SGF (Semantic Graphs Fusion): A Knowledge-based Representation of Textual Resources for Text Mining Applications

Morteza Jaderyan
Department of Computer Engineering, Bu Ali Sina Uinversity, Hamedan, Iran
m.jaderyan92@basu.ac.ir
Hassan Khotanlou*
Department of Computer Engineering, Bu Ali Sina Uinversity, Hamedan, Iran
khotanlou@basu.ac.ir

Received: 05/May/2019 Revised: 29/Sep/2019 Accepted: 29/Nov/2019

Abstract
The proper representation of textual documents has been the greatest challenge in text mining applications. In this paper, a knowledge-based representation model for text analysis applications is introduced. The proposed functionalities of the system are achieved by integrating structured knowledge in the core components of the system. The semantic, lexical, syntactical and structural features are identified by the pre-processing module. The enrichment module is introduced to identify contextually similar concepts and concept maps for improving the representation. The information content of documents and the enriched contents are then fused (merged) into the graphical structure of a semantic network to form a unified and comprehensive representation of documents. The 20Newsgroup and Reuters-21578 datasets are used for evaluation. The evaluation results suggest that the proposed method exhibits a high level of accuracy, recall and precision. The results also indicate that even when a small portion of the information content is available, the proposed method performs well in standard text mining applications.

Keywords: Semantic document representation; Ontology; Knowledge base (KB); Semantic network; Information fusion.

1- Introduction

The text mining techniques are heavily dependent on the generated representation of text documents; their performance is highly affected by it. In most text mining applications and techniques, a specific model is utilized to represent the information content. Most text mining systems employ simple representation models such as "Bag-of-Words" model to represent the contents. These models combined with an "exact term matching" method are used for retrieving the most relevant information to user preferences. However, such representation models suffer from serious drawbacks and limitations which are documented in [1, 2, 3]. Some of the most serious drawbacks and limitations of these systems are: (1) the inherent ambiguity of natural language, (2) synonymy and (3) polysemy. One solution is the integration of ontology and KBs and using knowledge-based representation models [4, 5]. These methods employ the structured knowledge of ontologies and knowledge bases (KBs) to overcome the ambiguity, to represent the content, to model the semantics and to develop text mining applications.

One of the most important aspects of semantic document representation is to introduce a mechanism for representing the contents, the semantics and also efficiently utilizing them in the intended applications. In this regard, filtering the relevant features and ignoring the irrelevant ones will be the main challenge. Introducing a semantic and knowledge-based document representation model using the graphical structure of semantic networks is the main idea of this paper. The semantic network representation generally consists of a number of interconnected nodes. The connecting links represent the semantic relations between the concepts.

The main contributions of this paper are: integrating the structured knowledge of ontology and KBs in every component of the proposed representation model, using semantic network for representing the contents of documents and the user preferences, introducing a knowledge-based approach for content enrichment and merging the document semantic networks with enriched concept maps to create a comprehensive representation of contents. Therefore, the idea presented in this paper can be summarized as follows. The semantic, lexical, structural and syntactical features of the documents are identified and extracted and the concepts are weighted. The content enrichment module identifies and extracts the concepts and semantic structures. In the next step, the semantic relations are established between concepts using the structured knowledge of ontology and KBs. Then, the identified concepts, semantic structures and the semantic relations between them are represented in a graphical structure of semantic networks. In the end, the concepts and the identified semantic structures are merged with semantic

* Corresponding Author

networks. It can be used in a variety of applications such as information retrieval, indexing, recommender system and information filtering and management.

The rest of the paper is structured as follows: In the second section, the related works are studied. In the third section, the structure of ontology, Wikipedia and WordNet is examined. In the fourth section, the proposed method of knowledge-based document representation is introduced. In the fifth section, the evaluation results are presented. In the sixth section, we will have the discussion and in the seventh section, the conclusion is presented.

2- Related works

In most text mining applications, the semantic and comprehensive representation of documents is the factor that guarantees the optimal performance and effectiveness of the implemented system. The information models determine how the documents should be represented. In this regard, text mining techniques can be classified into three categories: (1) techniques that employ information models, (2) techniques that employ intelligent learning models and (3) techniques that exploit the structured knowledge of ontology and KBs to represent the content. The probabilistic and vector space models (VSMs) are among the most popular and widely used information models for document representation [6]. The language models [7] and the Bayesian network models [8] are considered to be probabilistic models. They use the probability and statistics principles to generate information models. Whereas, the vector space models [9] utilize a vector form for representing the content of documents. In [10], the authors address the problem of text classification by considering Sentence-Vector Space Model (S-VSM) and Unigram representation models. A neural network based representation is then used to capture the semantic information.

Most VSMs-based information models are based on Salton et al.'s researches [9]. In recent years, there has been several studies [1, 4, 11, 12, 13], exploring the idea of employing the structured knowledge of ontology and KBs for constructing semantic information models. These models exploit the structured knowledge of ontologies and KBs to represent the content of documents. In these studies, using the structured knowledge, for creating information models, has shown promising results in this field. The Natural Language Processing (NLP) -based [14], rule-based [15], ontology-based [16] and fuzzy-based [17] are among the knowledge-based content representation models.

In [11, 13], using ontologies and KBs in semantic information indexing and retrieval is studied. In these studies, semantic networks are used for representing the information content. In [12], a personalized method for document search and retrieval is introduced. In this paper, the documents are represented by mapping the concepts to a graph-like structure. The relations between concepts are established using a web-based ontology called ODP [18]. In [19], a method for document indexing in engineering domain is introduced. A domain ontology is employed to represent the content of documents in the form of semantic networks. The Wikipedia is also used for representing content in text mining applications. In [20], the authors introduce a Wikipedia semantic matching approach for text document classification. In order to model the text semantics, documents are represented as concept vectors in Wikipedia semantic space. In [21], the authors introduce a two-level representation model (2RM) for representing text data. At the syntactic level, a document is represented as a term vector (tf-idf) and the Wikipedia concepts, related to the identified terms in syntactic level, are used to represent the document in semantic level, and a multi-layer classification framework (MLCLA) is then used to generate the output.

In [22], a graph-based feature extraction is used to extract meaningful features. The documents are represented as graphs and a weighted graph mining algorithm are applied to extract frequent sub-graphs. The sub-graphs are then further processed to produce feature vectors for classification.

Machine learning techniques can be also used for document representation. The authors in [23] propose the bag-of-concepts method as a document representation method. The proposed method creates concepts through clustering word vectors generated from word2vec. It uses the frequencies of these concept clusters to represent document vectors. Discourse analysis is a collection of Natural Language Processing tasks that are designed to identify linguistic structures and contextual information from textual resources. The extracted linguistic structures are identified at different levels, so that they can be utilized to implement NLP applications such as text analysis, Question Answering and text summarization. One of the most important discourse analysis systems is discourse parser system that is used to represent the structure of a document by a tree-based structure. The key similarities and differences between our approach and the concept of discourse analysis are:

- Both approaches build a tree-based representation of textual resources which are used to capture the semantics and the relations between linguistic elements.
- The resulting representation from a discourse parser is based on a tree-like structure. However, the proposed representation model exploits the graphical structure of semantic networks.
- A discourse analysis system is designed to create a formal representation of linguistic context. On the other hand, the proposed approach exploits the structured knowledge of ontology and KBs to compute and model the conceptual relations between extracted features.
- The discourse analysis uses rhetorical relations such as contrast, explanation and cause to define the semantics. However, the proposed representation model exploits the ontology-based relations to represent the textual resources.

In recent years, there is a growing interest in integrating model-based, learning-based and Knowledge-based approaches for document representation [24]. In [25], a novel framework for incorporating knowledge bases (KBs) into the neural network is introduced. In this method, a raw text is conceptualized and represented by a set of concepts using a knowledge base. The neural network is then used to transform the conceptualized text into a vector, in which both the semantics and the content information are encoded.
In most text mining applications, the ontology and KBs are used either to compute the similarity of documents or to represent the content of documents. However, in this paper, the structured knowledge of ontology and KBs are integrated with core components of the proposed framework (pre-processing, enrichment and representation). Also, most similar approaches rely solely on the content of a document to identify most similar documents to user preferences. In this paper, the ontology and KBs are used to infer contextually similar concepts and semantic structures.

