Journal of Computer and Knowledge Engineering, Vol. 3, No. 2, 2020.(1-20)
DOI: 10.22067/cke.2020.39286

Knowledge-Based Semantic Information Indexing and Management Framework: Integration of Structured Knowledge and Information Management Systems*

Research Article

Morteza Jaderyan[1] Hassan Khotanlou[2]

Abstract: One of the most challenging aspects of developing information systems is the processing and management of large volumes of information. One way to overcome this problem is to implement efficient data indexing and classification systems. As large volumes of generated data comprise of non-structured textual data, developing text processing, management and indexing frameworks can play an important role in providing users with accurate information according to their preferences. In this paper, a novel method of semantic information processing, management and indexing is introduced. The main goals of this study is to integrate structured knowledge of ontology and Knowledge Bases (KBs) in the core components of the method, to enrich the contents of the documents, to have multi-level semantic network representation of textual resources, to introduce a hybrid weighting schema (salient score) and finally to propose a hybrid method of semantic similarity computation. The structured knowledge of ontology and KBs are integrated from all aspects of the proposed method. The obtained results indicate the accuracy and optimal performance of the proposed framework. The obtained results suggest that using knowledge-based models leads to higher performance and accuracy in identifying and classifying documents according to user preferences; however, if learning-based models are not provided with sufficient amount of training data, they cannot yield satisfying results. The results also demonstrate that the complete integration of ontology and KBs in information systems can significantly contribute to a better representation of documents and evidently superior functionality of information processing, management and indexing systems.

Keywords: Ontology, Knowledge Base, Semantic Indexing, Knowledge-based Information System, Semantic Network Representation.

1. Introduction

Semantic information indexing and classification system deals with finding the most suitable representation of the documents and the best approaches to differentiate between the relevant and irrelevant documents in any given information domain. The representation model specifies how the documents and queries should be represented. Usually, a defined similarity metric determines the most relevant documents to a given information domain. The majority of information indexing and classification systems use a very simple representation model for documents and queries called *the bag-of-words model* which consists of a collection of single-word linguistic units. These models usually employ *the exact-term-matching methods* to search for the most relevant documents. Such a representation model suffers from serious limitations which are documented in several research papers [1-7]. Most of these limitations are present due to the inherent ambiguity in the content and the incapability of these models to represent the context of documents. They also suffer from other problems, such as *synonymy* and *polysemy*; therefore, it is hard to describe user's precise information needs via just keywords. So far, several methods have been introduced to overcome these limitations and problems, and knowledge-based approaches are among such methods. These methods utilize the structured knowledge of ontology and KBs to produce a semantic representation of the documents and user queries and also to draw a comparison between them using semantic similarity methods.

The knowledge-based methods [8-11] employ the structured knowledge of ontology and KBs to compute the true contextual meaning of words, semantic indexing and to identify the semantics in the information systems. In *sematic indexing* (i.e., the semantic representation of documents), the purpose is to extract or derive features and semantic structures that can describe the information content of the documents. Therefore, the main challenge is to determine a methodology for identifying the majority of relevant concepts and semantic structures while ignoring the irrelevant ones.

One of the most significant aspect of the proposed method is the semantic network representation of textual resources. The semantic network generally consists of a number of connected nodes (representing the concepts/words in the document.). These nodes are connected via edges. The connecting links between nodes in a semantic network represent the different relations between the concepts/words. The main idea is to extract every piece of useful and significant information about the information content from structured knowledge sources and generate a comprehensive representation of documents. The proposed method can be used in a number of IP&M-related applications, such as semantic indexing of the documents, document classification, topic spotting, personalized information filtering and recommender systems.

Two major factors play an important role in the novelty of the proposed system. Firstly, considering the synergy relationship between the different components of a text

* Manuscript received April, 28, 2019; accepted. September, 9, 2020.
[1] PhD. Candidate, RIV Lab., Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran.
[2] Corresponding Author. Professor, RIV Lab., Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran.
Email: khotanlou@basu.ac.ir

processing, indexing and management system, the idea of integrating structured knowledge into all components of the proposed system is presented. Secondly, there is a logical relationship between different components of the proposed system. In other words, a multi-layer module designed for feature extraction can identify the information structures contained in textual documents, and then content enrichment modules are introduced based on the extracted features. These modules, based on the extracted features, attempt to enrich and identify relevant information structures. The extracted features and enriched information structures are integrated into a graphical model (semantic networks). Also, the semantic similarity computation module calculates the similarity between documents and user preferences based on the extracted features, the identified information structures, and the generated representation model. Therefore, all modules and the processes embedded in them are designed in an articulate manner to implement a text-based information processing, indexing and management system. Such characteristics play an important role in the novelty of the proposed method.

However, the novelty of the present paper is explained in the following:

- The integration of structured knowledge of ontology and KBs in every component of the proposed semantic information indexing and management system.
- Utilizing the semantic networks for comprehensive
- Multi-level representation of the documents and user queries while introducing a hybrid weighting schema to identify the most significant concepts for creating the semantic network.
- Proposing a hybrid and multi-layer method of semantic similarity computation.

The paper is structured as follows: in the second section, the related works are explored. In the third section, the research objectives are declared, the hierarchical and taxonomic structures of the top-level ontology, Wikipedia and WordNet are examined and the proposed method of semantic information indexing and management is introduced. In the fourth section, the evaluation results are offered and finally, in the fifth section, the conclusion is presented.

2. Related Works

Three important, yet different criteria, will determine what kind of information indexing and management method can be used in a text mining application:

- What kind of information model should be employed?
- Should we assume semantic relations between concepts/words?
- Should we utilize structured KBs such as ontology?.

The information models determine how the textual resources should be represented and how the similarity between representation models of documents should be measured, so that the most similar documents to user preferences are identified. The probabilistic models and the Vector Space Models (VSMs) are among the widely used information models [3]. For example, the language models [1,12] and the Bayesian network models [13,14] are considered among the probabilistic models. The vector space models [15] represent the textual resources in a vector form, and the similarity between them is usually calculated using the cosine similarity measure. As the majority of the traditional information management model do not disambiguate the concepts/words and use basic feature extraction techniques, they are very easy to implement. However, they exhibit relatively low precision and poor performance. In one study [16], the authors introduce a hybrid Sentence-Vector Space Model (S-VSM) and Unigram representation models for the text document. However, in recent years, numerous studies [1, 8, 17-21] have utilized the graph-based methods for the information indexing and management in which a domain/Top-Level ontology is often used to represent textual resources and their contextual semantics in the form of a graph.

2.1 Learning-based Information Systems

Intelligent learning models are also used in the field of text mining. The bag-of-concepts method was introduced [22] as an alternative document representation method. The proposed method creates concepts through clustering word vectors generated from word2vec and uses the frequencies of these concept clusters to represent document vectors. In another study [23], the Metzler and Croft's MRF (Markov Random Field) model [24] is employed to construct the information model and a supervised learning method called regression rank [25] to improve the performance of the Markov information model. Also, the integration of machine learning techniques and knowledge-based methods has proved to be quite beneficial. For example, in one study [26], a novel framework for incorporating KB into the neural network is introduced to produce a high quality representation of text. The most important shortcoming of learning-based methods is their reliance on domain data for training a classification model. As the proposed framework is reliant on multi-domain structured knowledge of ontology and KBs, it can achieve better performance and accuracy.

2.2 Model-based Information Systems

As mentioned earlier, assuming semantic relations between concepts/words determines what kind information indexing and management method can be used for a text mining application [13,27,28]. The majority of the traditional methods are based on the Bag of Words (BoW) models. The underlying assumption in these models is that a document can be represented by a set of not connected concepts or words (i.e. no relation is defined between concepts/words) [15]. These information models usually need an additional term weighting schema; therefore, selecting a proper weighting schema has a profound effect on the accuracy and precision of the model. In one study [29], the importance of employing a suitable weighting schema for information retrieval-related applications is emphasized. However, since these models do not take into account the (semantic) relations between the concepts/words, unsatisfactory results are often obtained.

To overcome the drawbacks of the BoW-based systems, term-dependence models are introduced. These models exploit the relations established between concepts/words. For example, in one study [30], a fuzzy-based method for considering the relation between index terms is introduced.

Also, a number of conducted studies [12,23,24] demonstrate that the term-dependence models exhibit far better performance than the Bag-of-words models. However, the biggest challenge of these models is the large number of training data needed to estimate the joint distribution of the documents and queries. Also, limited domain information might lead to unsatisfactory performance. Utilizing the structured knowledge of ontology and KBs can help the term dependency-based retrieval methods overcome this limitation.

In the majority of the studies, using the WordNet Synonym sets are recommended to model the semantics (i.e., meaning) in the documents [31,32]. When coupled with an efficient Word Sense Disambiguation (WSD) technique, these systems exhibit good performance and accuracy. As the proposed system in this paper integrates structure knowledge and semantics of ontology and KBs in every component, it can easily overcome the shortcomings of model-based information systems.

2.3 Information Systems with emphasis on information extraction techniques

From another perspective, one can distinguish between the information indexing and management systems based on the information extraction module they use. Until now, different methods of extracting informative features are introduced; however, the main difference between them arise from two main factors: (1) the structure of auxiliary knowledge sources employed to extract features, (2) details of the extracted features. The natural language processing (NLP)-based methods are usually domain-independent and are used to extract semantic, syntactic and morphological features from documents and are usually computationally expensive [8,33-37]. However, in order to overcome the computational obstacles of these methods, the rule-based information extraction methods are introduced. These methods construct the extraction rules either manually or automatically. The automatic rule-based methods [8, 33, 38, 39, 40, 9] exhibit far better performance in domain-specific applications. Considerable number of manual rule-based methods are proposed for semantic annotation [41,42] and information extraction [43-46]. In this regard, the regular expressions are used to extract features and information from textual documents. Etzioni et al. [42] employed domain-independent rules to find the information and features that help system identify the correct class of documents. The use of ontology for information and feature extraction has also been investigated as such. In another study [41], a domain ontology is used to implement a semantic annotation and information extraction framework.

