Journal of Theoretical and Applied Information Technology
31st August 2018. Vol.96. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECTIVE IDEA MINING TECHNIQUE BASED ON MODELING LEXICAL SEMANTIC

[1]MOSTAFA ALKSHER, [2]AZREEN AZMAN, [3]RAZALI YAAKOB, [4]EISSA M. ALSHARI, [5]RABIAH ABDUL KADIR, [6]ABDULMAJID MOHAMED

[1,2,3] Department of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Malaysia

[4]Faculty of Computer Science, IBB University, Yemen

[5]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, Malaysia

[6]Faculty of Computer Science, Sebha University, Sebha, Libya

E-mail: [1]alksher@yahoo.com, [2]azreenazman@upm.edu.my, [3]razaliy@upm.edu.my, [4]alsharia202@gmail.com, [5]rabiahivi@ukm.edu.my, [6]abdulmajid.h@gmail.com

ABSTRACT

Automatic extraction of hidden ideas from texts is extremely important that would help decision makers to identify and retrieve significant information, which possibly used to solve current problems. However, adequate measurements need to be utilized to verify candidate ideas. In existing idea mining measurement research, a well-balanced measurement is used to measure the distribution of the number of known and unknown terms from the idea text and the context text to find useful ideas within a text pattern. The existing models do not take into consideration the relationships between these terms which may share one or more semantic component. This leads to a limited characterization of potential ideas. Therefore, this paper proposes an improvement to the idea mining model by considering the semantic relationships among terms based on synonyms by using the WordNet. The effectiveness of the proposed model is evaluated on a dataset consisting of 50 randomly selected abstracts of scientific articles. Based on the results, the proposed model showed an improvement in the performance of the idea mining model where an increase of 28.4% is achieved.

Keywords: *Idea mining, Information retrieval, WordNet, text pattern, text mining.*

1. INTRODUCTION

Innovation has become the key to the success of many organizations or nations in order to be competitive in the real world. It is driven by the capability of its member or citizen to generate an interesting idea and making it work. Brainstorming has been used as an idea generation technique for decades [1]. However, it is both expensive and challenging creative process to discover interesting ideas in order to solve a problem or to assist in decision-making. In the process, textual resources such as scientific publications and the Web have been utilized as the source for the idea [2, 3].

A technique called *idea mining* was introduced in [2] to identify interesting idea from the text. It has been successful in many related applications such as in [4, 5, 6, 7]. The technique is inspired by the concept of the technological idea [8], in which an idea comprises of either an unknown solution to a known problem or a known solution to an unknown problem. In *idea mining*, the problem is simplified to finding a piece of text that consists of either known and unknown terms or noun phrases within the same context in the text. The context has been modelled as a snippet of the text called *text pattern* in [2], or a sentence in [9], or an edge between clusters of the conceptual graph [10].

Thorleuchter *et al*. proposed an *idea mining* technique in [2] based on the framework of information retrieval [11]. In order to identify idea an , the text is first divided into a set of *text patterns* by using a special windowing technique. The problem of *idea mining* is simplified as assigning scores to each *text pattern* and ranked them in a decreasing order based on the scores. As such, the top-ranked *text pattern* can be identified to contain the interesting idea from the text.

In the paper, the authors suggested an idea mining measurement model that includes identifying terms in the *text pattern* as *known* and *unknown* terms, and the measurement will take into account the frequency of those terms as well as the balance distribution of them within the *text pattern*. In order

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to accomplish that, the authors used another collection of text as the context to identify the idea, and the text is also divided into another set of *text patterns*. As such, the *text patterns* from the text containing the idea (*idea text*) are compared to those *text patterns* from the other collection of text (*context text*) to identify the *known* and *unknown* terms within the *idea text*. As shown in Figure 1, *known* terms are those terms common in both *text patterns*, while *unknown* terms only appear in the *idea text*.

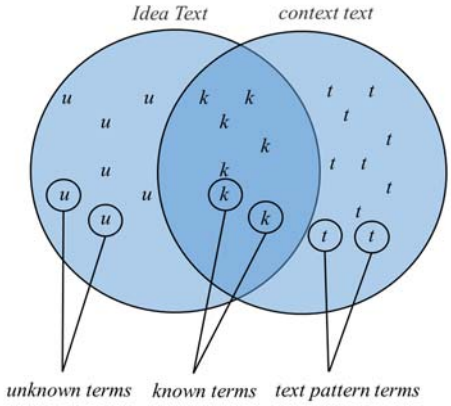

Figure 1:The combination of known and unknown terms visualized as a Venn diagram

Unfortunately, the proposed model does not take into consideration that there could be a semantic relation between those *known* and *unknown* terms. For instance, those terms may share a similar meaning through synonyms such that it should belong to the same *known* terms set. As the model depends largely on the correct classification of those terms, it will affect the performance of the model in identifying the idea from the text.

Therefore, this paper proposes an improvement to the *idea mining* model proposed in [2] by considering the semantic relationships among the *known* and *unknown* terms within the *text pattern*. In particular, this paper focuses on lexical semantic among terms based on synonyms by using the WordNet. It is argued that by modeling the semantic relationships between terms within the *text, the pattern* will improve the performance of the *idea mining* model.

The remaining of the paper is structured as follows. A review of related work is presented in the following section. Section 3 elaborates the proposed *idea mining* model that incorporates lexical semantic relationships between terms. Then, the experimental setup and the assumptions for the evaluation of the model are discussed in details in Section 4. After that, the results are analyzed in Section 5. Finally, the conclusion of the work is given in Section 6.

2. RELATED WORK

2.1 Knowledge creation

Text-based knowledge creation plays an important role in scientific discovery. Many useful studies have contributed to the organizing of valuable information and the identification of facts in textual data that will solve practical problems. In studies by [12, 13], word processing tools such as information retrieval (IR) and information extraction (IE) were used to support the creation process.