3- The Structure of Ontology and KBs

One of the most important features of the proposed method is the integration of ontologies and knowledge bases in every component of the method. Therefore, it is necessary to examine their features and information structures.

3-1- OntoWordNet Top-Level Ontology

Each concept in the ontology is organized as synonym set so that the contextually similar concepts can be identified. This facilitates content enrichment [26]. The classes are organized in the form of a sequence. This sequence defines synonym concepts that bear similar meaning in different contexts. OntoWordNet defines three important semantic relations: Superclass, Subclass, Synonymy and Part_of relations.

3-2- WordNet

WordNet[27] models a semantically enhanced lexicon for English language and consists of synsets. The synset organizes a set of synonym concepts. Every synset consists of several senses (different meanings of a concept).

3-3- Wikipedia

Wikipedia data are available for academic use through D.I.S.C.O project [28]. The Wikipedia consists of two sets of data [28, 29]: first-order word vector: which contains words that occur together in Wikipedia and second-order word vector: which contains words that occur in similar contexts.

4-The proposed Method

The proposed method integrates the structured knowledge of KBs into the document representation model. Incorporating the extracted semantics and informational structures into the representation model is the main idea of this paper. Such a representation model brings three important benefits: It exploits extracted information content and the enriched semantics to create a comprehensive representation model, It can be used as a multi-purpose information model in a variety of text mining applications and facilitates the process of matching documents to user preferences. Figure 1, shows the overview of the proposed method. It consists of three modules: the semantic document processing, the content enrichment and the semantic representation module.

4-1- The Semantic Document Processing Module

It performs a number of pre-processing operations to extract four types of features (semantic, morphological, syntactical and structural). This is the first step toward constructing a multi-level representation of documents. The following pre-processing operations are performed: stop-word removal, bi-gram and Uni-gram processing, Part of Speech (POS) tagging [30], lemmatization [31], named-entity recognition [32, 33] and shallow parsing of sentences [34, 35]. Each operation is designed to extract specific type of features. The features are then weighted using the CF-IDF [36]. Let $D = \{d_1, d_2, \dots\dots, d_n\}$ be the set of documents and $d_i = \{t_i^1,\ t_i^2,\ t, \dots\dots, t_i^n\}$ be the document vector d_i, after the weighting method is applied, $w_i = \{w_i^1,\ w_i^2,\ w_i^3, \dots\dots, w_i^n\}$ is the set of weights assigned to each member of d_i.

4-2- The Content Enrichment Process

The enrichment module enables system to infer useful knowledge and extract informative features from a set of concepts. This component exploits the structured knowledge of ontology and Wikipedia to discover additional informative features that might have been left out. This module can also be used to find concepts that can improve the information content of a given document.

4-2-1- Enrichment by OntoWordNet Ontology:

The notion of "concept map" graph is employed for enriching the content of documents. Each concept is mapped to a class in the OntoWordNet ontology. The concept and the corresponding class are then converted to the concept map. The concept maps are used to annotate the corresponding concepts in a semantic network. At first, for each concept, the corresponding classes of the ontology are extracted. Considering the structure of the ontology, a concept map consists of a concept and a set of corresponding classes. The links between the concept and the ontology classes are the "equivalent" property and the "subclass" relation. The concept maps are represented by a sub-ontology using OWL/XML schema. Such a representation would allow us to merge the generated concept maps with document semantic networks (see section 4.4). An example

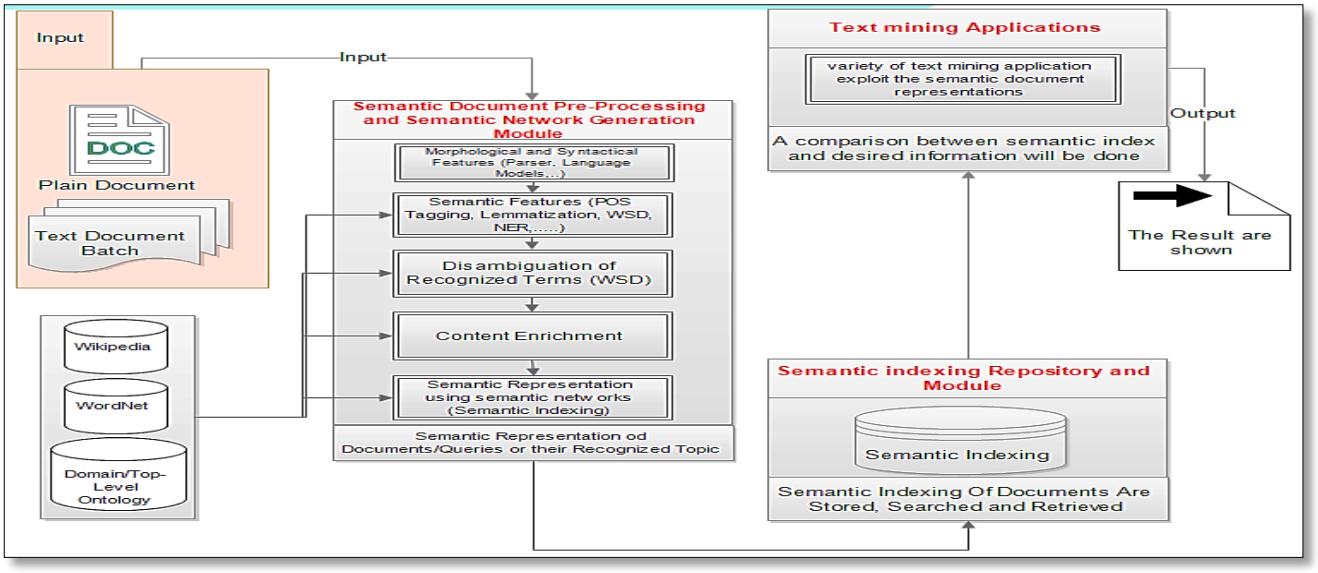

Figure 1- Overview of the proposed knowledge-based representation method

of a concept map for the concept "news story" is illustrated in Figure 2. The concept maps help the system discover commonalities between document semantic networks and user preferences. Also, the semantic structure of concept maps is vital in constructing a multi-level representation of documents. The superclass and the equivalent concepts are then weighted and appended to the document vector.

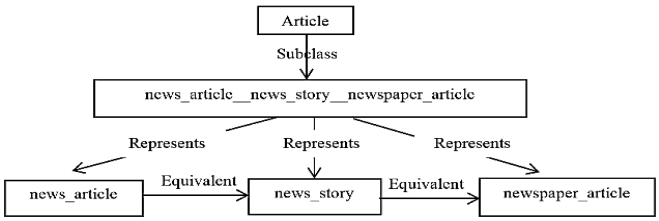

Figure 2- A representation of a generated Conceptual Map

4-2-2- Enrichment Using Wikipedia KB:

In this paper, the Wikipedia second-order word vector is used to enrich the document content. For each concept in the concept vector, co-occurring and contextually similar concepts are retrieved and appended to document vector.
The enriched concepts (contextually similar, co-occurring, superclass and equivalent concepts) are weighted and appended to the document vector. Since the new concepts are inferred indirectly from ontology and KBs, their assigned weight will be lower. The weighting equation is estimated using a subset of evaluation data.

$$Weight_{Related\ Conceot(ontology\ and\ Wiki)} = Weight_{Initial\ Concept} * 0.8 \quad (1)$$

Let $ec_i = \{t_i^{e1}, t_i^{e2}, t_i^{e3}, \dots\dots, t_i^{en}\}$ be the set of enriched words/concepts for document d_i, then after appending the enriched words/concepts to the original document vector, $d_i = \{t_i^1, t_i^2, t_i^3, \dots\dots, t_i^n, t_i^{e1}, t_i^{e2}, t_i^{e3}, \dots\dots, t_i^{en}\}$ is the extended document vector.

