JOURNAL OF PURE AND APPLIED MICROBIOLOGY, Aug. 2015. Vol. **9**, p. 79-86
SPECIAL ISSUE ON RECENT RESEARCH CHALLENGES IN BIO-MEDICAL APPLICATIONS

Semantic Medical Web: A Hyponym Based Entailment Rules Formedical Web Page Classification

P. Madhubala[1] and K. Murugesan[2]

[1]Tagore Institute of Engineering and Technology,
Department of Computer Science and Engineering, Salem, Tamilnadu, India.
[2]Sree Sastha Institute of Engineering and Technology,
Department of Electronics and Communication Engineering, Chennai, Tamilnadu, India.

(Received: 03 February 2015; accepted: 25 April 2015)

Lack of large semantic variability patterns is a major obstacle to progress in semantic inferencefor medical data available online. Prominent inference knowledge representation includes entailment rules. Large-scale inference based knowledge systems have initiatedwork on automatic paraphrase and entailment rules acquisition. This work identifiesHypernym of medical terms and clubs them with entailment rule acquisition. A hyponym word tree in the document is created and used with the dependency tree. Features extraction is achieved through weighted TF-IDF where word weight is computed based on hyponyms present in a radix tree. The proposed system was evaluated using k-Nearest neighbour (kNN) algorithm with good results.

Key words: Hypernyms, Hyponym, TF-IDF, k-Nearest neighbour, Rule acquisition, Ontology.

Increased use of medical related web sites, recommender systems can suggest matching participants. Here, a two-way interaction model is considered where users, called senders forward message to each user, called recipients. The latter reply positively/negatively[1]. The recommendation method in this model suggests a candidate recipients group likely to reply positively to senders.Adaptive Web concentrates on personalized services tailored to meet individual users specific needs (user context including user characteristics, device, and environment properties) regarding content recommendation, navigation adaptation, and presentation customization[2]. Semantic Web envisioned by Tim Berners-Lee planned to improve the present Web through addition of a semantic metadata layer, allowing advanced machine information processing and ultimately improving application interoperability, data integration, sharing, and availability.

Knowledge is essential to Semantic Web applications with explicit description of concepts in a discourse domain[3]. Knowledge acquisitions in present approaches are grouped into two types, ontology acquisition and rule acquisition. The focus is about gathering knowledge automatically from the web consisting of unstructured data. Knowledge acquisition based on ontology is more popular research area compared to rule acquisition[4]. Nevertheless, small rules were implied in web pages previously and have the potential to acquire them from Web pages similar to ontology learning[5]. For example,a text's upper part in a web page explaining return policies of books purchased online is a representative of online bookstores. The lower part reveals text acquired rules.

Ontology learning[6] refers to extracting conceptual knowledge from different sources to

* To whom all correspondence should be addressed.
E-mail: madhubala.web@gmail.com

build ontology. Ontology learning approaches need knowledge from the Web, as it offers valuable information for domains. The foundation to get knowledge structures is intensive use of incremental methodology to construct terms taxonomies, associated web resources and discovering non-taxonomic relationships relevant to desired domain[7]. This method's important characteristic is the entire process being performed automatically in a totally unsupervised and domain independent way directly from the Web.

A hyponym describes things more specifically. Proper nouns are hyponyms. Victoria Falls is a hyponym for the concept of a waterfall and Ford the hyponym for a concept car. Similar to a hypodermic needle ending in a very fine point, hyponyms pointedly focus the search[8]. Hypernyms refer to broad general concepts. Car or airplanes are hypernyms for precise terms like Toyota Camry or Boeing 747. If not enough results is got in a web search, the search can be broadened by the use of hypernyms.Common points between entailment rule acquisition and the proposed approach acquire inference rules from text/Web.

This work identifies web Hypernyms (superordinate words) in medical related web pages and clubs with entailment rule acquisition. This work is structuredwith section two reviewing previous knowledge acquisition based work. Section 3 describes methodologies used in this work. Experimental results of the proposed approach are discussedin Section 4. Section 5 presents the conclusion.

Related work

Anapproach to locate hypernym relations between terms of specific knowledge domain was introduced by Rios-Alvarado, *et al*., [9] which combined WordNet to retrieve a term's most representative hypernyms. Identification of hypernym/hyponymy relations between terms is imperative to build taxonomy. Earlier works used specific lexical patterns or identified new patterns. Efficient algorithms to identify key concepts based on least common-ancestor and greatest common descendant were proposed by Ranwez, *et al*.,[10].A method for automatic hypernym/hyponymy relations acquisition from a Turkish raw text was proposed by Yildizand Yildirim[11].