2.2 Idea generation technique

Idea generation is generally the starting point of innovation. Generating ideas, therefore, is a well-known topic that is related to creativity in psychology and cognitive science [14]. Liu *et al.* in [9] established an idea generation approach to automate the brainstorming process via machine learning. Besides that, there are several other approaches dealing with the creation of new ideas. By referring to these approaches, the adequate rationale for the idea mining process is proposed consists of three steps:

1. *Preparation of collecting document;* In the first step, the idea mining approach focused on the provided textual information by the user that contained a description of the problem.
2. *Extraction of text patterns from idea text and context text:* In the second step, the user has to provide further textual information which may contain an idea that probably can solve the problem. Therefore, with an automatic process, text patterns appeared in a very large number of overlapped text should be extracted (see section 3.2). Text patterns are then compared to the problem description by using a specific idea mining measure. With this measure, text patterns can be classified as an idea.
3. *Evaluation of text patterns:* In the third step, all extracted text patterns will be compared to the problem description by using a specific idea mining measure and evaluated for their usability and usefulness.

2.3 Idea definition

Recently, researchers have observed an increased interest in the issues of idea formulation. [9] addressed significant issues in idea generating.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The first issue is based on idea concept, which is not well-defined. However, experts find it difficult to evaluate, since there is no specific concept of the idea. The second issue, the composing of new ideas are constructed upon domain knowledge that is problematic to formalize. Literature shows a variety of idea definitions, but there is no precise and formal definition of the concept of an idea. However, based on an investigation by [15], a formal definition of the concept of the idea as a solution to the problem to improve the service or the process within the organization is presented. The production of ideas, whether about processes, products, or other relevant phenomena, considered to be both novel and useful [16].

The idea is defined as an association between two or more entities [17], or a pair of attributes called problem and solution, both represented by noun phrases [9]. The researchers in [2, 3, 6, 18] referred the definitions of the idea as a combination of several words appearing together in a textual pattern (purpose and a corresponding means) represented as term vectors, where all terms from the *idea text* (*known*) should not occur in text pattern from the *context text*. *known* terms refer to the terms that appear in both *idea text* and *context text*. While the *unknown* terms refer to the terms that only appear in the *idea text* and has no matches in the document collection. According to philosophy [19], there will be no idea if all terms in text pattern from the *idea text* are *unknown* because there is no relation to the *context text*.

2.4 Idea mining approaches

In recent years, a large number of models such as the TF-IDF weighting model [20], text extraction model [21] and idea mining model [5] have been proposed to estimate the probability of a query term in a document based on the distribution or the statistics of the term, such as the term frequency, the collection term frequency, and the term's similarity. However, most of the models use the bag of words approach to extract the potential text needed.

Several studies investigating the extraction of ideas have been carried out. The general idea mining approach is proposed in [2], who introduced the concept of idea mining as an automatic process of mining new ideas from the unstructured text. They used similarity-based measures to investigate how one can extract new and useful ideas from unstructured text from research proposals. This approach aims to extract the ideas from the *idea text* provided by the user and evaluate them based on their ability to solve the problems described in the document collections. It models *known* and *unknown* terms as to be well balanced based on their co-occurrence in text pattern. The balancing measure of *known* and *unknown* terms is considered to be the backbone of mining new idea field. The above finding is consistent with the study by [22] who proposed a method using patent databases for the fundamental idea search represented by object-condition-action attributes for solving a creative problem.

In another study, the authors proposed an idea web extraction approach from online domain [3]. This work is similar to the baseline model [2] and can be simplified as a text mining process of finding dissimilarity of text terms between a problem description and a problem solution idea. Furthermore, [4, 23] analyzed the impact of textual information from e-commerce websites on their commercial success. In addition, [9] extracted noun-phrases within the titles and abstracts of the publications using Part-of-speech (POS) tagging algorithm. They assumed that noun-phrases are sufficient to represent candidate ideas using n-grams model. Their approach extracted the idea candidates by classifying the idea components into *problem* and *solution* phrases. Then, pairing them with a co-occurring known-idea triple using a collaborative filtering algorithm.

A search of the literature revealed a study on idea discovery through data synthesis (*events and their relations*), which proposed a dynamic process of idea discovery to turn data into scenario maps by cluster the eliciting human insights [10]. The study by [24] also proposed a semi-automatic text mining technique to retrieve useful text patterns from ideas posted on crowdsourcing application. This technique is consistent with the proposed measure of [2] that used a well-balanced measurement to generate ideas for new product development.

The research to date tends to focus on classification rather than extraction [25] .The authors proposed a classification method to automatically detect text as an idea or none idea in online communities using machine learning and text mining techniques. In their study, linear support vector machine was used to classify the text pattern of 3,000 texts, and human experts were employed to evaluate the extracted texts with the promising results of idea classification.

Similar terms and documents may not be semantically similar. In order to refine this estimation, [26, 27] predefined the text as classes that can be interpreted as a textual pattern which represents the characteristics of textual documents, such as, a scientific publication. With the same objective, [28] identified textual patterns to represent

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

weak technological signals from the Internet. In order to determine the effects of idea mining, [6] combined the idea indication (weak signal analysis) with idea mining which is used to filter the results and weak signal analysis using semantic clustering (LSI). It has been reported that this work has a limit to recognize the low information content which needs and extending further cross-cutting relations using different semantic clustering instead of using LSI.

2.5 Lexicon syntactical of text identification

In term of extracting the overlapping relations between terms that are not appearing in the preprocessing stage, a lexicon syntactical approach for the idea mining is proposed. WordNet is an online semantic dictionary, lexical database, for the English language [29, 30] developed at the University of Princeton [31] and continued to be maintained. The version used in this study is WordNet 3.0, which contains 155287 words organized in 117659 synsets for a total of 206941 word-sense pairs; in compressed form, it is about 12 megabytes in size [32]. WordNet provides many types of relationships among concepts. A synonym is WordNet's basic relation because WordNet uses sets of synonyms (synsets) to represent word senses. Synsets are related to each other with terminological relations such as synonyms relations [33]. Fellbaum reported that synsets are an asymmetric relation between word forms [33].