4-3- The Word Sense Disambiguation of Concepts

Semantic network representation of documents depends on accurate modelling of semantics and relations between features. In this regard, all features need to be cleared of ambiguity. In order to handle the word ambiguity issue in text documents, a method of word sense Discrimination, inspired by [37], is introduced. The underlying assumption of this method is that similar senses occur in similar contexts. In other words, by comparing the collective contextual features of a concept with the information content of each possible sense, we can induce its true contextual meaning. This method relies on the structured knowledge of Wikipedia and WordNet. To this end, the following procedures are performed:
(1) A ± 7 context window around the desired concepts in the message is created. Also, the first-order word vector for each member of the context window is retrieved and appended to context window. The window and the appending vectors create a "context vector" for each concept, (2) all possible senses of the concept, their usage example in a sentence and their brief definition for each sense is extracted from WordNet. This will form a "sense vector" for each sense. The first-order word vector for each member of a sense vector is also retrieved and appended to the corresponding sense vector. Finding the similarity between each sense and the context vector determines the contextual similarity between them and (3) a combination of cosine [38] and Jaro-Winkler [38] measures are used to calculate the similarity score as follows.

$$Sim(Sense_{Vctor}, context_{Vector}) = \frac{1}{2}(Cosine_{Sim}(Sense_{Vctor}, context_{Vector}) + Jaro\_winkler_{Sim}(Sense_{Vctor}, context_{Vector})) \quad (2)$$

The sense vector with highest similarity score is selected as the correct sense vector and the corresponding sense is used to annotate the concept. The output is a set of weighted concepts that are annotated by their true contextual meaning.

4-4- Semantic Document Representation Module

Various models have been proposed to represent the information content of textual resources. Such models include machine learning-based models such as Word embedding model, vector space models, and models based on ontology and structured knowledge bases [11, 4, 16]. One of the most important issues with vector space models and Word embedding models is their inability to model meaning in the information content of documents. On the other hand, the proposed method, which is based on the structured knowledge of the ontology, enables the system to identify semantic relations between words, extract the latent semantics of documents, and ultimately, map the extracted information structures and inferred background knowledge into a semantic network, without losing the semantics in the process. Therefore, the proposed model is a far better choice than vector space and machine learning models. Moreover, the ontology and structured knowledge bases provide useful information such as semantic relations between concepts, vectors representing co-occurring and contextual similarity relations between words/concepts; which makes them the perfect choice for content modelling.

In this paper, semantic networks are used for document representation. The underlying assumption about the representation model is that the information content would be better represented by a percentage of concepts rather than all the concepts. The CF-IDF weighting method is used to determine what percentage of concepts are optimal. Ontology-defined relations are then used to link the concepts in the graphical structure of semantic network.

The enriched concepts and semantic structures play an important role in creating a fully-connected semantic networks. The semantic network generation process consists of two phases: selecting the top n% concepts and generating the semantic networks by linking the concepts.

4-4-1- The Semantic Network Generation Process:

The first step toward creating a semantic network representation of documents is to establish relations between concepts. To this end, the top-n% of concepts are projected onto OntoWordNet ontology. A number of separated concept clusters are then formed. The main reason for this phenomenon is that concepts, which can link the separated cluster, are not identified or they are left out. In summary, the semantic network generation process is carried out as follows: (1) the extracted features and enriched content are weighted using the CF-IDF method and the top-and% of concepts are selected, (2) the proposed semantic network generation algorithm links the concepts together one by one using ontology-defined semantic relations and (3) the liaison concepts connect the separated concept clusters. Figure 3 illustrates how semantic networklinks the concepts and how the liaison concepts link the separated concept clusters. Also, Figure 4 illustrates the proposed semantic network generation algorithm. As shown in Figure 3, after projecting the concepts onto the

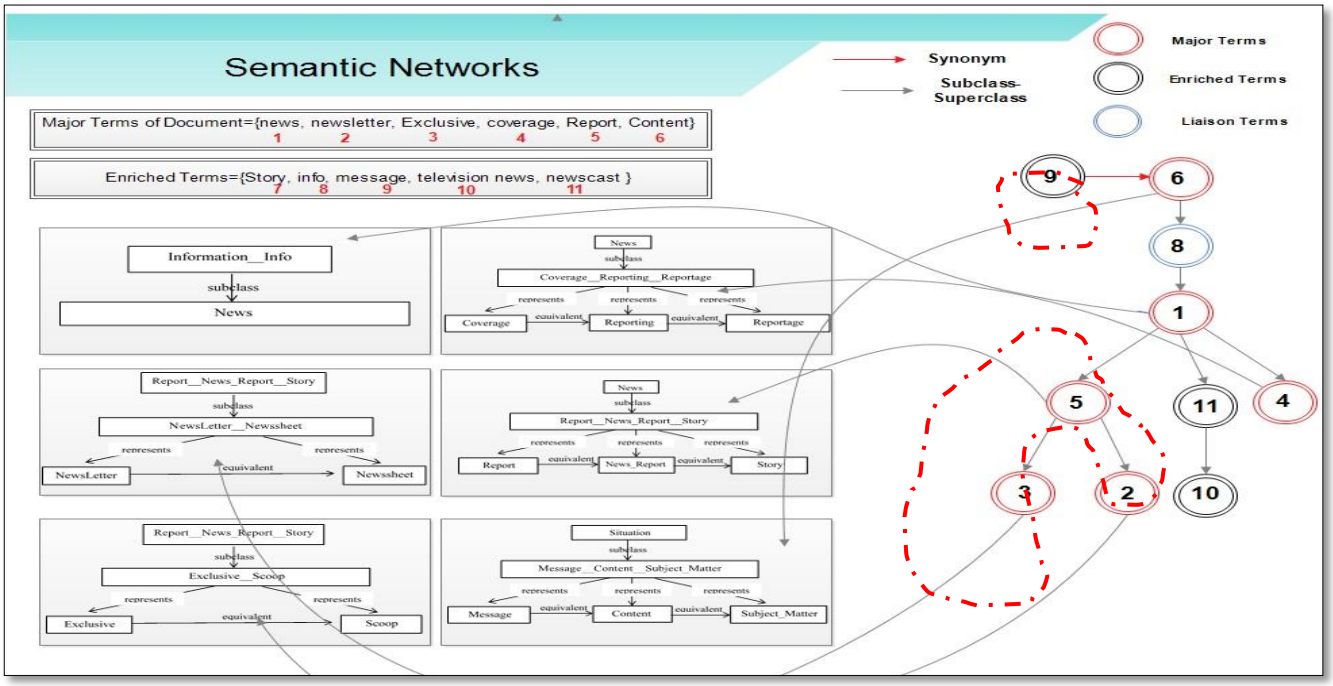

Figure 3- Semantic network and enrichment in connecting the concepts

OntoWordNet ontology, two separated concept clusters are formed. By analyzing the OntoWordNet ontology, it can be understood that the concept "info__information" is the "liaison" concept for connecting the two separated clusters. After enriching, the concept "info__information" is appended to the document semantic network and the connection between the two separated clusters is established. Also, the concepts "television_news" and "newscast" act as the "liaison" concepts for connecting the constructed semantic network with concepts in the deeper hierarchical structure of the ontology.

The semantic networks will be represented as a sub-ontology using OWL/XML schema. Such a representation not only makes the generated semantic networks machine-readable, it also enables the system to merge semantic networks with concept maps. This will create a unified and comprehensive representation of documents.

4-4-2- Merging the Enriched Information with Document Semantic Network:

Incorporating the enriched information and semantic features into the semantic networks will help the system create a fully-connected and comprehensive representation model. Therefore, the assumption is that, merging information from different knowledge sources will result better precision for the system. The following principles are used as a guideline to merge the semantic network with concept maps:

1) The document semantic network is selected as the "stable ontology". The stable ontology is the more preferred ontology. In case of merging classes, the name of the class in the stable ontology will become the merged class name. Also, if there is a conflict when classes are merged, the stable ontology will be preferred.

2) The class names in both ontologies are scanned to find lexically identical, or linguistically similar the class names. Several factors can be considered to determine the level of similarity between class names, namely synonymy of classes, common sub-strings and common prefix/suffix.

3) In order to merge two classes: If the name of the classes is identical, either the classes will be merged or one of classes will be removed. If the name of the class is linguistically similar, a link between tween two classes will be created. The label of this link will be "similar to". The class in the "stable ontology" will be linked to other ontology's class.

4) If a class sub-graph in "stable ontology" is similar to a class sub-graph in the other ontology, they are merged.

5) Automatic updates will be done and the steps *2*, *3* and *4* will be repeated until the ontologies are fully merged.

Figure 5 illustrates the process of merging semantic networks with concept maps.