2.4 Ontology-based Information Systems

The ontology-based methods exploit the structured knowledge of ontology to implement a semantic framework for integrating knowledge in information indexing and management systems [48, 49, 11]. In one study [49], an ontology-based approach for integrating knowledge of domain ontologies in information extraction and retrieval systems is introduced. A detailed study of the information extraction, indexing and retrieval systems is presented in other studies [50, 51]. Meanwhile, utilizing ontology-based methods for semantic information indexing and management is another alternative for considering the term-dependency. In such methods, the relation between concepts/words are inferred using the graphical structure of the ontology. In the next step, the relations between concepts/words are identified and employed to compute the similarity/relatedness between the documents and user preferences. The structured KBs such as ontologies, Wikipedia and WordNet, are widely used in information indexing and management applications [17-19].

In one study [18], a personalized method of textual document search and retrieval according to user profiles is introduced in which the documents are retrieved and ranked according to a graph-based distance measure. The relations between the concepts are extracted using a web-based ontology called ODP [52]. In another study [53], a knowledge-based recommender system based on the integration of ontology and sequential pattern mining (SPM) for e-learning resource recommendation is introduced. The ontology is used for domain knowledge modelling and representation, and SPM is utilized for detecting the learners' sequential learning patterns.

Researchers [54] have also utilized domain ontology to establish semantic relations between the concepts/words and to construct the semantic networks [54]. As such, the relations between concepts are weighted according to a specific weighting schema, and then the documents are ranked and displayed according to their similarity to the user queries.

The major problem with such systems is that they do not consider the synergy relationship between the different components of an information system. In this paper, the integration of structured knowledge and KBs in all components of the system is proposed to overcome this problem.

2.5 Knowledge-based Information Systems

The Wikipedia is also used for text mining applications and representation of the textual resources. The proposed method [55] represents each document as a concept vector in the Wikipedia's semantic space to model the text semantics. Then, several heuristic selection rules are defined to quickly pick out related concepts from the Wikipedia's semantic space. Then, the similarity between documents are computed to classify the documents.

Also, the personalized retrieval and ranking methods are gaining interests in recent years. These methods facilitate the rapid access and accurate retrieval of the textual documents [18,52,54,56]. The most similar/related documents to the user preference are identified based on the similarity of user preferences and document contents. The user preferences are easily obtained by analyzing the usage data and user's previously accessed documents.

Like ontology-based information system, knowledge-based systems do not consider the synergy relationship between the different components of an information system. Therefore, in order to overcome this limitation, the integration of structured knowledge of ontology and KBs in every component of the proposed method is considered.

The following table summarizes the related methods in indexing and information retrieval, their underlying model and their characteristics.

Table 1**.** Related methods and their characteristics

Methods	Personalization	Ontology-based	Model	Features
Kara et al. [27]	No	Yes	Graph-based, Keyword-based	Term-Dependency assumption, Scalable, Domain-Specific, ontology-based
Lasse [35]	No	No	Language Model	No Assumption of Term-Dependency, General Domain
Hahm et al. [23]	Yes	Yes	Graph-based, Keyword-based	Term-Dependency assumption, Domain-Specific, ontology-based
Metzler et al. [47]	No	No	Language Model	Term-Dependency assumption
Li et al. [38]	No	Yes	Graph-based	Term-Dependency assumption, ontology-based
Nefti et al. [27]	No	No	Fuzzy-based	Term-Dependency assumption
Daoud et al. [13]	Yes	Yes	Graph-based, Keyword-based	Term-Dependency assumption, General Domain, ontology-based
Li et al. [78]	No	No	Intelligent Learning model	Knowledge-based, conceptualized vector space
Proposed Method	Yes	Yes	Graph-based, Enriched Keyword-based, Language Model	Term-Dependency assumption, Scalable, General Domain, ontology-based, knowledge-based

3. Proposed Method

This section can be divided into three subsections: 1) research objectives, 2) the structures of ontology and KBs integrated into the proposed framework, 3) the specification and characteristics of the proposed information processing and management framework.

3.1 Research Objectives

The general objective of this paper is to develop a multi-purpose framework for collecting information from different knowledge sources and modelling the extracted semantic, lexical and syntactical features in a multi-level representation using the graph-like structure of semantic networks. In this regard, the specific objectives of this research are:

1. To describe a multi-purpose text mining framework which integrates ontology and KBs for developing a multi-level representation of textual resources using machine-readable semantic networks.
2. To describe a mechanism in which the information content of textual resources is enriched for better representation.
3. To describe a hybrid multi-layer semantic similarity module for identifying resources that satisfy users' information needs.
4. To assess and analyze the performance and effectiveness of the proposed framework in semantic information indexing and management applications.
5. To evaluate the effect of the enrichment module on the overall performance of the proposed method.
6. To evaluate the effect of the representation module and semantic similarity mechanism and its components on the overall performance of the framework.

3.2 The Structure of Ontology and KBs

The ontology and KBs play a crucial role in identifying the semantics and context. Therefore, familiarizing with the hierarchical and taxonomical structure of these KBs helps us figure out what kind of semantic structures can be identified and extracted from textual resources.

3.2.1 OntoWordNet Top-Level Ontology

The OntoWordNet ontology (OWL alignment of the WordNet ontology with DOLCE-Lite Plus Ontology library) is an essential component of the proposed system. Every concept of the ontology is organized as synonym set, so that the contextually similar (or equivalent) concepts can be retrieved. This will facilitate the enrichment of the contents [57].

3.2.2 WordNet

WordNet is an ontological lexicon for the English language. The purpose is to model a semantically enhanced lexicon for the English language. The main structure of WordNet consists of Synsets. The synset organizes a set of synonym concepts. More details about WordNet are available in the literature [58].

3.2.3 Wikipedia

Wikipedia and BNC data which have been used in this research are available for academic use through D.I.S.C.O project. Both data are structured the same way. In one study [59], the manner in which the data are created is described. Both data structures consist of two sets of data: (1) first-order word vector, which contains words that occur together in Wikipedia and BNC corpus, and (2) second-order word vector, which contains words that occur in similar contexts.

3.3 The Proposed Information Processing and Management Framework

The proposed method generates a semantic graphical representation (semantic network) of document contents and user profile and calculates the semantic similarity between them. The constructed user profile is used to personalize the information indexing and management process. Fig. 1 illustrates the overview of the proposed system. The proposed system consists of two separate processes: (1) the semantic information indexing, and (2) the semantic information management. On the other hand, the proposed method consists of three major components: 1) semantic network generation module, 2) content enrichment module,

and 3) semantic similarity/relatedness computation module. As shown in Fig. 1, the documents and user queries act as the system input. Several pre-processing operations are performed on inputs and all the concepts undergo the word disambiguation (WSD) process. First, assuming none of the documents in the repository are indexed (by semantic networks), every document in the repository are indexed by their keywords. These simple indexes are then stored in a database or repository. In the next step, a Boolean matching model (known for its rapid and accurate pattern matching) [61, 62] is built. As soon as a query is made by a user, it is converted into a Boolean search expression and a set of documents from repository which are fully or partly matched with the Boolean expression, and then they can be identified and extracted as such. Every retrieved document will be represented by a semantic network. The proposed hybrid semantic similarity module is used to determine which of the retrieved documents are the most similar to user query.

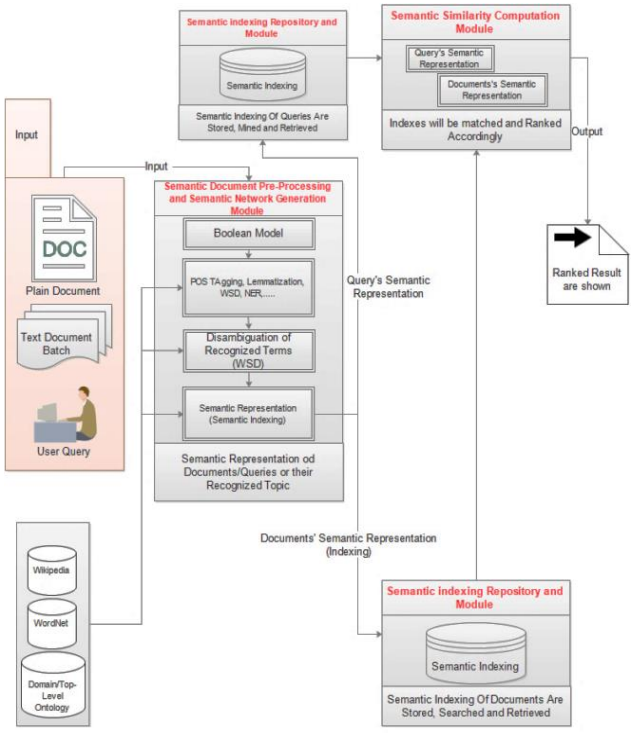

Figure 1. Overview of Semantic indexing and Retrieval System

Figure 2 depicts the document index matching and retrieval process. It should be noted that all the document semantic networks can be constructed offline. Also, the process of updating user profile semantic network using the previously accessed documents can be performed offline regularly and according to a pre-specified schedule. On the other hand, the process of calculating the semantic similarity between document semantic networks and user profiles is performed online and imposes negligible operational burden on the system. Also, the constructed semantic networks are stored in a database called index repository. The user profile's semantic network is stored in a repository to facilitate regular updating process. In the following sections, the details surrounding the proposed method is discussed.

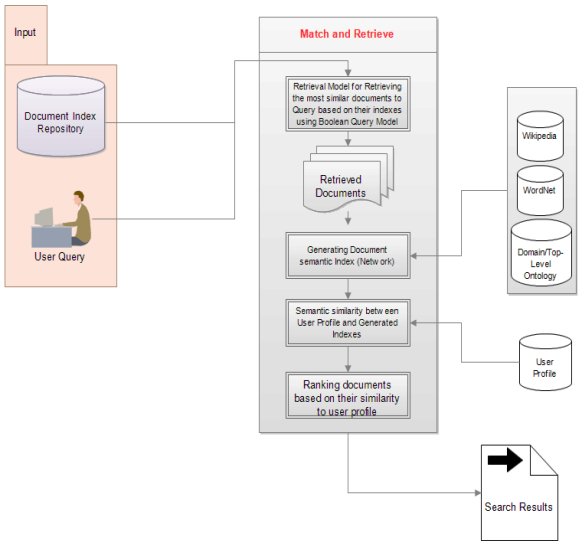

Figure 2. Document index Matching and Retrieval

3.3.1 The semantic document pre-processing

The following pre-processing techniques are performed on the contents of the documents:(1) Stop-Word Removal, (2) Uni-gram and Bi-gram processing of English Words, (3) Stemming concepts/ words [63,64], (4) Part of Speech Tagging [65], (5) Lemmatization of the concepts/words [65, 66], (6) Named-Entity Recognition [67,68], (7) Bi-gram Authentication. The authenticity of Bi-gram concepts is validated using Wikipedia KB, which is performed by searching for the frequency of Bi-grams in Wikipedia. Once the authentication operation is over, the rejected Bi-grams are removed.