Decision tree technology and its use in Web information extraction were studied by Hong-ye[12]. According to datasets, by information extraction, an agricultural products market decision tree was constructed by C4.5/C5.0 algorithm, which updated data constantly and generated understandable rules. The experiment proved the feasibility of using decision tree based Web information extraction.

A decision tree learning approach over Document Object Model (DOM) based features aimed at cleaning uninformative sections to extract informative content in three classes: title, main content and additional information was presented by Uzun, *et al*.,[13]. The proposed approach, different from earlier studies proved the learning model for extraction of main content on DIV and TD tags achieving 95.58% accuracy in cleaning uninformative sections and extracting informative content. 0.96 f-measure was obtained for the extraction of the main block.

A refined IDF schema called Channel Distribution Information (CDI) IDF, was proposed by Xu, *et al*.,[14]. Meaningless terms and top terms could be identified according to statistical figures. Compared to traditional TF-IDF, CDI TF-IDF increased F-measure 3.00%.

Various TF or TF-IDF weighting based term weights were proposed by Bartík[15]. Modification was visual areas based where text and their visual properties appear as web pages constantly increase. In the proposed method, textual/visual information found suitable web page content representation.

Recognition task was first defined, and then a new recognition framework to extract web entities was proposed by Liu and Zhang[16]. The author first proposed a sentence similarity based k-nearest neighbor classifier to discover sentences having web entity activity. Then, based on dependency parsing feature, heuristic rules set extracted information from sentences. Experiments proved the method's feasibility and effectiveness adaptable to multi-domains.

A solution using a k-nearest-neighbour approach to classify documents returned by search engines by building classifiers using data from collaborative tagging systems was proposed by Yeung, *et al*.,[17]. Traditional Web search engines use keyword-based approach. When user submitted keyword is ambiguous, search result consists of documents related to the keyword's

differing meanings, where the user will be interested in only one. Search results by Google showed the proposed method, classifying documents with high precision.

Fuzzy k-NN network based classification was presented by Zhang, *et al*.,[18] where the process of web classification, TF-IDF selected web content features to increase accuracy and suit real world which used membership grades. Experiments proved classification performance to be better than k-NN and Support Vector Machine (SVM).

Two classification algorithms for document classification,submitted to BioCreative II.5 PPI Classification Challenge Task were described and evaluated by Ambertand Cohen[19]. The task was to design classifiers to identify documents with PPI-related information in primary literature, evaluating them against each other. One of the proposed system, which was judged the best system, used a new approach to k-nearest neighbor classification described here comparing its performance with two SVM based classification systems, one also evaluating the challenge task.

Consideration of top-k classification approach as a word belonging to many concepts was proposed by Warintarawej, *et al*.,[20]. Labelled-words were separated into syllables, and feature selection selected discriminative syllables. A syllable frequency and mutual information performed with Naive Bayes classifier and K-nearest neighbor (KNN). The results revealed top-k classification model analyzing a new word by relating it to many concepts.Classification of textual data based on content was also explored by Wajeedand Adilakshmi[21].

A method to network harmful KNN based information and improving classification efficiency through training samples elimination to prevent misclassification was presented by Yan and Yang[22]. Experiment proved that proposed system's precision and recall were enhanced and reduced classification time-consumption.

An improvedKNN based text classification performance in P2P communication paradigm,P2PKNNCwas proposed by Yu and Yu[23]. P2PKNNC executed k nearest neighbor(s) queries in a distributed metric structure influencing part from neighbors classifying input document regarding disturbance degree. Experiments showed the algorithm ensured great performance improvement on imbalanced corpora.

METHODOLOGY

In this work, the medical web documents are pre-processed, features are extracted and on acquiring rules, kNN is used to classify the document. The stop words and stemming are the steps in pre-processing. In this work, Hypernyms are identified in the web page and clubbed with entailment rule acquisition. A tree of hyponym words that are available in the document is created and used with dependency tree. Features are also extracted using weighted TF-IDF where the weight of the word is computed based on the number of hyponyms present in the radix tree. Performance is evaluated using k-Nearest neighbour algorithm.Figure 1 shows the flowchart for the proposed methodology.