Input: documents D={D_1, D_2,…D_n}, concepts in each document D'={t_1, t_2,…, t_n}

- Loop: for each concept in D
  - Loop: until D' is empty
    - Condition: if semantic network is empty
      - Append the first concept to semantic network.
      - Delete the first Concept from D'
    - End of Condition
    - Min_Node= the minimum of nodes between concepts in hierarchical structure of ontology and KB
    - Loop: for each t_i that already exists in the semantic network
      - Loop: for each t_j in the D'
        - Condition: if the distance between t_i and t_j is less than Min_Node
          - Source= t_i
          - Destination= t_j
          - Min_Node= the minimum distance between t_i and t_j
        - End of Condition
      - End of Loop
    - End of Loop
    - Add "Destination" to semantic network
    - Remove the "Destination" from D'
    - Condition: if Min_Node is equal to 1
      - Connect t_i and t_j via superclass/subclass relation
    - Condition: if Min_Node is greater than 1
      - For each edge between t_i and t_j
        - Add the endpoint concept of the respective edge to semantic network
    - End of Condition
    - End of Condition
  - End of loop
- End of Loop

Output: the generated semantic network for the D'

Figure 4- The semantic network generation Algorithm

4-5- Employing the Semantic Graph Representations for Text document Ranking and Classification

In the final step, in order to demonstrate the effectiveness of the proposed representation model, the documents semantic networks are utilized for ranking and classifying documents according to user preferences. To this end, a hybrid semantic scoring function is introduced. The proposed function estimates the similarity between a document semantic network and user preferences based on four criteria: common information content, common semantic relations, the shortest path between concepts in the hierarchical structure of the ontology and lexical commonalities between concepts. The most similar documents to user preferences are ranked, classified and displayed to the user. It should be noted that the hybrid scoring function is an "ad-hoc" approach. It is designed for document ranking and classification tasks. The following methods are tailored to find similarity based on lexical, semantic and structural features of documents. The following depicts how semantic networks are formalized:

$SN(d_i) = [(d_i), \cup\ rel_i]$.	a generated semantic network for document d_i
$rel_i = \{(t_j, rel, t_k) \mid t, t_k \in (d_i)\}$	a semantic relation between a subject t_j and an object t_k in document d_i
UP	A user profile

4-5-1- Measuring the Commonalities in Information Content and Semantic Features

This method measures the commonality based on the notion of information content (IC) of the Least Common Subsumer (LCS) [27] in WordNet. IC is a measure of the specificity of a concept, and the LCS of concepts A and B is the most specific concept that is an ancestor of both A and B. This method considers the information content (IC) of the LCS concept as the most significant factor in computing the semantic similarity. High IC commonality between two concepts indicates that two concepts are semantically similar. This method is called "normalized Jiang and Conrath measure" [39].

$$IC_{Score}(A,B) = j\&c(A,B) = 1 - \left(\frac{\left[IC_{nrm}(A) + IC_{nrm}(B) - 2 * IC_{nrm}\left(LCS(A,B)\right)\right]}{2}\right) \quad (3)$$

This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score.

4-5-3- Measuring the Common Semantic Relations

To this end, two measures are introduced: Explicit_relation: measures the amount of shared information content between the relations in two semantic networks and Implicit_relation: measures how much a document semantic network resembles the user preferences.

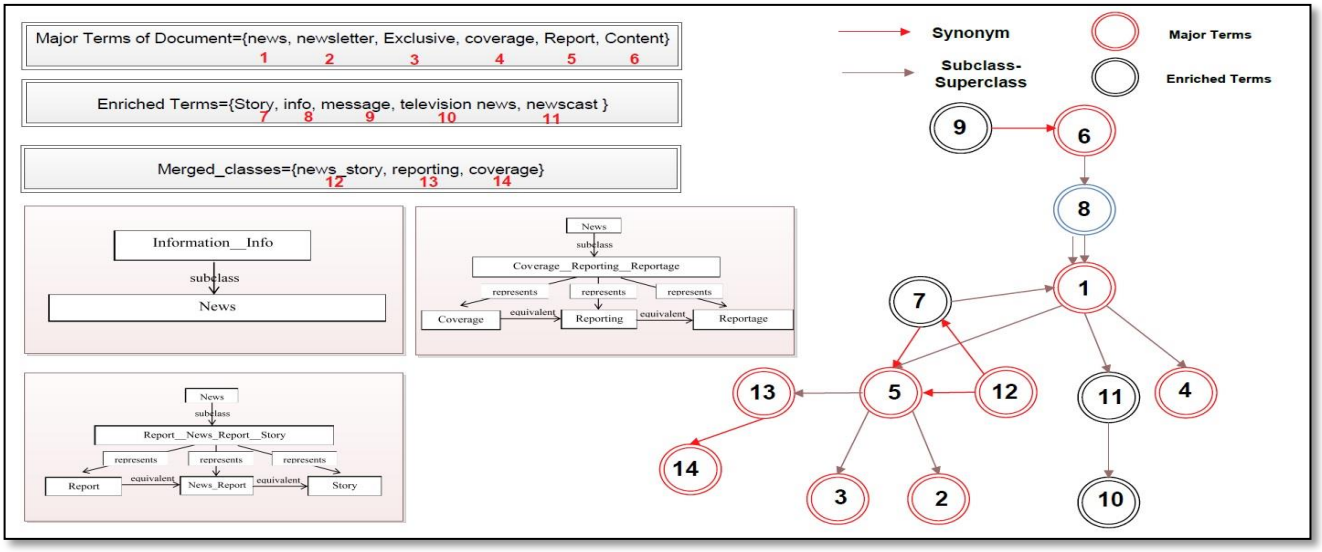

Figure 5- Merging concept maps with the document semantic networks

$$Score_{Explicit}\left(\cup\left(t_j, rel, t_k\right) \in SN(d_i)\right) = \frac{\sum_{all\ the\ triplets} Score\_term_{\text{explicit}(t_j,rel,t_k)}}{number\ of\ triplets\ in\ the\ semantic\ net.}$$

$$Score\_Term_{\text{explicit}(t_j,rel,t_k)} = \begin{cases} \delta_{exp}, & t_j, t_k\ are\ in\ user\ profile \\ 1-\delta_{exp} & o.w. \end{cases} \quad (4)$$

$$Score_{implicit}\left(\cup\left(t_j, rel, t_k\right) \in SN(d_i)\right) = \frac{\sum_{all\ the\ triplets} Score\_relation_{\text{implicit}(t_j,rel,t_k)}}{number\ of\ triplets\ in\ the\ semantic\ net.}$$

$$Score\_relation_{implicit}\left(\left(t_j, rel, t_k\right)\right) = \begin{cases} \delta_{imp}, & (t_j, rel, t_k)\ is\ in\ user\ profile \\ 1-\delta_{imp} & o.w. \end{cases} \quad (5)$$

Where, δ_{exp} and δ_{imp} are thresholds between [0, 1]. Also, A and B denote the subject and object of relations. These methods generate a number between [0, 1] indicating the similarity score.

4-5-2- Calculating the Path between Concepts in the Hierarchical Structure of WordNet

This method calculates the shortest path between concepts in the hierarchical structure of WordNet. This measure is called Wu and Palmer [27].

$$Path_{Score}(A,B) = WordNet\_Path\_based(A,B) = max\left[\frac{2 * Depth(LCS(A,B))}{Length(A,B) + 2 * Depth(LCS(A,B))}\right] \quad (6)$$

Where, $Depth(\mathrm{A})$ calculates the depth of concept A and $ength(A,B)$ calculates the shortest path between A and B. This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score.

