ISSN: 2454-8421, Volume 1, Issue 1, July-Dec, 2015. Page 44-49



# Database optimization and Novelty Mining<sup>B</sup> of News articles

Shweta Taneja, Charu Gupta, Ankita Mohan Saxena ,Jatin Rijhwani ,Sanya Malhotra

Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, Rohini, Delhi shweta\_taneja08(@yahoo.co.in, charu\_2287(@yahoo.com

Abstract- With rapid advances in Information Technology, the normal way for people to obtain information has changed. However, the current available search engines, like Google, cannot tell whether a newly posted article contains fresh content or not, as compared to all the previous posted articles. Thus, people may sometimes waste time reading articles which are about old or have known information. The solution to this problem is Novelty mining; it is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, text mining, natural language processing, information retrieval, and knowledge management.. In this paper, we have illustrated the various steps involved in the miming of the dataset. The objective of this research was to find an optimal way to scan through large databases and detecting relevant information efficiently. The results show that proposed novelty mining framework can detect novelty on a set of news articles with very high accuracy.

Key words — Novelty Mining; Database optimization; Preprocessing; Information retrieval; Indexing

### I. INTRODUCTION

In today's information age, it is easy to store large amounts of data. However, although the amount of data available to us is continuously growing, our ability to gather this information and use it remains constant. Imagine the time savings if we are only presented with novel information to read, while the old or redundant information is filtered out. Thus, novelty mining [1] helps to extract novel information out of a huge set of text documents. The term novelty (derived from Latin word Novus for "new") is the quality of being new, or following from that, of being striking, original or unusual. In novelty mining, users are able to send different documents to be tested for its relevance and novelty. Due to the millions of data in the database, the insertion and selection of data have to be kept at optimum.

A novelty mining system [2] is able to discover novel, yet relevant information based on context and reader's preference. It is helpful in personal newsfeeds, information filtering, as well as many other fields where duplicate

information may be returned to the users. In general, a nonmining system consists of three main parts, have preprocessing, classification and novelty mining. Firstly, the documents are input into the system for preprocessing the models will be built by using various machine learning the system will determine algorithms. Then, the system will determine relevant documents for a given topic and filter out the non-relevant to classification stage Finally documents in the classification stage. Finally, based historical articles, the system will determine whether the top article is novel or not. The contributions of this paper are twofold. Firstly, to design and develop the optimized techniques for SQL SERVER 2005 database for retrieval relevant information, which has not been well-studied being and secondly, to study the novelty mining system when involves pre- processing as its first phase followed classification and novelty mining techniques to detect and data from a dataset. This paper spans across the three man emerging research areas of databases that include databases indexing and information retrieval by query processing me processing of dataset and knowledge management.

This paper is organized as follows. In the first section he introduction about the motivations for the research at development of novelty mining system is presented. In it second section literature review of various optimization at novelty mining systems is described. The third secur comprises of the framework that we have proposed for a entire Novelty Mining system. In sections four details about dataset used i.e. Reuters 21578 is explained. In section five a six the experiments conducted and subsequent performance evaluation is shown. Finally, at the end of this paper at conclude and give suggestions for future work in this field.

## II. RELATED WORK

The major contribution in the field of optimization at novelty mining is by Flora S. Tsai. Other authors have as contributed in this area. In [1], authors have explored is importance of novelty mining and database optimization technique on a dataset of business blogs, with a very his accuracy. Previous research on novelty detection have stressed on the task of finding novel material, given a set documents on a certain topic. Authors in [2] studied the m

part task defined by TREC 2002 novelty track that is firstly, finding the relevant sentences from the documents and then identifying the novel sentences from the collection of relevant ones. The research here shows that the former step appears to be more difficult part of the task. In [3], authors have analysed web logs posts for various categories of cyber security threats related to detection of cyber attacks, cyber crittle and terrorism. They have used Latent Semantic models such as Latent Semantic Analysis (LSA) and Probabilistic LSA, to detect keywords from cyber security web logs. LSA is also discussed in another paper [6]. In another work [5], anothers have proposed experimental results on APWSJ data set. They have shown that Document to Sentence(D2S) framework outperforms standard (document level) novelty detection in terms of redundancy-precision (RP) and redundancy-recall (RR). However they have suggested that D2S shows a strong capability to detect redundant information. Also, in [8] authors aim to explore the performance of redundancy and novelty mining in the business domain. They have adopted the mixed metric approach which combines symmetric and asymmetric metrics

Different researchers have contributed in the area of database optimisation, but either they have focused on B-Trees or indexing techniques by LSA method. None has given attention to pre processing and optimisation using indexes. In our paper, we have proposed a framework which converts unstructured data of news articles to a structured form (tables) and there after indexing is performed and performance comparison is observed. This will also form basis for our future work of novelty mining, keeping in mind the constraints and challenges in natural text.

