An Efficient Information-Rich Representation Scheme for Information Access and Knowledge Acquisition

نظام لتمثيل المعلومات الغني الفعال للحصول على المعلومات واكتساب المعرفة

Asmaa M. El-Said and Hesham A. Arafat

الملخص

النموالهائل في عدد الوثائق النصية المنتجة يوميا تحتاج للتنمية الفعالة لاستكشاف وتحليل واكتشاف المعرفة من هذه الوثائق النصية. أنظمة التعدين النص وإدارة التقليدية تستخدم أساسا وجود أو عدم وجود كلمات رئيسية لاكتشاف وتحليل المعلومات المفيدة من الوثائق النصية. ومع ذلك، عدد مرات تكرار الكلمة وتردد توزيعها لا تساهم فى إلتقاط المعنى وراء الكلمات، مما يؤدي إلى الحد من القدرة على تعدين النصوص. وتقترح هذه الورقة خطة تمثيل رواية من النهج القائم على الفهم الدلالي للوثائق النصية . ويستند هذا النهج على المفاهيم الدلالية لتمثيل النص في الوثائق، لاستنتاج تبعيات غير معروفة والعلاقات بين المفاهيم في النص، لقياس ارتباط بين الوثائق والنصوص وتطبيق عمليات التعدين باستخدام التمثيل وتدبير الصلة. نظام التمثيل يعكس العلاقات القائمة بين المفاهيم ويسهل قياسات دقيقة الصلة التي تؤدي إلى أداء أفضل للتعدين. يتم إجراء تقييم تجريبي واسع النطاق على مجموعات البيانات الحقيقية من مختلف المجالات، مما يدل على أهمية النهج المقترح.

Abstract

Tremendous growth in the number of textual documents has produced daily requirements for effective development to explore, analyze, and discover knowledge from these textual documents. Conventional text mining and managing systems mainly use the presence or absence of key words to discover and analyze useful information from textual documents. However, simple word counts and frequency distributions of term appearances do not capture the meaning behind the words, which results in limiting the ability to mine the texts. This paper proposes a novel representation scheme of a semantic understanding-based approach to mine textual documents. This approach is based on semantic notions to represent the text in documents, to infer unknown dependencies and relationships among concepts in a text, to measure the relatedness between text documents and to apply mining processes using the representation and the relatedness measure. The representation scheme reflects the existing relationships among concepts and facilitates accurate relatedness measurements that result in a better mining performance. An extensive experimental evaluation is conducted on real datasets from various domains, indicating the importance of the proposed approach.

Keywords

Linguistic processing, Text analysis, Text mining, Knowledge acquisition, information access, Interactive data exploration and discovery.

1 Introduction

As the sheer number of textual documents available online increases exponentially, the need to manage these textual documents also increases. This growth of online textual documents plays a vital role in exploring information and knowledge

Received: 26 January, 2015 – Accepted: 26 February, 2015

[1]. The massive volume of available information and knowledge should be discovered, and the tasks of managing, analyzing, searching, filtering, and summarizing the information in documents should be automated [1-2]. Four main aspects pertain to most of the textual document mining and managing approaches: (a) representation models [3], (b) relatedness measures [4], (c) mining and managing processes [2], and (d) evaluation methods.

Selecting an appropriate data representation model is essential for text characterization including; text mining and managing, dictating how data should be organized, and what the key features. The “relatedness measures“ are used to determine the closeness of the objects in the representation space, while the "mining and managing processes" are the algorithms that describe the steps of a specific task to fulfill specific requirements. The evaluation methods are used to judge the quality of the mining process results that are produced [1].

At a certain level of simplicity, such mining and managing operations as gathering, filtering, searching, retrieving, extracting, classifying, clustering, and summarizing documents seem relatively similar. All of these operations make use of a text representation model and a relatedness measure to perform their specific tasks. Dealing with Natural Language (NL) documents requires an adequate text representation model to understand them [5]. Accordingly, the trend toward reliance on a semantic understanding-based approach is necessary. Knowledge-rich representations of text combined with accurate semantic relatedness measures are required. This paper introduces an efficient Semantic Hierarchy/Graph-Based Representation Scheme (SHGRS) based on exploiting the semantic structure to improve the effectiveness of the mining and managing operations. The semantic representation scheme is a general description or a conceptual system for understanding how information in the text is represented and used. The proposed representation scheme is an essential step for the actual information access and knowledge acquisition. The main contributions of this paper are summarized as follows:

1. Proposing an efficient representation scheme called SHGRS that relies on an understanding-based approach.
2. Exploiting the knowledge-rich notion to introduce an efficient relatedness measure at a document level.
3. Conducting extensive experiments on real-world datasets to study the effectiveness of the proposed representation scheme along with the proposed relatedness measurethat allow more effective document mining and managing processes.

The rest of the paper is organized as follows; section 2 gives a brief overview of the related work and basic concepts while section 3 illustrates the details about the proposed framework for constructing an efficient text representation scheme. Section 4 reports the experimental results and discussions. Conclusions and suggestions for future work are given in section 5.

2 The Basic Concepts and Related Works

In an effort to keep up with the tremendous growth of the online textual documents, many research projects target the organization of such information in a way that will make it easier for the end users to find the information they want efficiently and accurately[1]. Related works here can roughly be classified into three two categories of studies: text representation model, semantic relatedness measures.

Firstly, the growing amount of recent research in this field focuses on how the use of semantic representation model is beneficial for text categorization [3], text summarization [6], word sense disambiguation [7, 8], and documents classification [9]. Although the text representation model is a well-researched problem in computer science, the current representation models could be improved with the use of external knowledge that is impossible to extract from the source document itself. One of the widely used sources of external knowledge for the text representation model is WordNet, a network of related words organized into synonym sets, where these sets are based on the lexical underlying concept. Furthermore, machine learning is not yet advanced enough to allow large-scale extraction of information from textual document without human input. Numerous semantic annotation tools [14, 15] have been developed to aid the process of human text markup to guide machines. In the semantic level of understanding of text documents [5], models such as the WordNet-based Semantic model, conceptual dependence model, semantic graph model, ontology-based knowledge model, and Universal Networking Language can be used. In [9], the WordNet-based Semantic model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. This model analyzes the terms and their corresponding synonyms and hypernyms on the sentence and document levels, ignoring dependencies among terms in the sentence level. In [10], the conceptual dependence model identifies, characterizes, and understands the effect of the existing dependencies among the entities in the model, considering only nouns as a concept.

The semantic graph model proposes semantic representation based on a graph-based structure with a description of text structure [5]. In the semantic graph model, the traditional method to generate the edges between two nodes in the graph is usually based on the co-occurrence and similarity measures of two nodes, and the most frequent word is not necessarily chosen to be important. In [11], ontology knowledge represented by entities connected by naming relationships and the defined taxonomy of classes may become a much more powerful tool in information exploration. Traditional text analysis and organization methods can also be enriched with the ontology-based information concerning the co-occurring entities or whole neighborhoods of entities [12]. However, automatic ontology construction is a difficult task because of the failure to support order among the objects and the attributes. In [13], Universal Networking Language (UNL) is used to represent every document as a graph with concepts as nodes and relations between them as links, ignoring the sequence and orders of concepts in a sentence at the document level.

