British Journal of Applied Science & Technology
17(4): 1-9, 2016, Article no.BJAST.27571
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

Classification of News Document in English Based on Ontology

Lailil Muflikhah[1*] and Aldi Sunantyo Ali Murdianto[1]

[1]Faculty of Computer Science, Brawijaya University, Malang, East Java, Indonesia.

Authors' contributions

This work was carried out in collaboration between both authors. The both authors designed the study, performed the statistical analysis, designed and implemented prototype of system.

Article Information

DOI: 10.9734/BJAST/2016/27571
Editor(s):
(1) Ana Pedro, Department of Education, University of Aveiro, Aveiro, Portugal.
Reviewers:
(1) S. K. Srivatsa, Prathyusha Engg College, Chennai, India.
(2) Hui Li, Icahn Medicine School at Mount Sinai, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history/16385

Original Research Article

Received 8th June 2016
Accepted 13th September 2016
Published 29th September 2016

ABSTRACT

Aims: This paper aims to propose ontology method of news document classification. The common method of document classification is based on morphology of term, without considering the meaning. It is impact to the number of term-document and computational time. Furthermore, the performance is decrease, even though the number of training data is increase.
Methodology: The main idea of ontology is to handle the similarity of terms that have different morphological form but the same meaning (synonym). The ontology is built using WordNet database to find similary of meaning among terms-document. The terms that have similar meaning are merged including their term frequency to be constructed in vector space model. After that, the unknown document is classified using cosine similarity measurement of the weight-term. The text document that is used is English news text in general topic, such as interest, money-fx, trade, and crude. The experiment is compared to the conventional method which is document classification without ontology.
Results: Classification of news document can be implemented using cosine similarity method based on ontology. The performance measure of this method including precission, recall and f-measure has increased eventhough the number of terms is reduced.

Keywords: Document classification; ontology; cossine similarity; WordNet.

**Corresponding author: E-mail: laililmf@gmail.com;*

1. INTRODUCTION

News is one of the tools to get information about something. It can be presented in written and oral form. The news which is presented in written form is presented through electronic and printed media. It is to be grouped by content such as sports news, economics, science, and so forth. The contents of the news is stored in the form of text, and is categorized into the group.

The text is stored in large data volume, therefore we need a method of organizing information for easy retrieval of the text. Text is unstructured data, and if the text is not organized then the information retrieval process would require a long process. The process of organizing the text is Text Mining. One of the text mining tasks is classification and organization of documents based on their contents or called text Categorization [1].

Some algorithms used in text Categorization are including the K-Nearest Neighbor algorithm, Naïve Bayes Classifier, and ID3. According to Yiming which, algorithm K-Nearest Neighbor (KNN) has a better performance compared with the C4.5 decision tree algorithm and Rocchio algorithm. The advantages of KNN algorithm is able to handle text classification where data used large dimension [2]. All these algorithms are depend on the number of data and various data to achieve the high performance. This algorithm involved to similarity or distance concept based on the morphology of word, such as the frequent of word. Cosine similarity method is one of the most popular measure to apply in text mining and information retrieval of text document [3,4].

Therefore, this research is proposed to classify text document using cosine similarity with consider to similarity of meaning, refer to the same concept [5]. To determine the similarity of meaning between the terms, it is added background knowledge to the text. One of the background knowledge that can be used is WordNet lexical database conceptually linking words and semantics [6].

Many recearch are conducted on focus ontology approach of document classification. Ontology based systems for Classification of Web pages, even though it is less productive to build a sophisticated classification [7]. Then, it was developed their ontology semi-automatically in domain economy, and the result cannot achieve high accuration due to their ontology is so not descriptive [8]. Also, research on ontology is applied to Classification email and news in digital format [9] and Classification system for electronic newspapers by Tenenboim, et al. [10]. However, the result cannot obtain high performance due specific domain and Furthermore, the research conducted on Text Categorization Using Word Net [11]. Domain based Clasification of Punjabi Text Documents using Ontology and Hybrid (using Naive Bayes) Based Approach and this method can obtain high performance [12].

The main principle of document classification is to input new document of the unknown class or category to training documents that have been known their categories. The training process is to determine the similarity between the testing data and any data training. They are similar if a set of terms are appear in both documents. The more the same term, the more similar two texts is. The determining process of document similarity has a weakness because if there is a test data which has different terms even though it has the same meaning. Therefore, this research is using ontology approach to classify the documents.

