http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (6): 1255-1261, 2010
ISSN 1812-5638

# Review of Techniques for Intelligent Novelty Mining

Flora S. Tsai
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

**Abstract:** The detection of novel information is an important research area which is becoming more critical as we become inundated with an overload of information. Novelty mining, or novelty detection, is the process of mining the novel yet relevant information of a given topic. This study describes recent techniques for detecting novel sentences and documents. In particular, the study focuses on intelligent novelty mining techniques which address the domain-specific problem of detecting novel information with specific regard to the user context. These techniques are able to leverage the use of novelty metrics, novelty decision and novelty feedback to improve the results of mining new information from text data.

**Key words:** Novelty mining, novelty detection, novelty scoring, metrics, novelty decision, feedback, user context

## INTRODUCTION

The vast amount of current information that are readily available online leads to the problem of information overload due to the large quantity of irrelevant and redundant information contained in these documents (Kwee and Tsai, 2009). This information can be in the form of blogs (Tsai and Chan, 2007), social networks (Tsai *et al*., 2009), mobile information content (Tsai *et al*., 2010a), databases (Ong *et al*., 2009) and even Web services (Yee *et al*., 2009). To effectively alleviate this problem, Novelty Mining (NM), or novelty detection, has been proposed to retrieve novel yet relevant information, based on the specific topic defined by a user. In past studies (Tsai and Chan, 2010; Zhang *et al*., 2002), novelty was defined as the opposite of redundancy. Given a set of relevant documents, any document which is very similar to any of its history documents is regarded as redundant. Therefore, most of the later contributions have been made to sentence-level novelty mining (Allan *et al*., 2003; Kwee *et al*., 2009; Tsai and Chan, 2010; Zhang *et al*., 2010). This study surveys intelligent novelty mining techniques for overcoming the information overload problem.

## NOVELTY MINING TECHNIQUES

Novelty mining detects novel information in relevant documents in a given topic. The major components in novelty mining are (1) novelty scoring, (2) novelty decision making and (3) performance evaluation, as shown in Fig. 1. A novelty score is a calculated value that determines the novelty of a document and depends largely on the novelty metric that is selected. A novelty metric is a function which defines a distance between points in space. Many metrics used in Information Retrieval (IR) have been adopted for novelty mining based on the assumption that novelty is the opposite of similarity or redundancy (Tsai and Chan, 2010). The novelty of any document is quantitatively measured by the novelty metric based on its history documents (the documents that the user has read) and represented by the novelty score. The final decision on whether a document is novel or not depends on whether the novelty score falls above or below a novelty decision point, which is the boundary value or threshold that determ.

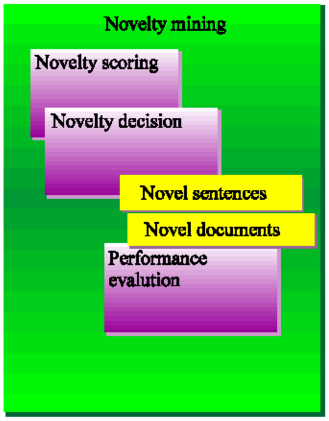

Fig. 1: Novelty mining components

**Novelty scoring:** For comparison of a relevant document to its history documents, several different geometric distance measures, or metrics, can be used, such as Manhattan distance and cosine distance metric. Depending on whether the ordering of the documents is taken into consideration, metrics can be either symmetric or asymmetric. Symmetric metrics, like cosine and Jaccard, yield the same result regardless the ordering of two documents. However, the results of asymmetric metrics, such as new word count and overlap, are based on the ordering of two documents.

In order to measure the degree of novelty directly, the similarity score is converted to the novelty score simply by (1-similarity score). There are two standard themes of comparison techniques in previous works. One of them is one-to-one comparison, where the current document is compared with each of the previous documents, then, the maximum of the redundancy (or similarity) scores obtained will be compared against a decision point (a) to finally decide whether the current document is redundant (Zhang and Tsai, 2009b). If the maximum redundancy score exceeds a, the current document is detected as redundant. The other theme is all-to-one comparison, where the current document is compared to the pool of all the previous documents, in order to generate the redundancy score. The simple similarity method and the overlap method are based on the one-to-one comparison paradigm, while the all-to-one paradigm is adopted for the simple pool method and the interpolated aggregate smoothing language model.