Table 1. Entailment Relation

Input	Correct	Incorrect
X change Y (DIRT)	(←) X modify Y	X adopt Y
	(←) X amend Y	X create Y
	(←) X revise Y	X stick to Y
X change Y (TEASE)	(→) X alter Y	X maintain Y
	(→) X affect Y	X follow Y
	(←) X extend Y	X use Y

Table 2. Performance measurement on Rules

Techniques Used	Precision	Recall
Entailment based rule acquisition	88.74	86.52
Proposed technique	94.62	92.34

Table 3. Performance measurement on Antecedent

Techniques Used	Precision	Recall
Entailment based rule acquisition	90.48	76.24
Proposed technique	95.2	89.4

Table 4. Performance measurement on consequent

Techniques Used	Precision	Recall
Entailment based rule acquisition	90.68	77.84
Proposed technique	96.44	87.68

Stop words

Stopwords are common words in text with little meaning and have no subject matter. Stopwords have varied impact on information retrieval effectiveness[24,25]. They also affect efficiency due to their nature and because they carry no meaning resulting in unproductive processing.

Stemming

Stemming transforms different inflections/same word derivations to one common "stem". Stemming can indicate removal of either prefix or suffix. A good stemmer should convert a word's different syntactic forms to its normalized form. Porter Stemmer[26] is a common compact method used for stemming. Text classification is possible through various learning classifier approaches like k-nearest neighbor, decision tree induction, Naïve Bayesian, SVM and latent semantic index[27].

Radix trees

Radix trees are not bound by the order in which data is inserted, and hence lead to balanced data structures. But, radix methods performance in some situations is poor: where keys have specific pattern, for e.g. all keys start with 000000, radix methods are slow and by themselves do not handle duplicate keys, though they are combined with other data structures to do so.

Radix trees include two node types: Inner nodes mapping partial keys to other nodes, and leaf nodes, which store values corresponding to keys. The best representation of an inner node is a 2^s pointers array. In tree traversal, an s bit chunk of key is the index into that array which determines next child node without additional comparison.

Radix trees have many properties distinguishing them from comparison-based search trees[28]:

· Radix trees height (and complexity) depends on keys length but generally not on number of elements in tree.

· Radix trees need no rebalancing operation as all insertion orders lead to same tree.

· Keys are stored lexicographically.

· A leaf node's path represents the key to that leaf. Hence, keys are stored implicitly and are reconstructed from paths.

An example of Radix tree for brain tumours is shown in figure 2. From this diagram, it is observed thatEpendymom is a hyponym of Intra parenchymatous and Intra parenchymatous is hyponym of brain tumor and so on. Also, Oligodendroglioma and Gangliocyto are co-hyponyms and its superordinate term is Intra paernchymatous.

Entailment Rules

An entailment rule 'L '! R' is a directional relationbetween two templates, L and R. Templates created can be mapped to textfragments with variables can be linear or parse sub trees[29]. Entailment rules help applications infer one text

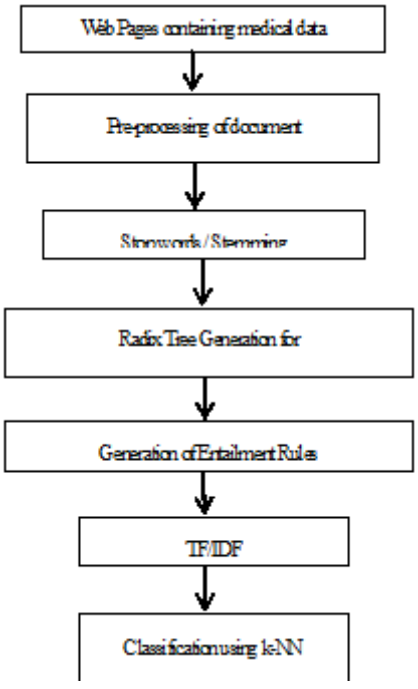

Fig. 1. Flowchart for Proposed Methodology

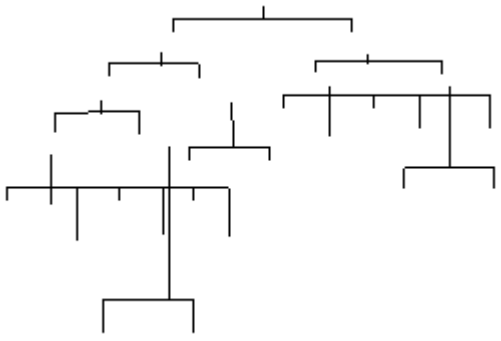

Fig. 2. Example of Radix tree for Brain Tumor

variant from another. An example of entailment relation is shown in Table 1.