This paper is organized as follows. The first is an introduction including the background of this research and the related work to the previous research. Then, the related literature review is explained in the second part. After that, methodology which consists of the steps to conduct this research. The experimental result and analysis are provided in fourth step. Finally, conclussion and future work are presented.

2. CLASSIFICATION OF DOCUMENTS

Classification of documents is a field of research in developing methods of information acquisition to determine or categorize a document into one or more groups which have been previously recognized automatically based on the contents of the document. Document classification aims to classify unstructured documents into groups that describe the content of the document such as news articles.

The main component of text classification is a preprocessing data, classifier construction, and document classification. Preprocessing the data is changing the representation of the document that will be used as training data, validation, and classification process. Classifier construction is a learning process of testing document to training documents. Therefore, the classification

of document is put into a certain class of test documents [13].

2.1 Porter-stemmer Algorithm

Stemming is one of the steps for preprocessing of document classification. It is purposed to get the root word in order to reduce the feature of document. Porter-Stemmer algorithm is a method for stemming of text document. It consists of five linear steps. In every step, there are several rules and conditions for the elimination of suffix. If the suffix rules is suitable to the terms, then conditions of the rule will be checked. If they meet conditions of the rule, then they will be fired. For examples, conditions of the rule that is the number of characters vowel followed by a consonant character in the stem must be greater than the rule to be executed [14].

2.2 Ontology

In this research, ontology is implemented after stemming process. The ontology has important role for heterogent information and knowledge sharing. In search engine, it retrieves the particular word in different syntax but it has similar meaning. Ontology is formal description of concept explicitly in domain, property and scope of concept. A concept in ontology has an object. It can be represented by class, property, facet and instances. Class is set of element with the same property. A class has subclass with the specific concept. Classes describe concept to the domain [15]. For example class of vehicle has subclass truck, car and train as shown in Fig. 1.

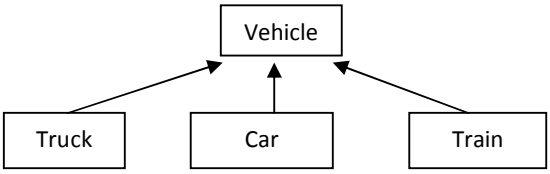


Fig. 1. Model of ontology: Class vehicle with instance

2.3 WordNet

Implementation of ontology to text classification can be used WordNet as an ontology-based. WordNet is lexical database for the English language. Nouns, verbs, adjectives, and adverbs are collected into a set of synonyms and called synsets. WordNet resembles a thesaurus, which collects words based on their meaning. In WordNet, a word represented a string of ASCII characters. WordNet consists of more than 118,000 words and more than 90,000 different senses [16]. In this study, the use of semantic relationship is synonymy. A set of words that has synonymous is combined in a synset.

2.4 Dictionary Construction

Dictionary Construction is a process of converting text documents into feature vectors. Term of feature vector is a term that has been through the process of stemming. Every feature in the vector is connected to the word in the dictionary [7]. The common method of dictionary construction is also called Inverted Index File. Inverted index consists of two parts, namely a term list and post list [17]. Term list is taken from training data and post list is the number of term in each document. The term will be mapped to WordNet based on their synset.

2.5 Feature Selection

Vector space which is built by the dictionary construction will have a great feature space dimensions due to the number of terms in the document. Each space is defined as a dimension feature vectors. The term dimensions will affect to the performance of the document classification. To solve this problem then it is reduced the vector dimensions without decreasing their performance result. Feature Selection is a process of selecting a subset of the features of the original ones [18]. One method of feature selection is unsupervised feature selection by term-document frequency threshold. Document frequency (DF) is the number of occurrences of certain terms in a document. This method calculates the frequency of each unique term for training documents and eliminate these terms if the feature frequency below the predetermined threshold. The assumptions used are terms which rarely appear will not affect the performance of the overall classification. DF threshold is the simplest technique for the reduction of words with the computational complexity versus linear with the number of training documents [18].

