NOVELTY DETECTION VIA TOPIC MODELING IN RESEARCH ARTICLES

S. Sendhilkumar[1], Nachiyar S Nandhini[2], G.S. Mahalakshmi[3]

[1]Department of Information Science and Technology,
Anna University, Chennai, Tamil Nadu
ssk_pdy@yahoo.co.in
[2]Department of Computer Science and Engineering,
Anna University, Chennai, Tamil Nadu
nachiyar.s.nandhini@gmail.com
[3]Department of Computer Science and Engineering,
Anna University, Chennai, Tamil Nadu
mahalakshmi@cs.annauniv.edu

ABSTRACT

In today's world redundancy is the most vital problem faced in almost all domains. Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training. The problem becomes more intense when it comes to "Research Articles". A method of identifying novelty at each sections of the article is highly required for determining the novel idea proposed in the research paper. Since research articles are semi-structured, detecting novelty of information from them requires more accurate systems. Topic model provides a useful means to process them and provides a simple way to analyze them. This work compares the most predominantly used topic model- Latent Dirichlet Allocation with the hierarchical Pachinko Allocation Model. The results obtained are promising towards hierarchical Pachinko Allocation Model when used for document retrieval.

KEYWORDS

Novelty detection, Topic modeling, LDA, hPAM, Novelty score, Concept maps

1. INTRODUCTION

This internet era has made searching for anything possible. Information on anything is available anytime because of advancement in information retrieval techniques. However information redundancy is the major problem being faced due to abundance. Novelty detection helps to mine out new information from varied sources. Many techniques and approaches have been proposed for handling structured and unstructured data. However, mining from semi-structured document like Research Articles, has captured interest of researchers in recent years. Text level novelty mining has been in concern since the conduct of TREC novelty track [So-boroff, 2005]. TREC novelty track 2002, 2003 and 2004 defines novelty as pro-viding new information that has not been found in any previously picked sentences.

David C. Wyld (Eds) : ICCSEA, SPPR, CSIA, WimoA, SCAI - 2013
pp. 401–410, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3542

2. LITERARY REVIEW

Novelty detection is the technique used to extract novel information from a set of relevant documents or from a same document in a given topic (query). Topic refers to a set of words that frequently occur together. Topic model aims at discovering abstract topics that occur in a document or collection of document. Topic models can be used to connect words with similar meanings using contextual clues.

Topic Modeling

A document can be viewed upon as a collection of words. Words that pertain to certain set of relevant words can be treated as topics; a document hence can be treated as random mixture of words with some probabilistic degree of distribution with them. The very first approach to model topics was using Latent Semantic Indexing (LSI) [Papadimitriou, 1998]. Later works showed the evolved version of LSI where the Probabilistic nature of documents was considered. The algorithm used was Probabilistic Latent Semantic Indexing (PLSI) [Hofmann, 1999]. In contrast to LSI, PLSI had a statistical foundation and defined a generative data model. Latent Dirichlet Allocation (LDA) views document as a collection of topics as in PLSI, the only difference being LDA assumes Dirichlet prior for topic distribution. Pachinko Allocation Model (PAM) documents the distribution of single set of topics in a graph and represents co-occurrences. In PAM each node represents a distribution over nodes in next lower level [Li et al., 2006]. Hierarchical Pachinko Allocation Model (hPAM) a variant of PAM encompasses the advantages of hierarchical LDA (hLDA) and four-level PAM. hPAM has every node (not only the lowest level) associated with a distribution of vocabulary [Mimno et al., 2007]. This is an extremely flexible framework for hierarchical topic modeling.