The output of this module is the document/user_profile vector. In the next step, all the detected Uni-gram and bi-gram concepts are weighted using the CF-IDF weighting method (a variant of TF-IDF) [69]. Accordingly, the weight of the concept c_i in document d_j is calculated using the following equation:

$$weight(c_i, d_j) = cf_{d_j}(c_i).ln(N/df) \quad (1)$$

Where N is the number of documents in repository, df (document frequency) is the frequency of the documents in which the concept c_i appears. Also, the local frequency of a concept like c_i which comprises of n words ($n \geq 1$) depends on the number of occurrences of concepts c_i and all its sub-concepts. Therefore, the weighting equation is formally rewritten as follows:

$$Cf(c_i) = count_{d_i}(c_i) + \sum_{sc \in Sub\_ConceptsI} \frac{Length\ (sc)}{Length(c_i)}.count_{d_j}(sc) \quad (2)$$

In this equation, $Length(c_i)$ represents the number of words in concept c_i and $Sub\_Concepts(c)$ deputes all the possible sub-concepts which are directly derived from c_i. 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 for document d_i, $w_i = \{w_i^1, w_i^2, w_i^3, \dots\dots, w_i^n\}$ is the document weight

vector after the weighting method is applied to all the documents.

3.3.2 The Word Sense Disambiguation of Concepts

To handle the word ambiguity issue in text documents, the method of word sense discrimination [70] was introduced. In this method, the underlying assumption is that similar senses occur in similar contexts. Hence, by using a semantic similarity method to compare all possible senses of a concept with the context it appears in, we can predict the correct sense of a given concept in the document. To this end, the following procedures are performed:

- A ± 7 window around the concept in the respective document is created. This will create a context vector for the corresponding concept. Also, the first-order word vector for each member of the context vector is retrieved and appended to. This window and the appending vectors create a context vector for the concept.
- The senses of the corresponding concept, an example of their usage in a sentence and their brief definition called gloss is extracted for each sense from WordNet. The Wikipedia-based first-order word vector for each sense is also retrieved. The collected information about a sense is aggregated and the sense vector is formed. As the first-order vector contains the co-occurrence words in similar contexts, the similarity found between the sense vector of each sense and the context vector determines the true contextual meaning of corresponding concept.
- A combination of cosine [71] and Jaro-Winkler [71] measures is used to calculate the similarity score as follows:

$$Sim(Sense_{Vctor}, context_{Vector}) = \frac{1}{2}\left(Cosine_{Sim} + Jaro_{winkler_{Sim}}\right) \quad (3)$$

- The sense vector with highest similarity score is selected as the correct sense of the concept. The correct sense number is used to annotate the concept.

3.3.3 The Enrichment process of extracted contents from documents

In most cases, the extracted features are not a good representation of the document context. The main purpose of this section is to identify concepts and semantic structures that can better describe the document context.

3.3.3.1 Enrichment using Wikipedia KB

External knowledge sources such as Wikipedia can be used to improve the representation of textual resources [72]. As mentioned earlier, Wikipedia KB contains a set of information called second-order word vector. This vector not only contains the co-occurrence words in Wikipedia but also words that are contextually similar and interchangeable in different contexts. Employing this vector to enrich textual resources is an interesting idea that is proposed in this paper. These vectors are searched to identify the co-occurring and contextually similar concepts to a given concept/word. The identified concepts are then weighted according to the following equation and are appended to the document vectors/user profiles.

$$Weight_{Related\ Conceot(second-order\ vector)} = Weight_{original\ Concept} * 0.8 \quad (4)$$

Because the new concepts are obtained indirectly and are inferred using Wikipedia, their assigned weight is lower than the original concepts. The weighting parameter is estimated using a subset of evaluation data.

3.3.3.2 Enrichment using OntoWordNet Top-Level Ontology

The OntoWordNet ontology classes are organized in the form of sequences. Every sequence defines a set of synonym concepts (the synonym concepts are separated by two consecutive underline "__" and the space between multi-word concepts are specified by an underline "_") contexts. The concept map consists of a concept and a set of related ontology classes. The links between the concept and related ontology classes are the *equivalent property* and the subclass/superclass relations. The equivalent property is transitive and reversible. The aforementioned procedure results in the creation of *concept maps* for each concept. Also, the concept maps play a vital role in constructing a multi-level representation of documents. It should be noted that the obtained concept maps are represented by a sub-ontology using OWL/XML schema to facilitate the process of annotating semantic networks with concept maps. An example of a concept map for the concept of news story is illustrated in Fig. 3.

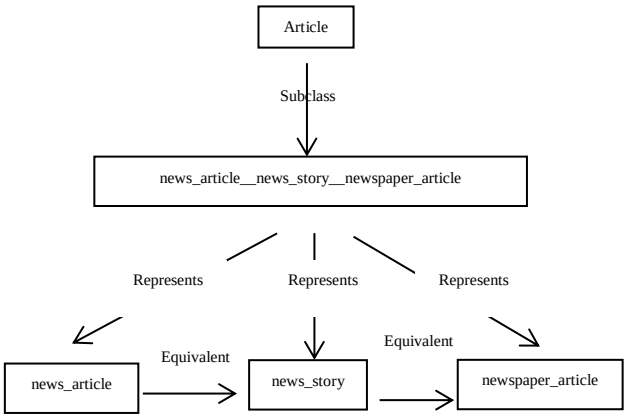

Figure 3. Representation of a generated conceptual maps

The superclass and equivalent concepts are weighted according to the following equation and are appended to the document vectors/user profiles. Since the new concepts are obtained indirectly and are inferred using top-level ontology structure, their assigned weight is lower than the original concepts. The weighting parameter is estimated using a subset of evaluation data.

$$Weight_{Related\ Conceot(ontology)} = Weight_{Initial\ Concept} * 0.8 \quad (5)$$

It should be noted that the semantic and graph-like structure of the concept maps are used in the semantic network generation phase to annotate the documents semantic networks and to infer new links between concepts.

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

3.3.4 The semantic network generation

In order to generate the semantic networks, a robust and semantic weighting method must be used to identify and select the most informative concepts in documents. The CF-IDF measure does not possess the required semantic characteristic for this purpose. In this paper, a new weighting schema called *salient score* is introduced to select the most prominent concepts. These concepts will participate in the semantic network generation process. In the next step, four important relations in the ontology (namely, synonymy, superclass, subclass and Part_of) are established between concepts to link the concepts to each other. The connected concepts are organized in a graph-like structure to form the semantic network. Also, the enriched contents play an important role in identifying the concepts that reflect the information content of documents. Therefore, the generated semantic network acts as a thorough and comprehensive abstract of the documents. Accordingly, the semantic network generation process consists of two important parts: (1) calculating the salient score of concepts in the documents, (2) connecting the concepts using ontology-defined relations. Figure 4 illustrates the process of generating semantic networks. First, the elements of a semantic network are discussed here.

3.3.4.1 The elements of a semantic network

The semantic network consists of a set of concepts and the relations connecting them. In this paper, two given concepts are connected to each other through one of the four relations of synonymy, superclass, subclass, and part of relations.

Definition of Concept: Concepts refer to a significant entity in the document.

Definition of Superclass/Subclass relation: Assuming that two concepts x_i and x_j are given, if concept x_i categorizes the concept x_j, then x_i is called superclass of x_j, and x_j is called subclass of x_i.

Definition of Synonymy relation: Assuming x_i is a concept in the document, if we can find a concept x_j and replace it , the informational content of the document does not change. Thus, it can be inferred that x_i and x_j are connected by Synonymy relation.

Definition of Part_of relation: The Part_of relation represents the part-whole relationship between the concepts. The Part_of relation is established between concepts x_i and x_j, if presence of x_j implies the existence of x_i. However, the presence of x_i does not indicate the presence of x_j. The Part_of relation is obtained by aligning DBpedia ontology [73] and its related NLP datasets with OntoWordNet ontology.

Theses relations can be represented in the form of ordered triplet [Subject, Object, Relation]. For example:

- Superclass relation: [Sports, Football, Superclass],
- Subclass relation: [Football, Sports, Subclass],
- Synonymy relation: [Sports, Athletics, Synonym],
- Part_of relation: [Halfback, Football, Part_of],
- Part_of relation: [Halfback, Sports, Part_of],
- Subclass relation: [Halfback, position, Subclass].

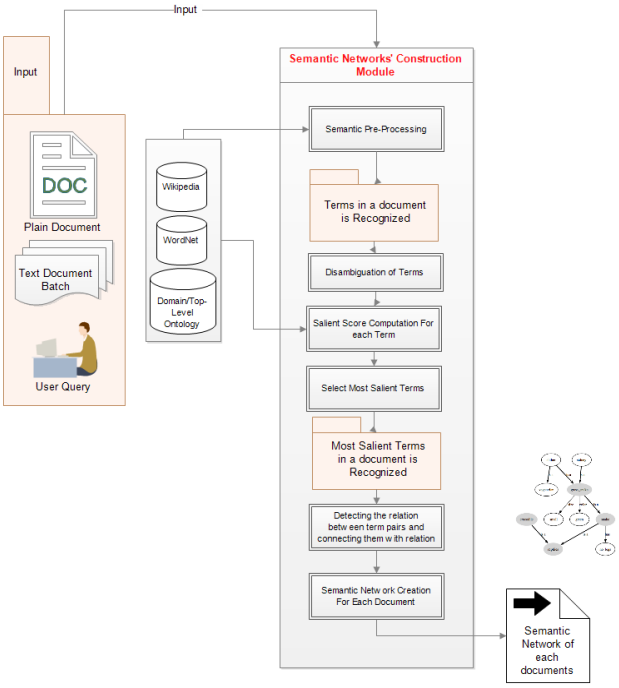

Figure 4. Semantic Network Construction Process

3.3.4.2 The salient score

Not all the concepts/words contribute to the information content of a document. Some are only used to impart the linguistic or formal expressions and they are usually meaningless. Therefore, it is better to discard the trivial concepts/words which have insignificant contribution to the context.