#### III. PROPOSED FRAMEWORK

The framework of Novelty Mining system is shown in figure 1. It is divided into four phases:-i. Pre-processing

ii. Database Optimization iii. Novelty Mining iv. Information Retrieval. The detailed explanation of these phases is given below.

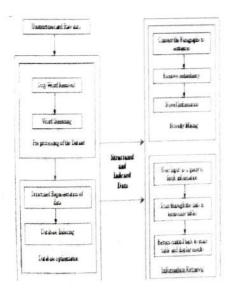


Fig. 1. Proposed Framework

#### A. PRE- PROCESSING

There are various pre-processing techniques that infer or extract structured representations from raw unstructured data sources. There are different operations under pre processing like stop word removal and word stemming. Stop Word Removal aims to remove stop words like 'is', 'an', 'the' etc. Word Stemming is the process of reducing inflected (or sometimes derived) words to their stem, basic root formgenerally a written word form. E.g. running-> run, Drinks-> drink, Mangoes-> mango

#### L) ALGORITHM USED FOR WORD STEMMING

We have used a modified form of Porter Stemmer Algorithm [10]. The Porter stemming algorithm (or 'Porter stemmer') is a process for performing stemming i.e., reducing the word to its root form. It is mainly used as a part of term pre-processing, that is usually done when setting up Information Retrieval systems. The algorithm stems the data using a set of rules. There are 60 rules in 6 steps of porter stemmer algorithm. These steps are:-

- 1. Removes plurals of the words.
- Turns terminal y to i when there is another vowel in the stem.
- Maps double suffixes to single ones, eg-'ization'. 'ational' etc
- 4. Deals with suffixes -full, -nests etc.
- 5. Takes off -ant, -ence etc.
- 6. Removes a final -e.

In our modified porter stemmer algorithm, we remove stop words like 'is', 'an', 'the' etc along with above suffix removal. We have used java as a programming language for implementing our algorithm. The benefit of implementing porter stemmer is to enhance search process in the large pool of data and moreover to increase the efficiency of the entire system.

#### **B. DATABASE OPTIMIZATION**

Database optimization is a technique to improve the query performance with indexing and statistics. It can be defined as the optimization of resources used to increase throughput and minimize contention, enabling the largest possible CPU workload to be processed. In our paper we have used indexing to optimize the dataset.

There are two types of indexes that have been built on the data namely clustered index and non clustered index. The two types of indexes are as explained below:

#### 1) NON-CLUSTERED INDEX

The data is present in random order [12], but the index specifies the logical ordering. The index keys are in sorted order, with the pointer to the record contained in the leaves of the tree. There can be more than one non-clustered index on a database table. Non-Clustered indexes have structures that are different from the data rows. A non clustered index key value o points to data rows that contain the key value. This is called as row locator. Its structure is determined on the basis of the type of storage of the data pages. A heap table [11] by definition is a table that doesn't have any clustered indexes.

Another case arises when no index is defined for a table at all. In that case the address of the first IAM page of the heap table itself is stored in the sysindexes table with indid = 0 as shown in figure 3. So, the full form of IAM is a limite misleading; it would be better called as SAM (Storage Allocation Map or Space Allocation Map).

## 2) CLUSTERED INDEX

Clustering modifies the data block [12] into a certain specific order to match the index. Therefore, only one clustered index can be created on a given database table. Clustered indices greatly increases the overall speed of retrieval, but usually only if the data is accessed sequentially in the same or reverse order of the clustered index.

Fewer data block reads are required as the physical records are in the sort order on disk, the next row item in the sequence is immediately before or after the last one, and so on. Some databases separate the data and index blocks into separate files, others put two completely different data blocks within the same physical file(s). An object is created where the physical order of rows is the same as the index order of the rows and the bottom (leaf) level of clustered index contains the actual data

## C NOVELTY MINING

Novelty mining is the identification of new or unknown information from a given set of text documents. It is useful in personal newsfeeds, information filtering, as well as many other fields where duplicate information may be returned to

The approach followed to perform Novelty mining in our project is Document to sentence. The document (or paragraph) of information is fragmented into sentences to remove duplicity or redundancy. Temporary tables are created on the fly to accomplish the task. Moreover the search is also carried out in these tables for patterns as specified by the user, but the information is displayed from the main table.