Sentence Level Novelty Detection

Novelty of a sentence is usually calculated with respect to the number of new words appearing in them. This involves two tasks namely relevant sentence extraction and novelty estimation which form the main concentration of TREC novelty track. Named Entity Recognition (NER) method helps in identifying the meaning of a sentence by recognizing some key characteristics [Zhang, 2009]. Vector Spaced Model (VSM) is used to rank documents. Term co-occurrence and term weights along with term sets are used as term indices to capture semantic relationship of terms that appear close to each other. S*et-based vector model* [Bruno Pôssas et al., 2005] refers a term set to a set of index terms of collection of documents. Cosine distance metric was used to compute novelty by assigning all non- stop words a value of 1 [Schiffman et al., 2005]. A Graph-based text representation model [Tomita et al., 2004] represents texts formally as Subject Graphs. Translation form text to subject graphs involved three steps: 1) term extraction from text, 2) term significance calculation, and 3) significance calculation for term-term association and making association vector. The similarity is then measured as a linear combination of inner products of term vectors and the association vectors. Later work involved creating feature vector for each tagged sentence and set of sentences that has already seen information [Michale Gamon, 2006]. These features captured the relationship between tagged sentences and set of background sentences. These sentences are then represented as *graphs* based on *21 graph features* and few *text rank features*. The novelty score for the sentence was computed based on *KL divergence, sentence graph* and *text rank.* Another major contribution for sentence level novelty detection using *overlap method* [Zhao et al., 2006], defines novelty as a combination of partial overlap (PO) and complete overlap (CO. The overlaps were measures using similarity, pool method and selected pool method.

Document Level Novelty Detection

Document level novelty detection is considered rarely useful, as nearly every document will have something new [Soboroff et al., 2005]. Most work on document level novelty detection treats document as a set of sentences. Novelty of the document is determined by sentence novelty. The main focus is on information filtering system to retrieve relevant document based on relevancy based recall, precision and utility metrics [Zhang et al., 2004]. Newness of a document is dependent on relevance of the document with those retrieved previously. An adaptive information filtering system was used to identify novel documents based on document classification as redundant, relevant and non-relevant [Zhang et al., 2002]. Novelty mining in multi-lingual document [Zhang et al., 2011] say Malay [Kwee et al., 2009] and Chinese [Zhang and Tsai, 2009] language is done at document level. The novelty of a document is then quantitatively represented by calculating cosine similarity and taking the difference between 1 and similarity score (1-cosine similarity). A new framework for document level novelty detection using *document-to-sentence (D2S)* annotation was proposed [Tsai et al., 2010]. The document was first segmented into sentence, novelty score for each sentence was determined and novelty score of whole document was predicted based on fixed threshold. A Fuzzy Cognitive Map approach considers a document as a collection of topics [Sendhilkumar et al., 2011]. The document was represented as fuzzy concept maps that had concepts and information of the domain which was compared with domain specific ontology.

3. METHODOLOGY

This paper proposes a new method to identify novelty of research article. The work focuses on aspects of novelty such as similarity, divergence and relevance. The entire work is divided into seven modules which involves functions: topic modeling, relevant document retrieval, clustering, document similarity measure, concept mapping, document segmentation and novelty estimation. The flow of work is as proposed in Fig 1.

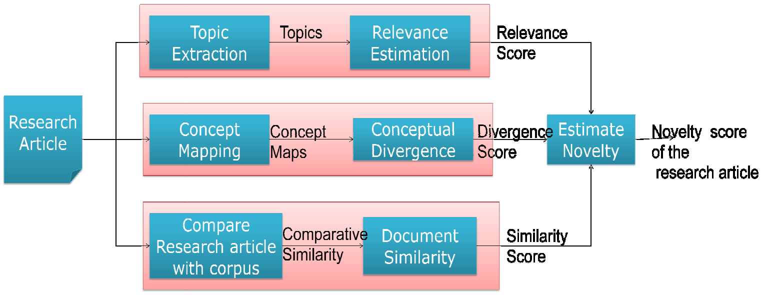

Fig. 1. Framework of proposed system

Topic Modeling

The Input Research Article is first pre-processed. The processed document is then topic modeled using Latent Dirichlet Algorithm and Hierarchical Pachinko Allocation model. Topic model has a major influence on novelty score as novelty is estimated in terms of the topics, super-topics and subtopics identified by LDA and hPAM. The performances of both methods were evaluated. hPAM showed a better results in terms of sensitivity and selectivity.

Document Retrieval

The topics obtained from the topic modeling were compared with a corpus. The documents in the corpus containing the topics, sub-topics and super-topics obtained from topic modeling along with their semantics were retrieved. Retrieval was done using LDA and hPAM.

Clustering

The documents retrieved were then clustered using their term frequencies. The clustering involved feature selection using TF/IDF. The documents with matching TF/IDF of topics were clustered. The number of clusters depends on input research article and the number of topics identified by topic model. The distance between the input research article and clusters were measured and the nearest cluster was identified. Clustering was done to reduce the number of documents and retain the most relevant ones.