In this paper, in order to identify the most informative concepts/words in the document, salient score is introduced, which has three different criteria: (1) structural criterion, (2) CF-IDF criterion, and (3) semantic criterion. In other words, the proposed weighting schema is a hybrid schema that integrates the term-based weighting, structural-based weighting and the knowledge-based weighting approaches.

The Structural Criterion: Let $D = \{d_1, d_2, \dots\dots, d_n\}$ be a set of retrieved documents for a given query, n be the number of the retrieved documents and $d_j = \{t_1^j, t_2^j, \dots\dots\dots, t_n^j\}$ be the extended document vector d_j. Also $sub(t_i^j)$, $sup(t_i^j)$, $partof(t_i^j)$ and $synonym(t_i^j)$ are sets of concepts/words in d_j that are connected to the concept t_i^j via subclass, superclass, Part_of and synonym relations, respectively. The structural criterion score for t_i^j is calculated by $str\_Score(t_i^j)$ as follows:

$$Str_{score}\left(t_j^j\right) = \begin{cases} \dfrac{|sub(t_i^j) + sup(t_i^j) + partof(t_i^j) + synonym(t_i^j)|}{\max(|sub(t_i\ )| + |up(t_i^j)| + |partof(t_i\ )| + |synonym(t_i\ )|)}, & t_i^j \text{ is a uni/bigram in document} \\ 0, & o.w. \end{cases} \quad (6)$$

This criterion assumes that for each concept/word in the document there is at least one concept in the document that is connected to t_i^j via subclass/ superclass/ Part_of/ synonym relations. As the number of subclass/ Part_of/ synonym concepts for t_i^j in the document increases, higher score for structural criterion will be yielded. If no subclass//Part_of/synonym for t_i^j is found in the document, the structural criterion score for t_i^j is zero.

The CF-IDF Criterion: The CF-IDF criteria and how the concepts in the document are weighted is fully discussed in section 5.1.

The Semantic Criterion: This criterion is calculated by the following equation and is denoted by $Sem_{Score}(t_i^j)$:

$$Sem_{Score}(t_i^j) = \frac{|\{t_k^j |\, Sem_{Distance}(t_i^j, t_k^j) \le \alpha\}| + |\{t_k^j |\, IC(t_i^j) \le \beta\}|}{2 * |d_j|}$$
$$(1 \le K \le |d_j|) \quad (7)$$

Where, t_i^j is the underlying concept in the document, t_k^j represents the elements of document vector and $Sem_{Distance}(t_i^j, t_k^j)$. This is also known as the semantic distance, which refers to the minimum number of nodes between two concepts t_i^j and t_k^j in the hierarchical structure of KBs. The $IC(t_i^j)$ is the information content of t_i^j computed using the WordNet and Penn Treebank. More information is available in some other studies [69, 37]. The parameters α and β are Thresholding parameters for the semantic distance and the information content. The illustrated equation means that when the context of a given document is about a particular information domain, the concepts/ words in the document are very similar in terms of the context, and they form a cluster in ontology/KBs as such. In other words, the concepts/words with high semantic distance from the context are considered insignificant and will have a lower semantic score as such.

In order to calculate the salient score, a linear and weighted combination of these Criterions is computed as follows:

$$Sal_{Score}(t_i^j) = w_1.Str_{Score}(t_i^j) + w_2.cf - idf(t_i^j) + w_3.Sem_{Score}(t_i^j) \quad (8)$$

In this equation, w_1, w_2 and w_3 are the weighting parameters for calculated scores between (0- 1) and their sum is equal to 1. These parameters are estimated using a subset of evaluation data in the evaluation stage. It should be noted the salient score is computed only for the concepts in documents. All the concepts in user profile will participate in user profile semantic network generation. In the next step, the top-n% of the concepts/words with highest salient score are selected to generate the document semantic network

3.3.5 The document semantic network generation

When the top-n% concepts/words are projected onto the OntoWordNet ontology, a number of separated clusters of concepts/words are formed in ontology because some of concepts/structures that can connect the separated clusters are left out. One of the main objectives of the proposed method is to identify the connective concepts/structures, so that a comprehensive, thorough and connected representation of the documents is formed. In order to generate the document semantic network, the proposed algorithm puts together the identified concepts/structures one by one, connects them using the aforementioned relations and then forms a connected graph. Concepts/structures essential for generating a fully connected semantic network and connecting the separated concept clusters are mostly identified during the content enrichment stage. These connecting/structures concepts are called *Liaison features*. This property of the proposed algorithm contributes heavily to the novelty of the proposed method. The proposed algorithm for generating semantic networks is illustrated in Figure 5. Also, Figure 6 depicts how the semantic network is formed and how the Liaison features connect the separated concept clusters.

The resulting semantic networks will be represented as a sub-ontology using the OWL/XML schema. Such a representation not only makes the semantic networks machine-readable but it also enables us to merge them with the generated concept maps.

Input: set of documents D={D_1, D_2,…D_n}, set of prominent 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 the semantic network.
      - Delete the first Concept from D'
    - End of Condition
    - Min_Node= the minimum number of nodes between concepts in the 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 the semantic network and 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 the semantic network
    - End of Condition
    - End of Condition
  - End of loop
- End of Loop

Output: the generated semantic network for the D'

Figure 5. The pseudo-code for the creatio of semantic network

As shown in Figure 6, after projecting concepts/words onto the OntoWordNet ontology, two separated clusters of concepts are formed in ontology. By analyzing the ontology, it can be understood that the concept "info__information" is

the Liaison concept for connecting the two separated clusters. By enriching the content of documents, using the ontology and Wikipedia-based approaches, the concept "info__information" is appended to the document semantic network and the connection between the two separated clusters is established. Also, the concepts "message", "story", "television_news" and "newscast" act as the Liaison concepts for connecting the already constructed document semantic network with concepts in the higher/deeper hierarchical structure of the ontology.

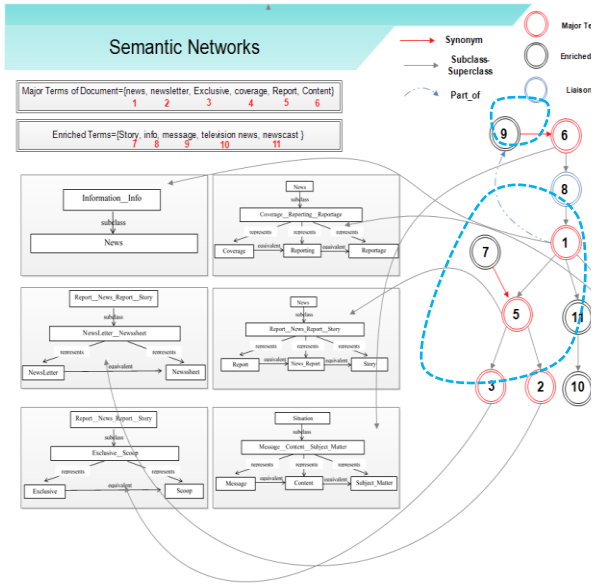

Figure 6. An example of Semantic Network

To provide a personalized experience, the user queries and user's previously accessed documents are analyzed to generate the user profile semantic network. Using the proposed algorithm, user queries and all the previously accessed documents are converted into semantic networks. The resulted semantic networks are then combined to create a connected graph. The graph represents a portion of the ontology that covers the informational preferences and priorities of the user. The process of personalized document identification is illustrated in Figure 7.

3.3.6 the semantic similarity of computation modules

In the proposed semantic information indexing and management method, the semantic similarity between a document semantic network and a user profile is computed based on three types of similarity measures. The first two types compute the similarity based on the established relations between concepts, while the last one computes the semantic similarity based on the commonalities in semantic features. The proposed semantic similarity measure relies heavily on the structured knowledge of ontology, Wikipedia and WordNet.

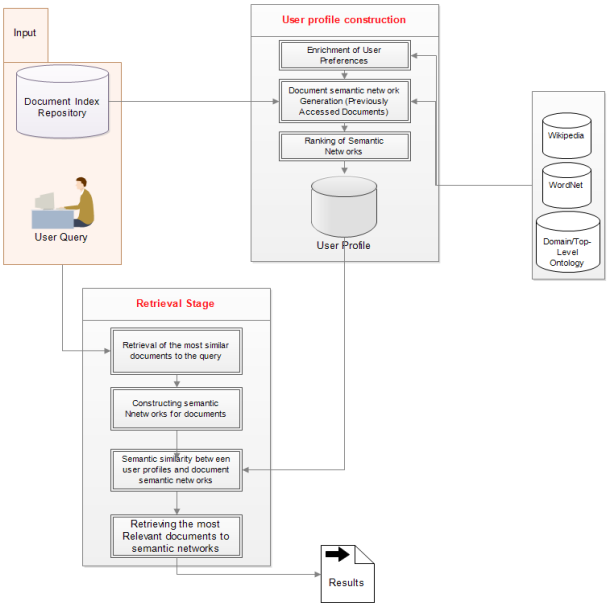

Figure 7. The process of personalized document retrieval using user profiles

3.3.6.1 The semantic relation robustness measure

This measure determines the robustness of established relations between concepts. When multiple relations have the same 'subject', this measure determines which relation is more effective in describing the 'subject'. It should be noted that 'subject' is the informative element in a relation between concepts. Assuming d_i and rel_i represent the document vector and the set of relations between them in a semantic network, respectively, a generated semantic network by $SN(d_i) = [(d_i), rel_i]$. Also, if it is assumed that (t_j, rel, t_k) represents a semantic relation between a subject t_j and an object t_k in document d_i, hence the set rel_i can be written as $rel_i = \{(t_j, rel, t_k) | t, t_k \in (d_i)\}$. In this case, the discriminatory power of a relation such as (t_j, rel, t_k) is obtained using the following equation:

$$score_{discriminity}(SN(d_i), UP) =$$

$$\frac{\sum_{all\ the\ triplets} score_{discriminity}((t_j, rel, t_k))}{number\ of\ triplets\ in\ the\ semantic\ net.}$$

$$score_{discriminity}\left((t_j, rel, t_k)\right) =$$

$$1 - \left(\frac{10 - \left| \ln\left(\frac{2 * (SR_{doc}(t_j) + 1) * (SR_{UP}(t_j) + 1)}{2 + (SOR_{doc}(t_j, t_k) + SOR_{UP}(t_j, t_k))} \right) \right|}{10} \right)$$

SR=$(Subject{-}relation)$, SOR=$(Subject{-}Object{-}relation$N$)$, UP=$(User\ Profile)$, doc=$(document)$

(9)

Where t_j and t_k represent the subject and object of a relation in semantic networks, $SR_{doc}(t_j)$ and $SR_{UP}(t_j)$ represents the number of the relations with t_j as the subject, $SOR_{doc}(t_j, t_k)$ and $SOR_{UP}(t_j, t_k)$ represents the number

of the relations with $t_j$as the subject and t_k as the object in documents and queries, respectively. Therefore, a relation such as $(t_j, rel, *)$ with high values of *SR* and low value of *SOR* is more robust in describing the subject t_j. It can be concluded that this relation is a descriptive relation for subject t_j. The documents that share highest number of descriptive relations with user query have the highest similarity score.