Secondly, the semantic relatedness measure has become an area of research as one of the hotspots in the area of information technology. Semantic relatedness and semantic similarity are sometimes confused in the research literature, and they are not identical [18,19]. The semantic similarity is a special case of semantic relatedness that only considers synonymous relationships and subsumption relationships.

There are several measures for semantic relatedness recently conducted. According to the parameters used, they can be classified into three major categories, including Distance-based methods (Rada and Wu&Palmer [18]) which selects the shortest path among all the possible paths between concepts to be more similar, Information-based methods (Resnik, Jaing and Conrath and Lin [18]) which considers the use of external corpora avoiding the unreliability of path distances and taxonomy, and Hybrid methods (T.Hong & D.smith and Zili [18]) which combines the first two measures.

In this paper the Information-based methods are focused. In [20], the semantic relatedness of any concept is based on a similarity theorem in which the similarity of two concepts is measured by the ratio of the amount of information needed to the commonality of the two concepts to the amount of information needed to describe them. The Information Content (IC) of their Lowest Common Subsumer (LCS) captures the commonality of the two concepts and the IC of two concepts themselves. The LCS is the most specific concept, which is a shared ancestor of the two concepts. The Pointwise Mutual Information (PMI) [23, 26] is a simple method for computing corpus-based similarity of words.

As clearly, there is extensive literature on measuring the semantic relatedness between long texts or documents [27], but there is less work related to the measurement of similarity between short texts [29]. Such methods are usually effective when dealing with long documents because similar documents will usually contain a degree of co-occurring words. However, in short documents, the word co-occurrence may be rare or even null. This is mainly due to the inherent flexibility of NL enabling people to express similar meanings using quite different sentences in terms of structure and word content.

To utilize the structural and semantic information in the document in this paper, a formal semantic representation of linguistic input is introduced to build an SHGRS scheme for the documents. This representation scheme is constructed through the accumulation of syntactic and semantic analysis outputs. A new semantic relatedness measure is developed to determine the relatedness among concepts of the document as well as relatedness between contents of the documents for long/ short texts.

3 The proposed shgrs

In this section, a new framework is introduced for constructing the SHGRS Scheme using multidimensional analysis and primary decision support. In fact, the elements in a sentence are not equally important, and the most frequent word is not necessarily the most important. As a result, the extraction of text Main Features (MFs) as concepts as well as their attributes and relationships is important. Combinations of Semantic Annotation [14,15] and Reinforcement Learning (RL)[17] techniques are used to extract MFs of the text and to infer unknown dependencies and relationships among these MFs. The RL fulfills sequential decision making tasks with long-run accumulated reward to achieve the largest amount of interdependence among MFs. This framework focuses on two key criteria: 1) how the framework refines text to select the MFs and their attributes, 2) what learning algorithm is used to explore the unknown dependencies and relationships among these MFs, and

Details of this framework process in two stages are given in Figure 2. The first stage aims to refine textual documents to select the MFs and their attributes with the aid of OpenNLP and AlchemyAPI. Hierarchy-based structure is used to represent each sentence with its MFs and their attributes which achieves dimension reduction with more understanding. The second stage aims to compute the proposed MFs Semantic Relatedness (MFsSR) that contributes to the detection of the closest-synonyms of the MFs and to inferring the relationships and dependencies of the MFs. Graph-based structure is used to represent the relationships and dependencies among these MFs which achieves more correlation through many-to-many relationships. The main proceedings of this framework can be summarized as follows:

1. Extracting MFs of sentences and their attributes to build the hierarchy-based structure for efficient dimension reduction.
2. Estimating a novel semantic relatedness measure with consideration of direct relevance and indirect relevance between MFs and their attributes for promising performance improvements.
3. Detecting Closest Synonyms of MFs for more disambiguation.
4. Exploring and Inferring Dependencies and Relationships of MFs for more understanding.

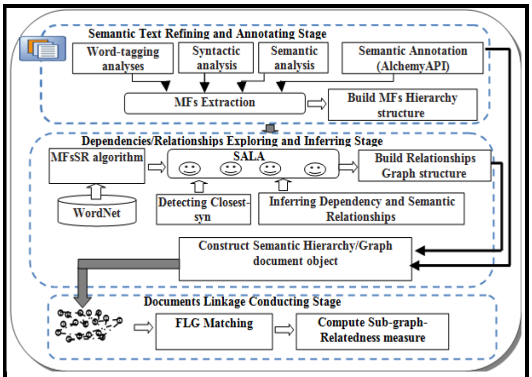
Semantic Text Refining and Annotating Stage
Word-tagging analyses
Syntactic analysis
Semantic analysis
Semantic Annotation (AlchemyAPI)
MFs Extraction
Build MFs Hierarchy structure
Dependencies/Relationships Exploring and Inferring Stage
MFsSR algorithm
SALA
Build Relationships Graph structure
WordNet
Detecting Closest-syn
Inferring Dependency and Semantic Relationships
Construct Semantic Hierarchy/Graph document object
Documents Linkage Conducting Stage
FLG Matching
Compute Sub-graph-Relatedness measure

Fig. 2. The proposed framework for constructing a Semantic Hierarchy/Graph Text Representation Scheme.

5. Representing the interdependence among MFs of sentences in graph-based structure so as to be more precise and informative.

3.1 Semantic Text Refining and Annotating Stage

This stage is responsible for refining the text to discover the text MFs with additional annotation for more semantic understanding and aims to:

1. Accurately parse each sentence and identifying POS, subject-action-object and named-entity recognition.
2. Discover the MFs of each sentence in the textual document.
3. Exploit semantic information in each sentence through detection attributes of the MFs.
4. Reduce the dimensions as much as possible.
5. Generate an Effective Descriptive Sentence Object (DSO) with a hierarchical sentence object automatically.

This stage can be achieved through these processes: first, the text NL is studied at different linguistic levels, i.e., words, sentence and meaning for semantic analysis and annotation [14-15]. Second, exploiting the information ("who is doing what to whom") clarifies dependencies between verbs and their arguments for extraction of MFs. Finally, building a MFs Hierarchy-base structure explains sentences MFs and their attributes.

With regard to extracting MFs and building a MFs hierarchy-based structure, the following points must be highlighted. First, there is a dire need to refine the text content by representing each sentence with its MFs instead of a series of terms to reduce the dimensions as much as possible. The OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. Furthermore, the AlchemyAPI extracts semantic meta-data from content, such as information on subject-action-object relation extraction, people, places, companies, topics, facts, relationships, authors, and languages. Based on the NL concept by OpenNLP and AlchemyAPI, the sentence MFs are identified as Subject "Sub", Main Verb "MV", Object "Obj" (direct or indirect object). In addition, such other terms remaining in the sentence as Complement "Com" (subject complement or object complement) and Modifiers (Mod) are considered as attributes to these MFs.