Despite prior findings, most studies have been conducted for many different purposes. However, the dominant research domain is to calculate the term's similarity rather than meanings [35]. In contrast, little attention has been paid to examine the possibility of detecting the representation of the text by identifying the relationship between terms. [36] used syntactic patterns to extract semantic relationships such as meronymy and synonymy from clinical documents. In other words, this study relied on syntactic patterns to facilitate the identification of the nouns and noun-phrases, which are the important terms in the documents. However, experimental results showed that not all syntactic patterns produced the same quality of semantic relationships due to the low frequency of the syntactic patterns in any of the documents although WordNet was used for extracting the noun-phrases. Furthermore, there is also other work that used nouns and noun-phrases to improve text extraction [37].

In another study, [38] developed a morphology system using WordNet to enhance creative ideation based on meronym/holonym relationship, whereas values are generated based on hypernym/hyponym relationship of WordNet. Ibekwe *et al.* [39] also used synonyms from WordNet to build a finite state automaton with syntactic patterns to tag sentences that used to summarize the scientific documents based on its category. However, the all mentioned work are all limited in specific types of sentences within documents.

In general, the proposed approach of integrating WordNet provides an effective means to generate a large number of alternatives, which may give an impression, and perception of the appropriate composition of the text extracted. Therefore, the processing of the existing idea mining framework will be modified according to the new proposed balancing model. However, existing approaches [2, 3, 5, 9, 24, 25, 40] have paid no attention to the relationships between terms from the characterization of potential ideas. The motivation of formulating a new measure is needed to extract potential ideas latent within a huge amount of text. It will help to improve idea identification performance and to effectively characterize more candidate ideas.

3. IDEA MINING FRAMEWORK

The aim of idea mining is to identify a piece of text that possibly contains an interesting idea. The framework was proposed by Thorleuchter *et al.* in [2] that consists of several steps such as preprocessing, *text pattern* creation, terms vector generation, *text patterns* similarity calculation and idea measurement, as shown in Figure 2. An improvement of idea mining proposed in this paper based on modeling lexical semantic is discussed in Section 3.5 and 3.6.

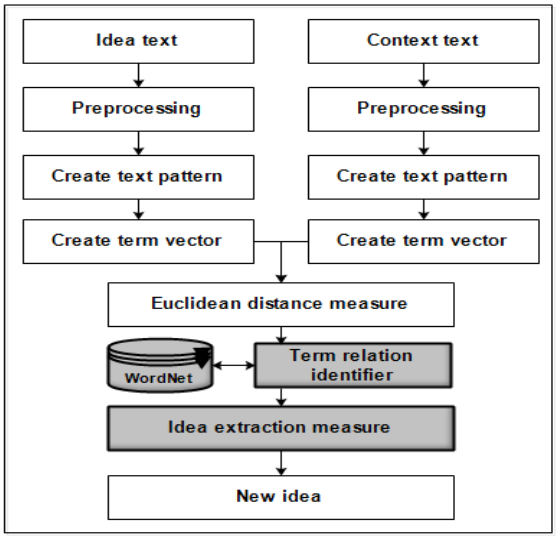

Figure 2: Framework of idea mining based on the lexical semantic model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1. Preprocessing

In the preprocessing step, both texts (*idea text* and *context text*) are transformed into an appropriate format for further processing [41]. First, the text is cleaned to remove scripting code, punctuation, as well as specific characters. Next, the texts are tokenized to split them into tokens (terms). The terms that appear only once or twice are also discarded as a result of Zipf's Law [42]. Those terms carry less meaning and removing them will reduce the size of the vocabulary. Finally, the well-known Porter stemmer algorithm [43] is applied to the tokens, and related terms with the same stem are grouped together for further processing.

3.2. Text Pattern Creation

In this step, both texts (*idea text* and *context text*) are split into a set of *text patterns*. The *text pattern* from the *idea text* represents the text phrases in which the interesting idea could be identified, while those from *context text* will be the context for the identification. Therefore, those *text patterns* from the *idea text* will be the candidates for the idea.

The algorithm for extracting *text pattern* from the text was proposed and elaborated extensively in [2], and it is adopted in this paper. In short, the text is scanned in order to identify stopwords and non-stop words. For each non-stopword appears in sequence, a set of other non-stopwords surrounded by a number of stopwords to its left and to its right are selected based on the length l as shown in Figure 3. In the example, the non-stopword being processed is 'extracting' and $l = 2$. From left, the non-stopword term 'process' is selected because it is surrounded by two stopwords which are 'as' and 'off', while the terms 'new', 'useful' and 'idea' are selected from right as they are surrounded by the stopwords 'and' and 'form'. Based on the previous study, the best l is set to 8 so that they contain all terms representing the idea [7]. Due to the potential overlap in the process, many duplicates *text patterns* will be generated, thus the duplicates are removed.