Document Similarity

The topics of input research article are compared with the documents in the corpus to identify similarity among document. This gives a measure of whether the topics discussed in the input research article are already been proposed or discussed by others. The similarity is calculated using cosine similarity as

$$\cos(s_i, s_j) = \frac{\sum_{k=1}^{l} v_i,k \times v_j,k}{|s_i|.|s_j|} \qquad \text{eq.(1)}$$

where s_i represents Sentence vector$(v_{i1}, v_{i2}, \dots, v_{il}\ )$, l denotes number of documents retrieved from the reference corpus and s_j is another sentence vector [Tsai et al., 2004].

Concept Mapping

The abstract of input research article and the documents retained by clustering are concept mapped. Concept mapping involves representing the documents as a set of concepts and relationships among them. Concepts include ideas, words, topics and new logical terms. Relationship represents the way in which the identified concepts are linked to one another. These are preserved in XML files. The concept maps obtained are compared and the KL divergence is measured between the concept maps.

Novelty Estimation

The work on novelty reported so far has treated novelty as a measure of similarity. Novelty was considered to be the inverse of cosine similarity [Zang and Tsai, 2009] .i.e.,

$$Novelty = 1 - \max_{1 \leq i \leq t-1} \cos(s_i, s_j) \qquad eq.(2)$$

This approach cannot be treated as an accurate measure of novelty since it computes only the similarity and the inverse can only be the dissimilarity measure of the document compared to the document under consideration. Another approach calculates novelty as a measure of conceptual divergence where the score obtained is treated as originality of the document [Sendhilkumar et al., 2012]. The novelty score of the article was given as

$$Novelty\ score(d|dt) = \frac{W(d|dt)}{W(d)} \qquad eq.(3)$$

where W(d) is the initial weight of the document. The score thus obtained focuses solely on conceptual divergence. This work proposes a new measure for novelty as in eq.4. Here novelty is treated as a combined representation of similarity, relevance and divergence.

$$Novelty = 1 - (Similarity + Relevance) + Divergence \qquad eq.(4)$$

The novelty thus obtained includes properties of input research article such as conceptual divergence, semantic relevance and contextual similarity.

4. RESULTS AND DISCUSSION

The input research article is converted into topic model using two algorithms LDA and HPAM. *SPECIFICITY and SENSITIVITY* is used as metric for evaluating hPAM and LDA. From Table.1 it can be inferred that though LDA returns more number of subtopics, the total count of subtopics in manual evaluation is less than 388.Which means that LDA wrongly identifies some topics as subtopics.

	hPAM				LDA		
Correctness Parameters	topic	super topic	sub topic	Average	topic	sub topic	Average
Correctly Identified	3	43	89	45	1	180	90.5
Correctly Rejected	3	98	46	49	69	4	36.5
Incorrectly Identified	1	31	52	28	4	179	91.5
Incorrectly Rejected	0	234	211	148.3	324	25	174.5
Number Identified	3	277	300	193.3	325	205	265
Number Rejected	4	265	263	177.3	73	204	138.5

Table.1. hPAM and LDA Correctness

Table 2 shows that the average of topic super topics and subtopics identified by hPAM provides better rate of precision, recall and F-score than the traditional LDA. Similarly the accuracy, specificity and fall-out rate of hPAM shows better results.

	hPAM				LDA		
Performance Parameters	topic	super topic	sub topic	Average	topic	sub topic	Average
True Positive Rate	1.00	0.16	0.30	0.48	0.00	0.88	0.34
False Positive Rate	1.00	0.76	0.47	0.74	0.01	0.88	0.34
Accuracy	0.86	0.35	0.34	0.51	0.18	0.47	0.32
True Negative Rate	0.75	0.76	0.47	0.66	0.95	0.02	0.14
Positive Predictive Value	0.75	0.58	0.63	0.65	0.20	0.50	0.50
Negative Predictive Value	1.00	0.30	0.18	0.49	0.18	0.14	0.17
False Discovery Rate	0.25	0.42	0.37	0.35	0.80	0.50	0.50

Table.2. Performance of hPAM & LDA

Though fall-out rate of hPAM for topics is higher, hPAM on average has lesser fall-out rate and higher accuracy. hPAM returns more specific results. The Fig 2 shows the comparative result generated by LDA and hPAM for document retrieval. hPAM shows more precision and recall in document retrieval for a given input research article.