4-5-4 Measuring lexical Commonalities between Concepts

For this purpose, Jaro-Winkler measure is used.

$$Lexical\ (A,B) = d_j + lp(1-d_j) \quad (7)$$

Where, d_j is the Jaro similarity score for Concepts A and B. Also, l_p is the length of the common prefix between two

concepts and acts as a control parameter. The following equation is used to compute the Jaro similarity score.

$$d_j = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m-t}{m}\right) & otherwise \end{cases} \quad (8)$$

In this equation, m is the number of matched characters and t is half the number of characters displacements between two concepts. This method computes the semantic similarity between all the possible pair of concepts in the document semantic networks and user preferences. It generates a number between [0, 1] indicating the similarity score. The overall similarity is calculated by a linear and a weighted combination of the each computed score as follows:

$$\begin{aligned} Sim_{Score}(SN(d_i), UP) &= \left(k_1 * Score_{Explicit}\left(\bigcup (t_j, rel, t_k) \in SN(d_i)^{\forall t_j \in d_i}_{\forall t_k \in d_i}\right)\right) \\ &+ \left(k_2 * Score_{implicit}\left(\bigcup (t_j, rel, t_k) \in SN(d_i)^{\forall t_j \in d_i}_{\forall t_k \in d_i}\right)\right) \\ &+ k_3.\frac{\sum_{\forall A \in d_i}^{\forall A \in d_i}\sum_{\forall B \in UP} IC(A,B).cf\_idf_{score}(A)}{\sum_{\forall A \in d_i} cf\_idf_{score}(A)} \\ &+ k_4.\frac{\sum_{\forall A \in d_i}^{\forall A \in d_i}\sum_{\forall B \in UP} Path(A,B).cf\_idf_{score}(A)}{\sum_{\forall A \in d_i} cf\_idf_{score}(A)} \\ &+ k_5.\frac{\sum_{\forall A \in d_i}^{\forall A \in d_i}\sum_{\forall B \in UP} Lexical(A,B).cf\_idf_{score}(A)}{\sum_{\forall A \in d_i} cf\_idf_{score}(A)} \end{aligned} \quad (9)$$

Where, k_1, k_2, k_3, k_4 and k_5 are the weighting parameters between [0, 1] and their sum is equal to 1. These parameters are estimated using a subset of the training data.

5-Evaluation

The proposed method is developed for text mining application. The *20Newsgroup* [40] and *Reuters-21578* [41] datasets are used to evaluate the performance of the proposed method. In the case of *20Newsgroup* dataset, the evaluation data are classified in 20 different newsgroups. However, some newsgroups are contextually related and can be further categorized into five broad categories or "Topics", namely computer, politics, science, religion and recreation. 4000 randomly selected documents are used to evaluate the proposed method (800 documents from each topic). Also, in the case of *the Reuters-21578 dataset*, 4000 documents from five categories of dataset, namely "earn", "acq", "interest", "trade" and "crude", are randomly selected (800 documents from each category).

5-1- Evaluating the Performance of the Proposed Method in Classifying and Ranking Documents

For each dataset, five tests are designed to evaluate the performance of the proposed method in classifying and ranking documents according to user preferences. We assume that the user preferences are exactly the same as the contents of the documents in one of the topics. In order to create a semantic representation of user preferences, two queries are created using the document in the respective topic. So, the documents in each topic are analyzed to identify the most frequent and informative concepts/words. Then, a list of candidate concepts/words is formed and presented to the experts to select the concepts/words that can describe the underlying topic the best and the queries for each topic are formed. In the next step, a semantic network representation of each topic is created. In other words, the queries are converted to semantic networks. The created semantic networks represent the user preferences in each test. The semantic networks are then used for classifying documents in their respective topics. In other words, the semantic networks are used for comparing the information content of each topic to the information content of documents. To this end, the hybrid semantic scoring function (introduced in section 4.5) is employed. In each test, the semantic similarity between each document and the semantic representation of the respective topic is measured. Then, each document is classified into the topic that most closely resembles it. The results of five tests are used to evaluate the system performance. To this end, the Mean Average Precision (MAP) score is used. The MAP value is the arithmetic mean of the average precision values for the individual information needs [30]. For a set of given queries q_i, the MAP value is calculated as follows:

$$MAP_i = \frac{1}{m}\sum_{k=1}^{m} Precision\,(R_k) \quad (10)$$

where m is the number of retrieved documents, R_k is the set of ranked results from top until the k-th document.

At first, the validity of the assumption made in section 4.4 is examined. To this end, different percentages of concepts are used to create the document semantic networks. The semantic similarity between the document semantic networks and the queries is then calculated. The performance of the proposed method is then evaluated using the average MAP score of the 10 queries. Also, the overall effect of the enrichment module on the accuracy and precision of the proposed method is evaluated. To this end, the performance of the proposed method with the enrichment module is evaluated against the performance of the proposed method without enrichment module. The results of these experiments are illustrated in Figures 6, 7, 8 and 9. As can be seen in Figures 6 and 7, the information content of the documents in "20newsgroup" dataset are better represented by Top-50% of the concepts. Also, when the semantic networks of documents in the "*Reuters-21578*" dataset are constructed by top-60% of the concepts, the system performs better. Also, the results in Figures 8 and 9 indicate that the enrichment process has a positive impact on the overall performance, even when a small portion of the information is available.

Also, the effect of merging concept maps with the semantic networks on the overall precision of the proposed method is evaluated. The results are depicted in Figure 10, 11. As evident from the results, in both datasets, when a small

percentage of concepts are used to generate semantic network, the effect of merging concept maps with the semantic networks is minimal. However, when the amount of available information content grows, the positive effect of merging increases. Therefore, the assumption, that merging information from different knowledge sources yields better precision also holds true. This would also imply that when more information about context is present, the information fusion would result in a higher performance and precision.

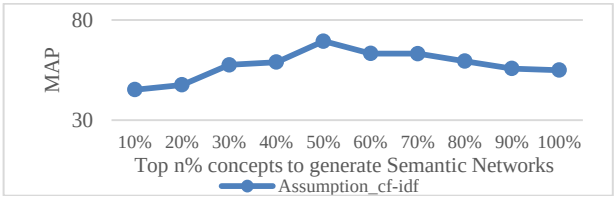

Figure 6- Evaluating the validity of assumption on 20newsgroup

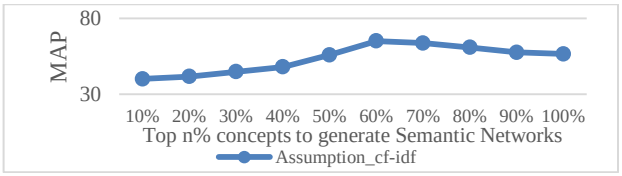

Figure 7- Evaluating the validity of assumption on Reuters-21578

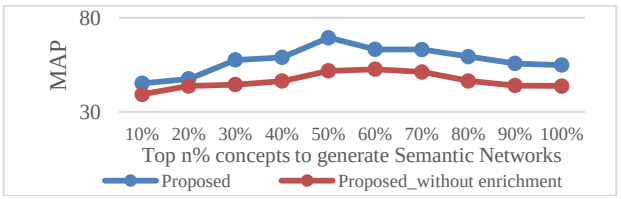

Figure 8- Evaluation of Enrichment Process on 20newsgroup

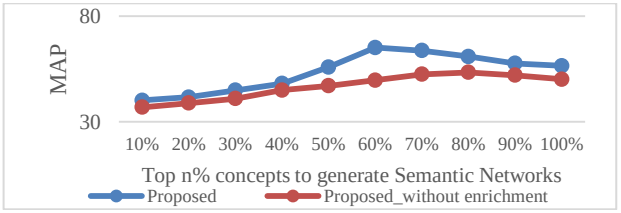

Figure 9- Evaluation of Enrichment Process on Reuters-21578

In the next step, the performance of the proposed method is compared with similar approaches. The first similarity is called the Vector Space Model (VSM)-based (Lucene) scoring function [42]. Also, the performance of the proposed method is compared with MCS-mcs document ranking and retrieval method proposed in [20].

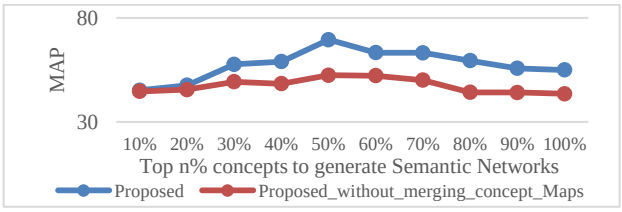

Figure 10- Evaluation the effect of merging concept maps with the semantic networks on the overall performance on 20newsgroup dataset

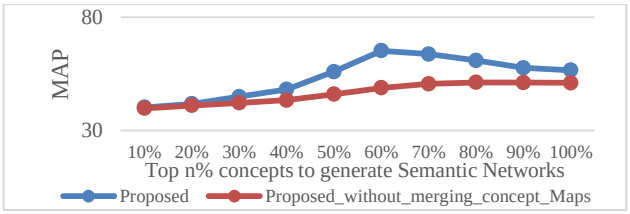

Figure 11- Evaluation the effect of merging concept maps with the semantic networks on the overall performance on the Reuters-21578

The parameters of the VSM-Based model and the MCS-mcs method have been tuned up to achieve the best possible results. The evaluation is carried out by calculating the average MAP score of designed queries for each topic. The results are illustrated in Figures 12 and 13.