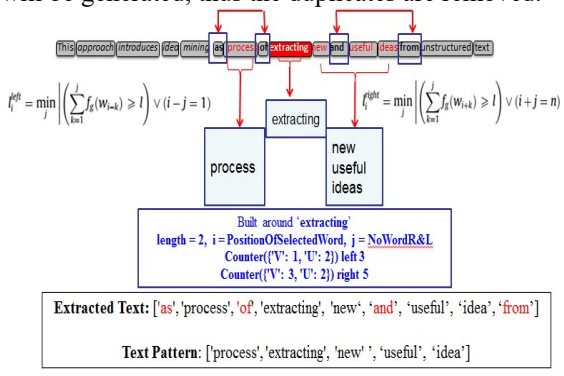

Figure 3: An example of how text pattern is created from text

3.3. Term vector creation

The *text patterns* discovered from the *idea text* are compared to the *text patterns* from the *context text* to find the top similar pairs. As such, term vectors in vector space model are created to represent every *text pattern*. In this paper, tf-idf weighting proposed in [44] is used. In essence, a term t in the *text pattern* d is weighted based on its frequency $tf_{(t,d)}$ in d and its inverse document frequency idf_t in the collection of *text patterns*.

$$tf - idf_{t,d} = tf_{t,d}.idf_t \quad (1)$$

3.4. Similarity measure

In order to compute the overlap between two vectors, a frequently used distance measure is the *Euclidian distance*, which is suited for idea mining implementation [45, 46]. In practice, this similarity measure considered as a proper predictor for vector (word) similarity calculations and it is very sensitive values frequency counts occur.

In more recent studies for mining ideas [3, 5, 18], the term vector from *idea text* are compared to all vectors from *context text* using similarity measures, and then only the top similar pairs for more than the threshold is considered for idea mining measurement presented in Section 3.6.

Traditionally, the extraction of a new idea depends on the *context text* where textual patterns representing known ideas occur [18]. This is further simplified as finding part of a text that potentially contains both terms (or words) that are *known*, refers to the intersected terms in *idea text* and *context text*, and *unknown*, refers to the terms solely in *idea text*. This is visualized in the Venn diagrams of Figure 1 in which the intersection of these documents shows the overlapping terms to produce the *known* and *unknown* terms which will be used for the extraction measurements.

3.5. Term relation identifier

The existing model is syntactical and does not consider lexical relations between words for identifying the *known* and *unknown* terms. In this paper, an approach to incorporate term relation in idea mining model is proposed. In order to detect the relations between terms in *idea text* with terms appear in both *idea text* and *context text*, the proposed model integrates WordNet with the processed *text patterns*. WordNet provides a tool to search dictionaries conceptually by grouping words into synonym sets (called synset). For this purpose, this work limits the relations between the *synsets* to synonym relationships. WordNet uses sets of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

synonyms (*synsets*) to represent the *text pattern* which contains a lexical relation that shares one or more semantic component. Accordingly, the most appropriate combinations are selected as a *SynWord* set, which are words that resulted from the relationship between *known* and *unknown* terms. In order To realize the processing, it is described as follows:

Definition 1. Let *K= idea text* ∩ *context text*, be the set of *known* terms. Let *U = idea text - K* , be the set of *unknown* terms in *idea text*. Let K_{syn} be the list of the synonym words of set K that will be extracted from WordNet. Let U_{syn} be the list of the synonym words of set U that will be extracted from WordNet. Then the description of *SynWord* is defined in Equation 2 as:

$$SynWord = (K \cap U_{syn}) \cup (U \cap K_{syn}) \quad (2)$$

Based on Definition 1, the new characterizations of *known* terms are set and defined as:

$$q = K - SynWord \quad (3)$$

and the *unknown* terms are set and defined as:

$$p = U - SynWord \quad (4)$$

The overlapped components after identifying the *SynWord* terms are visualized in the Venn diagrams of Figure 4 to produce the terms *known*, *unknown* and *SynWord*.

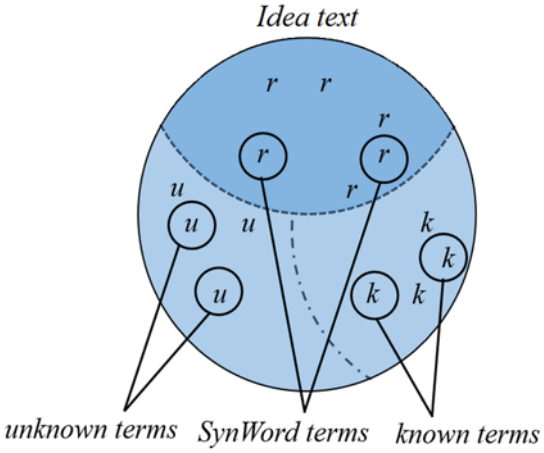

Figure 4: Venn diagram of the co-occurrence relation terms based on the intersection between known and unknown

3.6. Lexicon relation idea mining measurement

In this section, the idea mining measurements proposed in [2] is investigated, which support users to extract ideas in text phrases based on *known* and *unknown* terms. Further, the proposed extraction measurement that considers the lexical relationships between terms which would produce more term's dimensions is presented in details.

3.6.1. Baseline approach

The recent technique [2] proposed a sub-measures of idea mining to compare the relationship between a set of pair texts based on the statements introduced in [8] outlined as if:

- the number of *known* and *unknown* terms is well balanced,
- *known* terms occur more frequently in the *context of text* than other terms,
- *unknown* terms occur more frequently in *idea text* than other terms, and
- specific terms occur, which are characteristic for a new idea.

The heuristic idea mining measurement took into account these logic statements and conceived them as parameters, which are discussed as follows:

First, the balancing between *known* and *unknown* terms is computed based on:

Definition 2. Let γ be defined as the number of terms from a *text pattern* of the *idea text*, let β be defined as a set of terms from a *text pattern* of the *context text*. Let $p = |\gamma|$ be defined as a the number of terms in *idea text*, and let $q = |\gamma \cap \beta|$ be defined as the number of terms existing in both *idea text* and *context text*. Then, the balancing between the *known* and *unknown* distribution terms m_b is defined as:

$$m_b = \begin{cases} \frac{2(p-q)}{p} & q \geq \frac{p}{2} \\ \frac{2q}{p} & q < \frac{p}{2} \end{cases} \quad (5)$$

Further, a set of parameter values has to be identified for the idea mining approach. These parameters can be distinguished into two categories:

