced by the concept which is related to. In this process, we use a matching method using some rules to treat the following cases.

1. The whole keyword label matches the whole concept label.
2. A part of keyword label matches a whole concept label.
3. A whole of the keyword label matches a part of concept label.

In the cases 1 to 3, we replace the keyword by the concept to represent the document and we use the keyword weight $nwk_{l,i}$ in the calculation of the concept weight $ACW_{m,i}$ (for more details see Intra-Concept Weight $ACW_{m,i}$ calculation in Section 3.2).

4. A part of keyword label matches a part of concept label. In this case, **if** the length of the matching part is greater than or equal to $\lambda\%$ of the concept or keyword label length **then** we process it as the above cases; **otherwise**, we do not replace the keyword by the concept. The $\lambda\%$ represents a tolerance percentage in the matching process. It is fixed according to the matching experiments.

In the cases 2 to 4, **if** the keyword label matches different concepts **then** we replace the keyword with the lower level concept in GALEN ontology. This will assure a high quality distance calculation process between the concepts.

5. **If** the keyword label does not match partially or fully any concept **then** it will be used with its weight directly in the classification process.

For example, we consider the keywords list {Heart, Heartbeat, Heartbreak, Heartburn, Heartland, and Heartrend} which has been extracted from Reuters documents. These keywords will be replaced in the keywords list of a document, where they appear by the corresponding concept "Heart" existing in the GALEN ontology. Also, the weight $wk_{n,i}$ of each keyword in the list will be used to calculate the weight $ECW_{m,i}$ of the concept "Heart".

Once we set the concepts that represent each document and calculate the weights of the concepts existing in the document, we use the support vector machine to classify the documents collection.

4.4. **Machine learning.** Support vector machine (SVM) [25] is one of the best classification algorithms used in the document classification area. SVM is a binary supervised learning classification method. It is based on the use of kernel function that allows optimum separation of the data called hyperplane. The SVM algorithm is originally a mono-class algorithm for determining if an element (qualified positive) or not (qualified negative) to a class. To resolve the multi-class classification problem, the mono-class classifier is merged, where, N classifiers are trained taking one class at a time as positive, and grouping the rest under a negative label. After training, a new point is assigned to that class for which the largest positive output is computed. This method uses a training data set to learn the model parameters used to classify a testing dataset. SVM Models can be classified to linearly separable cases and non-linearly separable cases. The simplest example of kernel function is the linear kernel. However, the SVM with polynomial kernel proved to be the best in the literature in documents classification [25]. Thus, it is used in this study to classify the selected documents in the ten RCV1-v2 classes.

4.5. **Performance measures.** In order to evaluate our semantic approach and compare it with the traditional approaches in documents classification, we used the most common performance measures in IR literature: Accuracy, Recall and Precision [26].
Accuracy measure. The accuracy measures the ability of the classifier to classify all documents that belong to class C and the other documents to the other classes. It can be calculated by the ratio of correct predictions to the total number of cases [26].

Table 2. Table of documents affiliation between classifier prediction and reality

	Document belongs to Class C	Document belongs to other Classes
Prediction in Class C	*TP*	*FN*
Prediction in other Classes	*FP*	*TN*

In Table 2, *TP* means true positives, *TN* true negatives, *FP* false positives, and *FN* false negatives.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Recall measure. The recall measures the ability of the classifier to classify all documents that belong to class C correctly. It is defined as the probability that a document belongs to class C or in other classes is correct [26].

$$recall = \frac{TP}{TP + FN} \tag{5}$$

Precision measure. The precision measures the ability of the classifier to not classify documents that belong to other classes as documents of the class C. It is defined as the probability that a document belongs to class C is classified correctly [26].

$$precision = \frac{TP}{TP + FP} \tag{6}$$

5. **Experimental Evaluation.** The main purpose of our concept weighting approach is to improve the documents classification. The performance of the documents classification methods is measured by the ability to increase the accuracy, precision and recall measures.