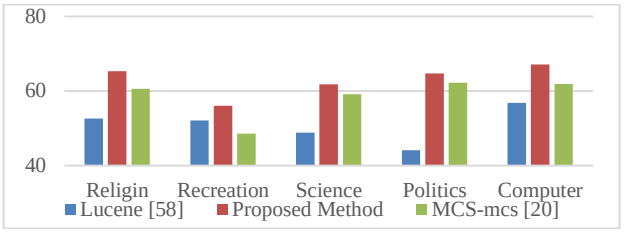

Figure 12- the comparison with similar approaches on 20Newsgroup

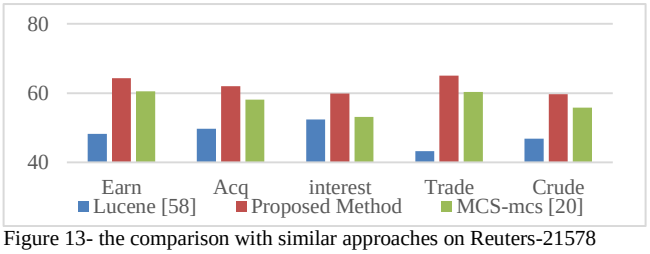

Figure 13- the comparison with similar approaches on Reuters-21578

The results suggest the proposed method outperforms all similar methods and exhibits better performance compared with MCS-mcs method. The results also suggest that the proposed method is effective in correctly classifying and ranking documents according to user preferences.

5-2- Evaluating the Performance of the Proposed Method in Identifying the Most Relevant Documents

The 20 Newsgroup and Reuters-21578 datasets are used. The Accuracy, Precision, Recall and F-measure are used for benchmarking. The evaluation data is identical to the previous experiments. In this stage, a single test for each topic is designed and the performance is evaluated. We assume that in each test, the user preferences are exactly the same as the contents of the documents in one of the topics. At first, the procedure described in section 5.1 is used for creating the semantic network representation of each topic (user queries). Then, the document semantic networks are constructed. For each test, 800 documents are labelled as relevant and 3200 documents are labelled irrelevant. Then, the similarity between document semantic networks and

each query is calculated. The system will classify each document in the most similar topic. According to the results, the document is either labelled with TP (True Positive), TN (True Negative), FP (False Positive) or FN (False Negative) labels. Finally, the performance is evaluated using the following measures. The evaluation results are illustrated in Tables 2 and 3.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{11}$$
$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 2- The evaluation results on 20 Newsgroup dataset

Test	Topics	Accuracy	Precision	Recall	F-measure
Test#1	Computer	*99.025%*	*97.62%*	*97.5%*	*97.56%*
Test#2	Religion	*98.55%*	*96.26%*	*96.5%*	*96.38%*
Test#3	Politics	*97.675%*	*93.27%*	*95.25%*	*94.25%*
Test#4	Recreation	*96.575%*	*89.7%*	*93.625%*	*91.62%*
Test#5	Science	*96.875%*	*90.52%*	*94.25%*	*92.35%*
Mean Performance		*97.74%*	*93.474%*	*95.425%*	*94.432%*

Table 3- The evaluation results on Reuters-21578 dataset

Test	Topics	Accuracy	Precision	Recall	F-measure
Test #1	Earn	*97.025%*	*91.47%*	*93.875%*	*92.66%*
Test #2	Acq	*96.875%*	*90.42%*	*94.375%*	*92.36%*
Test #3	Interest	*96.3%*	*89.27%*	*92.625%*	*90.92%*
Test #4	Trade	*97.55%*	*93.44%*	*94.375%*	*93.917%*
Test #5	Crude	*96.025%*	*89.13%*	*91.25%*	*90.18%*
Mean Performance		*96.755%*	*90.746%*	*93.3%*	*92.0074%*

As shown in Tables 2 and 3, the proposed method performs well in identifying the most relevant documents. However, the precision values in some of the topics are lower than the mean precision. After careful study of the documents in these topics, we have concluded that high levels of distinction between the topics and some of the documents and also a high degree of overlap between these documents and other topics are the reasons. The mean performance of the proposed method compared with similar text analysis methods are illustrated in Tables 4 and 5.

Table 4-The evaluation results of MCS-mcs and Lucene on 20Newsgroup

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
MCS-mcs	*95.98%*	*89.916%*	*90%*	*89.958%*
Lucene (VSM-based)	*93.935%*	*86.238%*	*82.875%*	*84.5212%*
Proposed Method	*97.74%*	*93.47%*	*95.43%*	*94.43%*

The illustrated results indicate that the proposed method outperforms other similar text analysis approaches. The multi-level representation brings several advantages that help the system outperform other methods. First, semantic, syntactical/structural and lexical features are incorporated into the semantic networks. Second, available information about the context and semantics from different knowledge sources are merged into the semantic networks.

Table 5-The evaluation results of MCS-mcs and Lucene on Reuters-21578

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
MCS-mcs	*95.72%*	*87.364%*	*91.9%*	*89.5774%*
Lucene (VSM-based)	*94.795%*	*86.26%*	*88%*	*87.118%*
Proposed Method	*96.76%*	*90.75%*	*93.3%*	*92.00%*

into the semantic networks. Second, available information about the context and semantics from different knowledge sources are merged into the semantic networks.

Next, we have decided to implement a number of learning methods for topic and text classification and evaluate these methods on Reuters and 20Newsgroup data. The evaluation results are compared with the proposed method. Three well-known machine learning algorithms for classification purposes are considered: Extreme gradient boosting [43], Random Forest [44], and Recurrent Neural Network (RNN)-Long Short Term Memory (LSTM Network) [45].

Extreme gradient Boosting (XGB): This learning model is very similar to the gradient boosting framework. It uses a linear model solver and a decision tree learning algorithm to learn the underlying data model and predict their labels. These models are decision tree-based ensemble models, and are used to reduce bias and variance of learning models.

Random Forest: Random Forest models are ensemble models. The building block of these models are decision tree method. This model generates a number of classification and regression trees (CART) with different samples and variables to learn the underlying data model.

Recurrent Neural Network-LSTM: In these models the output of their activation functions are propagated in two directions (from input to output and from output to input). This feature creates a loop in the network architecture, which acts as "memory" for neurons. This memory allows the neurons to remember what was learned. But when the number of data and consequently the number of layers increases, learning and adjusting the parameters of the earlier layers becomes even more difficult. To overcome this problem, a new type of RNN network called LSTM was developed [45]. In LSTM networks, information flows through a mechanism called "cell state". The LSTM network consists of a set of memory blocks called cells. Memory blocks are tasked with the remembering of information and memory manipulations are done through three very important mechanisms called "Gates". Forget Gate: This gate is responsible for removing redundant and unimportant information from cell state. Input Gate: it adds information to the cell state. This gate helps the system remember only the important information. Output Gate: this gate is tasked with selecting important information from the current cell state and showing it out as the output [46, 47].

In addition to the extracted standard features, LSTM is trained using the Word_Embeddings features [46, 47]. These features model the contents of documents using a dense vector representation model. In this model, the position of a word in vector space is learned from textual

documents. The learning is based on the co-occurrence words in the context. The Word_Embeddings features can be trained using the input data but it is recommended to use a pre-trained one such as Glove, FastText, or most importantly Word2Vec.

The Performance of the Learning-based approaches and the proposed method on Reuters-21578 and 20Newsgroup datasets are illustrated in Tables 6 and 7. Comparing the results suggest that the proposed method, in most cases, achieves better Recall and Accuracy results. However, machine learning methods produce better or comparable precision results compared with the proposed method. Machine learning methods achieve solid results on 20Newsgroup and Reuters-21578 datasets. However, the LSTM Network method has achieved disappointing results compared with the proposed method. It can be concluded that the knowledge-based methods are better in terms of semantic modeling than the Deep Learning methods. It can be concluded that during the numerical transformation of semantic features, a part of semantics will be lost. This can be a contributing factor in scoring disappointing results.