**Novelty decision setting:** After obtaining the novelty score of the incoming document, the system will make a final decision on whether a document is novel or not based on the novelty decision point. If the novelty score of the document is above the novelty score decision point a, the document is considered as novel.

**Performance evaluation:** The F-score is a popular evaluation measure that is used for evaluating the results of novelty mining (Zhang *et al*., 2010), as well as information retrieval for measuring search, blog classification and query classification performance. The F-score (also F1 score or F-measure) considers both the precision and recall to compute the score: P is the number of correct results divided by the number of all returned results and R is the number of correct results divided by the number of results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0, which means, the higher the F-score is, the more accurate the test is. Thus, F-score should be maximized.

Precision, recall and F-score are used to evaluate how well the system performs in detecting novel documents. On the other hand, to find out how well the system can detect non-novel or redundant documents, Redundancy-Precision (RP), Redundancy-Recall (RR) and redundancy F-score (RF) can be used instead (Tsai and Zhang, 2010).

Based on all the topics' P, R, F, RP, RR and RF, the average performance for documents can be obtained by calculating the arithmetic mean of these scores.

**Sentence-level novelty mining:** Many studies related to sentence-level novelty mining, or novelty mining, originated from the TREC Novelty Tracks (Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004). The novelty track was introduced in the eleventh Text REtrieval Conference (TREC) in 2002. This track was designed to investigate systems' abilities in locating new and relevant information from a given document set which is categorized into topics. Note: although the TREC Novelty Track uses the term novelty detection, the use of novelty mining (Kwee and Tsai, 2009; Zhang and Tsai, 2009a; Zhang *et al*., 2010) is preferred in order to avoid confusion with novelty detection in outlier detection (Hoffmann, 2007; Markou and Singh, 2003), which is based on one-class classification and is very different to novelty detection of text.

The novelty track document set was developed using the AQUAINT collection. The task of novelty mining was clearly defined as follows: given a topic and an ordered set of documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen (Soboroff, 2004). To fulfill the task, there are essentially two steps. The first is identifying relevant sentences, which is essentially a passage retrieval task. The second step is detecting those relevant sentences containing enough novel information.

The following four tasks allowed the participants to test their approaches to novelty mining. These four tasks were used both in 2003 and 2004.

- **Task 1:** Given the set of documents for the topic, identify all relevant and novel sentences
- **Task 2:** Given the relevant sentences in all documents, identify all novel sentences
- **Task 3:** Given the relevant and novel sentences in the first 5 documents ONLY, find the relevant and novel sentences in the remaining documents
- **Task 4:** Given all relevant sentences from all documents and the novel sentences from the first 5 documents, find the novel sentences in the remaining documents

As the main objective is to identify all novel sentences, this survey focused on task 2. F-score (F) was the primary performance measure that was used to evaluate the systems' novelty mining capabilities in this task.

Table 1: Statistics of benchmark sentence-level data

| Data | Topics | Relevant sentences | Novel sentences | Average novelty (%) |
|---|---|---|---|---|
| TREC2003 | N1-N50 | 15557 | 10226 | 65.7 |
| TREC2004 | N51-N100 | 8343 | 3454 | 41.1 |

## Novelty benchmark datasets

**TREC sentence-level novelty data:** Currently, the possible public datasets for sentence-level novelty mining are from the TREC Novelty Track 2002 to 2004, including three datasets, TREC 2002, 2003 and 2004 Novelty Track data. The TREC 2002 Novelty Track data is usually excluded because it contains too few redundancies and 23 out of the 50 topics consist of all relevant sentences marked as novel. Therefore, a baseline algorithm that marks every sentence as novel will perform almost perfectly for this type of dataset. Moreover, this dataset is unrealistic in the real world, where there are many redundancies in sentences and documents.

TREC 2003 and 2004 Novelty Track data are developed from the AQUAINT collection. The news providers of the document set are Xinghua English (XIE), New York Times (NYT) and Associated Press Worldstream (APW). Both relevant and novel sentences from the National Institute of Standards and Technology (NIST) TREC document set are selected by TREC's assessors. The statistics of these two datasets are summarized in Table 1, where the novelty percentage is the percentage of novel sentences in the dataset.