3.3.6.2 The semantic relation effectiveness measures

These metrics measure how much the semantic network of a document is effective in covering the information content of the user profile based on semantic network.

The explicit measure calculates the amount of shared information content between the occurring relations in the documents and the user preferences and how much the semantic network of a document is similar to a user profile. This measure is computed as follows:

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

$$Score_{Term_{\text{explicit}}}(0) = \begin{cases} \delta_{exp}, & t_j, t_k\ are\ in\ user\ profile \\ 1-\delta_{exp} & o.w. \end{cases} \quad (10)$$

In this equation, δ_{exp} is a threshold between [0, 1]. When the *subject* t_j and *object* t_k of a relation in the document semantic network appears in the user profile, a high similarity score is assigned.

The implicit measure evaluates the document semantic network and how much it resembles the semantic network representation of the user preferences:

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

$$Score_{relation_{implicit}}\left(\left(t_j, rel, t_k\right)\right) = \begin{cases} \delta_{imp}, & (t_j, rel, t_k)\ is\ in\ user\ profile \\ 1-\delta_{imp} & o.w. \end{cases} \quad (11)$$

Where δ_{exp} is a threshold between [0, 1]. When a relation with *subject* t_j and *object* t_k appears in the user profile, a high similarity score is assigned.

3.3.6.3 Semantics-based measures

Semantic features of textual resources are the most informative portion of information content. Computing the amount of commonalities and/or differences in semantic features between two semantic networks can be a good indicative of similarity between them.

WordNet-based semantic similarity measure: This method is based on the notion of Information Content (IC) of the Least Common Subsumer (LCS) [58]. 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. Higher commonalities in semantic features indicate higher similarity score. This method is called normalized Jiang and Conrath measure and is calculated as follows [74]:

$$WordNet_{Score}(A, B) = 1 - \left(\frac{\left[IC_{nrm}(A) + IC_{nrm}(B) - 2 * IC_{nrm}\left(LCS(A, B)\right)\right]}{2}\right) \quad (12)$$

Wikipedia-based semantic similarity measure: It computes the semantic similarity between the two concepts based on the commonalities and differences in their respective second-order and first-order vectors. For this purpose, Lin's information theoretic measure [31,40] is utilized:

$$Wiki_{Score}(A, B) = \frac{\sum_{(rel,\ w')} freq(A,\ *_{rel},\ *_w) + freq(B,\ *_{rel},\ *_w)}{\sum_{(rel,\ w')} freq(A, *_{rel}, B) + \sum_{(rel,\ w')} freq(B, *_{rel}, A)}$$

$$*\ rel = \{co - occurrence\ rlation, contextually\_similar\ relation\}$$

$$*\ w = \{concepts\ in\ either\ the\ document\ or\ user\ profile\} \quad (13)$$

Where *A* and *B* are concepts in document and user query respectively. The $freq()$ function calculates the frequency of *A* or/and *B* in these relations. According to Lin's information theoretic measure, the similarity between concept A and B is related to the commonalities and differences between them. Higher level of commonality means higher similarity score for two concepts. This measure is somewhat similar to latent semantic analysis, especially the one applied in [75].

For WordNet-based and Wikipedia-based measures, the notion of semantic similarity between two concepts is used to compute the similarity between the user profiles semantic network and the semantic network of documents. These two methods compute the semantic similarity between all the possible pair of concepts in the document semantic network and the user profile semantic network and generate a number between (0-1), which indicates the similarity score.

Finally, to compute the overall semantic similarity, a linear and weighted combination of these measures is used as follows:

$$Similariyu_{Score}(SN(d_i), UP) = \left(k_1 * score_{discriminity}(SN(d_i), UP)\right) + \left(k_2 * Score_{Explicit}\left(\bigcup\left(t_j, rel, t_k\right) \in SN(d_i)^{\forall t_j \in d_i}_{\forall t_k \in d_i}\right)\right)$$

$$+\left(k_3 * Score_{implicit}\left(\bigcup(t_j, rel, t_k) \in SN(d_i)^{\forall t_j \in d_i}_{\forall t_k \in d_i}\right)\right)$$

$$+\left(k_4 * \frac{\sum_{\forall \boldsymbol{A} \in d_i}^{\forall \boldsymbol{A} \in d_i} Sal_{Score}(A) * WordNet_{Score}(A,B)}{\sum_{\forall \boldsymbol{A} \in d_i} Sal_{Score}(A)}\right)$$

$$+\left(k_5 * \frac{\sum_{\forall \boldsymbol{A} \in d_i}^{\forall \boldsymbol{A} \in d_i} Sal_{Score}(A) * Wiki_{Score}(A,B)}{\sum_{\forall \boldsymbol{A} \in d_i} Sal_{Score}(A)}\right) \quad (14)$$

Input: the preliminary simple indexes D={D_1, D_2,…D_n}, user queries and previously accessed documents

- Loop: for each previously accessed documents by the user
  - Construct the semantic network and combine it with other networks and construct user profile
- End of Loop
- Loop: for each document in D
  - Retrieve the most similar preliminary indexes to the user query.
- End of Loop
- Loop: for each retrieved document
  - Construct the semantic network
- End of Loop.
- Compute the semantic similarity between document semantic networks and user profile semantic network.
- Rank the document according to the similarity of their semantic network to user profile semantic network.

Output: Retrieved documents based on the user preferences.

Figure 8. The pseudo-code for semantic indexing and Retrieval system

Where k_1, k_2, k_3, k_4 and k_5 are the weighting parameters between (0-1), while their sum is equal to 1. In the final step, documents are ranked according to their similarity with the user profile and the results are displayed to the user. Figure 8. Illustrates the proposed semantic information indexing and management algorithm.

4. Evaluation

4.1 The experiments and the data collection process

In order to evaluate the proposed method, a series of 3 experiments with different parameters are considered. The first experiment evaluates the accuracy of the semantic network modelling of modules and the semantic similarity of computation modules. It also measures their effect on the overall performance of the proposed method. The second experiment evaluates the efficiency of the system in identifying the most similar documents with regard to the user preferences. The third experiment evaluates the effectiveness of the proposed method in predicting the correct topic classification of documents.

Table 2. The topics and their constituent classes in the 20 newsgroup dataset

#	Topics	The Constituent classes
1	Computer	Comp.graphics Comp.os.windows.misc Comp.sys.ibm.pc.hardware Comp.sys.mac.hardware Comp.windows.x
2	Recreation	Rec.autos Rec.motorcycles Rec.sport.hockey Re.sport.baseball
3	Science	Sci.crypt Sci.electronics Sci.med Sci.space
4	MISC	Misc.forsale
5	Politics	Talk.politics.gun Talk.politics.mideast Talk.politics.misc
6	Religion	Talk.religion.misc Alt.atheism Soc.religion.christian

One of the most important and widely used text dataset in the field of text mining and related applications is the *20Newsgroup* dataset [76]. This dataset consists of 19997 news articles and web pages categorized in twenty different classes (or newsgroups). In the last update (released in 2008), several existing duplications were removed; thus, the number of unique documents in the dataset was reduced to 18827. Consequently, the number of unique concepts/words which occurs more than once in the dataset is equal to 71830. Since some of the classes in the dataset are contextually related to each other, the documents can be classified in broader categories called *topics*. Table 2 illustrates the topics and their constituent classes in the *20Newsgroup* dataset.

4.2 The evaluation of the semantic network generation and semantic similarity modules

For performance evaluation, a set of 4000 documents of *20Newsgroup* was selected randomly. These documents were categorized in five different topics. Out of 4000, 800 documents were categorized in the “sciences” topic, 800 were in the “computer”, 800 documents were in “politics”, 800 documents were in “religion”, and 800 eighty documents were in the “recreation” topic.

Furthermore, 10 different tests were designed to evaluate the performance and precision of the proposed method. In other words, for each topic in dataset, two experiments are designed. These tests are designed to evaluate two important component of the system: (1) evaluating the semantic network generation of modules, and (2) evaluating the semantic similarity of modules. Also, the performance of the proposed method is compared to other similar approaches.

The details of the designed test are described here. For each test, 800 documents out of 4000 are selected randomly from the respective topic. The remaining 3200 documents are selected from the other four topics which are completely irrelevant to the respective topic.

For each topic, two different queries are extracted from the documents in the respective topics. In order to do this,

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 based on this analysis and presented to the experts. The experts select the concepts/words that can describe the underlying topic the best and the preliminary queries for each topic are formed. In the next step, all the queries are enriched. The main objective of these tests is to evaluate the semantic indexing and management capabilities of the proposed system when faced with different queries. In addition, two documents, which are deemed most similar to the subject of underlying topic by the experts, are selected to act as user's previously accessed documents.

In this paper, Mean Average Precision (MAP) is used to evaluate the performance of the proposed approach. The MAP value is the arithmetic mean of the average precision values for each query [54]. Meanwhile, MAP has been known for its good discrimination and robustness [77]. For a given queryq_i, the map value is calculated as follows:

$$MAP_i = \frac{1}{m}\sum_{k=1}^{m} Precision\ (R_k) \quad (15)$$

Where m is the number of the retrieved documents, R_k is the set of ranked retrieval results from the top until the k-th retrieved document. Table 3 depicts the queries for each test.