Second, the automatic annotation is essential for each of the MFs with additional information for more semantic understanding. Based on the semantic annotation by OpenNLP and AlchemyAPI, the sentence MFs are annotated with Feature Value, Part-Of-Speech, Named-entity recognition, and a list of Feature Attributes. The list of Attributes is constructed from the remaining terms in the sentence (complement or modifier) relying on the grammatical relation. Each attribute is annotated with Attribute Value, Attribute Type which is complement or modifier, Attribute POS, and Attribute Named-entity recognition.

Finally, the Hierarchy-based structure of the textual document is represented as an object [16] containing the hierarchical structure of sentences with the MFs of the sentences and their attributes. This hierarchical structure maintains the dependency between the terms on the sentence level for more understanding which provides sentences fast access and retrieval. The summation of the feature Attributes is used to measure the contribution of the MF in the sentence that is called MF score.In the Text Refining and Annotating (TRN) algorithm, the textual document is converted to an object model with the hierarchical DSO. The TRN algorithm uses OpenNLP and AlchemyAPI tools for MFs Extraction and the hierarchy model construction.

Algorithm 1. TRN algorithm

```
Input: List of Sentences of The document di.
Output: List_DSO /*list of DSO objects and its MFs , attributes./*
Procedure
{ D ← New document
  List_DSO ← Empty list{list of sentences object of document};
  for each sentence si in document D do
    AlchemyAPI(si); /* This function calls AlchemyAPI to determine all MF and its NER. /*
    for each feature mfj € {mf1,mf2,...mfn} in DSO do
      get_MF_Attributes(mfi); /* This function determines attributes of each MF. /*
      OpenNLP (si); /* This function calls OpenNLP to determine POS and NER for MF and attributes. /*
      compute_MF_score(mfi); /* This function determines the score of each MF. /*
    end
    List_DSO.add(DSO);
End}
```

3.2. Dependencies/Relationships Exploring and Inferring Stage

This stage is responsible for exploring how the dependencies and relationships among MFs have an effect and aims to:

1. Represent the interdependence among sentence MFs in a graph-based structure known as the Feature Linkage Graph (FLG).
2. Formulate an accurate measure for the semantic relatedness of MFs.
3. Detect the closest synonyms for each of the MFs.
4. Infer the relationships and dependencies of MFs with each other.

This stage can be achieved in three processes: first, building a FLG represents the dependencies and relationships among sentences. Second, an efficient MFsSR measure proposed to contribute to the detection of the closest synonyms and to infer the relationships and dependencies of the MFs. This measure considers the direct relevance and the indirect relevance among MFs. The direct relevance is the synonyms and associated capabilities (similarity, contiguity, contrast, and causality) among the MFs. These association capabilities indicate the relationships that give the largest amount of interdependence among the MFs, while the indirect relevance refers to other relationships among the attributes of these MFs. Finally, the unknown dependencies and relationships among MFs are explored by exploiting semantic information about their texts at a document level.

In addition, a Semantic Actionable Learning Agent (SALA) plays a fundamental role to detect the closest synonyms and to infer the relationships and dependencies of the MFs. Learning is needed to improve the SALA functionality to achieve multiple-goal. Many studies show that the RL agent has a high reproductive capability for human-like behaviors [21]. As a result, the SALA performs an adaptive approach combining thesaurus-based and distributional mechanisms. In the thesaurus-based mechanism, words are compared in terms of how they are in the thesaurus (e.g., Wordnet), while in the distributional mechanism, words are compared in terms of the shared number of contexts in which they may appear.

3.2.1 Graph-base Organization Constructing Process

The relationships and dependencies among the MFs are organized into a graph-like structure of nodes with links known as FLG. The graph structure FLG represents many-to-many relationships among MFs. The FLG is a directed acyclic graph FLG (V, A) that would be represented in a collection of vertices and a collection of directed arcs. These arcs connect pairs of vertices with no path returning to the same vertex (acyclic). In FLG (V, A), V is the set of vertices or states of the graph, and A is the set of arcs between vertices. The FLG construction is based on two sets of data. The first set represents the vertices in a one-dimensional array Vertex (V) of Feature_Vertex objects (FVOs). The FVO is annotated with Sentence Feature key, Feature value, Feature closest synonyms, Feature Associations and Feature Weight. Where SFKey combines sentence key and feature key in DSO, the sentence key is important because that facilitates accessing the parent of the MFs. The Fval object has the value of the feature; the Fsyn object has a list of the detected closest synonyms of each MF and a list of the explored feature associations and relationships. In associations list, each object has Rel_Typ between two MFs that indicates the value of the relation type linkage between the two MFs (1=similarity, 2=contiguity, 3=contrast, 4=causality and 5=synonym) and Rel_vertex index of the related vertex. The Feature Weight object has accumulated the associations and relationships weights clarifying the importance of the Fval in the textual document. The second set represents the arcs in a two-dimensional array Adjacency (V,V) of Features_link Weights (FLW), which indicates the value of the linkage weight in an adjacency matrix between related vertices.

3.2.2 The MFs Semantic Relatedness (MFsSR) Measuring Process

A primary motivation for measuring semantic relatedness comes from the NL processing applications such as information retrieval, information extraction, information filtering, text summary, text annotation, text mining, word sense disambiguation, automatic indexing, machine translation and other aspects [2-3]. In this paper, one of the Information-based methods is utilized by considering IC. The semantic relatedness that has been investigated concerns the direct relevance and the indirect relevance among MFs. The MFs may be one word or more, and thereby the MF is considered as a concept. The IC is considered as a measure of quantifying the amount of information a concept expresses. Traditionally, the semantic relatedness of concepts is usually based on the co-occurrence information on a large corpus. However, the co-occurrences do not achieve many matching in a corpus, and it is essential to take into account the relationships among concepts. Therefore, development of a New Information Content method based on the co-contributions instead of the co-occurrences of the proposed MFsSR is important.

The New Information Content (NIC) measure is an extension of the information content measure. The NIC measures based on the relations defined in the WordNet ontology. The NIC measure uses hypernym/hyponym, synonym/antonym, holonym/meronymy, and Entail/Cause to quantify the informativeness of concepts. For example, a hyponym relation could be "bicycle is a vehicle" and a meronym relation could be "a bicycle has wheels". NIC is defined as a function of the hypernym, hyponym, synonym, antonym, holonym, meronymy, entail, and cause relationships normalized by the maximum number of MFs/attributes objects in the textual document using Equation (1).

$$\mathrm{NIC(c)} = 1 - \frac{(\log(\mathrm{Hype\_Hypo(c)}+\mathrm{Syn\_Anto(c)}+\mathrm{Holo\_Mero(c)}+\mathrm{Enta\_Cause(c)}+1))}{\log(\mathrm{max\_concept})} \quad (1)$$

Where Hype_Hypo (c) returns the number of FVOs in the FLG related to the hypernym or hyponym values, Syn_Anto(c) returns the number of FVOs in the FLG related to the synonym or antonym values, Holo_Mero (c) returns the number of FVOs in the FLG related to the holonym or meronymy values, and Enta_Cause (c) returns the number of FVOs in the FLG related to the entail or cause values. The max_concept is a constant that indicates the total number of MFs/attributes objects in the considered text. The max concept normalizes the NIC value, and hence the NIC values fall in the range of [0, 1].