Web Page Classification has two steps: 1) Pre-processing and 2) Classification. Pre-processing depends on factors like feature selection that plans to reduce webpages complexity to minimize time to build a classifier model.

Term Frequency- Inverse Document Frequency (TF-IDF)

The tf*idf[30,31] metric includes two components: inverse document frequency (idf) and term frequency (tf) which are multiplied when computing tf*idf. The first word-based feature is term frequency (tf) which measures importance of word tj in document, i.e. mood, dj with ni,j occurrences of word in document dj , divided by sum of number of occurrences of words in document dj[32].

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad ...(1)$$

k-Nearest Neighbor (k-NN)

k-Nearest Neighbor (k-NN) is the simplest among machine learning algorithms and is an instance based learning method[33]. k-NN algorithm is a similarity-based learning algorithm has been effectively used forInformation retrieval[34]. In a test document, k-NN algorithm finds k nearest neighbors among training documents. Each neighbor document's similarity score to test document is used as categories weight of neighbor document.

k-NN is simple but effective for classification motivating improvement of the algorithm regarding efficiency and accuracy and preventing bias from dominating classes[35]. There

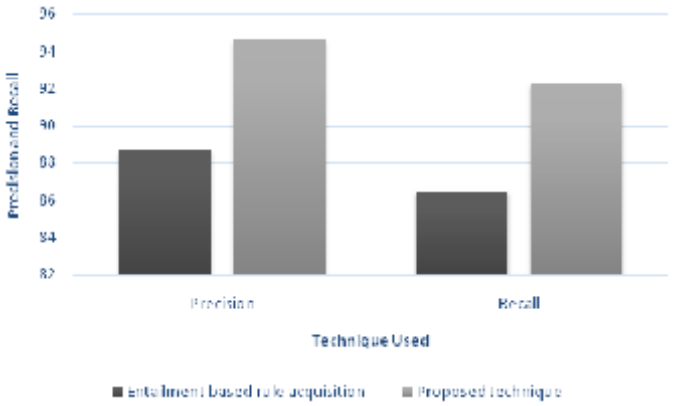

Fig. 3. Performance measurement on Rules

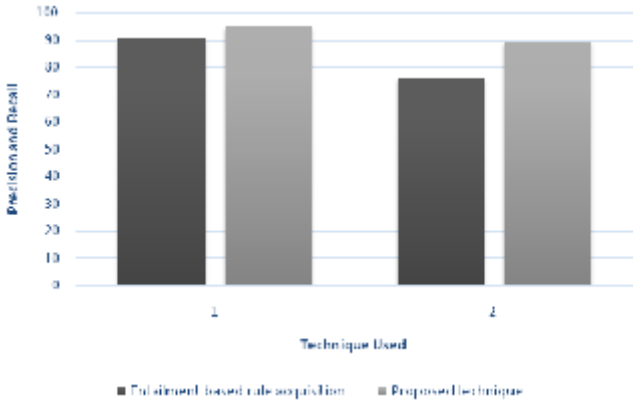

Fig. 4. Performance measurement on Antecedent

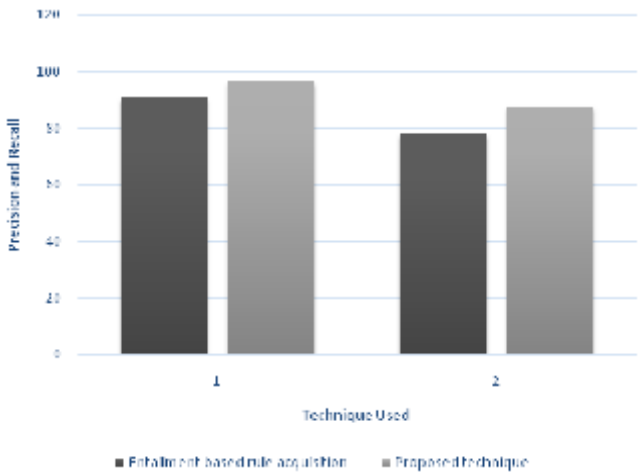

Fig. 5. Performance measurement on consequent

are three factors affecting k-NNalgorithm including 1) similarity/distance measure to find k-NN, 2) k, which is number of nearest neighbor, and 3) decision rule to identify a class for test document from k-NN. An object is classified by majority neighbours vote; the aim being to be assigned to most common class amongst k nearest neighbors. Creation of the neighbourhood is by finding distance of test instance with various cluster centroids and calculating mean distance of total distance. So, a circle of radius 'r' is formed around test instance, where r is mean distance of test instance to various cluster centroids and is given by equation (2)[37]:

$$radius(r) = \frac{1}{c}\sum_{1}^{c}\left(\sum_{i=1}^{n}((a_i.x_i)^2 - (a_i.y_i)^2)\right) \quad ...(2)$$

Where,
c is number of clusters formed,
n is total attributes,
a_i is attribute weight of corresponding attribute, and
, are values of attributes corresponding to i.