4.2.1 Evaluating the semantic network generation process

When document semantic networks are generated, the underlying assumption for computing the salient score for each concept is that the information content of the documents can be represented more efficiently by a portion of concepts that are most informative. In order to verify this assumption, different percentages of the salient concepts are utilized to generate the semantic networks. Then, the semantic similarity between the document semantic networks and the user profiles is measured. In the end, the precision and performance of the proposed method is evaluated using the average MAP score of the 10 queries. Also, in order to demonstrate that the salient score achieves much better results than the CF-IDF weighting method, the performance of the proposed method with the salient score and the proposed method with CF-IDF instead of the salient score are compared and evaluated. The results are illustrated in Figures 9 and 10.

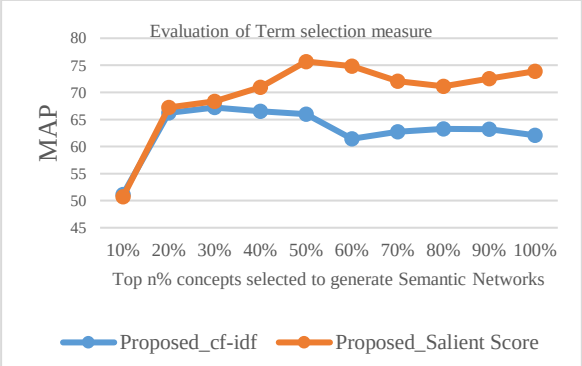

Figure 9. Evaluation of term selection measure

Table 3. Queries for each test

#	# of related documents	The constituent keywords of each query	The appended concepts after the enrichment (Expanded Query)	Class
1	800	Bible, Christian, belief, theist, heaven, faith, History	Gospel, Quran, ideology, doctrine, agnostic, hell, soul	Religion
2	800	God, Atheism, religion, Jesus, devil	Worship, Christ, faith, Christianity, theism	Religion
3	800	Car, bike, auto, , vehicle, Brake, oil , driver	Truck, engine, automobile, steer, gear, fuel, rider	Recreation
4	800	season, hit, pitch, team, score, catcher, baseball	Softball, varsity, pitcher, baseman, Yankees, infielder	Recreation
5	800	Space, launch, technology, orbit, satellite, research milky way	Spacecraft, NASA, system, engineering, innovation, space shuttle, radar, transponder	Science
6	800	Health, patient, medical clinical, disease, diagnosis	Mental health, education, nutrition, symptom, treatment	Science
7	800	Gun, police, violence, rifle, drug, victim	Conflict, racism, riot, weapon, ammunition, law enforcement	Politics
8	800	Military, war, government, building, assault, crowd	Civilian, air force, naval, troop, invasion, army	Politics
9	800	Software, graphic, render, shader, display, interface	Computer, open source, processor, polygon, image, VGA	Computer
10	800	Computer, system, hardware, device, storage, driver	Processor, computer, decoder, server, data, disk, application	Computer

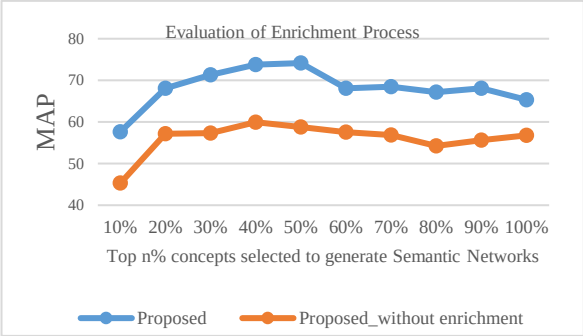

Figure 10. Evaluation of enrichment process

In the next step, we evaluate whether the enrichment process have any effect on the generated semantic networks and consequently on the accuracy of the final results. Therefore, the performance of the proposed method with enrichment module and the proposed method without enrichment module are compared and evaluated.

As illustrated in Figure 11, the underlying assumption holds true again and the documents can be represented more efficiently by the top 50% of the prominent concepts. Also, the proposed method coupled with the enrichment module performs far better than the proposed method without the enrichment module. As mentioned earlier, one of the most important components of the proposed method is the enrichment module. The content enrichment component helps the system identify the Liaison features that link the separated concept clusters, which forms a connected semantic network.

4.2.2 Evaluating the performance and the efficiency of the semantic similarity measure

In this section, a series of tests are designed to evaluate the performance and efficiency of the proposed semantic similarity module. Five tests for different settings are prepared to determine the effectiveness of different components of the proposed hybrid similarity measure and their effect on overall precision. This is done by computing the average MAP scores of the queries. The settings of the designed tests are illustrated in Table 4.

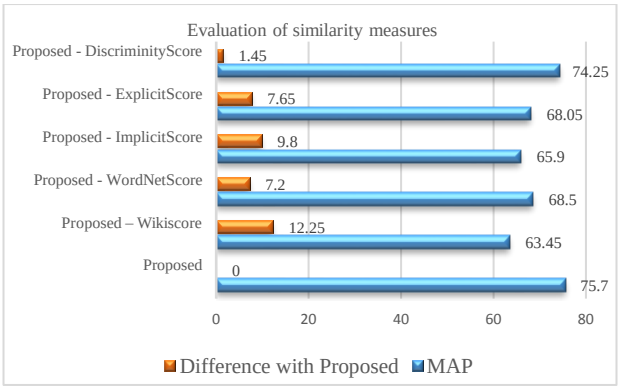

Figure 11. Evaluation of relevance/similarity measures

The illustrated results suggest that among the contributing components of the similarity module, the Wikipedia-based components (Wikiscore) have the greatest effect on the precision and efficiency of the proposed method. The relation-based components (explicit score and implicit score), WordNet-based components (WordNet score) and structural component (Discriminity score) are in the next places, respectively. The results were somewhat expected as the semantic constructs and relation-based structure of ontology, WordNet and Wikipedia KBs make them perfect tools for computing the semantic similarity.

Table 4. The settings of the designed test for evaluating the effectiveness of the proposed hybrid similarity measure

#	Experiment	Description of the Experiment
1	Proposed	The complete process of semantic similarity
2	Proposed – Wikiscore	The process of semantic similarity without Wikiscoreth
3	Proposed - WordNetScore	The process of semantic similarity without WordNetScore
4	Proposed - ExplicitScore	The process of semantic similarity without ExplicitScore
5	Proposed - ImplicitScore	The process of semantic similarity without ImplicitScore
6	Proposed - DiscriminityScore	The process of semantic similarity without DiscriminityScore

4.2.3 Exploring other existing Approaches

Two different approaches are selected here; we have implemented these approaches based on the description of their ranking formula. The first approach is called *Lucene scoring function* [61], which uses a combination of vector space model and Boolean matching model to rank and retrieve the most similar documents to user preferences. More details about Lucene scoring function is depicted in [61, 54]. The second approach is proposed by Daoud. et al. [18]. This method introduces a personalized ontology-based ranking approach. The documents and user profiles are represented by graph structures and the relations between the concepts are established using a web ontology. Then, a graph-based distance measure computes the similarity between document graphs and user profile graph. More details about this ranking method is depicted in [18]. We have also already attempted to compare the performance of the proposed method with the relation-based ranking approach in another study [54]. However, since the utilized ontology is a modified one which is not available to the public, the comparison was not possible in this study. The comparison results are obtained by computing the average MAP score for 10 queries. The results are illustrated in Figure 12.

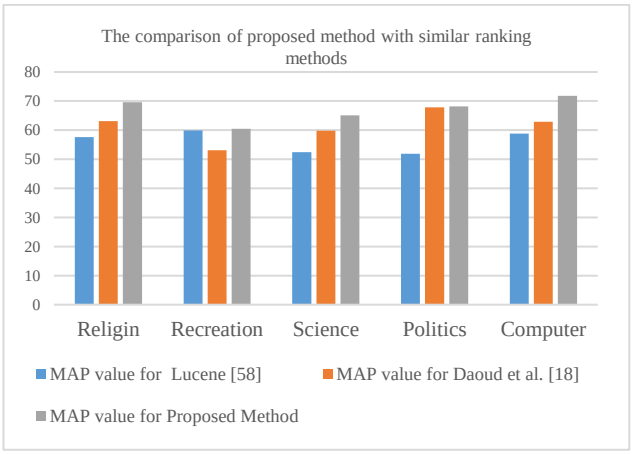

Figure 12. Comparison of the proposed method with similar methods

As illustrated in Figure 12, the proposed method outperforms its similar methods of ranking and retrieval and also exhibits high performance and precision. An interesting fact about the illustrated results is the relatively good performance of the Lucene method in "Recreation" Topic. The recorded results suggest that the Lucene Ranking method performs well, especially when most of the documents can be retrieved using only the keywords in the documents.

4.3 Evaluating the performance of the system in topic classification

In the next step, the capability of the proposed method in topic classification of documents is assessed. To this end, five tests are prepared. Each test is designed to evaluate the capability of the proposed method in classifying documents in the correct topic. In order to do this, 10000 documents are selected randomly from the *20Newsgroup* dataset. Out of the 10000, 2000 document are from the "computer" topic, 2000 from the "religion" topic, 2000 from the "politics" topic, 2000 from the "recreation" topic and the rest are from the "science" topic. The first test assesses the performance of the proposed method in classifying documents from the "computer" topic. In this test, the selected documents from the "computer" topic are labelled "relevant" and the 8000 remaining documents from the other topics are labelled "irrelevant" to the "computer" topic. The second, third, fourth and fifth tests are designed to evaluate the performance of the proposed method in classifying the documents from the "religion", "politics", "recreation" and "science" topics respectively. In order to generate the user profiles that reflect the information content of each topic, the same procedure applied in section 6.2 was also used here. The designed queries for each topic are illustrated in Table 5.

It should be noted that only the Wikipedia-based enriched concepts are displayed in this table and other enriched information are not displayed.

The evaluation mechanism in this step is describe below. First, for each topic in the dataset, the semantic network of user profile is created according to the user queries and user's previously accessed documents. Next, the semantic similarities between the semantic networks of the documents and the user profiles are computed. This process results in a set of five similarity scores for each document in the test dataset; each score indicatd the degree of similarity between a document and one of the topics in the dataset. In the next step, each document is classified in terms of the topic with the highest semantic similarity score. According to the obtained results, the documents are classified and labelled as TP (true positive), TN (true negative), FP (false positive) and FN (false negative). Finally, in order to measure the performance of the system in terms of topic classification, the following measures are taken.