Tables 3 and 4 show the evaluation results of our semantic documents classification approach using ten classes belonging to four categories of the RCV1-v2 corpus. Accuracy, recall and precision measures are calculated for each class of the ten classes using a traditional "keyword-based" document classification approach, and for the GHEA class using our semantic document classification approach.

Table 3 shows that the results of the classification accuracy, of the class which belongs to the same category, are almost close together. Also, we note that the classification results of documents that belong to the category GCAT are relatively weak when compared to other categories. This is due to the number of classes in each category, for example, there are four classes in the GCAT category which means that documents that belong to the four classes have many keywords in common. Accordingly, this makes the task of the classifier more difficult. In other categories which contain two classes, this phenomenon is less apparent, as more as classes existing in a category, the more hard is the classification process, and vice versa.

Table 3. Comparison of classification accuracy, precision and recall rates when using keywords to categorize documents of ten selected categories of Reuters Corpus

Category/Class Label		Traditional approach "keyword-based"		
		Accuracy (%)	Recall (%)	Precision (%)
GCAT	GHEA	73.34	71.03	74.15
	GWEA	69.72	68.21	71.30
	GSPO	77.46	75.4	81.72
	GPOL	79.50	77.4	80.10
ECAT	E12	83.22	81.20	85.44
	E14	89.50	87.4	93.40
CCAT	C12	91.80	90.1	92.93
	C17	89.70	87.3	91.60
MCAT	M11	87.50	80.40	90.10
	M14	90.70	89.20	91.80

Table 4. Comparison of classification accuracy, precision and recall rates when using traditional versus semantic approach to categorize documents on Health category and the other nine selected categories of Reuter Corpus

		GHEA	**Other classes**
Accuracy (%)	Traditional approach "keyword-based"	73.34	84.35
	Semantic approach "concept-based"	94.80	–
	Gain	21.46	–
Recall (%)	Traditional approach "keyword-based"	71.03	81.85
	Semantic approach "concept-based"	90.55	–
	Gain	19.52	–
Precision (%)	Traditional approach "keyword-based"	74.15	86.48
	Semantic approach "concept-based"	98.20	–
	Gain	24.05	–

Table 4 shows an improvement in the accuracy of documents classification based on concepts (Semantic approach) then based on keywords (Traditional approach).

The use of the GALEN ontology improves the quality of identification and classification of documents (the gain in the accuracy, recall and precision measures of 21.7%, 19.52% and 24.05% respectively). This could be explained by the fact that the concepts are better representative of documents than simple keywords.

Thereby, the weight of concepts calculated in our approach allows a more accurate understanding of the documents meanings, which enhances their classification.

6. **Conclusion and Future Works.** In this paper, we have proposed a semantic approach for documents classification, which is one of the main problems of information retrieval domain. Based on the idea that the presence of concept or concepts in the same paragraph, section, document, or document set, provides important information to better extract knowledge and understand the semantic content of the document, and therefore, improves its classification, we have proposed the use of domain ontology to classify documents. In our approach, we have replaced the traditional keywords by the concepts and their relationships to capture some aspect of the semantic meanings existing in the documents. The concept weight calculation process takes into account: 1) the concepts apparent importance in the documents (measured through the frequency of keywords which are in relation with the concept), and 2) the concepts meaning importance in the documents (measured through its semantic relationships with other concepts). The proposed approach has been evaluated on the RCV1-v2 Reuters collection by considering a subset of documents distributed on ten classes, which are categorized in four categories. The evaluation results showed that our semantic approach for documents classification performs better than the traditional keyword-based documents classification.

Future research will focus on testing our approach on all RCV1-v2 document collection using other ontologies for other areas. Also, we aim to test the impact of relationships types (e.g., has-a, is-a, and part-of) that connect the concepts on ability of our approach to enhance documents classification.

REFERENCES

[1] E. M. Voorhees and L. P. Buckland, The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings, *NIST Special Publication*, 2012.