Attribute weighting keeps imbalanced class distribution at bay lowering creation of a large neighborhood. Though mean distance might not represent the best solution through trial and error, it represents a good fit for neighborhood selection issues.

y_i

EXPERIMENTAL RESULTS

For evaluation of the proposed methods, Barnes and Noble web pages are used. The experiments are conducted under three scenarios. They are Performance measurement on Rules, Performance measurement on Antecedent and Performance measurement on consequent.

From figure 3, the precision of the proposed methodincreased by 6.21% and recall of the proposed method increased by 6.3%.

From figure 4 the precision of the proposed method increased by 4.96% and recall of the proposed method increased by 14.72%.

Similarly from figure 5 that the precision of the proposed method increased by 5.97% and recall of the proposed method increased by11.22%.

CONCLUSION

Automatic medical web page classification assigns a document to a relevant category. Classification system should take advantage of categories, hierarchal structure in medical terms. Hyponymy is transitive and asymmetrical. For e.g. economic is of social science hyponymy, but social science a hypernym of economic. This work proposed a novel technique of Semantic integration of medical terms with entailment. This work used a radix tree to construct

hyponyms from hypernyms. The identified Hypernyms are clubbed with entailment rule acquisition using weights from the radix tree constructed. The obtained bag of features is converted to feature vectors using TF-IDF.Results show that the proposed method achieves higher precision and recall when compared to entailment based rule acquisition.

REFERENCES

1. Kim, Y. S., Mahidadia, A., Compton, P., Krzywicki, A., Wobcke, W., Bain, M., &Cai, X. (2010). Automated Recommendation Rule Acquisition for Two-Way Interaction-based Social Network Web Sites.
2. Barla, M., Tvarozek, M., &Bielikova, M. Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems. *Computing and Informatics*, 2012; **28**(4): 399-428.
3. Soumya.S.,H.Swathi. Automatic Repeated Rule Acquisition from Similar Web Sites using Rule Ontology. *International Journal of Computer Applications* (0975 – 8887), 2013; **66**(16).
4. Kumar, E. D., Kumar, M. V., & Deepa, P. A Framework for Ontology Based Repeated Rule Acquisition from Similar Web Sites.
5. Kumar, G. S., & Mohan, K. K. M. An Effective Repeated Rule Acquisition Using Rule Ontology from Similar Web Sites based on Genetic Algorithm. *International Journal of Science, Engineering and Technology Research*, 2013; **2**(2), pp-325.
6. Singh, A., &Muruganantham, B. Automatic Rule Extraction from Frequent Rules in Similar Network Sites. *International Journal of Computer Trends and Technology*, **4**(3).
7. Sanchez, D., & Moreno, A. A methodology for knowledge acquisition from the web. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 2006; **10**(6): 453-475.
8. What Are Hyponyms, and Hypernyms? How Can These Nyms Help You Focus A Search? 21st Century Information Fluency Project. IMSA.
9. Rios-Alvarado, A. B., Lopez-Arevalo, I., & Sosa-Sosa, V. (2011, October). Discovering hypernyms using linguistic patterns on web search. In Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on (pp. 302-307). IEEE.
10. Ranwez, V., Ranwez, S., &Janaqi, S. (2012). Sub-ontology extraction using hyponym and hypernym closure on is-a directed acyclic graphs.
11. Yildiz, T., &Yildirim, S. (2012, July). Association rule based acquisition of hyponym and hypernym relation from a Turkish corpus. In Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on (pp. 1-5). IEEE.
12. Hong-ye, C. Method of web information extraction based on decision tree. In Information Technology and Applications, 2009. IFITA'09. *International Forum on* 2009; **1**: pp. 664-666.
13. Uzun, E., Agun, H. V., &Yerlikaya, T. (2012, April). Web content extraction by using decision tree learning. In Signal Processing and Communications Applications Conference (SIU), 2012 20th (pp. 1-4). IEEE.
14. Xu, M., He, L., & Lin, X. (2010, March). A Refined TF-IDF Algorithm Based on Channel Distribution Information for Web News Feature Extraction. In Education Technology and Computer Science (ETCS), 2010 Second International Workshop on (Vol. 2, pp. 15-19). IEEE.
15. Bartík, V. (2010, August). Text-Based Web Page Classification with Use of Visual Information. In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on (pp. 416-420). IEEE.
16. Liu, H., & Zhang, C. (2012, April). A Web entity activity recognition approach based on k-nearest neighbors classifier. In Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on (pp. 848-852). IEEE.
17. Yeung, C. M. A., Gibbins, N., &Shadbolt, N. (2008, December). A k-nearest-neighbour method for classifying web search results with data in folksonomies. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on (Vol. 1, pp. 70-76). IEEE.
18. Zhang, J., Niu, Y., &Nie, H. (2009, December). Web Document Classification Based on Fuzzy k-NN Algorithm. In Computational Intelligence and Security, 2009. CIS'09. International Conference on (Vol. 1, pp. 193-196). IEEE.
19. Ambert, K. H., & Cohen, A. M. (2012). K-information gain scaled nearest neighbors: a novel approach to classifying protein-protein interaction-related documents. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(1), 305-310.
20. Warintarawej, P., Laurent, A., Pompidor, P., Cassanas, A., & Laurent, B. (2011, August).