With the consideration of direct relevance and indirect relevance of MFs, the proposed MFsSR can be stated as follows in Equation (2).

$$\mathrm{SemRel} = \lambda\left(\frac{2*\mathrm{NIC\ (LCS(\ c1,c2))}}{\mathrm{NIC(c1)+\ NIC(c2)}}\right) + (1-\lambda)\left(\frac{2*\mathrm{NIC\ (LCS(\ att\_c1,att\_c2))}}{\mathrm{NIC(att\_c1)+\ NIC(att\_c2)}}\right) \quad (2)$$

where λ €[0, 1] decides the relative contribution of direct and indirect relevance to the semantic relatedness, and because the direct relevance is assumed to be more important than the indirect relevance, λ €[0.5, 1].

3.2.3 The Closest-Synonyms Detecting Process

In MFs Closest-Synonyms detection plausible ection, the SALA carried out the first action to extract the closest synonyms for each MF, where SALA defies automatic discovery of similar meaning words (synonyms). Due to the important role played by a lexical knowledge base in the closest synonyms detection, SALA adopts the dictionary-based approach to the disambiguation of the MFs. In the dictionary-based approach, the assumption is that the most plausible sense to assign to multiple shared words is that sense that maximizes the relatedness among the chosen senses.

In this respect, SALA detects the synonyms by choosing the meaning whose glosses share the largest number of words with the glosses of the neighboring words through lexical ontology. Using a lexical ontology such as WordNet allows the capture of semantic relationships based on the concepts and exploiting hierarchies of concepts besides dictionary glosses. One of the problems that can be faced is that not all synonyms are really related to the context of the document. Therefore, the

MFs disambiguation is achieved by implementing the Closest-Synonym Detection (C-SynD) algorithm.

Algorithm 2. C-SynD algorithm

```
Input: Array Feature_Vertex Objects (FVO), Wordnet
Output: List of Closest-Synonym of FVO objects /*The detected closest synonym to set Fsyn value. /*
Procedure
  {SRL← Empty list {list of each synonym and their relationships};
   F_SRList ← Empty list {list of SRL (all synonyms and their relationships lists)};
   /*Parallel.foreach used for parallel processing of all FVO. /*
   Parallel.ForEach(FVO, fvo=> {
   Fsyn← Empty list;
   F_SRList= get_all_Syn_Relations_List(fvo); /* function to get all synonyms and their relationships lists from wordnet /*
   /* compute semantic relatedness of each synonym and their relationships with the fvo object then get score of each list/*
        for (int i = 0; i < F_SRList.Count(); i++)
         { score[i] =sum( function_Get_SemRel(F_SRList[i], fvo)); /*function to compute the score of each synonym list, this score contributes in accurate filtering the returned synonyms and selecting the closest one/*
         }
         Fsyn=maxth(score[]);/*function to select SRL with highest score according to the specific threshold /*
}); // Parallel For}
```

In the C-SynD algorithm, the main task of each SALA is to search for synonyms of each FVO through wordnet. Each synonym and their relationships are assigned in a list called Syn_Relation_List (SRL). Each item in the SRL contains one synonym and their relationships such as Hypernyms, Hyponym, Meronyms, Holonymy, Antonymy, Entail, and Cause. The purpose of using these relations is to eliminate the ambiguity and polysemy because not all of the synonyms are related to the context of the document. Then, all synonyms and their relationship lists are assigned in a list called F_SRList. The SALA starts to filter the irrelevant synonyms according to the score of each SRL. This score is computed based on semantic relatedness between the SRL and every FVO in the Feature_Vertex array as follows in Equation (3).

$$\text{Score}(\ \text{SRL}_k\ ) = \sum_{i=0}^{m} \sum_{j=0}^{n} \text{SemRel}(\text{SRL}(i).\text{item}, \text{FVOj})) \quad (3)$$

where n is the number of FVOs in the Feature_Vertex array with index j, m is the number of items in SRL with index i, and k is index of SRLk in F_SRList.

The lists of synonyms and their relationships are used temporarily to serve the C-SynD algorithm. The C-SynD algorithm specifies a threshold value of the score for selecting Closest-Synonyms. SALA applies the threshold to select the Closest Synonyms with the highest score according to the specific threshold as in Equation (4).

$$\text{Closest} - \text{Synonyms} = \max_{th}(\text{F\_SRList.SRList.score}()) \quad (4)$$

Then, each SALA retains the Closest Synonyms in an object Feature closest synonymsof each FVO, and SALA links the related FVOs with a bi-directional effect. The SALA receives each FVO for detecting links among the other FVOs according to closest synonyms object values and Associations object values. The SALA assigns the FLW between two FVO to their intersection location of the FLG.

3.2.4 The MFs Relationships and Dependencies Inferring Process

In inferring MFs relationships and dependencies, SALA aims to explore the implicit dependencies and relationships in the context such as human relying on four association capabilities. These association capabilities are *similarity, contiguity, contrast (antonym), and causality* [21] which give the largest amount of interdependence among the MFs. These association capabilities are defined as follows:

Similarity: for nouns, is the sibling instance of the same parent instance with is-a relationship in wordnet, indicates hypernym/hyponym relationships.

Contiguity: for nouns, is the sibling instance of the same parent instance with part-of relationship in wordnet, indicates holonym/meronymy relationship.

Contrast: for adjectives or adverbs, is the sibling instance of the same parent instance with is-a relationship and with antonym (opposite) attribute in Wordnet, indicates a synonym/antonym relationship.

Causality: for verbs, indicates the connection of the sequence of events by a cause/entail relationship. Where a cause picks out two verbs, one of the verbs is causative such as (give), and the other is called resultant such as (have) in Wordnet.

To equip SALA with decision-making and experience learning capabilities to achieve multiple-goal RL, this study utilizes the RL method relying on Q-Learning to design a SALA inference engine with sequential decision making. The objective of **SALA** is to select an optimal association to maximize the total long-run accumulated reward. Hence, SALA can achieve the largest amount of interdependence among MFs through implementing the Inference Optimal Association Capabilities with the Q-Learning (IOAC-QL) algorithm.

RL specifies what to do but not how to do it through the reward function. In sequential decision making tasks, an agent needs to perform a sequence of actions to reach goal states or multiple-goal states. One popular algorithm for dealing with sequential decision making tasks is Q-learning [21]. The promising action can be verified by measuring the relatedness score of the action value with the remained MFs using a reward function. The formulation of the effective association capabilities exploration is performed by estimating an action-value function. In single goal reinforcement learning, these Q-values are used only to rank order the actions in a given state. The key observation here is that the Q-values can also be used in multiple-goal problems to indicate the degree of preference for different actions. The available way in this paper to select a promising action to execute is to generate an overall Q-value as a simple sum of the Q-values of the individual SALA. The action with the maximum summed value is then chosen to execute.