2.6 Weighting TFIDF

Weighting feature is a process of giving weight to each *term.* One of the most weighting method is

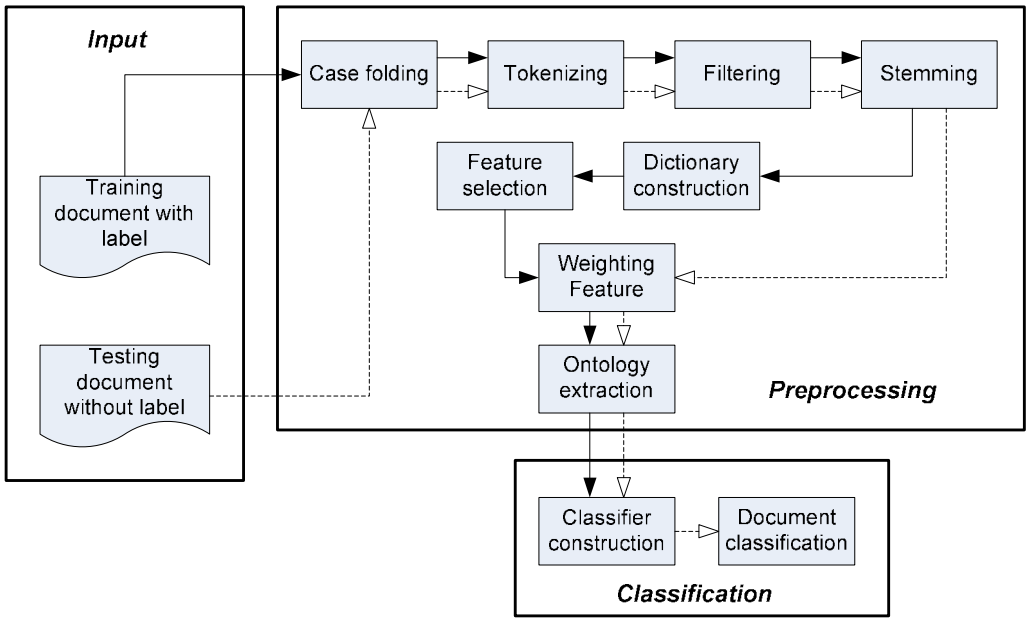


Fig. 2. Block diagram of document classification

the Term Frequency-Inverse Document Frequency (TFIDF). TFIDF is multiplying TF and IDF values and can be formulated as follows: [19].

$$w_{ij} = tf_{ij} \,.\, idf_i = tf_{ij} \,.\, log_2\left(\frac{N}{df_i}\right) \quad (1)$$

Where, the weight of term i in document j, and tf_{ij} is the frequency of term i in document j, while N is the number of documents and a document frequency its value.

2.7 Vector Space Model and Cosine Similarity

Term and weighting term are represented by a vector space model. Each document is represented as a feature vector of term occurrences in all documents. Vector operations, dot products, are performed on the vector space model. A set of documents can be seen as a set of vectors in a vector space, which means that there is one axis for each term. To measure the similarity of two documents d1 and d2, it is used the cosine similarity equation as follows: [17]

$$CosSim(q, d_j) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(w_{ij} \,.\, w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}} \quad (2)$$

Where,

$CosSim(q, d_j)$= *Cossine similarity* test document q to train the document j.

W_{ij} = Weight of term i in document j train
W_{iq}= Weight of term i in document q test
t = The number of *terms*

The numerator of the above equation is the *dot product* (*inner product*) of the test documents and training documents, while its denominator is the result *eucledian distance* between the vectors.

2.8 Evaluation

The performance of system is depend on the accuray of document classification. The evaluation measurements are recall, precision, and F1-Measure [20]. Recall is the number of documents classified correctly by the system divided by the number of documents that should be recognized by the system. Precision is the number of classified documents correctly by the system divided by the total number of performed classification by the system. F1-measure is a value that represents the whole system performance and is an combination between recall and precision values.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (5)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (6)$$

$$F1\ measure = \frac{2\ x\ recall\ x\ precision}{recall + precision} \quad (7)$$

Table 1. Compatibility of classification result

		Classification result from expert	
		Yes	**No**
Class	Yes	True positive	False positive
	No	False negative	True negative