A Cursor makes it possible to perform complex logic in SQL. A cursor can be viewed as a pointer to a row. It can only refer one row at a time.

Two Cursors have been used in this project for the following purposes:

- (1) For removing redundancy/duplicity.
- (2) For implementing search in the temporary tables.

Cursor is a server side tool. It is giving row-wise solution to the result set.

# D. INFORMATION RETRIEVAL

Retrieval of information is also an integral part while designing a system, so as to provide the user relevant information according to the query input by him. If the results obtained are relevant and correct, then the system developed is said to be efficient. The various steps for retrieval of information from the structured tables are as follows:

(1) User inputs the request through query.

- (2) Search is carried out in temporary tables through
- (3) Transfer of control from temporary tables to the man
- (4) Output generated from the main table

#### IV. DATASET

#### A. REUTERS 21578

We have used Reuters 21578 dataset [9] in our work Tra documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. We have used a subset of the complete dataset for our study. The dataset is divided into in categories namely companies, exchanges, organizations people, exchanges and topics.

#### B. TOOL USED

The database software that we have used is MICROSOFT SQL SERVER 2005.Microsoft SQL Server 2005 is a relational database management system developed by Microsoft SOL Server 2005 (formerly codenamed "Yukon") was released a October 2005. It included native support for managing XMI data, in addition to relational data.

# V. EXPERIMENTS CONDUCTED

The various phases under the project such as pre processing, database optimization and novelty mining were carried out as under. Performance evaluation is a key step to examine ; project. We have evaluated our work and calculated the efficiency of our work. But before discussing the various cases of execution, we first give an overview of the work done and then its relevant efficiency.

# Step 1: Creation of tables

a se pro trata tor trata torona		The set of the set
		and a second second
in the second se	12.5. 12.200	
there are a state of the state	P Bland of Landson	
d rieses	the set of the last	
C & The Laborator		
a Willier advance.		
· J med	and a second sec	
a menti		
4 - 4 - 4 - Anna - A		
- 199 - 19-1		
# *** rgm.etc		
a minate and		
a war of the second		
5 mm		

Figure 2: Creation of tables

The above figure shows the query executed to create the table in SQL Server for converting unstructured raw data to structured tabular form. The different attributes for the data are chosen keeping in mind the various categories present m the actual dataset.

Step 2: Insertion of data into the tables



## Figure 3 : Insert query

The above figure shows the query which is executed to enter the data into the tables. The various attributes are assigned vales accordingly.

Se 1	in the One I	teen took Window Comments, Man.				-		-	
	the state of the state	020405301 +.							
- 1	21 LETHE	- 1 teres - 178 2 5 17	植物	0.5	0.22.64	· .			
-	NO. NO.	hery/at Schally							1944
	an () =+	per e 10)							
2.6	The Sector								
	41.04	m.	-	-	www.	-	-	-	
	IN COMPANY AND INC.	10.000 (25.9)119 at 1.	in.	410		mail		\$ 30	MART & Non-Address
	10.10110-0.00	served in the second second second second second	4.5	Arrist.	custility.	\$1.01			stilled an inner on this caller
	appropriate and the	BREASHING BREEDOWN, RECURSO ON PROVIDE		410	where the second	at to	1.1	+4.	MARIANELIA MART NO.
	100-1-10-10 Max - Box	THEY REPORT THE CARDING STREET, AND THE	8.51	10.1		12.44	4.1	640	tensor area bruinder metrica
	HARVOR HEALTH	NO FRAME, 41-144, 4127-11471-1141-1	10	8.1	Carl .	817	N	11.0	ittention in monthly
	19. 01012334-20.	Advert 120, 485 phills	-	10.11		No.14	Print 1	40.5	inserantin section.
		DAVERS TO DESCRIPTION OF DEPOSITE OF PARTY	NO.	Note		New	B.1.	- 44	Midentel Michigh Scient PL
	INCOMPACT AND	1915/00 (01014 1 48) 40 (41 (1 0)	418			100	æ/ -	*1.	targeters - which has not \$ 2.0
	to contraction	TRUE MERCHANNEL - 10 12 March - 1, 12 March	1.00	1443	10.00	41.4	A	0.00	item & dearch inferration
ν.		TO MORE CREATED TO AND TRADE AT THE TAX THE	-	P 11		**:-1	×	-	To be a frank of a weather that
	11110-01151 WINF	data proceeding of the second second second	-	1411	16.0	P. 11	A	+ *	Long M LT. ( Long B
κ.	Wair Mich.	appropriate transport to a reply an	10	18-11		16.61	4	1.2.	laterati E lance At at de.
	104 YORG THE LIVE	456 19105 941				10.7.4	PE1	1474	Second Annual St. De Hot I
6	Solar Report of the	LARGENTS SHOTTER WITH CONTRACT.	81.h.;	(44)	UNITARIAN PARTY	A	m-1		
		same agreeded also considered	-		an whe	2011	*	- 10.4	
۰.	101010-0111-11	same and a second secon	41.0	19-11	10.040	100	No. 1	16.8.1	
	101111-0111-010	TOTAL TOTAL AND THE AND TOTAL CONF	**	$\psi_{11}$		20,41	N.(.)	- 14	trater and an other second s
÷.)	-KA(16) (0.1 - 17)	wallow allocate		1916	-	w	4		
		100004(100,01000)		46.1	- 80	44.42	44	- 1	LOCAL ROOM AND LODGE ADDRESS -