Table 6-The evaluation results of XGB, Random Forest and RNN-LSTM on 20Newsgroup

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	*92.0518%*	*96.014%*	*92.484%*	*94.1322%*
Random Forest	*95.764%*	*95.76%*	*95%*	*95.372%*
RNN-LSTM	*91.708%*	*95.43%*	*94.156%*	*94.786%*
Proposed Method	*97.74%*	*93.47%*	*95.42%*	*94.43%*

Table 7-The evaluation results of XGB, Random Forest and RNN-LSTM on Reuters-21578

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	*91.376%*	*92.998%*	*92.852%*	*92.862%*
Random Forest	*93.98%*	*93.526%*	*92.6584%*	*93.046%*
RNN-LSTM	*90.778%*	*94.926%*	*93.47%*	*94.19%*
Proposed Method	*96.75%*	*90.75%*	*93.3%*	*92.99*

5-3- Evaluating the reliability of the proposed method in identifying the correct topic classification

In the final stage, the goal is to examine the reliability of the proposed method in predicting the correct topic classification of documents. The assessment of the reliability of the proposed method is carried out through "Hypothesis testing". For this purpose, 4,000 documents from the "*20newsgroup*" and "*Reuters-21578*"dataset are randomly selected. In the selected collection of data, there are 800 documents representing each of the five topics. The method of evaluating the reliability of the proposed method in predicting the correct topic classification is described for one of the topics and the evaluation for other topics is done in the same way. Assuming that the user preferences are similar to the content of the documents in the "computer" topic, the semantic representation of user preferences is created using the procedure explained in Section (5.1). In the next step, the documents in "Computer" topic are assigned the label "1" and the documents in other topics are assigned the label "-1". Next, the semantic similarity between document semantic networks and the semantic network representation of user preferences is computed using the introduced document ranking and classification method (see section 5.5). If the similarity of a given document to the "Computer" topic is higher than other topics, the prediction label "1" is assigned to this document, otherwise the prediction label "-1" is assigned. The assigned prediction labels act as the topic prediction for each document. In other words, if the true label of each document is equal to its prediction label, the document is classified in its correct topic, otherwise the topic classification of the document is incorrect.

5-4- Hypothesis Testing for Evaluating the Reliability in Predicting the Correct Topic Classification of:

For this purpose, the two-sample t-test is performed. The optimal value (correct prediction label) for documents relevant to "Computer" topic is "1" and the optimal value of irrelevant ones is "-1". The mean and sample standard deviation of the computed prediction labels is *-0.6005* and *0.7997*, respectively. The purpose of two-sample t-test is to test whether the means of two different populations, the population of true labels and prediction labels are equal or not. The two-sample t-test does not assume the equality of variances. Let the null hypothesis be as follows:

H_0: The data of both populations come from independent random samples of normal distribution with equal means. In other words, the propose method is reliable in predicting the correct topic classification.

H_1: The null hypothesis is rejected. In other word, the proposed method is not reliable in predicting the correct topic classification and results may have been obtained by random chance in sample selection.

The significance level is 5% (0.05). In order to assess whether the null hypothesis should be accepted or rejected, first we need to calculate the t-value as follows:

$$t = (\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (12)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1 and s_2 are the sample standard deviation, and n_1 and n_2 are the sample size. Table 8. Shows the results on the "Computer" topic. The illustrated results suggest that the proposed method is reliable in predicting the correct topic classification of documents (the Null hypothesis are accepted in all cases) and the results have not been obtained by random chance. Finally, the reliability of the proposed method in predicting

the correct topic classification of documents in the Reuters-21578 dataset is assessed. The results are shown in Table 9.

Table 8- The results of Hypothesis Testing on "*20newsgroup*" dataset

Test	Topics	Mean	STD	*p*-value	Null Hypothesis
Test #1	Computer	*-0.6005*	*0.7997*	*0.9777*	*Accepted*
Test #2	Religion	*-0.5990*	*0.8008*	*0.9554*	*Accepted*
Test #3	Politics	*-0.5919*	*0.8064*	*0.6361*	*Accepted*
Test #4	Recreation	*-0.5825*	*0.8129*	*0.3319*	*Accepted*
Test #5	Science	*-0.5835*	*0.8122*	*0.3601*	*Accepted*

Relevant/irrelevant documents: 800/3200, significance level=5%

Table 9- The results of Hypothesis Testing on "*Reuters-21578*" dataset

Test	Topics	Mean	STD	*p*-value	Null Hypothesis
Test #1	Earn	*-0.5895*	*0.8079*	*0.5592*	*Accepted*
Test #2	Acq	*-0.5825*	*0.8129*	*0.3319*	*Accepted*
Test #3	Interest	*-0.5850*	*0.8111*	*0.4051*	*Accepted*
Test #4	Trade	*-0.5960*	*0.8031*	*0.8234*	*Accepted*
Test #5	Crude	*-0.5905*	*0.8071*	*0.5970*	*Accepted*

Relevant/irrelevant documents: 800/3200, significance level=5%

The results demonstrate the reliability of the proposed method in predicting the correct topic classification of documents in the Reuters-21578 dataset.

6- Discussion

One of the most promising aspects of the proposed method is the fusion of information from different sources in the semantic network. The results indicate that the fusion of information will result in better precision; making our assumption about the information fusion true. Since the information come from different knowledge sources, the generated semantic networks are comprehensive. They cover all the available information about the context. The semantic networks coupled with the enrichment module have a positive impact on the performance of the proposed method. As it is evident from the results, the semantic network yields the best results when we employ the top-50% and top-60% of the concepts for "*20newsgroup*" and "*Reuters-21578*"datasets, respectively. It suggests that the proposed representation model would impose less computational burden on the system. Also, the incorporation of enrichment module into the proposed method has a direct effect on generating fully-connected semantic networks. Fortunately, the results are still satisfactory. Comparing the results of the proposed method with the results of machine learning methods is promising. The proposed method provides better accuracy and recall values than these methods. However, machine learning methods achieve better precision values. The surprising thing about the results is lower than expected performance of the LSTM Network method compared to other methods. The reason for such results is that this method is designed specifically for the image processing tasks and also that a part of the semantic is lost during the numerical transformation of semantic features.

7- Conclusion

In order to overcome the lack of semantics and inherent ambiguity associated with textual resources, the structured knowledge of ontology and KBs is integrated in every component of the proposed method. Coupling the content enrichment with the semantic network generation module contributes to the novelty of the proposed method. In the first stage of evaluation, the validity of the assumption, that documents are better represented by the top-n% of the concepts, is assessed. The evaluation results suggest that using the top-50% and top-60% of the concepts for generating the document semantic networks yield the best results for the system. Examining the effect of the content enrichment module on the overall performance shows that this module has a positive effect in improving the performance and precision of the proposed method. Also, the proposed method yields far better results compared with VSM-based and MCS-mcs methods. Creating a unified and comprehensive representation of the documents, by merging concept maps with the semantic networks, is one of the most important contributions of this paper. The results shows that, when sufficient information is available about the information content of the documents, merging concept maps with documents semantic network will improve the performance and precision of the proposed system. Also, the effectiveness of the proposed method in identifying the most relevant information to user preference is assessed. Also, the results illustrate that the proposed method compared with well-known machine learning methods exhibits better or comparable performance. The evaluation results also suggest that the proposed method is reliable and effective in predicting the correct topic classification of documents. The proposed method can be employed in most text mining applications that require semantic representation of the documents, especially when limited information is available.

References

[1] M. Fernández, I. Cantador, V. López, D. Vallet, Pablo Castells, E. Motta, "Semantically enhanced Information Retrieval: An ontology-based approach", Web Semantics: Science, Services and Agents on the World Wide Web 9, 434–452, 2011.

[2] M. R. Bouadjeneka, H. Hacidc, M. Bouzeghoubd, "Social networks and information retrieval, how are they converging? A survey, a taxonomyand an analysis of social information retrieval approaches and platforms", Information Systems, Vol. 56, 1-18, 2016.

[3] B. Steichen, H. Ashman, V. Wade, "A comparative survey of Personalized Information Retrieval and Adaptive Hypermedia techniques", Information Processing and Management, Vol. 48, 698–724, 2012.