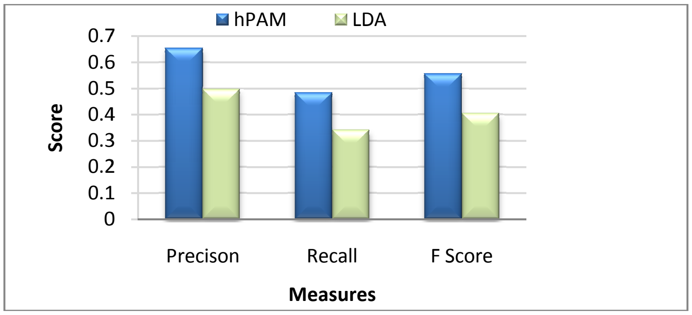

Fig. 2. Performance of hPAM & LDA in relevant document retrieval

The results from hPAM are clustered and documents from the nearest cluster are retrieved using k-nearest neighbor algorithm. The graph in Fig.3 shows the number of clusters obtained with hPAM.

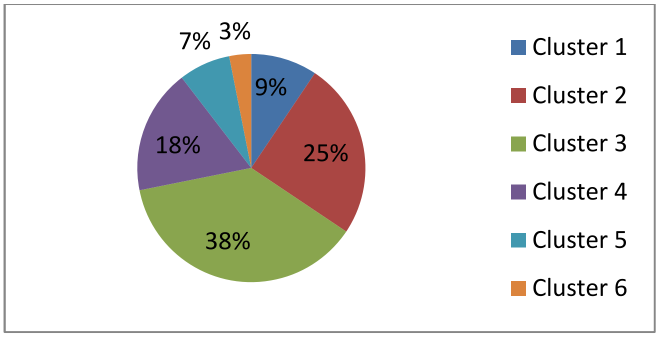

Fig. 3. Clustering with hPAM

The novelty score obtained by the system is compared with novelty using cosine similarity and expert opinion of the research article and the results obtained are as in Fig 4.

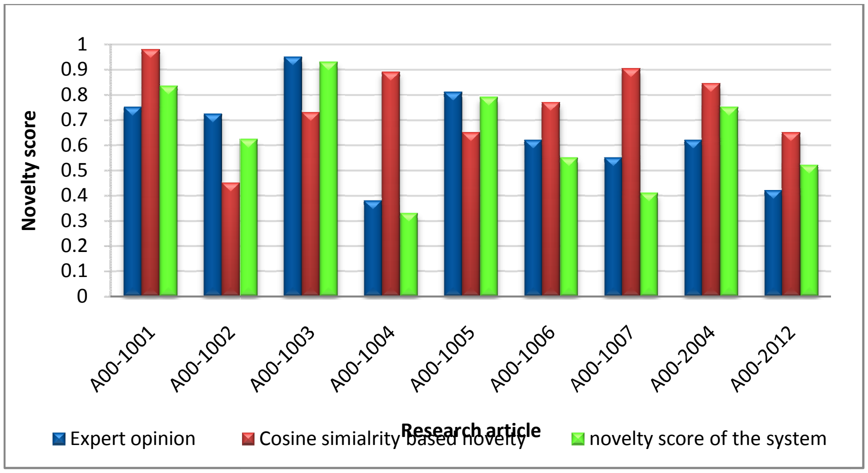

Fig. 4. Novelty Score of the system

The novelty score of the proposed system shows a correlation co-efficient of 0.8852 with the expert opinion. This shows that the novelty score of the proposed system is 88.52% relevant to the expert opinion. The following analysis were made

- The relevant documents retrieved by using hPAM showed better performance than LDA. But the use of hPAM is limited to the kind of input that can be categorized at different levels. Performance of hPAM for generic unstructured document is yet to be determined.

- The clustering mechanism used in the proposal involves term frequency based approach. A better algorithm for clustering research articles based on bibliometrics can be implemented to optimize the result obtained.

- The concept maps in this work included concepts and generic relations among them. The concepts in the document were stemmed to its root word, which lead to some misinterpreted relations among concepts obtained. Few concepts and relationships were lost due to stemming. The possibilities of using POS tagging is to be analyzed.

- The novelty score of the system shows results in close correlation with expert evaluation; however this fails to account the writing style of the author which plays a major role in novelty mining.