Figure 4 : Table contents

The above figure shows the content of the table. This is only a portion of the complete database. The table consists of 255 records in all pertaining to different categories.

#### Step 3: Creation of indexes

Following figure 2 is the list of indexes that we have created on out database:-

] Results	J Massa	Qes	
PRIM	ARY	1	
Index	name	index_description	index_keys
ciation		clustered located on PRIMARY	topic_date
Same and	and an owner the second s	nonclustered located on PRIMARY	companies
	comparies	nonclustered located on PRIMARY	exchanges
	exchanges	nonclustered located on PRIMARY	លក្ន
nclat	0.6	nonclustered located on PRIMARY	people
	people	nonclustered located on PRIMARY	places
i nolet,	places	nonclustered located on PRIMARY	topics
/ ndist	topic	nonclustered located art this	

Figure 5: list of indexes on the table

As it can be seen from the above figure, we have made a total of 7 indexes on our dataset (table). One of them is a clustered index while all the others are non clustered index. The significance of using such indexes has already been discussed.

Step 4: Applying queries on the dataset for comparison

The various execution plans for different queries are given in a tabular form for easy understanding and comparison. The numerical values depicted in the table are in the form of CPU cycles required to perform the task. As can be noticed from table 1, the execution times in case of "clustered index seek" are lowest. The query execution is optimum in this case. Also the values for "index seek" and "table scan" are similar, this is due to the fact that in both the cases the indexes are not used.

Step 5: Converting paragraph ( or document ) level to sentence level.

theat (\$ low to a	and the
	Aust fair Aller un retri de
trate 1 54	13 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
A G most	1 them 1 4 4 5 1 9 1 0 6 0 1 1
4. 5	The second se
10 100 100 100 10 100 100	MSSI & Lances Gardin C. C.
1 1.0	
	NF and
in a grant of	and the A.C. was republicated user of the scores. Analytic traff are all large consequences also represent the boots of the
maintiagion	he & annual future of all of me
814001-8181	an early to part white strate ( from ( ) ) happ if things range and in the range ( )
Distanti in World	1.0040+1
0130010000	P directory region and
- anticketers	fen fundeten som inkant och at unsperiforssändet och erverligtet
a appropried	wants with our found all one had not been all the new set of the s
ALCO DO FROM	advarda y se service aven de la construcción
SPACE AND ADDRESS OF	I conception on the discretion
	na one ser her verbesselbesende standers of electronic meaning in the mean
	The provide an increase of a descent all and a special devices in special or produce solvey. We can use where
1 BOODSHEET	<ul> <li>and taking sectors only only and into out or post of Press one of sectors append.</li> </ul>
	Finance instruction on the instruction of the in
	For texas on texas or do very departs of our particular our solution in a second diff. "I still for allows a difference difference and texas."
a material a	(1996-1999年1月1日)(1997年1月1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日)(1997年1日))
	franziska z ustrani od
F REALIZING MILET	f den gevene i de file en l'Ales.
	1 valide and a l
A PERMIT	1. Complete cap are spin-cill advects in Appleto in Start Adda in 12

Figure 6: Sentence level fragmentation of data

The output depicted above in the figure 6 is obtained after executing the first code which is written to convert the paragraph level data to sentence level. This step is carried out in order to apply novelty mining techniques on the dataset at the sentence level rather than document level. The above output is of a temporary table which has only two attributes namely title and the text.