**APWSJ document-level novelty data:** For document-level novelty mining, the APWSJ novelty data (Zhang *et al*., 2002) is the public dataset that consists of news articles from Associated Press (AP) and Wall Street Journal (WSJ). There are 50 topics from Q101 to Q150 in APWSJ. Previous studies (Zhang and Tsai, 2009a) excluded the 5 topics (Q131, Q142, Q145, Q147, Q150) which lack human redundancy assessments (all documents were novel). The assessors provide two degrees of judgements on non-novel documents, absolute redundant and somewhat redundant. Most experiments adopt the strict definition used by Zhang *et al*. (2002) where only absolute redundant documents are regarded as non-novel. There are 11896 documents on 50 topics. After sentence segmentation, these documents have 319616 sentences in all. The APWSJ data contains a total of 10839 (91.1%) novel documents and 1057(8.9%) non-novel documents.

**TREC 2003/4 document-level novelty data:** For document-level novelty mining, another two datasets were created from the sentence-level TREC 2004/2003 Novelty Tracks, namely document-level TREC 2004 and document-level TREC 2003 (Tang *et al*., 2010; Tsai and Zhang, 2010). In order to obtain the documents, the sentences were first combined into documents according to their document id.

Because we already have the ground truth for the novelty of each TREC sentence, we can easily calculate the NovelRate for each document, which is the actual percentage of novel sentences in that document. If we set a low NovelRate decision point, most documents in this dataset are considered to be novel. This means that this is a dataset with a high percentage of novel documents. When setting different NovelRate decision points, we can observe performances on datasets with different percentages of novel documents.

Table 2: Statistics of BizBlogs document-level data

| Category | # Blogs | # Novel blogs | Novelty (%) |
|---|---|---|---|
| 1 (Product) | 392 | 379 | 96.7 |
| 2 (Company) | 264 | 251 | 95.1 |
| 3 (Marketing) | 267 | 260 | 97.4 |
| 4 (Finance) | 346 | 321 | 92.8 |

**BizBlogs document-level novelty data:** The novelty BizBlogs dataset (Liang *et al*., 2009), a novelty annotated version of the BizBlogs07 (Chen *et al*., 2007) dataset of business blogs, can also be used in novelty mining experiments. The dataset contains 1269 blog posts in four categories: product, company, marketing and finance (Chen *et al*., 2007). The product category deals with description or review of specific company products, as well as other product-related news. The company category deals with news or other information of the corporations, organizations or business. The marketing category includes blogs talking about the marketing, sales and advertising strategies of a company. The finance category relates to financing, funding and loans, financial statements, cash flow and credit information (Chen *et al*., 2007).

Table 2 summarizes the percentage of novel blogs for the four categories as well as the overall blogs. As seen from the table, the majority of the blogs were assessed as novel, which implies that there is little overlap in the content of the blogs for this particular dataset.

## INTELLIGENT NOVELTY MINING TECHNIQUES

Considering the diverse and changing scenario in the real world, this section describes techniques for intelligent novelty mining by bridging the gap between the existing novelty mining methods and user performance requirements. Intelligent novelty mining addresses the domain-specific problem of mining novel information from text data with specific regard to the user context and aims to balance the technical significance and business concerns to create techniques that are useful in real-world scenarios. These techniques aim to adapt to the users' desired level of novel information and human interaction. By addressing the issues of intelligent novelty mining, the techniques are useful from both the technical and business perspectives.

**Metrics and models for novelty scoring:** In a past study, a thorough comparative study was performed on different types of novelty metrics, symmetric (i.e., cosine and Jaccard similarities) and asymmetric metrics (i.e. new word count and overlap) (Tsai *et al.*, 2010b, c). Complementary behavior was observed in the symmetric and asymmetric novelty metrics and a new framework of novelty measurement, a mixture of both types of novelty metrics, was proposed. The experimental results showed the superior performance of this new framework under different performance requirements and for data with different percentages of novel sentences (Tsai *et al.*, 2010c). This method effectively avoided the significant performance drop compared to using individual metrics in either high-recall or high-precision novelty mining algorithms. Furthermore, this new framework is convenient to be applied for the novelty mining algorithms with a series of individual novelty metrics. Moreover, because it does not require any prior information from data, it is very suitable to the real.