3. METHODOLOGY

In general, the system is divided into two main processes including document preprocessing and classification as in Fig. 2. Input system consists of two types of documents, such as training document and testing documents. They are applied through several steps, starting from document preprocessing including case folding, tokenizing, filtering and stemming. The next step is ontology extraction. The terms are matched to other words based on synonymous in WordNet database. The terms which have less frequency will be ommited but the number of terms will be combined. The next step is feature selection using inverted index. After that, the feature weighting of training and testing of document-term is used TF-IDF. The weights are used to build a classifier process. Classifier is constructed based on cosine similarity measurement.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The data set for this experiment is news document in text format (.txt) and English language. The data set is taken from Reuters-21578. Each document has one category. There are four categories including interest, money-fx, trade, and crude. The number of training data is 100 documents in each four categories and differrent from training documents. As illustration, a data set of news document is shown in Fig. 3 and as a result of ontology extraction from 20 news document collection is shown in Table 2. The term-document is illustrated on the 'term-1', term-2', 'term-3', which have synset on synonymous as they are available in WordNet.

In feature selection, terms which have the number of frequencies (DF) less than their synonym will be ommited. However, the number of term will be compiled to them.

Table 2. The result of ontology extraction from 20 news document

Term-1	Synonymous-1	Term-2	Synonymous-2	Term-3	Synonymous-3
Trillion	Billion	Joint	Stick	Affair	Matter
Prospect	Outlook, view	Rate	Pace	Tradition	Custom
Net	Profit	Bid	Tender	Trader	Dealer
Side	Face, position	Plan	Program, design	Total	Amount
Red	Loss	Step	Pace	Number	Turn
Base	Fundament, stand	Flow	Current, period	Gain	Profit
Drop	Fall	System	Scheme	Award	Honor
Lead	Star, steer	Favor	Favour	Area	Field
Band	Set, lot	Care	Aid, caution, concern,fear	Break	Fault, shift, rift
Sign	Mark,signal	Demand	Need	Button	Push
Crude	Petroleum	Control	Restraint	Subject	Field, matter
Crude	Oil	Cost	Price	Buck	Dollar
Leap	Jump	Attempt	Effort	Level	Point
Bill	Account, note, peak	Question	Head, doubt	Credit	Cite
Offset	Start	Exploit	Effort	Risk	Danger
Help	Aid,assist	Stock	Fund, origin	Hurt	Harm
Report	Account	Drive	Effort	Surplus	Excess
Clear	Open	Post	Position, place	Feet	Base, fundament
Fourth	Quarter	Broker	Factor	Accord	Pact
Cut	Swing	Output	Yield	Record	Book
Hike	Rise, boost	Call	Claim	Upper	Speed

```
DIAMOND SHAMROCK (DIA) CUTS CRUDE PRICES NEW YORK, FEB 26 - Diamond
Shamrock Corp said that effective today it had cut its contract prices for
crude oil by 1.50 dlrs a barrel. The reduction brings its posted price for
West Texas. Intermediate to 16.00 dlrs a barrel, the copany said. "The
price reduction today was made in the light of falling oil product prices
and a weak crude oil market," a company  spokeswoman said. Diamond is the
latest in a line of U.S. oil companies that have cut its contract, or
posted, prices over the last two days citing weak oil markets.
```

Fig. 3. A data set of news document with 'crude' category

In order to know performance of the proposed classifier method, this research is used various number of training data. In this experiment, the number of training data is used as 20, 40, 60, 80, and 100 with four topics (crude, money-fx, interest,trade). The data is taken five times randomly in each experiment. It is applied to compare classifiction of news document with ontology and without ontology. .The performance measurement are including precision, recall, and f-measure as well as accuracy.. The first, it is applied to classifier method (cossine similarity) without ontology and as a result, it is shown in Table 3.

Ther best performace of document classification is by using the number of training documents 20 with four topics. The more number of training documents does not show the higher performance value is.

The second experiment, it is applied to the proposed method using ontology based to classify news document. The experimental result is shown as in Table 4.

The best performance is also obtained which is used the number of 20 training document. The more number of training documents does not show the higher performance value is.

To know the cause factors of performance, it is applied to other scenario which refers to content of document either different topic or similar topic (category). When it is classified to two different category, such as crude, moneyfx, the result is shown as in Tables 5 and 6.

How, when it is classified to the two similar category such as trade, moneyfx, then the result is shown as in Tables 7 and 8.