Step 6: Applying Novelty Mining on the dataset.

In the next step the application of novelty mining technique is

transition in	State of the state of the state				
W2012034N0.K04	and the second second				
and the second s				11. 11 11. 11. 11.	
and the second s	10.74 1-4.147	-14-187 N. ( 7 17 27 122 282 1 41	 1-114	later Angeler Angeler Angeler Angeler Angeler	

carried out. In this step the redundancy or the duplicacy in the data if present is removed by using the code developed in the

form of cursors. Each sentence in the dataset is compared to the pool of existing data already present in the data and thus redundant data is ommitted and never added to this temporary table. The purpose of removing redundancy is to increase the effectiveness of search queries as the data needs not to be checked in redundant data again and again, as this consumes time and thereby degrading performance.

Step 7: Applying a search query on the dataset prior to conversion of document to sentence.

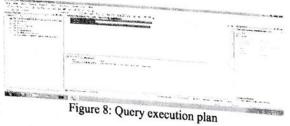
ki bă da Çaş haşai tak si Jaraa Litili Dis grad Sili anar itili Ballanar itili	51003.	509. A.
Hanner & S. S. S. S. S. Son Call Down Call Down C. S. C. Strand C. C. Son Call Down Call Down C. S. Strand Transmission S. S. Strand Transmission S. S. Strang S. Strang	<ul> <li>The distribution of the second se second second sec</li></ul>	
<ul> <li>a betra baya</li> <li>a betra baya</li> <li>a betra baya</li> <li>a betra baya</li> <li>b betra baya&lt;</li></ul>		
	g Jong J Margan Markan (Marka) - San Angel (Marka) - Angel (Ma	
	2	

Figure 7: Query execution

As can be seen from the above figure 7 that the search of patterns is not applicable to the data directly as the data in the table is of XML form. And pattern search or text phrase search is not possible in the case of Xml attribute type. So we need to cast this data to another type such that we are able to perform search queries on the text attribute as well. The alternative solution to this problem is described in the next step.

Step 8: Casting the XML data to NVARCHAR.

The figure 8 illustrates the casting of XML data attribute to a NVARCHAR type data. The purpose of this conversion is to apply text search or pattern searches to this attribute. As depicted by the figure a search query is executed searching for the presence of phrase 'COCOA' in the dataset. And at the bottom of the figure the successful execution of the query is shown. The detailed execution times have been shown by another figure 9 shown below.



Chastered Inde	** Sean
Scenning a clustered index, enti-	uty or only
Physical Operation Logical Operation Estimated I/O Cost Estimated CPU Cost Estimated Operator Cost Estimated Subree Cost Estimated Number of Rows	Clustered Index Scan Clustered Index Scan 0.0416433 0.000437 0.042081 (993, 0.04208
Istimated Row Size	25
Ordered	40501
Node ID	Faise
Object	
(dataset) (dbo) (dataset) (cistind	[Ho_ xH]
Output List	
(dataset) (dbo).(dataset) title. (d	ataset) (rikes)
(dataset).text	(internet)

Figure 9: Detailed execution plan

Step 9: Application of second code to perform search in

1. **** 12 12 12 2 3 4 4	1 2 4 5 5 7 .	Renne .
To by Adams	A set of a set of the	
<ul> <li>File Sales Sales</li> <li>File Sales ingdes</li> <li>Jonne</li> <li>Jonne</li> <li>Sales</li> <li>Jonne</li> <li>Sales</li> <li>Jonne</li> <li>Jakova</li> </ul>	na anti-anti-francesa Anti-anti-paulota (assa Anti-anti-paulota)	07
dig formgen og Tig het hanne formen (1) Ve forser foget hander forboar og (1)	and a second sec	
	THE T	

Figure 10: Query execution plan

The above figure 10 shows the application of the search in XML form data. The highlighted text in the figure is the query which is executed to display the relevant data. This method is also an alternative to the search procedure described in the previous step. The difference in this approach is that the search operation is carried out only on non redundant data; thereby overall execution time is reduced if the dataset is having redundancies present in the text. The execution plan of the query is shown above which shows the actual working of the query and transfer of control amongst the tables to display relevant data to the user.