Heuristic 1 defines the action variable and the optimal policy.

Heuristic 1:(Action Variable). Let a_k symbolize the action that is executed by SALA at round k.

$\mathbf{a_k = \pi^*(s_k)}$ Therein,

- a_k € action space { similarity; contiguity; contrast and causality}
- $\pi^*(s_k)$=arg $\max_a Q^*(s_k, a_k)$

Once the optimal policy (π *) is obtained, the agent chooses the actions using the Maximum Reward.

The reward function in this paper measures the dependency and the relatedness among MFs. The learner is not told which actions to take as in most forms of machine learning. Rather, the learner must discover which actions yield the most reward by trying them [21].

The optimal association actions are achieved according to the IOAC-QL algorithm. This algorithm starts to select the action with the highest expected future reward from each state. The immediate reward, which the SALA gets to execute an action from a state **s**, plus the value of an optimal policy, is the Q-value. The highest Q-value points to the greater chance of that action being chosen. First, initialize all the Q(*s, a*) values to zero. Then, each SALA performs the following steps in Heuristic 2.

Heuristic 2:

- At every round k, select one action $\mathbf{a_k}$ from all the possible actions (similarity, contiguity, contrast and causality) and execute it.
- Receive the immediate reward $\mathbf{r_k}$ and observe the new state $\mathbf{s_k}$**'**.
- Get the maximum Q-value of the state $s_\mathbf{k}$' based on all the possible actions $\mathbf{a_{k=}}$ $\pi^*(s_k)$=arg $\max_a Q^*(s_k, a_k)$.
- Update the $Q(s_\mathbf{k}, a_\mathbf{k})$ as follows in Equation (5).

$$Q(s_k, a_k) = (1-\alpha)Q(s_k, a_k) + \alpha\left(r(s_k, a_k) + \gamma \max_{a\prime \in A(s_k)} Q(s'_k, a'_k)\right) \quad (5)$$

where $Q(s_k,a_k)$ is worthy of selecting the action(a_k) at the state(s_k). $Q(s_k',a_k')$ is worthy of selecting the next action(a_k') at the next state(s_k'). The $r(s_k, a_k)$ is a reward corresponding to the acquired payoff. $A(s_k')$ is a set of possible actions at the next state(s_k'). $\alpha(0 < \alpha \leq 1)$ is learning rate. $\gamma(0 \leq \gamma \leq 1)$ is discount rate.

During each round after taking the action a, the action value is extracted from Wordnet. The weight of the action value is measured to represent the immediate reward, where the SALA computes the weight $AVW_{i,k}$ for the action value i in document k as follows in Equation (6).

$$Reward(r) = AVW_{i,k} = AVW_{i,k} + \sum_{j=1}^{n}\left(FW_{j,k} * \mathrm{SemRel(i,j)}\right) \quad (6)$$

Where r is the reward of the selected action a_i, $AVW_{i,\ k}$ is the weight of Action_Value i in document k, $FW_{j\ ,k}$ is the weight of each feature object FVO_j in document k, and SemRel(i, j) is the semantic relatedness between the Action_Value i and feature object FVO_j.

The most popular weighting scheme is the normalized word frequency TFIDF [22], used to measure $AVW_{i,\ k}$ and $FW_{j,\ k}$. The $FW_{j,\ k}$ in Equation (6) is calculated likewise $AVW_{i,\ k}$ is based on Equations (7) and (8).

ActionValue Weight (AVW$_{i,k}$) is a measure to calculate the weight of the action_value that is the scalar product of action_value frequency and inverse document frequency as in Equation (7). The Action_Value Frequency (F) measures the importance of this value in a document. Inverse Document Frequency (IDF) measures the general importance of this action_value in a corpus of documents.

$$AVW_{i,k} = \frac{\mathrm{F_{ik}}}{\sum_\mathrm{k} \mathrm{F_{i,k}}} * \mathrm{IDF_i} \quad (7)$$

where $F_{\mathrm{i,k}}$ represents the number of this action_value i co-contribution in document d_k, normalized by the number of co-contributions of all MFs in document d_k, normalized to prevent a bias towards longer documents. IDF is performed by dividing the number of all documents by the number of documents containing this action_value defined as in Equation (8).

$$IDF_i = \log\left(\frac{|\mathrm{D}|}{|\{\mathrm{d_k}: \mathrm{v_i} \in \mathrm{d_k}\}|}\right) \quad (8)$$

where |D| is the total number of documents in the corpus, and $|\{d_k : v_i € d_k\}|$ is the number of documents containing action_value V_i.

Thus, in the IOAC-QL algorithm implementation, the SALA selects the optimal action with the highest reward. Hence, each SALA retains effective actions in the association List object of each FVO. The SALA then links the related FVOs with the bidirectional effect using sets the FLW value of the linkage weight in the adjacency matrix with the Reward of the selected action.

Algorithm 3. IOAC-QL algorithm

```
Input: Feature_Vertex Object(FVO), Wordnet
Output: The optimal action for FVO relationships;
  Procedure
{ intialize Q[,];
  /*implement q-learning algorithm to get the action with the highest Q-value. /*
  ak € action_space[]={similarity, contiguity, contrast, causality}
   /*Parallel.foreach used for parallel processing of all actions. /*
  Parallel.for(0; action_space.count(); k =>
  { R[k]= Get_Action_Reward(a[k]); /*Function to implement equation 7. /*
     $Q[i,k]=(1-\alpha)* Q[i,k]+ \alpha*(R[k] + \gamma* maxa(Q[i,k+1]))$
   }); // Parallel.For
  return Get_highest_Q-value_action(Q[,]);/*select action with the maximum summed value of Q-values /*
  } /*execute Equations (7) to get reward value. /*
Get_Action_Reward(ak)
  { AVWk=(freq(ak.val)/FVO.Count())*log(N/na);
     for (int j = 0; j < FVO.Count(); j++)
     { FWj=(freq(FVO[j])/FVO.Count())*log(N/nf);
        SemRel(i,j )=Get_SemRel(ak.val, FVO[j])
        Sum+= FWj* SemRel(i,j ); }
   return AVWk= AVWk + Sum;  }
```

4 Evaluation results

This evaluation is especially vital, as the aim of building the SHGRS is to use them efficiently and effectively for the further mining process. To explore the effectiveness of the proposed SHGRS scheme, examining the correlation of the proposed Semantic Relatedness measure compared with other previous relatedness measures is required. The impact of the proposed MFsSR for detecting the closest synonym is studied and compared to the PMI [26, 30] and independent component analysis (ICA) [31] for detecting the best near-synonym. The difference between the proposed discovering semantic relationships or associations IOAC-QL, implicit Relation Extraction Conditional Random Field (CRFs) [24, 32] and Term frequency and inverse cluster frequency (TFICF) [33] is computed.