[4] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, F.N. Alpaslan, "An ontology-based retrieval system using semantic indexing", Information Systems, Vol. 37, 294-305, 2012.

[5] A. N. Jamgade, and J. K. Shivkumar, "Ontology based information retrieval system for Academic Library." International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE, 2015.

[6] T. Roelleke, "Synthesis Lectures on Information Concepts, Retrieval, and Services", Morgan & Claypool Publishers, 2013.

[7] Z. Hengxiang, J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval", SIGIR Forum, Vol. 51, 268-276, 2017.

[8] K.M. Kim, J.H. Hong, S.B. Cho, "A semantic Bayesian network approach to retrieving information with intelligent conversational agents", Information Processing & Management,Vol.43,225–236, 2007.

[9] Y. Bassil, P. Semaan, "Semantic-Sensitive Web Information Retrieval Model for HTML Documents", European Journal of Scientific Research, Vol. 69, 1-11, 2012.

[10] S.N. B. Bhushan, A. Danti, "Classification of text documents based on score level fusion approach", Pattern Recognition Letters, Vol. 94, 118-126, 2017.

[11] F. Ramli, S. A. Noah, T. B. Kurniawan, "Ontology-based information retrieval for historical documents", 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP), 2016.

[12] M. Daoud, L. Tamine, M. Boughanem, "A personalized search using a semantic distance measure in a graph-based ranking model", Journal of Information Science, Vol. 37, 614–636, 2011.

[13] D. Laura, A. Kotov, and E. Meij, "Utilizing knowledge bases in text-centric information retrieval", In Proceedings of ACM International Conference on the Theory of Information Retrieval., 2016.

[14] M. Banko, O. Etzioni, "The tradeoffs between open and traditional relation extraction", in Proceedings of ACL-08: HLT, Association for Computational Linguistics, 2008.

[15] B. Mitra, N. Craswel, "Neural Text Embeddings for Information Retrieval", In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17, 2017.

[16] F. Gutierrez, D. Dejing, F. Stephen, W. Daya, Z. Hui. "A hybrid ontology-based information extraction system", Journal of Information Science, Vol. 42, 798-820, 2016.

[17] Y. Gupta, A. Saini, A.K. Saxena, "A new fuzzy logic based ranking function for efficient Information Retrieval system", Expert Systems with Applications, Vol. 42, 1223-1234, 2015.

[18] M. Daoud, L. Tamine, M. Boughanem, "Towards a graph based user profile modeling for a session-based personalized search", Knowledge and Information Systems Vol. 21, 365-398, 2009.

[19] G-J. Hahm, J-H. Lee, H-W. Suh, "Semantic relation based personalized ranking approach for engineering document retrieval", Advanced Engineering Informatics, Vol. 29, 366-379, 2015.

[20] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, G. Xu, "An efficient Wikipedia semantic matching approach to text document classification", Information Sciences, Vol. 393, 15-28, 2017.

[21] J. Yun, L. Jing, J. Yu, H. Huang, "A multi-layer text classification framework based on two-level representation model", Expert Systems with Applications, Vol. 39, 2035-2046, 2012.

[22] C. Jiang, F. Coenen, R. Sanderson, M. Zito, "Text classification using graph mining-based feature extraction", Knowledge Based Systems, vol. 23, 302-308, 2010.

[23] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation.", Neurocomputing, vol. 266, 336-352, 2017.

[24] W. Jin, Z. wang, D. zhang, J. Yan, "Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification.", Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017.

[25] Y. Li, B. Wei, Y. Liu, L. Yao, H. Chen, J. Yu, W. Zhu, "Incorporating Knowledge into neural network for text representation", Expert Systems With Applications, In Press - Accepted Manuscript, 2017.

[26]<http://www.loa.istc.cnr.it/DOLCE.html#OntoWordNet>, "Laboratory for applied ontology - DOLCE", last visited on 19 Feb 2013.

[27] L. Meng, R. Huang, J. Gu, "A review of semantic similarity measures in wordnet," International Journal of Hybrid Information Technology, vol. 6, 1-12, 2013.

[28] P. Kolb, "DISCO: A Multilingual Database of Distribution-ally Similar Words", In Proceedings of 9th Conference in Natural Language, 2008.

[29] B. T. McInnes, T. Pedersen, "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text", Journal of Biomedical Informatics, Vol. 46, 1116-1124, 2013.

[30] S. Pyysalo, "Part-of-Speech Tagging", In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology, Springer, 2013.

[31] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky.,"The Stanford CoreNLP Natural Language Processing Toolkit", In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations, 2014.

[32] J. Hakenberg, "Named Entity Recognition", In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology, Springer, 2013.

[33] B. Mohit, "Named Entity Recognition. In: Zitouni I. (eds) Natural Language Processing of Semitic Languages", Theory and Applications of Natural Language Processing, Springer, 2014.

[34] J. Vilares, M. A. Alonso, M. Vilares, "Extraction of complex index terms in non-English IR: A shallow parsing based approach", Information Processing & Management, Vol. 44, 1517-1537, 2008.

[35] S.K. Saritha, R.K. Pateriya, "Rule-Based Shallow Parsing to Identify Comparative Sentences from Text Documents", In: Shetty N., Prasad N., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications, Springer, 2016.

[36] M. Baziz, M. Boughanem, S. Traboulsi, "A Concept-based Approach for Indexing in IR", in the proceedings of INFORSID05, 2005.

[37] C. Biemann, S. P. Ponzetto, S. Faralli, A. Panchenko, and E. Ruppert, "Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation.," in EACL, 2017.

[38] W. Cohen , P. Ravikumar, S. Fienberg, "A comparison of string distance metrics for name-matching tasks", American Association for Artificial Intelligence, 73-78, 2003.

[39] Lang, K. "The 20 Newsgroups data set, version 20news-18828", [last update on Jan 14, 2017], [Online] Available: <http://www.qwone.com/~jason/20Newsgroups>.

[40] N. Seco, T. Veale, J. Hayes., "An Intrinsic Information Content Metric for Semantic Similarity in WordNet", In Proceedings of European Chapter of the Association for Computational Linguistics, 2004.

[41] W. Zhang, X. Tang, T. Yoshida, "TESC: An approach to TExt classification using Semi-supervised Clustering", Knowledge-Based Systems, Vol. 75, pp. 152-160, 2015.

[42] S. Langer, J. Beel, "Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned.", 5th International Workshop on Bibliometric-enhanced Information Retrieval, 2017.

[43] R. Song, S. Chen, B. Deng, and L. Li, "eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices", In Proceedings of WAIM, Vol. 9658, pp. 43–54, 2016.

[44] Q. Wu, Y. Ye, H. Zhang, M. Ng and S. Ho, " ForesTexter: An efficient random forest algorithm for imbalanced text categorization", Knowledge-Based Systems, Vol. 67, pp.105-116, 2014.

[45] G. Rao, W. Huang, Z. Feng and Q. Cong, "LSTM with sentence representations for document-level sentiment classification", Neurocomputing, Vol. 308, pp.49-57, 2018.

[46] C . Olah, "Understanding LSTM Networks", [last update on Aug 27, 2015], [Online] Available: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/s>, [Retrieved on Nov 01, 2018].

[47] P. Srivastava, "Essentials of Deep Learning : Introduction to Long Short Term Memory", [last update on Dec 10, 2017], [Online] Available: < https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/ >, [Retrieved on Nov 01, 2018.

Morteza Jaderyan received his Master degree in computer engineering from Shahid Chamran University of Ahvaz, Iran. He is currently in pursuit of Ph.D. degree from Bu-Ali university of Hamadan. His main Research Interest is Artificial Intelligence, Information Retrieval and Management Systems, Semantic Web and Web Engineering, Knowledge Management Systems, Mobile Robotics, Machine learning and Intelligent Systems.

Hassan Khotanlou received the B.E. and M.E. degrees in Computer Engineering from Shiraz University in 1998 and the Ph.D. degree in Computer Engineering (Machine Vision) from Telecom ParisTech in 2008. Since 2008, He has been with the Bu-Ali Sina University and currently He is an Associate Professor of Computer Engineering and Head of the Robot Intelligence and Vision (RIV) Research Group. His current research interests include evolutionary computation, Medical Image Processing, Statistical Pattern Recognition, Deep Learning and NLP.