Table Scan Scan rows from a table			
in a table		Clustered Inde	
Physical Operation		Scarving a chosered index entit	n acan
Logical Operation	Table Scan		ed or cult a mile
Estimated I/O Cost	Table Scan	Physical Operation	Outpet index (col
	0.0032035	Logical Operation	Outmet Index Scar
Estimated CPU Cost	0.0000851	Estimated I/O Cost	10416411
Estimated Operator Cost	0.0049584 (9%)	Estimated CPU Cost	0.0304364
Columnated Culture e	0.0040504	Estimated Operator Cost	1.5425799(78%)
Columnated Minute	0.0049584	Estimated Subtree Cost	53420796
Estimated Row Size Ordered	6	Estimated Number of Rows	254
Urrerad	29.6	Estimated Row Size	4463 i
Node ID	Faise	Ordered	Spie
	6	Node ID	2
Object			
		Object	
[tempdb].[dbo].##TEMP_NEW		[dataset].[dbc].[dataset] [cistind	a ari
Output List		Output List	
[tempdb].[dbo].[##TEMP_NEW]	title	[dateset] [dbu] (dateset) title (da  dateset] text	alaset](dbc)

re 11: Detailed execution plan

# VI. CONCLUSION AND FUTURE WORK

The proposed work uses a large dataset of news articles. An efficient way to optimize database has been proposed with indexing technique. The experimental results obtained show that the work optimizes the database with the execution time in clustered index seek and as can be noticed is lowest out of all other attributes. Also the values of Index seek and Table scan are similar, as both do not consider indexes. With the proposed work the effectiveness of optimization has been studied experimentally. Further investigation to the topic reveals that novelty mining with database optimization can give good results.

The results obtained from the experiments conducted show that the execution time in case of sentence level search as well as cast search is similar. This is due to the fact that the dataset does not contain redundancies Moreover the Document to sentence conversion is also successfully carried out with the help of the proposed algorithm. Thus the proposed optimization and novelty mining algorithms are efficient. Mining of documents for novel information is successfully accomplished by removing redundancy or duplicity from the data.

#### VII. ACKNOWLEDGEMENTS

We take this opportunity to express our sincere thanks and deep gratitude to all those who extended their wholehearted cooperation and have helped us in completing this work successfully. We express our sincere thanks to Mr. Suyash Gupta, DBA (HCL Technologies) for his encouragement and valued suggestions.

#### VIII. REFERENCES

- A.T. Kwee, and F. S. Tsai, "Database Optimization for Novelty Mining of business blogs", Elsevier Expert Systems with Applications vol. 38, pp. 11040-11047, 2011
- [2] J. Allan, C. Wade, and A. Bolivar, "Retrieval and Novelty Detection at the Sentence Level", SIGIR'03, ACM 1-58113-646-3/03/0007., Toronto, Canada, July 28-August 1, 2003
- [3] F. S. Tsai and K. L. Chan, "Detecting Cyber Security Threats in Weblogs Using Probabilistic Models", Springer – Verlag Berlin Heidelberg, LNCS 4430, pp. 46-57, 2007.
- [4] H. Cui ,M.Y. Kan and T.S. Chua, "Unsupervised Learning Of Soft Patterns For Generating Definitions From Online News", ACM 1-58113-844-X/04/0005, WWW 2004, May 17-22, 2004.
- [5] F. S. Tsai, Y. Zhang, "D2S: Document-to-sentence framework for novelty detection", Received: 18 June 2009 /

Revised: 22 July 2010 / Accepted: 11 December 2010© Springer-Verlag London Limited 2010

- [6] S. Deerwester, S. T. Dumais, R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, vol. 41, pp. 391, 1990
- [7] F. S. Tsai and K. L. Chan (2010), "Redundancy and novelty mining in the business blogosphere", The Learning Organization, vol. 17, Jss: 6, pp. 490 – 499, Emerald Article.
- [8] R. Feldman, The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data, Israel James Sanger ABS Ventures, Waltham, Massachusetts.
- [9] The dataset source: http://kdd.ics.uci.edu/databases/reuters21578
- [10] Porter-Stemmer algorithm for preprocessing of dataset.
- [11] <u>http://msdn.microsoft.com/en-us/library/aa964133(v=SQL.90).aspx</u>
- [12] http://en.m.wikipedia.org/wiki/Database\_index