Classifying Words: A Syllables-based Model. In Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on (pp. 208-212). IEEE.

21. Wajeed, M. A., &Adilakshmi, T. (2011, September). Different similarity measures for text classification using KNN. In Computer and Communication Technology (ICCCT), 2011 2nd International Conference on (pp. 41-45). IEEE.
22. Yan, X., & Yang, J. (2011, July). Research on Filtering System of Harmful Information on Network Based on K Nearest Neighbor Algorithm. In Control, Automation and Systems Engineering (CASE), 2011 International Conference on (pp. 1-4). IEEE.
23. Yu, X. P., & Yu, X. G. (2007, August). Novel text classification based on k-nearest neighbor. In Machine Learning and Cybernetics, 2007 International Conference on (Vol. 6, pp. 3425-3430). IEEE.
24. El-Khair, I. A. Effects of stop words elimination for Arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences,* 2006; **4**(3): 119-133.
25. Negm, N., Elkafrawy, P., Amin, M., & Salem, A. B. M. Investigate the Performance of Document Clustering Approach Based on Association Rules Mining. *International Journal of Advanced Computer Science & Applications,* 2013; **4**(8).
26. Carlberger, J., Dalianis, H., Hassel, M., &Knutsson, O. (2001, May). Improving precision in information retrieval for Swedish using stemming. In the Proceedings of NODALIDA (pp. 21-22).
27. Yun-tao, Z., Ling, G., & Yong-cheng, W. An improved TF-IDF approach for text classification. *Journal of Zhejiang University Science A,* 2005; **6**(1), 49-55.
28. Wang, X. L., Zhao, H., & Lu, B. L. (2011, September). Enhanced K-Nearest Neighbour Algorithm for Large-scale Hierarchical Multi-label ClassificationÆ". In Proceedings of the Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification, Athens, Greece (Vol. 5).
29. Leis, V., Kemper, A., & Neumann, T. The Adaptive Radix Tree: ARTful indexing for main-memory databases. ICDE 2013.
30. Soucy, P., &Mineau, G. W. Beyond TFIDF weighting for text categorization in the vector space model. *In IJCAI* , 2005; **5**: pp. 1130-1135.
31. Forman, G. (2008, October). BNS feature scaling: an improved representation over tf-idf for svm text classification. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 263-270). ACM.
32. Van Zaanen, M., &Kanters, P. (2010, August). Automatic Mood Classification Using TF* IDF Based on Lyrics. In ISMIR (pp. 75-80).
33. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (pp. 986-996). Springer Berlin Heidelberg.
34. Miah, M. Improved k-NN Algorithm for Text Classification. In DMIN (pp. 434-440) 2009.
35. Suguna, N., &Thanushkodi, K. An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues,* 2010; **7**(2): 18-21.
36. Gupta, M. Dynamic k-NN with Attribute Weighting for Automatic Web Page Classification (Dk-NNwAW). *International Journal of Computer Applications,* 2012; **58**(10): 34-40.