4.1 Evaluation Measures

Evaluation measures are subcategorized into text quality-based evaluation, content-based evaluation, and co-selection-based evaluation. The first category of evaluation measures is based on text quality using aspects such as grammaticality, non-redundancy, referential clarity and coherence. For content-based evaluations, measures such as similarity, semantic relatedness, longest common subsequence and other scores are used. The third category is based on co-selection evaluation using precision, recall, and f-measure values. In this paper, the content-based and the co-selection-based evaluations are used to validate the implementation of the proposed scheme over the real corpus of documents.

4.1.1 Measure for content-based evaluation

The relatedness or similarity measures are inherited from probability theory and known as the correlation coefficient [25]. The correlation coefficient is one of the most widely used measures to describe the relatedness r between two vectors, X and Y.

Correlation Coefficient r:

The correlation coefficient r is a relatively efficient relatedness measure, which is a symmetrical measure of the linear dependence between two random variables. Therefore, the r value between sequences X = {x_i: i = 1, . ., n} and Y = {y_i: i = 1, .., n} is defined as in Equation (9).

$$r = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\left(\sum_{i=1}^{n} {X_i}^2\right) * \left(\sum_{i=1}^{n} {Y_i}^2\right)}} \qquad (9)$$

Acceptance Rate AR:

Acceptance rate is a proportion of correctly predicted similar or related sentences compared to all related sentences as in Equation (10). High acceptance rate means that recognizing almost all similar or related sentences.

$$AR = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{10}$$

Accuracy Acc:

Accuracy is a proportion of all correctly predicted sentences compared to all sentences as in Equation (11).

$$Acc = \frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN})} \tag{11}$$

Where TP, TN, FP, and FN stand for true positive (the number of pairs correctly labeled as similar), true negative (the number of pairs correctly labeled as dissimilar), false positive (the number of pairs incorrectly labeled as similar), and false negative (the number of pairs incorrectly labeled as dissimilar).

4.1.2 Measure for co-selection-based evaluation

For all the domains, a precision P, recall R, and F-measure are utilized as the measures of performance in the co-selection-based evaluation. These measures may be defined via computing the correlation between the extracted, correct, and wrong closest synonyms or semantic relationships/dependencies. Let TP denote the number of correctly detected closest synonyms or semantic relationships explored, let FP be the number of incorrectly detected closest synonyms or semantic relationships explored, and let FN be the number of correctly but not detected closest synonyms or semantic relationships explored in a dataset. The F-measure combines the precision and recall in one metric and is often used to show the efficiency. Precision, Recall, and F-measure are defined as follows in Equation (12), Equation (13), and Equation (14).

$$\mathrm{Recall(R)} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{12}$$

$$\mathrm{Precision(P)} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \tag{13}$$

$$\mathrm{F-measure(F)} = \frac{2(\mathrm{Recall} * \mathrm{Precision})}{(\mathrm{Recall}+\mathrm{Precision})} \tag{14}$$

4.2. Evaluation Setup (dataset)

Content-based and co-selection-based are used to validate experimenting over the real corpus of documents, the results are very promising. The experimental setup consists of some datasets of textual documents as detailed in Table 2.

Table 2. The experimental setup Datasets details

DS	**DS Name**	**Description**
DS1	Miller and Charles	M&C consists of 30 pairs of nouns extracted from the WordNet.
DS2	Microsoft Research Paraphrase Corpus (MRPC)	The corpus consists of 5,801 sentence pairs collected from newswire articles, 3,900 were labeled as relatedness by human annotators. The training subset (4,076 sentences of which 2,753 are true).
DS3	British National Corpus (BNC)	BNC is a 100-million-word text corpus of samples of written and spoken English with the near-synonym collocations. Only 2.61% of our near-synonyms do not occur; and only 2.63% occur between 1 and 5 times.
DS4	SN (Semantic Neighbors)	SN relates 462 target terms (nouns) to 5910 relatum terms with 14.682 semantic relations (7341 are meaningful and 7341 are random).
DS5	BLESS	BLESS relates 200 target terms (100 animate and 100 inanimate nouns) to 8625 relatum terms with 26.554 semantic relations (14.440 are meaningful (correct) and 12.154 are random).
DS6	TREC	TREC includes 1437 sentences annotated with entities and relations at least one relation. There are three types of entities: Person 1685, Location 1968 and Organization 978, in addition there is a fourth type Other 705. There are five types of relations: Located In 406, Work For 394, OrgBased In 451, Live In 521 and Kill 268.
DS7	IJCNLP 2011-New York Times(NYT)	NYT contains 150 business articles from NYT. There are 536 instances (208 Positive, 328 Negative) with 140 distinct descriptors in NYT dataset.
DS8	IJCNLP 2011-Wikipedia	Wikipedia personal/social relation data set previously used in Culotta et al. There are 700 instances (122 Positive, 578 Negative) with 70 distinct descriptors in Wikipedia dataset.

4.3. Evaluation Results

This section reports on the results of three experiments conducted using the evaluation datasets outlined in the previous section. The SHGRS is implemented and evaluated based on concept analysis and annotation as sentence-based in experiment 1, and document-based in experiment 2.

4.3.1 Experiment1: Comparative study (Content-based evaluation)

This experiment shows the necessity to evaluate the performance of the proposed

MFsSR based on a benchmark dataset of human judgments, so the results of the proposed semantic relatedness would be comparable with other previous studies in the same field. In this experiment, they attempted to compute the correlation between the ratings of the proposed semantic relatedness approach and the mean ratings reported by Miller and Charles (DS1 in Table 2). Furthermore, the results produced are compared against eight other semantic similarities approaches, namely, Rada, Wu and Palmer, Rensik, Jiang & Conrath, Lin, Hong & Smith and Zili Zhou [18].

Table 3 shows the correlation coefficient results between nine componential approaches and the Miller and Charles ratings mean. The semantic relatedness for the proposed approach outperformed all the listed approaches. Unlike all the listed methods, in the proposed semantic relatedness, different properties are considered. Furthermore, the good correlation value of the approach also results from considering all available relationships between concepts and the indirect relationships between attributes of each concept when measuring the semantic relatedness. In this experiment, the proposed approach achieved a good correlation value with the human-subject rating reported by Miller and Charles. Based on the results of this study, the MFsSR correlation value has proven that considering the direct/indirect relevance is specifically important for at least 6% improvement over the best previous results. This improvement contributes to the achievement of the largest amount of relationships and interdependence among MFs and their attributes at the document level.

Table 3. The results of the proposed MFsSR compared to the previous relatedness measures.