- Parameters used to identify a *text pattern* (e.g., length, stop words and non-stop words).
- Parameters used to calculate the idea mining measure (e.g. m_b, m_{f_k},m_{f_u},m_c, $\hat{a}$, and z), where $\hat{a}$ is used as a threshold for the classification decision, and the percentage z to distinguish frequent words from non-frequent words. Thus, The optimal value of $\hat{a}$ is set to 60% for the total calculation of idea mining measure in [2].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The overall calculation of the idea mining m_{idea} combining all parameter's values as:

$$m_{idea} = \begin{cases} m_b + m_{fk} + m_{fu} + m_c & (p \neq q) \\ 0 & (p = q) \end{cases} \quad (6)$$

where,
m_b as a measure for the balance of *known* and *unknown* terms distribution,
m_{f_k} as the number of frequently *known* terms occur in *idea text* over the number of all *known* terms in *context text*,
m_{f_u} as a measure for frequently of *unknown* terms occur in the *idea text*,
m_c a set of characteristic terms (e.g, quicker, better, higher etc.).

Finally, this measure computes based on the aforementioned sub-measures to extract ideas depending on the balancing between the bi-vectors. Therefore, an improvement to the *idea mining* model [2] by considering the semantic relation between the *known* and *unknown* terms within the *text pattern* is discussed in the following section.

3.6.2. The proposed model: WordNet-based relation extraction

In spite of the function of m_b in Equation 6 that calculates the balancing of only *known* and *unknown* vectors, however, this measure is restricted for bi-vectors. To overcome this limit, a new balancing measurement is developed and introduced to compute the multi-vectors by adding *SynWord* vector. As shown in Equation 9, the new balancing measure is proposed for identifying the balancing between the multi-vectors as (*known*, *unknown* and *SynWord*).

$$m_{b_r} = 1 - \frac{1}{3}\sum_{i=1}^{3} \frac{|q_i - pw_i|}{pw_i} \quad (7)$$

where,
m_{b_r} as a measure of (*known, unknown* and *SynWord*) balancing vectors,
q_i is a number of *known, SynWord or unknown* vectors in a *text pattern* of the *idea text* set,
p is a number of terms in *idea text*,
w_i is the weighting ratio of the selected *known, SynWord* or *unknown* vectors
m_{f_s} as a measure for the most frequent *SynWord*

Finally, the proposed idea mining measure aggregates all the sub-measures as:

$$m_{idea_r} = \begin{cases} m_b + m_{f_k} + m_{f_u} + m_{f_s} + m_c & (p \neq q), \\ 0 & (p = q). \end{cases} \quad (8)$$

4. EXPERIMENTAL SETUP

The relevant ideas can be generally ranked by the assumed relevance and presented as a new useful idea. Performance measures evaluate computed rankings based on the expert's feedback and then allow comparing different filtering or recommendation strategies [47]. In order to evaluate the results of the proposed model, first, the effect of excluding *SynWord* terms as in Equation 2 is investigated. The related terms, which are not considered as *known* and *unknown* terms are excluded from the computation to know the effect of the relation. As such, these related terms are considered as noises to the model, most recently mentioned as *SynWordExc* in this paper, where [2] is applied to identify the effect of removing the related terms.

In contrast, instead of treating *SynWordExc* terms as noise for the model, they would be treated as a different class to be used for modeling the semantic balancing between *known*, *unknown* and *SynWord* terms, in order to characterize more attributes that would help to effectively compose more candidate ideas using Equation 8.

For this study, the performance measure commonly used in information retrieval such as an *average precision* (AP) is adopted [48]. It is being used in this research to evaluate the precision of the top *k* results of each *idea text* from each paper. Additionally, the *mean average precision* (MAP) is utilized, which is designed to compute the average precision over sorted result lists (rankings) [49]. The further measure that is applied for ranking is the *normalized discounted cumulative gain* (NDCG), which is used to measure the graded relevance judgments rather than the binary relevance as in the mentioned two measures [50, 51].

First, in the case of binary relevance, in order to evaluate the results of the model, the *precision* is calculated based on how many documents are relevant to the query among the returned documents, it is being calculated based on *true positives* (tp), *false positives* (fp) and *true negatives* (tn), as shown in the below formula that is obtained from a confusion matrix in Table 1.

where,

$$precision = \frac{tp}{tp + fp} \quad (9)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Confusion matrix to calculate precision values

	Predicated positive	Predicated negative
Actual positive	tp	fn
Actual negative	fp	tn

For each paper, AP is defined as the average of the P@n values for all relevant documents:

$$AP = \frac{1}{|R|} \cdot \sum_{i=1}^{n} p\,re_{(i)} \cdot rel_{(i)} \quad (10)$$

where $|R|$ is the total number of relevant documents, $pre_{(i)}$ is the precision of the top 10 queries, and $rel_{(i)}$ is an indicator function equaling 1 if the item at rank n is a relevant document, zero otherwise:

$$rel(i) = \begin{cases} 1, & \text{if } i \text{ doc is relevant} \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Interestingly, the main advantage of MAP measure is to consider whether all of the relevant items tend to get ranked highly. The output of the idea mining measures is ordered, and only the top 10 results are considered, then the precision for the top 10 is computed. This is known as Precision at 10, or P@10. So in this research, the P@10 extracted ideas would be considered particularly the relevant texts that scored high in the top and computed as:

$$MAP = \frac{1}{|Q|} \cdot \sum_{Q_i \in Q} A\,P_{(Q_i)} \quad (12)$$

where Q is the number of queries.

However, AP and MAP measures can only handle cases with the binary decision as (*relevant or not relevant*), the proposed methodology is applied to other evaluation measures and demonstrate how NDCG (*normalized discounted cumulative gain*) is estimated with multiple scales of judgments. The reason for applying the discount is that the probability of viewing a document decreases with respect to its rank. Thus, NDCG@10 is a used for weighting the higher ranked top 10 list to have more effect on the resulting score. It is computed as follows,

First, the *cumulative gain* (CG) simply adds the ratings up to a specified rank position:

$$CG_p = \sum_{i=1}^{p} r\,ating(i) \quad (13)$$

where the $rating_{(i)}$ is the graded relevance level of the document retrieved at rank *i*.