Table 3. The performance of classification using cosine similarity without ontology

Number of training docs	Precision	Recall	F-measure	Accuracy
20	0.958	0.95	0.949	0.95
40	0.772	0.75	0.739	0.75
60	0.696	0.700	0.668	0.7
80	0.857	0.799	0.781	0.8
100	0.567	0.7	0.609	0.567

Table 4. The Performance of classification using cosine similarity (with ontology)

Number of training docs	Precision	Recall	F-measure	Accuracy
20	0.958	0.95	0.949	0.95
40	0.864	0.799	0.784	0.8
60	0.864	0.799	0.784	0.8
80	0.834	0.75	0.706	0.75
100	0.567	0.65	0.574	0.65

Table 5. The performance of classification without ontology in different category

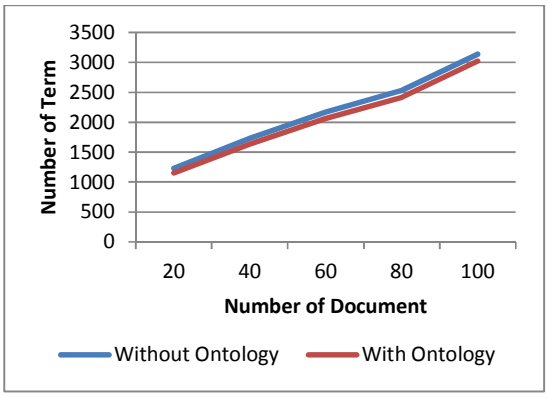
Number of documents	Precision	Recall	F-measure	Accuracy
20	0.954	0.95	0.94	0.95
40	0.954	0.95	0.94	0.95
60	1	1	1	1
80	1	1	1	1
100	0.853	0.85	0.849	0.85

Table 6. The performance of classification with ontology in different category

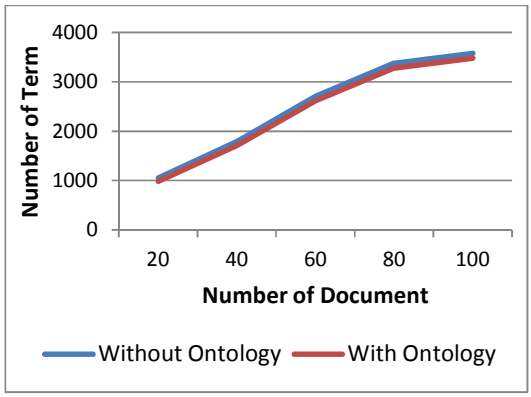
Number of documents	Precision	Recall	F-measure	Accuracy
20	0.954	0.95	0.949	0.95
40	1	1	1	1
60	1	1	1	1
80	0.954	0.95	0.949	0.95
100	0.853	0.85	0.849	0.85

Table 7. The performance of classification without ontology in similar category

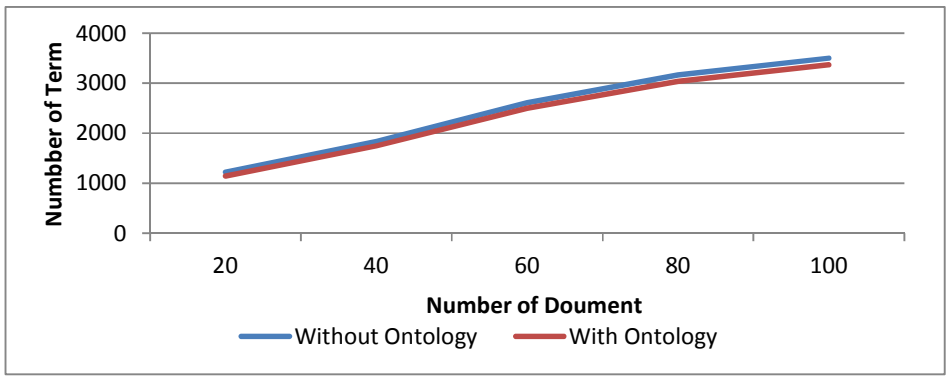
Number of documents	Precision	Recall	F-measure	Accuracy
20	0.853	0.85	0.849	0.85
40	0.853	0.85	0.849	0.85
60	0.853	0.85	0.849	0.85
80	0.853	0.85	0.849	0.85
100	0.9	0.9	0.9	0.9



(a)



(b)



(c)

Fig. 4. The number of terms for document classification (a) Mixed category; (b) Different category; (c) Similar category

Table 8. The performance of classification with ontology in similar category

Number of documents	Precision	Recall	F-measure	Accuracy
20	0.884	0.85	0.846	0.85
40	0.884	0.85	0.846	0.85
60	0.884	0.85	0.846	0.85
80	0.853	0.85	0.849	0.85
100	0.916	0.9	0.898	0.9

In overall, the performance of measure has increased when it is used the data set with different category. There is no ambiguity of the words and the number of document does not effect to their performance. Another word that the performance is not depend on the number of term-document. However, it is depend on the various term with different meaning.