5. CONCLUSION

The experiment shows that hPAM is better to topic model research article as it shows better performance in terms of accuracy, precision and recall for retrieval of relevant document. The novelty score of the proposed system is in close proximity to the expert evaluation than the traditional similarity based novelty. This work includes originality (inverse of similarity) as a parameter to define novelty. This approach is not fully quantitative as it considers the semantics of concepts in the research article. Qualitative approach for research articles involving sentence importance and sentence contribution to novelty is to be focused in further implementations. The graph based approach for novelty will be implemented as an extension of this work and its effect on novelty will be determined in future. Section wise novelty estimation can be focused as research article has certain sections that contain information listed previously like Introduction and Survey sections. Measures can also be taken to extract the regions that contribute to the novelty score of the research article, which will enable researchers, reviewers and other readers to exactly identify the newness in the article without having to read the entire article. The novelty score obtained can also be used to qualitatively rank the articles.

Acknowledgments

This work was funded by the Center for Technology Development and Transfer, Anna University Chennai under Approval no. CTDT-1/2360/RSS/2011 for Innovative project by Young faculty members under research support scheme for Dr. S. Sendhilkumar, Assistant Professor (Senior Grade), Department of Information Sciences and Technology, Anna University.

REFERENCES

[1] Agus T Kwee, Flora S Tsai and Wenyin Tang, “Sentence level novelty detection in English and Malay”, Lecture notes in computer Science, Advances in Knowledge Discovery and Data Mining, Vol.5476, pp. 40-51, Springer, 2009.

[2] Alexander Ypma, and Robert P. W. Duin, “Novelty detection using self-organizing maps”, In Progress in Connectionist-Based Information Systems, volume 2, pages 1322–1325. Springer , London, 1997.

[3] Barry Schiffman, and Kathleen R. McKeown, “Context and Learning in Novelty Detection”, ACM, Proceedings of the Conference on Human Language Technology and Empirical Me-thods in Natural Language Processing, 2005.

[4] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt, “Support Vector Methods for novelty detection”, S.A. Solla, T.K. Leen and K-r Muller (eds.),pp.582-588, MIT Press 2000.

[5] Bruno Pôssas, Nivio Ziviani, Wagner Meir, Jr., and Berthier Riberio – Neto, “Set-based vector model: An efficient approach for correlation-based ranking”, Journal - ACM Transactions on Information Systems (TOIS), Vol. 23, Issue 4, pp. 397-428, October 2005.

[6] C.M. Bishop, “Novelty Detection and Neural Network Validation”, IEE Proceedings of V international symposium on Image Signal Process, Vol.141, No.4, 1994.

[7] F. S. Tsai, and Y. Zhang, “D2S: Document-to-sentence framework for novelty detection," Association for Computing machinery (ACM), vol. 29, no. 2, pp. 419-433, November 2011.

[8] Flora S Tsai, Wenyin Tang, and Kap Luk Chan: “Evaluation of novelty metrics for sentence-level novelty mining”, Journal of Information Sciences, vol.180, pp. 2359–2374, Elsevier, February 2010.

[9] Flora S. Tsai , “Review of Techniques for intelligent novelty mining”, Information Technology Journal, Vol.9, issue 6, pp 1255-1261, 2010.

[10] Flora S. Tsai, “Review of Techniques for Intelligent Novelty Mining”, Information Technology Journal, vol 9, issue 6, pp 1255 - -1261, 2010.

[11] Guilherme A. Barreto, and Rewbenio A. Frota, “A Unifying methodology for the evaluation of neural network models on novelty detection tasks”, Journal of Pattern Analysis Application, Springer, 2012. DOI: 10.1007/s10044-011-0265-3.

[12] Ian Soboroff and Donna Harman, “Novelty Detection : The TREC Experience” , Proceedings of HLT/EMNLP,October 2005

[13] James Allan, “Introduction to Topic Detection and Tracking”, Topic Detection and Tracking, Information Retrieval Series, vol.12, pp.1-16, 2002.

[14] James Allan, Courtney Wade, and Alvaro Bolivar, “Retrieval and Novelty Detection at the Sen-tence Level”, ACM, SIGIR ’03, August 2003.

[15] Jian Zhang, Zoubin Ghahramani, and Yiming Yang, “A Probalisitic Model for Online Document Clustering with Application to Novelty Detection”, Neural Information Processing System (NIPS), Vol.17,2004

[16] Junji Tomita, Hidekazu Nakawatase, and Megumi Ishii, “Calculating Similarty Between Texts using Graph-based Text Representation Model”, Proceedings of thirteenth AVM international conference on Information and Knowledge Management,pp. 248-249, ACM, 2004.