Measure		Relevance Correlation with M&C
Distance-based measures	Rada	0.688%
	Wu & Palmer	0.765%
Information-based measures	Resnik	0.77%
	Jiang & Conrath	0.848%
	Lin	0.853%
Hybrid measures	T.Hong & D.smith	0.879%
	Zili Zhou	0.882%
Information /Feature-base measures	The proposed MFsSR	0.937%

Table 4 summarizes the characteristics of the MRPC dataset (DS2 in Table 2) and presents comparison of the Acc and AR values between the proposed Semantic Relatedness measure MFsSR and A. Islamand D. Inkpen [28]. Different relatedness thresholds ranging from 0 to 1 with interval 0.1 are used to validate the MFsSR with A. Islamand D. Inkpen. After evaluation, the best relatedness thresholds of Acc and AR are 0.6, 0.7 and 0.8. These results indicate that the proposed MFsSR surpasses A. Islamand D. Inkpen in terms of Acc and AR.

Table 4. The results of the comparison of the accuracy and acceptance rate between the proposed Semantic Relatedness measure and A. Islamand D. Inkpen.

MRPC dataset	Relatedness threshold	Human judgment (TP + FN)	A. Islam and D. Inkpen		The proposed MFsSR	
			Acc	AR	Acc	AR
Training subset (4,076)	0.1	2,753 true	0.67	1	0.68	1
	0.2		0.67	1	0.68	1
	0.3		0.67	1	0.68	1
	0.4		0.67	1	0.68	1
	0.5		0.69	0.98	0.68	1
	0.6		0.72	0.89	0.68	1
	0.7		0.68	0.78	0.70	0.98
	0.8		0.56	0.4	0.72	0.86
	0.9		0.37	0.09	0.60	0.49
	1		0.33	0	0.34	0.02

The results of each approach listed below were based on the best Acc and AR through all thresholds instead of under the same relatedness threshold. This improvement in Acc and AR values is due to the increase in the numbers of pairs predicted correctly after considering direct/indirect relevance. This relevance takes into account the closest synonym and the relationships of sentences MFs.

4.3.2 Experiment 2: Comparative study of closest-synonym detection and semantic relationships exploration

This experiment shows the necessity to study the impact of the MFsSR for inferring unknown dependencies and relationships among the MFs. This impact is achieved through the closest synonym detection and semantic relationship exploration algorithms, so the results would be comparable with other previous studies in the same field. Throughout this experiment, the impact of the proposed MFsSR for detecting the closest synonym is studied and compared to the PMI approach and ICA approach for detecting the best near-synonym. The difference among the proposed discovering semantic relationships or associations IOAC-QL, the implicit Relation Extraction CRFs approach and TFICF approach is examined. This experiment attempts to measure the performance of the MFsSR for the closest synonym detection and the IOAC-QL for semantic relationship exploration algorithms. Hence, more semantic understanding of the text content is achieved by inferring unknown dependencies and relationships among the MFs. Considering the direct relevance among the MFs and the indirect relevance among the attributes of the MFs in the proposed MFsSR constitutes a certain advantage over previous measures. However, most of the time, the incorrect items are due to a wrong syntactic parsing from the OpenNLP Parser and AlchemyAPI. According to the preliminary study, it is certain that the accuracy of parsing tools' effects is on the performance of the MFsSR.

Detecting the closest synonym is a process of MF disambiguation resulting from the implementation of the MFsSR technique and the threshold of the synonyms scores through the C-SynD algorithm. In the C-SynD algorithm, the MFsSR takes into account the direct relevance among MFs and the indirect relevance among MFs and their attributes in a context that gives a significant increase in the recall without disturbing the precision. Thus, the MFsSR between two concepts considers the information content with most of wordnet relationships.

Table 5 illustrates the performance of the C-SynD algorithm based on the MFsSR compared to the best near-synonym algorithm using the PMI and ICA approaches. This comparison was conducted for two corpora of text documents that are BNC, and NS (DS3, and DS4 in Table 2). As indicated in Table 5, the C-SynD algorithm yielded highest average precision, recall, and F-measure values than PMI approach by 25%, 20%, and 21% on SN dataset, respectively, and also by 8%, 4% and 7% on BNC dataset. Furthermore, the C-SynD algorithm also yielded highest average the precision, recall, and F-measure values than ICA approach by 6%, 9% and 7% on SN dataset, respectively, and also by 6%, 4% and 5% on BNC dataset. The improvements achieved in the performance values are due to the increase of the number of pairs predicted correctly and the decrease of the number of pairs predicted incorrectly through implementing MFsSR. The important observation from this table is the improvements achieved in recall which measures effectiveness of the C-SynD algorithm. Achieving better precision values are clear, with a high percentage in different datasets and domains.

Table 5. The results of the C-SynD algorithm based on the MFsSR compared to the PMI and ICA for detecting the closest synonym.

Ds	PMI			ICA			C-SynD		
	P	R	F	P	R	F	P	R	F
SN	60.6%	60.6%	0.61%	79.5%	71.6%	75.3%	85%	80%	82%
BNC	74.5%	67.9%	71%	76 %	67.8%	72%	82%	71.9%	77%

Inferring the MFs relationships and dependencies is a process of achieving a large number of interdependences among MFs resulting from the implementation of the IOAC-QL algorithm. In the IOAC-QL algorithm, the Q-learning is used to select the optimal action (relationships) that gains the largest amount of interdependences among the MFs resulting from the measure of the AVW.

Table 6 illustrates the performance of the IOAC-QL algorithm based on AVWs compared to the implicit relationship extraction based on CRFs and TFICF approaches, which was conducted over the TREC**,** Bless, the NYT, and the Wikipedia corpus (DS5, DS6, DS7, and DS8 in Table 2). The data in this table shows increases in precision, recall, and F-measure values due to the increase in the number of pairs predicted correctly after considering direct/indirect relevance through the expansion of closest synonym.

Table 6. The results of the IOAC-QL algorithm based on AVWs compared to CRFs and TFICF approaches.

Ds	CRFs			TFICF			IOAC-QL		
	P	R	F	P	R	F	P	R	F
TREC	75.08%	60.2%	66.28%	89.3%	71.4%	78.7%	89.8%	88.1%	88.6%
BLESS	73.04%	62.66%	67.03%	73.8%	69.5%	71.6%	95.0%	83.5%	88.9%
NYT	68.46%	54.02%	60.38%	86.0%	65.0%	74.0%	90.0%	74.0%	81.2%
Wikipedia	56.0%	42.0%	48.0%	64.6%	54.88%	59.34%	70.0%	44.0%	54.0%

In addition, the IOAC-QL algorithm, the CRFs and the TFICF approaches are beneficial to the extraction performance, but the IOAC-QL contributes more than CRFs and TFICF. Thus, IOAC-QL considers similarity, contiguity, contrast, and causality relationships between MFs and its closest-synonyms, while the CRFs and TFICF consider is-a and part-of relationships only between concepts. The improvements of the F-measure were achieved through an IOAC-QL algorithm up to 23% for TREC dataset, 22% for Bless dataset, 21% for NYT dataset and 6% for Wikipedia dataset, approximately from the CRFs approach. Furthermore, the IOAC-QL algorithm also yielded highest F-measure values than TFICF approach by 10% for TREC dataset, 17% for Bless dataset and 7% on NYT dataset, respectively.