Then, the *discounted cumulative gain* (DCG) at rank position p (DCG@p) discounts each rating based on its position in the results and formed by:

$$DCG_p = rating_1 + \sum_{i=2}^{p} \frac{rating_{(i)}}{log_2(i+1)} \quad (14)$$

And finally, the normalization is accomplished by dividing the query's DCG with ideal DCG (IDCG) values, which is the DCG of the best possible results based on the given ratings:

$$IDCG_p = rating_1 + \sum_{i=2}^{|REL|} \frac{rating_{(i)}}{log_2(i+1)} \quad (15)$$

where $|REL|$ is the number of the best ratings up to position p, and $|REL| \leq p$.

Then, the NDCG normalizes the gain to a number ≤ 1 at any rank position. To summarize, the NDCG for a given query can be defined as:

$$NDCG_p = \frac{DCG_p}{IDCG_p} \quad (16)$$

4.1. System implementation

In order to show the applicability and validity of our proposed approach, we implemented an automated system to support the new idea mining processes. The suggested idea – Lexical-based – extraction is developed using Python programming particularity using NLTK[1] (Natural Language Processing Toolkit) package and WordNet package in python.

An implementation of the three models was conducted for predicting the MAP performance. The following section describes the results to compare and evaluate the trade-off between the baseline, *SynWordExc,* and *SynWord* models.

4.2. Dataset

The system automatically identified new ideas from 50 randomly selected abstracts of scientific articles from different domains (e.g., Social, medical and technological domains), each abstract corresponds to text to identify the *idea text*. This data set is limited to English articles because most of the relevant scientific articles use English as the standard language. The scientific paper usually

[1]www.nltk.org

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

depends on further publications formed in references [52]. Thus, for each abstract, all reference's abstracts of each paper as a *context text* for that *idea text* are extracted. Then the idea mining techniques in [3] are applied to identify the ideas. The text produced by the technique would be ranked in a list of text patterns from *idea text*.

Finally, the top 10 ranked list is selected to be evaluated by human experts. Alksher *et al.* proposed a generic framework of idea evaluation process by intervening the human judgment, who were authoring these papers or familiar with, by identifying ideas from these *text patterns* manually without using the idea mining techniques [53]. Human experts assessed and rated 500 extracted ideas from 50 abstracts to manually identify the idea components. The experts check and score the top ranked ideas to be defined as the ground truth for the evaluation, then a statistical evaluation to analyze the validity and reliability of the extracted ideas is used. The characteristics of this dataset are depicted in table 4.2.

Table 2: The characteristics of 50 abstracts dataset

No of paper's abstracts	50
No of abstracts of *context text*	1602
Average number of abstracts per papers	32
No of query *text pattern*	2038
No of *text pattern* from *context text*	65657
No of search *text pattern* in total	67695
Average number of queries per paper	41

4.3. Parameter selection

In this paper, the parameters of the idea mining measure $m_b, m_{f_k}, m_{f_u}, m_{f_r}, m_c, z$ are determined. Therefore, the parameters are heuristically determined as (40,20,20,20,0 and a percentage of 30%) respectively, as well as the parameter for the length of the text patterns (l =8). Further, the top-10 extracted text that represent the potential ideas by the idea mining measures are selected in this research.

5. EXPERIMENTAL RESULTS

In order to assess the performance of the system, the relevant ideas retrieved during the evaluation were matched against the relevant selection of the expert's evaluation. AP, MAP, and NDCG measures were used to evaluate idea relevancy at the top 10 (P@10) ideas of each paper level. Table 3 clearly shows the variations between the models for 50 document papers. From this table, it can be seen that both the *SynWordExc* and *SynWord* models are better than the baseline model at almost all papers. This indicates that when queries are processed using *SynWord* extraction, they yield ideas that are more relevant compared to queries which are not processed.

Table 3: The performance measurement of SynWord model

	Baseline	*SynWordExu*	*SynWord*
P@10	0.644	0.720 (Δ9%)	0.758 **(Δ22.4%)**
MAP	0.702	0.817 (Δ11.5%)	0.940 **(Δ28.4%)**
NDCG	0.846	0.870 (Δ4%)	0.929 **(Δ11.1%)**

Based on the table, it is obvious that the proposed model outperforms the other models. In addition, the maximum MAP of the proposed model is 0.92 as compared to 0.70 for the baseline, which is an increase of 28.4%. In addition, the result is better than the *SynWordExc* (0.81). The *SynWord* model shows better performance with an average precision (0.758) compared to *SynWordExc* model with an average precision (0.70) and the benchmark model with a low average precision (0.64). The performance of the proposed model outperforms the baseline model with an increase of almost 22.4%, from 0.64 to about 0.76. These results were based upon data from Table 4 in Annexure 1.

In addition to the previous metrics, this research adopted the normalized discounted cumulative gain (NDCG) to determine the top first ranking. Table 3 reported the best average NDCG to the *SynWord* model (0.92) which leads to high ranking quality results, and outperform the benchmark model (0.84) with an increase of almost (11.1%).

Performing comparative analysis, *SynWord*, and *SynWordExc* indicate that including word relation outperformed better than the existing approach. Figure 5 shows the clear trend of increasing the value of *SynWord* for idea mining.