Furthermore, this research also investigate the number of feature (term-document). As we know that by using ontology, there are class and subclass. It means there are many various words has similar meaning (synonym). Therefore, the number of terms has reduced as showed in Fig. 4. It is about 4-5% reduction of term-document.

The number of reduced terms is not significant. There are many terms after preprocessing are not found in dictionary, so that they donnot have synset. The problem is in stemming process which cannot optimally get the correct root word as illustration in Table 9.

Table 9. Several unknown term (after steeming) in the dictionary

Original term	After steeming
Customer	Custom
Repurchases	Repurchas
Federal	Feder
Intervene	Interven
Temporary	Temporari
Indirectly	Indirectli
Agreements	Agreem

5. CONCLUSION

Cosine Similarity based on ontology method can be implemented to clasify the news document in English. This method actualy can reduce feature of document representation but it increases computational time due to retreiving process into dictionary for particular word. However, overall for the performance measure is increase.

6. THE FUTURE WORK

The system has been built on this research still has shortcomings, especially in steming process. It showed that the number of term reduction. Therefore, it is need to improve the steeming process using another method.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ben-Dov M, Feldman R. Text mining and information extraction, Chapter 38; 2001.
2. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. USA; 2002.
3. Yates RB, Neto BR. Modern information retrieval. Addison Wesley,New York; 1999.
4. Larsen B, Aone C. Fast and effective text mining using linear time document clustering in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999.
5. Rosso P, Ferretti E, JimEnez D, Vidal V. Text categorization and information retrieval using WordNet senses. National University of San Luis, Argentina; 2004.
6. Hotho A, Staab S, Stumme G. Wordnet improves text document clustering. University of Karlsruhe, Germany; 2003.
7. Prabowo R, Jackson M, Burden P, Knoell HD. Ontology-based automatic classification for the web pages: design, implementation and evaluation. Proceedings of the 3rd International Conference on Web Information Systems Engineering; 2002.
8. Song MH, Lim SY, Kang DJ, Lee SJ. Automatic classification of web pages based on the concept of domain ontology. Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05). 2005;645-651.
9. Taghva K, Borsack J, Coombs J, Condit A, Lumos S, Nartker T. Ontology based classification of email. Proceedings of Information Technology: Coding and Computing; 2003.
10. Tenenboim L, Shapira B, Shoval P. Ontology-based classification of news in an electronic newspaper. Proceedings of the International Conference on Intelligent Information and Engineering Systems; 2008.
11. Elberrichi Z, Rahmoun A, Bentaalah MA. Using WordNet for text categorization. The International Arab Journal of Information Technology. 2008;5:1.
12. Nidhi, Gupta V. Domain based classification of Punjabi text documents using ontology and hybrid based approach. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), COLING 2012, Mumbai. 2012;109–122.

13. Guo G, Bell DA, Wang H, Bi Y. An kNN model-based approach and its application in text categorization. University of Ulster Newtownabbey; 2004.
14. Porter MF. An algorithm for suffix stripping. Cambrigde; 1980.
15. Noy F. Natalya, McGuinness L. Deborah. ontology development 101: A guide to creating your first ontology. Standford University, Stanford, CA, 94305.
16. Miller GA. WordNet: A lexical database for English, Communication of the Acm. 1995; 38:11.
17. Manning CD, Raghavan P, Schütze H. An introduction to information retrieval. Cambridge University Press, Cambridge, England; 2009,
18. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In International Conference on Machine Learning. 1997;412–420.
19. Soucy P, Mineau GW. Beyond TFIDF weighting for text categorization in the vector space model. Canada; 2003.
20. Sebastiani F. Machine learning in automted text categorization. Computing Surveys. 2002;34(1):1–47.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/16385