5 Conclusion

This paper proposed a SHGRS to improve the effectiveness of the mining and managing operations in textual documents. Specifically, a three-stage approach was proposed for constructing the scheme. First, the text MF with their attributes are extracted, and hierarchy-based structure is built to represent sentences by the MFs and their attributes. Second, a novel semantic relatedness computing method was proposed for inferring relationships and dependencies of MFs, and the relationship between MFs is represented by a graph-based structure. Future work will focus on conducting other case studies to processes such as gathering, filtering, retrieving, classifying, and summarizing information.

References

[1] F.i.T , M A.G , A C.K and S A.B ,"Knowledge discovery in online repositories: A text mining approach" , european jornal of science research issn 1450-216x vol.22 no.2 Pp.241-250. 2008.

[2] Hotho A., Nürnberger A., and Paaß, G."A Brief Survey of Text Mining". Journal for Computational Linguistics and Language

Technology. Vol. 20, pp. 2005.

[3] Bernotas M., Karklius K., Laurutis R., and Slotkiene A., "The peculiarities of the text document representation, using ontology and tagging-based clustering technique". Journal of Information Technology and Control Vol 36, pp 217 – 220, 2007.

[4] Y. H. Li, Z. Bandar, and D. McLean. “An Approach for Measuring Semantic Similarity Using Multiple Information Sources”, IEEE Trans. Knowledge and Data Eng., vol.15, no.4, pages 871-882, July/Aug., 2003.

[5] Khaled S, Otman B, Mohamed K,"Document Mining Based on Semantic Understanding of Text" , Springer Progress in Pattern Recognition, Image Analysis and Applications ,Lecture Notes in Computer Science Volume 4225, pp 834-843, 2006

[6] Feldman R., Sanger J.”The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data”, Cambridge Univ. Pr. 2006.

[7] Navigli, "Word sense disambiguation: A survey", ACM Computing Surveys, Vol. 41, No. 2, Article 10, February 2009.

[8] Wenlei. Mao, Wesley. W. Chu, "The phrase-based vector space model for automatic retrieval of free-text medical documents". Data & Knowledge Engineering, 2007.

[9] Shady S. ,"A WordNet-Based Semantic Model for Enhancing Text Clustering." ICDM Workshops, page 477-482. IEEE Computer Society, 2009.

[10] Shady S, Fakhri K, and Mohamed K, " An Efficient Concept-Based Mining Model for Enhancing Text Clustering" ,IEEE Transactions on Knowledge and Data Engineering Volume 22 Issue 10, Pages 1360-137, October 2010.

[11] El Moukthtar Zi, Hicham B , Abdelaziz M, Brigitte T , "Ontology-Based Knowledge Model for Multi-View KDD Process", lnternational Journal of Mobile Computing and Multimedia Communications (IJMCMC) 4, 3 21-33, 2012.

[12] C.M and F.G ,"Knowledge-based interactive postmining of association rules using ontologies" ieee transaction on knowledge and data engineering ,vol 22,no.6,june 2010.

[13] Choudhary B., and Bhattacharyya P. "Text clustering using Universal Networking Language representation". In Eleventh International World Wide Web Conference. 2003.

[14] Dingjia L, Zequan L, Qian D ,"A Dependency Grammar and WordNet Based Sentence Similarity Measure", Journal of Computational Information Systems 8: 3 1027–1035, 2012.

[15] Amal Z, Michel G and Benoit O, "Semantic Analysis using Dependency-based Grammars and Upper-Level Ontologies", International Journal of Computational Linguistics and Applications, 1(1-2): 85-101, Bahri Publications 2010.

[16] Cartic R, Krys J, and Amit P, "A Framework for Schema-Driven Relationship Discovery from Unstructured Text", ISWC 2006, LNCS 4273, pp. 583 – 596, 2006.

[17] S.T. yuan, and Y.C. Chen, “Semantic ideation learning for agent-based E-brainstorming”, IEEE Trans. Knowledge Data Eng, vol. 20, no 2, pp. 261-275, February -2008.

[18] Giuseppe Pirró, Jérôme Euzenat , "A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness", The Semantic Web – ISWC 2010 , pp 615-630 , 2010.

[19] Zhenjiang Lin, Michael R. Lyu, Irwin King, "MatchSim: a novel similarity measure based on maximum neighborhood matching", Springer Knowl. Inf. Syst. Volume 32 Issue 1, Pages 141-166, July 2012.

[20] D. Lin, “An information-theoretic definition of similarity”, Proceedings of the 15th International Conference on Machine Learning, July 24-27; Madison, Wisconsin, USA 1998.

[21] Watkins, Christopher J.C.H. and Dayan, Peter ,”Technical Note: Q-Learning, Machine Learning “,8:279-292, 1992.

[22] Lan, M., Tan, C. L., Su. J., and Lu, Y. “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 31 (4), pp. 721 – 735, 2009.

[23] DIANA I, "A Statistical Model for Near-Synonym Choice", ACM Transactions on Speech and Language Processing, Vol. 4, No. 1, Article 2, Publication date: January 2007.

[24] Aron C, Andrew M, Jonathan B, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text", In: Proceedings of HLT-NAACL

2006, pp. 296–303, 2006.

[25] G.A. Miller and W.G. Charles, “Contextual correlates of semantic similarity”, Language and Cognitive Processes, 1–28, 1991.

[26] Lushan H, Tim F, Paul M, Anupam J, and Yelena Y,"Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy", IEEE Transactions on Knowledge and Data Engineering (TKDE), June 01, 2013.

[27] J. O’Shea, Z. Bandar, K. Crockett, and D. McLean, “A comparative study of two short text semantic similarity measures,” KES-AMSTA 2008, LNAI 4953, pp. 172–181, 2008.

[28] A. Islamand and D. Inkpen, “Semantic text similarity using corpus-based word similarity and string similarity,” ACM Transactions on Knowledge Discovery from Data, vol. 2, 2009.

[29] J. Oliva, J. Serrano, M. del Castillo, and A. Iglesias, “SyMSS: a syntax-basedmeasure for short-text semantic similarity,” Data and Knowledge Engineering, vol. 70, no. 4, pp. 390–405, 2011.

[30] Liang-Chih Y, Hsiu-Min S, Yu-Ling L, Jui-Feng Y, and Chung-Hsien W, "Discriminative training for near-synonym substitution",COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics Pages 1254-1262 ,2010.

[31] Liang-Chih Y and Wei-Nan C,"Independent component analysis for near-synonym choice", Elsevier Decision Support Systems 55 146–155, 2013.

[32] Limin Yao Sebastian Riedel Andrew McCallum, "Collective Cross-Document Relation ExtractionWithout Labelled Data", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1013–1023, MIT, Massachusetts, USA, 9-11 October 2010.

[33] Minxin S, Duen-Ren L and Yu-Siang H, "Extracting semantic relations to enrich domain ontologies", J Intell Inf Syst (2012) 39:749–761, 2012.

Internet Sites

[34] http://wordnet.princeton.edu.

[33] http://opennlp.sourceforge.net/

[34] http://www.alchemyapi.com/