[2] M. Lan, C. L. Tan, J. Su et al., Supervised and traditional term weighting methods for automatic text categorization, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.31, no.4, pp.721-735, 2009.

[3] G. Chowdhury, *Introduction to Modern Information Retrieval*, Facet Publishing, 2010.

[4] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

[5] F. Debole and F. Sebastiani, Supervised term weighting for automated text categorization, *Text Mining and Its Applications*, pp.81-97, 2004.

[6] M. Motwani and A. Tiwari, Comparative study and analysis of supervised and unsupervised term weighting methods on text classification, *International Journal of Computer Applications*, 2013.

[7] Ó. Corcho and A. Gómez-Pérez, A roadmap to ontology specification languages, *Proc. of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, pp.80-96, 2000.

[8] E. Mena and A. Illarramendi, *Ontology-Based Query Processing for Global Information Systems*, Springer Science & Business Media, 2012.

[9] C. C. Aggarwal and C. Zhai, A survey of text classification algorithms, *Mining Text Data*, pp.163-222, 2012.

[10] Z. Erenel, H. Altincay and E. Varoglu, A symmetric term weighting scheme for text categorization based on term occurrence probabilities, *Proc. of the 5th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, pp.1-4, 2009.

[11] C. Fellbaum, *WordNet*, Wiley Online Library, 1998.

[12] M. Fernández, I. Cantador, V. López et al., Semantically enhanced information retrieval: An ontology-based approach, *Web Semantics: Science, Services and Agents on the World Wide Web*, vol.9, no.4, pp.434-452, 2011.

[13] M. C. Lintean, C. Moldovan, V. Rus et al., The role of local and global weighting in assessing the semantic similarity of texts using latent semantic analysis, *Proc. of FLAIRS Conference*, 2010.

[14] H. H. Tar and T. T. S. Nyunt, Ontology-based concept weighting for text documents, *World Academy of Science, Engineering and Technology*, vol.81, pp.249-253, 2011.

[15] R. Arun, V. Suresh, C. V. Madhavan et al., On finding the natural number of topics with latent Dirichlet allocation: Some observations, *Advances in Knowledge Discovery and Data Mining*, pp.391-402, 2010.

[16] J. Cao, T. Xia, J. Li et al., A density-based method for adaptive LDA model selection, *Neurocomputing*, vol.72, no.7, pp.1775-1781, 2009.

[17] T. Slimani, Description and evaluation of semantic similarity measures approaches, *arXiv preprint arXiv:1310.8059*, 2013.

[18] J. Ge and Y. Qiu, Concept similarity matching based on semantic distance, *Proc. of the 4th International Conference on Semantics, Knowledge and Grid*, pp.380-383, 2008.

[19] Y. Doen, M. Murata, R. Otake et al., Construction of concept network from large numbers of texts for information examination using TF-IDF and deletion of unrelated words, *Proc. of the 15th International Symposium on Soft Computing and Intelligent Systems*, pp.1108-1113, 2014.

[20] *OpenGALEN*, http://www.opengalen.org/.

[21] D. D. Lewis, Y. Yang, T. G. Rose et al., RCV1: A new benchmark collection for text categorization research, *J. Mach. Learn. Res.*, vol.5, pp.361-397, 2004.

[22] A. Shukla and R. Tiwari, *Intelligent Medical Technologies and Biomedical Engineering*, IGI Global, 2010.

[23] P. L. Whetzel, N. F. Noy, N. H. Shah et al., BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, *Nucleic Acids Research*, vol.39, no.suppl 2, pp.W541-W545, 2011.

[24] S. Karbasi and M. Boughanem, Document length normalization using effective level of term frequency in large collections, *Advances in Information Retrieval*, pp.72-83, 2006.

[25] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[26] Wikipedia, *Precision and Recall*, http://en.wikipedia.org/wiki/Precision_and_recall.