The evaluation results are illustrated in Table 7 and 8 below.

Table 5. The queries needed to retrieve the most similar documents for each topic

#	# of documents	The constituent concepts of each query	The appended concepts after the enrichment (Expanded Query)	class
1	2000	Bible, Christian, belief, theist, heaven, faith, History God, Atheism, religion, Jesus, devil	Gospel, Quran, ideology, doctrine, agnostic, hell, soul Worship, Christ, faith, Christianity, theism	Religion
2	2000	Car, bike, auto, , vehicle, Brake, oil , driver season, hit, pitch, team, score, catcher, baseball	Truck, engine, automobile, steer, gear, fuel, rider Softball, varsity, pitcher, baseman, Yankees, infielder	Recreation
3	2000	Space, launch, technology, orbit, satellite, research, milky way Health, patient, medical, clinical, disease, diagnosis	Spacecraft, NASA, system, engineering, innovation, space shuttle, radar, transponder Mental health, education, nutrition, symptom, treatment	Science
4	2000	Gun, police, violence, rifle, drug, victim Military, war, government, building, assault, crowd	Conflict, racism, riot, weapon, ammunition, law enforcement Civilian, air force, naval, troop, invasion, army	Politics
5	2000	Software, graphic, render, shader, display, interface Computer, system, hardware, device, storage, driver	Computer, open source, processor, polygon, image, VGA Processor, computer, decoder, server, data, disk, application	Computer

Table 6. Evaluating the performance of the system

Selected As	True Label	
	Relevant	Irrelevant
Relevant	True Positive	False Positive
Irrelevant	False Negative	True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (17)$$

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (20)$$

Table 7. The evaluation results in terms of TP, TN, FP, and FN

Test	Topics	TP	TN	FP	FN
Test #1	Computer	1936	7958	42	64
Test #2	Religion	1959	7909	91	41
Test #3	Politics	1903	7926	74	97
Test #4	Recreation	1877	7784	216	123
Test #5	Science	1782	7634	366	218

Table 8. The evaluation results in terms of accuracy, precision, recall and F-measure

Test	Topics	Accuracy	Precision	Recall	F-measure
Test #1	Computer	98.94%	97.88%	96.8%	97.34%
Test #2	Religion	98.68%	95.56 %	97.95%	96.74%
Test #3	Politics	98.29%	96.26%	95.15%	95.7%
Test #4	Recreation	96.61%	89.68%	93.85%	91.72%
Test #5	Science	94.16 %	82.96 %	89.1%	85.92 %
Mean Performance		97.34%	92.47%	95.09%	93.48%

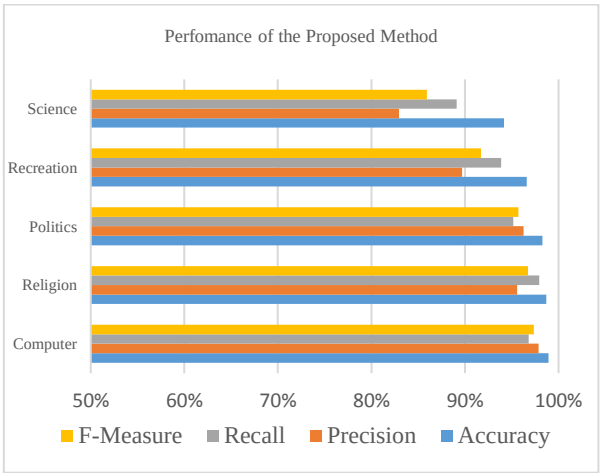

Figure 13. The evaluation results

Table 9. The results of hypothesis testing for each topic in the test dataset
(The number of relevant/irrelevant documents in each topic: 2000/8000, the significance level=5% (0.05))

Test	Topics	Observed Mean	Observed Standard deviation	p-value	Null Hypothesis
Test #1	Computer	-0.6040	0.7974	0.9109	Accepted
Test #2	Religion	-0.5940	0.8049	0.8673	Accepted
Test #3	Politics	-0.5980	0.8019	0.9555	Accepted
Test #4	Recreation	-0.5740	0.8193	0.4729	Accepted
Test #5	Science	-0.5660	0.8248	0.3497	Accepted

As it is illustrated, the proposed method exhibits high accuracy, recall and precision rate in the classifying documents in the "computer", "religion" and "politics" topics. Also, the lowest accuracy, precision and recall rates are recorded in the classifying documents in the "science" and "recreation" topics (although the recorded recall and accuracy rates are quite suitable for text mining applications). After investigating the documents in these topics, it was found that there was a relatively clear distinction in the content of these topics. Therefore, we came to the conclusion that the distinction in the documents might be the reason for these results. In other words, when the documents of a topic discuss subjects related to other topics, lower results (esp. lower precision rate) might be expected.

4.4 The evaluation of the effectiveness and reliability of the proposed method in predicting the correct topic classification of documents

In the last step, the effectiveness of the proposed method in predicting the correct topic classification of the documents in each topic is assessed. This is carried out through hypothesis testing. In this step, 10000 documents from *20Newsgroup* dataset are randomly selected. The test dataset consists of 2000 documents per topic. The manner in which the hypothesis testing is conducted will be explained for one topic and the hypothesis testing for other topics will be conducted the same way. Assuming that the user preferences is closely related to the content the "computer" topic, the semantic network of the user profile is created using two documents from this topic. These documents reflect the information content of the "computer" topic very well. In the next step, the documents in the "computer" topic are assigned to the label "1" and the documents in other topics are assigned the label "-1". In the next step, the semantic similarity found between the documents and the user profiles of each topic is computed. If the similarity of a given document to the "computer" topic is higher than the other topics, the prediction label "1" is assigned to this document, otherwise the prediction label "-1" is assigned to this document. 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 as the correct topic, otherwise the topic classification of the document is incorrect.

Hypothesis testing to evaluate the effectiveness of the proposed method in predicting the correct topic classification of document in the "computer" topic: In order to conduct the hypothesis testing on the randomly sampled test data, the two-sample t-test is performed. It should be noted that the optimal value (correct prediction label) for documents relevant to the "computer" topic is 1 and the optimal value of irrelevant ones is -1. The mean and sample standard deviation of the computed prediction labels for test documents is found to be -0.6040 and 0.7974, respectively. The purpose of two-sample t-test is to test whether the means of two different populations, namely the population of true labels and the population of prediction labels, are equal or not. The two-sample t-test does not assume the equality of variances. The null hypothesis is formulated in the following:

H_0:

The data of both populations come from independent random samples of normal distribution with equal means without assuming that the populations have also equal variances (i.e., the proposed method is capable and effective in predicting the correct topic classification of the documents in the "computer" topic).

H_1:

The null hypothesis is rejected. That is, tthe proposed method is not capable and effective in predicting the correct topic classification of the documents and the results may have been obtained by random chance in the sample selection process.

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (23)$$

In this equation, $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1 and s_2 is the sample standard deviation, and n_1 and n_2 include the sample size. The following results are obtained for the hypothesis testing on the "computer" topic:

p-value = 0.9109. As the *p*-value is greater than the significance level, the null Hypothesis must be accepted. In other words, both populations come from a normal distribution with equal means. This suggests that the proposed method is in fact capable and effective in predicting the correct topic classification of the documents in the "computer" topic.

The result of the second, the third, the fourth and the fifth tests are illustrated below in Table 9.

The illustrated results in Table 9 suggest that the proposed method is both effective and reliable in identifying the correct topic classification of documents in each topic (i.e., the null hypothesis is accepted in all cases), and the results have not been obtained by random chance during the

sample selection process.

5. Conclusion

In this paper, a novel method of semantic information indexing and management is introduced. The proposed method is developed by integrating the structured knowledge of ontology and KBs (esp. Wikipedia and WordNet) in every component of the proposed method. The documents and user profiles are represented by semantic network graphs. The main characteristics of the proposed method are the semantic, ambiguity-free and multi-level representation of the contents. In addition, the properties of semantic networks are applied to identify the documents similar to the user preferences. As mentioned earlier, the main contribution and novelty of the proposed method include (1) the integration of the structured knowledge in every component of the system, (2) utilizing the semantic networks for a unified and multi-level representation of textual resources, (3) introducing a hybrid weighting schema called the salient score, and (4) proposing a hybrid semantic similarity measurement.

The proposed method is evaluated in three stages using the *20Newsgroup* dataset. In the first stage, different components of the proposed system and their effect on the overall performance and efficiency is evaluated. The evaluation results suggest that (1) employing a portion of the most prominent concepts (top-50% with the highest salient score) to generate the semantic networks achieves the highest accuracy and precision, (2) the salient score (weighting schema) achieves better results compared with the CF-IDF weighting method. In other words, employing a semantic and relation-based weighting schema results in higher precision compared with a term/frequency-based weighting method, (3) the enrichment module have a significant effect on the overall precision and performance of the proposed method and (4) among the contributing components of the hybrid similarity measurement, the semantic components ($Wiki_{score}$ and $WordNet_{score}$) and relation-based components (explicit scores and implicit scores) have the greatest effect on the precision and efficiency of the proposed method. Then, the capability of the proposed method in topic classification of documents is assessed. The proposed method is evaluated using accuracy, precision, recall and F-measure metrics. The results suggest that the proposed method exhibits high accuracy, recall and precision rate and performs well in topic classification of documents. In the last evaluation stage, the effectiveness and reliability of the proposed method in predicting the correct topic classification is assessed through hypothesis testing. The evaluation results suggest that the proposed system is both effective and reliable in classifying documents in their correct topic and that the results have not been obtained by random chance during the sample selection process.

As for avenues for future research, it is suggest that the direct integration of deep learning-based word embedding models such as seq2seq and word2vec into the proposed information processing, indexing and management system be investigated. One of the important features of the proposed system is the indexing of text resources by semantic networks which can be used to index web resources; therefore, future studies can aim at exploring knowledge-based approaches to improve vector space models and use improved models in machine learning methods.