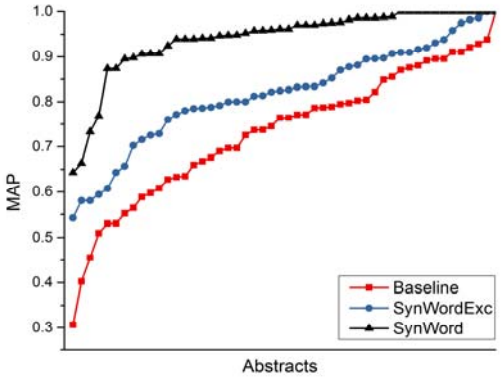

Figure 5: The MAP results for the proposed model compared to the baseline model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Considering these results, we can observe that employing the related words to identify the idea achieved better results than another idea classifier. Hence, it can be inferred from the results that it is very satisfactory. Although this improvement is not very high, it is significantly better than the baseline model at $p \leq 0.0017$ according to a t-test.

6. CONCLUSION

This paper investigates the problem of existing idea mining approaches due to the less consideration of the relationships between terms. This work proposes a new lexicon-syntactical approach that overloads the existing idea characterization. This approach extends the existing baseline approach by modeling the semantic relationships between the textual attributes that most likely comprise potential ideas. The effect of excluding relation terms that considered as noises to the model is investigated.

The results of this investigation show a meager increase of about (11.5%) compared to the baseline model. The proposed approach takes into account the relationships between the terms and treats them as a different class. The results of this study show an encouraging improvement and outperform the existing model with an increase of (28.4%). These results demonstrate that this model compares well to other automated methods and is a useful approach to idea mining.

Further investigations are needed to solve the duplication of the idea text within the idea text and context text. Further research might explore the effect of text position in order to identify the idea within the article paper.

ACKNOWLEDGMENT:

This work is supported by the Ministry of Higher Education Malaysia under the FRGS Grant (FRGS/1/2015/ICT04/UPM/02/5).

REFERENCES:

[1] B. M. Kudrowitz, D. Wallace, Assessing the quality of ideas from prolific, early-stage product ideation, Journal of Engineering Design 24 (2) (2013) 120–139.

[2] D. Thorleuchter, D. Van den Poel, A. Prinzie, Mining ideas from textual information, Expert Systems with Applications 37 (10) (2010) 7182–7188.

[3] D. Thorleuchter, D. Van den Poel, Web mining based extraction of problem solution ideas, Expert Systems with Applications 40 (10) (2013) 3961–3969.

[4] D. Thorleuchter, D. Van den Poel, A. Prinzie, Analyzing existing customers websites to improve the customer acquisition process as well as the profitability prediction in b-to-b marketing, Expert systems with applications 39 (3) (2012) 2597–2605.

[5] D. Thorleuchter, D. Van den Poel, Extraction of ideas from microsystems technology, Advances in Computer Science and Information Engineering (2012) 563–568.

[6] D. Thorleuchter, D. Van den Poel, Idea mining for webbased weak signal detection, Futures 66 (2015) 25–34.

[7] D. Thorleuchter, D. Van den Poel, Identification of interdisciplinary ideas, Information Processing & Management 52 (6) (2016) 1074–1085.

[8] D. Thorleuchter, Finding new technological ideas and inventions with text mining and technique philosophy, in: Data Analysis, Machine Learning and Applications, Springer, 2008, pp. 413–420.

[9] H. Liu, J. Goulding, T. Brailsford, Towards computation of novel ideas from corpora of scientific text, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 541–556.

[10] H. Wang, Y. Ohsawa, Idea discovery: A scenario-based systematic approach for decision making in market innovation, Expert Systems with Applications 40 (2) (2013) 429–438.

[11] C. D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.

[12] D. R. Swanson, A second example of mutually isolated medical literatures related by implicit, unnoticed connections, Journal of the American Society for Information Science 40 (6) (1989) 432.

[13] M. Weeber, H. Klein, L. de Jong-van den Berg, R. Vos, et al., Using concepts in literature-based discovery: Simulating swanson's raynaud–fish oil and migraine–magnesium discoveries, Journal of the Association for Information Science and Technology 52 (7) (2001) 548–557.

[14] C. M. Crawford, New products management, Tata McGraw-Hill Education, 2008.

[15] C. Riedl, N. May, J. Finzen, S. Stathel, V. Kaufman, H. Krcmar, et al., An idea ontology for innovation management, International Journal on Semantic Web and Information Systems (IJSWIS) 5 (4) (2009) 1–18.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] R. C. Litchfield, L. L. Gilson, P. W. Gilson, Defining creative ideas: Toward a more nuanced approach, Group & Organization Management 40 (2) (2015) 238–265.

[17] D. Swanson, Literature-based discovery? the very idea, Literature-based discovery (2008) 3–11.

[18] D. Thorleuchter, S. Herberz, D. Van den Poel, Mining social behavior ideas of przewalski horses, in: Advances in Computer, Communication, Control and Automation, Springer, 2011, pp. 649–656.

[19] G. Rohpohl, Das ende der natur, Naturauffassungen in Philosophie, Wissenschaft und Technik (1996) 143–163.

[20] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (11) (1975) 613–620.

[21] V. Baghela, S. Tripathi, Text mining approaches to extract interesting association rules from text documents, International Journal of Computer Science Issues 9 (3) (2012) 545–552.

[22] D. M. Korobkin, S. A. Fomenkov, S. G. Kolesnikov, A. B. Golovanchikov, Technical function discovery in patent databases for generating innovative solutions, in: Multi Conference on Computer Science and Information Systems, Vol. 2016, 2016, p. 241.

[23] E. K. Ozyirmidokuz, M. H. Ozyirmidokuz, Analyzing customer complaints: A web text mining application, International Conference on Education and Social Sciences.

[24] T.-C. Dinh, H. Bae, J. Park, J. Bae, A framework to discover potential ideas of new product development from crowdsourcing application, arXiv preprint arXiv:1502.07015.

[25] K. Christensen, S. Nørskov, L. Frederiksen, J. Scholderer, In search of new product ideas: Identifying ideas in online communities by machine learning and text mining, Creativity and Innovation Management 26 (1) (2017) 1730.