[17] Le Zhao, Min Zhang, Shaoping Ma, “The Nature of Novelty Detection”, ACM, vol.9, no.5, pp. 521-541, Nov 2006.

[18] Li-Tung Weng, Yue Xu, Yuefeng Li, and Richi Nayak, “Improving Recommendation novelty based on Topic Taxonomy”, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology workshops, IEEE Computer Society, 2007.

[19] Marcelo Keese Albertini, and Rodrigo Fernandes de Mello, “A Self-Organizing Neural Network for Detecting Novelties”, Proceedings of 2007 ACM symposium on Applied Computing, pp. 462-466, ACM, 2007.

[20] Michael Bendersky and Oren Kurland, “Utilizing passage-based language models for document retrieval”, Proceeding ECIR’08Proceeinggs of the IR research, 30th European Conference on Advances in IR,pp.162-174, ACM, 2008.

[21] Michael Gamon, “Graph-Based Text Representation for Novelty Detection”, ACM, Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, pp.17-24, 2006.

[22] Ming-Feng Tsai, Ming-Hung Hsu, and Hsin-His Chen, “Similarity Computation in Novelty Detection”, National Institute of Science and Technology(NIST), 2004.

[23] Mohammed Al-Kabi, Niveen Z. Halalsheh, and Heider A. Wahsheh, “Arabic News: Topic and Novelyt Detection”, Proceedings of 3rd International Conference on Information and Communication Systems, Article no.7, ACM, 2012.

[24] Nicola Stokes and Joe Cathy, “First story detection using composite document representation”, ACL, Proceedings of first international conference on Human Language Technology re-search, pp 1-8. 2001.

[25] R.T. Fernndez, “The effect of smoothing in language models for novelty detection”, ACM, pp. 17-24, 2006

[26] Sendhilkumar S., Mahalakshmi G.S., Harish S., Karthik R., Jagadish M. and Dilip Sam, Assessing Novelty of Re-search Articles using Fuzzy Cognitive Maps, First International Symposium on Intelligent Informatics ISI 2012, Springer, 2012.

[27] Sugato Basu, Raymond J. Mooney, Krupakar V. Pasupuleti, and Joydeep Ghosh, “Evaluating the novelty of Text-Mined Rules using Lexical Knowledge”, Proceedings of seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp.233-238, ACM, 2001.

[28] Victor Lavrenko, James Allan, Edward De Guzman, Daniel LaFlamme, Veera Pollard, and Ste-phen Thomas, “Relevance Models for Topic Detection and Tracking”, Proceedings of second International Conference on Human Language Technology Research, pp.115-121, 2002.

[29] Xiaoyan Li, and W. Bruce Croft, “Novelty Detection Based on Sentence Level Patterns”, Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp 744-751, ACM, 2005.

[30] Xiaoyan Li, W.Bruce Croft, “An Information-Pattern-Based approach to novelty detection”, Journal of Information Processing and Management, vol.44, issue 3, pp: 1159-11881, El-sevier 2008.

[31] Y. Zhang, J. Callan, and T. Minka, “Novelty and redundancy detection in adaptive filtering,” SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 81-88, 2002.

[32] Yi – Hung Liu, Yan-Chen Liu and Yen-Jen Chen, “Fast Support Vector Data Descriptions for Novelty Detection”, IEEE Transactions on Neural Networks, vol. 21, no.8,pp.1296-1313, 2010.

[33] Yi Zhang and Flora S. Tsai, “Chinese Novelty Mining”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.1561-1570, ACL, 2009.

[34] Yi Zhang, and Flora S. Tsai, “Combining Named Entities and Tags for Novel Sentence Detec-tion”, ACM, ESAIR’09, Spain, 2009

[35] Yi Zhang, Flora S. Tsai and Agus Trisnajaya Kwee, “Multilingual sentence categorization and novelty mining”, Journal of Information Processing and Management, vol. 47, issue.2011, pp.667-675, Elsevier, 2011.

[36] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin, "Topic Conditioned Novelty Detec-tion", Proceedings of eighth ACM SIGKDD international Conference on Knowledge Discovery and Data mining, pp. 688-693, 2002.

[37] Ying-Ju Chen, and Hsin-His Chen, "NLP and IR approaches to monolingual and multi lingual link detection", Proceedings of the 19th international conference on Computational Linguis-tics, vol.1, pp.1-7, Association of Computer Linguistics, 2002.