Finally, it should be noted that the high performance and accuracy of the proposed method can be interpreted as a new and effective method in semantic information indexing and management; however, it does not mean that the proposed method is the best information indexing and management framework ever developed. It just can be considered as a successful implementation of an information management system, which is perhaps suitable for semantic information indexing and management and the other related applications.

References

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

[2] Liu, B., "Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag Berlin Heidelberg, 2007.

[3] Bouadjeneka, M. R., Hacidc, H., Bouzeghoubd, M., "Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms", *Information Systems*, vol. 56, pp. 1–18, 2016.

[4] Baeza-Yates, R. A., Ribeiro-Neto, B., "Modern Information Retrieval", 2nd edition, Addison-Wesley Longman Publishing Co., 2010.

[5] Belkin, N. J., "Some(what) grand challenges for information retrieval", SIGIR Forum, vol. 42, p. 47–54, 2008.

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

[7] Kolomiyets, O., Moens, M-F., "A survey on question answering technology from an information retrieval perspective", *Information Sciences*, vol. 181, pp. 5412–5434, 2011.

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

[9] Jayaratne, M., Haththotuwa, I., Arachchi, C. D., Perera, S., Fernando, D., Weerakoon, S., "iSeS: Intelligent semantic search framework", *In Proceedings of 6th Euro American Conference on Telematics and Information Systems (EATIS)*, 2012.

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

[11] Bansal, R., Jyoti, B. K. K., "Ontology-Based Ranking in Search Engine", In: Aggarwal V., Bhatnagar V., Mishra D. (eds), *Big Data Analytics. Advances in Intelligent Systems and Computing*, vol. 654, pp. 97-109, 2018.

[12] W.B. Croft, J. Lafferty, J., "Language Modeling for Information Retrieval", Kluwer Academic Publishers, 2013.

[13] Crestani, F., de Campos, L., Fernandez-Luna, J., Huete, J., "Ranking structured documents using utility theory in the Bayesian Network retrieval model", *Lect. Notes Comput. Sci.*, vol. 2857, pp. 168–182, 2003.

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

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

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

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

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

[19] Uthayan, K. R., Anandha Mala, G. S., "Hybrid Ontology for Semantic Information Retrieval Model Using Keyword Matching Indexing System", *The Scientific World Journal*, vol. 2015, pp. 1-9, 2015.

[20] Tarus, J. K., Niu, Z., Yousif, A., "A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining", *Future Generation Computer Systems*, vol. 72, pp. 37-48, 2017.

[21] Mirończuk, M., Protasiewicz, J., "A recent overview of the state-of-the-art elements of text classification", *Expert Systems with Applications*, vol. 106, pp. 36-54, 2018.

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

[23] Lease, M., "An Improved Markov Random Field Model for Supporting Verbose Queries", *In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, 2009.

[24] Metzler, D., Croft, W.B., "A Markov random field model for term dependencies", *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR*, ACM Press, 2005.

[25] Lease, M., Allan, J., Croft, W. B., "Regression Rank: Learning to Meet the Opportunity of Descriptive Queries", *In Proceedings of the 31st European Conference on Information Retrieval (ECIR)*, 2009.

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

[27] Pérez-Agüera, J. R., Arroyo, J., Greenberg, J., Iglesias, J. P., Fresno, V., "Using BM25F for semantic search", In Proceedings of the 3rd International Semantic Search Workshop on – SEMSEARCH, ACM Press, 2010.

[28] Pinheiro de Cristo, M. A., Calado, P. P., de Lourdes da Silveira, M., Silva, I., Muntz, R., Ribeiro-Neto, B., "Bayesian belief networks for IR", *International Journal of Approximate Reasoning*, vol. 34, pp. 163–179, 2003.

[29] Zhang, J., Yuan, H., "A comparative study on collectives of term weighting methods for extractive presentation speech summarization", *In Proceedings of IALP: International Conference on Asian Language Processing*, 2015.

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

[31] Lastra-Díaz, J. J., García-Serrano, A., "A new family of information content models with an experimental survey on WordNet", *Knowledge based systems*, vol. 89, pp. 509–526, 2015.

[32] Wei, T., Lu, Y., Chang, H., Zhou, Q., Bao, X., "A semantic approach for text clustering using WordNet and lexical chains", *Expert Systems with applications*, vol. 42, pp. 2264–2275, 2015.

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

[34] Ferruci, D., Lally, Uima, A., "an architectural approach to unstructured information processing in the corporate research environment", *Natural Language Engineering*, vol. 10, pp. 327–348, 2004.

[35] Etzioni, O., Cafarella, M. J., Downey, D., maria Popescu, A., Shaked, T., Soderland, S., Weld, D. S., Yates, A., "Unsupervised named-entity extraction from the web": an experimental study", *Artificial Intelligence*, vol. 165, pp. 91–134, 2005.

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

[37] Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D., "Semantic annotation, indexing, and retrieval", *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, pp. 49–79, 2004.

[38] Mooney, R. J., Bunescu, R., "Mining knowledge from text using information extraction", SIGKDD Explorations Newsletter, vol. 7, pp. 3–10, 2005.

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

[40] Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y., "Learning to Harvest Information for the Semantic Web", In Proceedings of the 1st European Semantic Web Symposium (ESWS-2004), 2004.

[41] Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L., Mylopoulos, J., "Cerno: light-weight tool support for

semantic annotation of textual documents", Data and Knowledge Engineering, vol. 68, pp. 1470–1492, 2009.

[42] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., Yates, A., "Web-scale information extraction in know-it-all: (preliminary results)", *In Proceedings of WWW '04: the 13th International Conference on World Wide Web*, ACM, 2004.

[43] Ramakrishnan, C., Kochut, K., Sheth, A. P., "A framework for schema-driven relationship discovery from unstructured text", *In Proceedings of International Semantic Web Conference*, 2006.

[44] Xu, C., Wang, J., Wan, K., Li, Y., Duan, L., "Live sports event detection based on broadcast video and web-casting text", *In Proceedings of the Fourteenth annual ACM international conference on Multimedia*, ACM, 2006.

[45] Saggion, H., Cunningham, H., Bontcheva, K., Maynard, D., Hamza, O., Wilks, Y., "Multimedia indexing through multi-source and Multilanguage information extraction: The MUMIS project", *Data and Knowledge Engineering*, vol. 48, pp. 247–264, 2004.

[46] Yang, Y., Li, L., "Research on sports game news information extraction", *In proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, 2007.

[47] Wessman, A., Liddle, S. W., Embley, D. W., "A generalized framework for an ontology-based data-extraction system", *In Proceedings of Fourth International Conference on Information Systems Technology and its Applications*, 2005.

[48] Gangemi, A., Catenacci, C., Battaglia, M., "Inflammation ontology design pattern: an exercise in building a core biomedical ontology with descriptions and situations", in D.M. Pisanelli (Ed.), Ontologies in Medicine, IOS Press, 2004.

[49] Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Baumann, S., Vembu, S., Romanelli, M., Buitelaar, P., Engel, R., Sonntag, D., Reithinger, N., Loos, B., Zorn, H.-P., Micelli, V., Porzel, R., Schmidt, C., Weiten, M., Burkhardt, F., Zhou, J., "DOLCE ergo SUMO: on foundational and domain models in the Smart-Web integrated ontology (SWIntO)", *Journal of Web Semantics*, vol. 5, pp. 156–174, 2007.

[50] Muller, H.-M., Kenny, E. E., Sternberg, P.W., "Textpresso: an ontology-based information retrieval and extraction system for biological literature", *PLoS Biology*, vol. 2, pp. 1984-1998, 2004.

[51] Tsinaraki, C., Polydoros, P., Christodoulakis, S., "Interoperability support between mpeg-7/21 and owl in ds-mirf", *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 219–232, 2007.

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

[53] Sun, S., Song, W., Zomaya, A. Y., Xiang, Y., Choo, K. K. R., Shah, T., Wang, L., "Associative retrieval in spatial big data based on spreading activation with semantic ontology", *Future Generation Computer Systems*, vol. 76, pp. 499-509, 2017.

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

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

[56] Liu, F., Yu, F., Meng, W., "Personalized web search for improving retrieval effectiveness", *IEEE Transaction on Knowledge and Data Engineering*, vol. 16, pp. 28–40, 2004.

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

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

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

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

[61] Langer, S., Beel, J., "Apache Lucene as Content-Based-Filtering Recommender System: 3 Lessons Learned", 5th International Workshop on Bibliometric-enhanced Information Retrieval, BIR2017, 2017.

[62] Zanger, D. Z., "Interpolation of the extended Boolean retrieval model", *Information Processing and Management*, vol.38, pp. 743–748, 2002.

[63] Moral, C., de Antonio, A., Imbert, R., Ramírez, J., "A survey of stemming algorithms in information retrieval", *Information Research: An International Electronic Journal*, vol. 19, pp. 2014.

[64] Bounabi, M., Moutaouakil, K. E., Satori, K., "A comparison of Text Classification methods Method of weighted terms selected by different Stemming Techniques", *In Proceedings of BDCA: international Conference on Big Data*, Cloud and Applications, 2017.

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

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

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

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

[69] Baziz, M., Boughanem, M., Traboulsi, S., "A Concept-based Approach for Indexing in IR", In Proceedings of INFORSID, 2005.

[70] Biemann, C., Ponzetto, S. P., Faralli, S., Panchenko, A., Ruppert, E., "Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation", In Proceedings of European Chapter of the Association for Computational Linguistics, 2017.

[71] Liu, B., "Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag Berlin Heidelberg, 2007.

[72] Malo, P., Siitari, P., Ahlgren, O., Wallenius, J., Korhonen, P., "Semantic Content Filtering with Wikipedia and Ontologies", *In Proceedings of the 2010 IEEE International Conference on Data Mining Workshops (ICDMW'10).* IEEE Computer Society, 2010.

[73] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M. van Kleef, P., Auer, S., Bizer, C., "DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", *Semantic Web Journal*, vol. 6, pp. 167-195, 2015.

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

[75] Kontostathis, A., Pottenger, W., "A Framework For Understanding Latent Semantic Indexing (LSI) Performance", *information Processing and Management, Special issue: Formal methods for information retrieval*, Vol. 42, 56-73, 2006.

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

[77] Manning, P., Raghavan, H., Schutze, "Introduction to Information Retrieval", Cambridge University Press, 2008.