[26] Y. Ko, J. Seo, Text classification from unlabeled documents with bootstrapping and feature projection techniques, Information Processing & Management 45 (1) (2009) 70–83.

[27] J. Finzen, M. Kintz, S. Kaufmann, Aggregating web-based ideation platforms, International Journal of Technology Intelligence and Planning 8 (1) (2012) 32–46.

[28] D. Thorleuchter, T. Scheja, D. Van den Poel, Semantic weak signal tracing, Expert Systems with Applications 41 (11) (2014) 5009–5016.

[29] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (11) (1995) 39–41.

[30] A. Kilgarriff, Wordnet: An electronic lexical database (2000).

[31] P. University, About wordnet (2015). URL http://wordnet.princeton.edu

[32] W. statistics, Wnstats(7wn) manual page, http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html, (Accessed on 11/03/2017).

[33] C. Reynaud, B. Safar, Exploiting wordnet as background knowledge, in: Proceedings of the 2nd International Conference on Ontology Matching-Volume 304, CEURWS. org, 2007, pp. 291–295.

[34] C. Fellbaum, Wordnet, in: Theory and applications of ontology: computer applications, Springer, 2010, pp. 231–243.

[35] R. Richardson, A. Smeaton, J. Murphy, Using wordnet as a knowledge base for measuring semantic similarity between words (1994).

[36] K. Liu, W. Chapman, G. Savova, C. Chute, N. Sioutos, R. S. Crowley, Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents, Methods of information in medicine 50 (05) (2011) 397–407.

[37] K. Barker, N. Cornacchia, Using noun phrase heads to extract document keyphrases, in: Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 2000, pp. 40–52.

[38] Y. Geum, Y. Park, How to generate creative ideas for innovation: a hybrid approach of wordnet and morphological analysis, Technological Forecasting and Social Change 111 (2016) 176–187.

[39] F. Ibekwe-Sanjuan, F. Silvia, S. Eric, C. Eric, Annotation of scientific summaries for information retrieval, arXiv preprint arXiv:1110.5722.

[40] P. Kruse, A. Schieber, A. Hilbert, E. Schoop, Idea mining–text mining supported knowledge management for innovation purposes, Americas Conference on Information Systems (AMCIS), 2013 Nineteenth Americas Conference.

[41] W. Fan, L. Wallace, S. Rich, Z. Zhang, Tapping the power of text mining, Communications of the ACM 49 (9) (2006) 76–82.

[42] J. Zeng, J. Duan, W. Cao, C. Wu, Topics modeling based on selective zipf distribution, Expert Systems with Applications 39 (7) (2012) 6541–6546.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[43] M. F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[44] G. Salton, J. Allan, C. Buckley, Automatic structuring and retrieval of large text files, Communications of the ACM 37 (2) (1994) 97–108.

[45] A. Hotho, A. Nurnberger, G. Paaß, A brief survey of text mining., in: Ldv Forum, Vol. 20, 2005, pp. 19–62.

[46] E. Emad, E. M. Alshari, H. Abdulkader, Boolean information retrieval based on semantic, in: International conference on intelligent computing and information, Vol. 8, 2013, pp. 87–93.

[47] C. Faloutsos, D. W. Oard, A survey of 47 retrieval and filtering methods, Tech. rep. (1998).

[48] E. M. Alshari, Semantic arabic information retrieval framework, arXiv preprint arXiv:1512.03165.

[49] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, S. Quarteroni, An introduction to information retrieval, in: Web information retrieval, Springer, 2013, pp. 3–11.

[50] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, Vol. 283, AddisonWesley Reading, 2010.

[51] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, T.-Y. Liu, A theoretical analysis of ndcg ranking measures, in: Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013), 2013.

[52] V. R. Khare, R. Chougule, Decision support for improved service effectiveness using domain aware text mining, Knowledge-Based Systems 33 (2012) 29–40.

[53] M. A. Alksher, A. Azman, R. Yaakob, R. A. Kadir, A. Mohamed, E. M. Alshari, A review of methods for mining idea from text, in: Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on, IEEE, 2016, pp. 88–93.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: The overall results of Mean Average Precision Values

Papers	Baseline	SynWordExu	SynWord
1	0.85	0.91	0.97
2	0.77	0.91	0.99
3	0.76	0.93	0.96
4	0.86	0.91	1.00
5	0.40	0.70	0.98
6	0.31	0.59	0.87
7	0.80	0.82	0.99
8	0.79	0.83	1.00
9	0.61	0.73	1.00
10	0.63	0.81	0.94
11	0.67	0.72	0.91
12	0.76	0.83	0.90
13	0.79	0.92	1.00
14	0.82	0.92	0.99
15	0.89	0.89	1.00
16	0.88	0.88	0.97
17	0.92	0.96	1.00
18	0.77	0.85	0.94
19	0.91	0.98	1.00
20	0.70	0.80	0.94
21	0.80	0.83	0.96
22	0.74	0.77	0.94
23	0.79	0.84	0.97
24	0.70	0.78	0.99
25	0.90	0.90	0.99
26	0.53	0.58	0.91
27	0.56	0.83	0.90
28	0.66	0.80	0.96
29	0.69	0.79	0.92
30	0.51	0.58	0.96
31	0.75	0.82	0.95
32	0.91	1.00	1.00
33	0.60	0.66	0.87
34	0.73	0.80	0.95
35	0.68	0.79	0.95
36	0.59	0.76	0.91
37	0.53	0.54	0.66
38	0.63	0.88	0.96
39	1.00	1.00	1.00
40	0.63	0.73	0.73
41	0.87	0.87	1.00
42	0.46	0.64	0.64
43	0.90	0.90	0.98
44	0.74	0.78	0.95
45	0.94	0.99	1.00
46	0.88	0.94	0.98
47	0.79	0.68	0.97
48	0.80	0.88	0.94
49	0.93	0.98	1.00
50	0.55	0.61	0.77
Average	70.25	81.75	94.08