Furthermore, another novelty model, document-to-sentence (D2S), was proposed for document-level novelty mining (Tsai and Zhang, 2010). This model can make document-level novelty mining more effective by adopting the techniques for the sentence-level. Experiments on document-level APWSJ data show that D2S can significantly improve the document-level novelty mining performance in terms of redundancy-precision and redundancy-recall, achieving higher redundancy-precision and redundancy-recall than the cosine benchmark. Through experiments on document-level TREC 2004 and TREC 2003 datasets, they also found that it was significantly better to use D2S when the percentage of novel documents is high.

Other studies utilized models for novelty scoring with named entity extraction and Part-of-Speech (POS) tagging (Ng *et al.*, 2007; Zhang and Tsai, 2009b). Ng *et al.* (2007) determined the novelty score of each sentence by using two different metrics, Unique comparison and Importance value. Unique comparison calculated the number of matched entities, whereas Importance value took into account the total weight of matched words that coexisted in both the test and history sentences. The results were promising when compared to the benchmark scores for TREC 2004. Zhang and Tsai (2009b) proposed a new mixed method that treated the sentence novelty score as the novelty score between entities of sentences and the novelty score between other significant words, which can improve the novel sentence detection performance.

**Adaptive novelty decision:** Tang and Tsai (2009) addressed the important problem of setting an adaptive decision point by utilizing the user's feedback over time, which has rarely been addressed in novelty mining. An algorithm was proposed which can be tuned according to different performance requirements varying from high-precision to high-recall, by combining with different optimization criteria.

In the experimental study, the novelty mining algorithm was tested on both document-level and sentence-level novelty mining data. With complete user feedback, the experimental results showed the promising performance of the adaptive decision setting for novelty decision making for a real-time novelty mining algorithm. In order to test the adaptive decision setting on a more practical level, the algorithm was tested with partial feedback. The experimental results indicate that the adaptive decision setting can work robustly with partial feedback from the user. The adaptive decision setting can therefore be employed for realistic situations of varying performance requirements (high-precision/recall) as well as varying degrees of user feedback.

In another set of experiments, the novelty mining algorithm employing the adaptive decision setting algorithm has been tested on the experimental datasets with complete user's feedback on data with low, medium and high novelty ratios (percentage of novel sentences/documents) (Tang *et al.*, 2010). The experimental results show that the adaptive decision setting is very effective in finding the best decision point in the novelty mining algorithm. Therefore, the original algorithm the adaptive decision setting has solved the important problem of adapting to different performance requirements of different topics and users, which is a significant issue for intelligent novelty mining.

**Users' context for novelty feedback:** Recently, there has been increased interests in exploiting contextual information for data mining. Contextual aspects may intervene in several steps of the novelty mining process and it is important to define how these aspects interact within the process and the forms that this can take. During the novelty mining task, context may be used as constraints to allow the development of more efficient algorithms. Context information and domain knowledge may also intervene in the post-processing step to help in the explanation of results.

An important issue in context-aware novelty mining is performance evaluation. It is commonly accepted that the traditional evaluation methodologies used in TREC may not always be suitable for considering the contextual dimensions in the novelty mining process. Indeed, laboratory-based or system oriented evaluation is challenged by the presence of contextual dimensions such as user interaction, profile or environment which significantly impact on the relevance judgments or usefulness ratings made by the end user. Therefore, we need to understand how to overcome the challenge of user-oriented evaluation and to design novel evaluation methodologies and criteria for contextual information retrieval evaluation. For example, in the TREC Novelty Track, the F score was used as the primary evaluation measure (Soboroff, 2004) which actually assumes that the user wants to keep a balance between returning the most novel information (i.e. high-precision case) and not missing any novel information (i.e., high-recall case).

A user's context in novelty mining is defined as his/her specific requirements, which can be one or more of the following:

**Level of novelty:** Different users have different definitions of novel information. For example, a user would regard a sentence with 50% novel information as a novel sentence while another user would only regard a sentence with 80% novel information as a novel sentence. The threshold of novelty scores should be higher for the user with a stricter definition for the novel sentence. As novelty mining is an accumulating system, more training information will be available for threshold setting, based on the user's feedback given over time.

**Performance requirement:** The user's performance requirements can tell us the specific concerns about the novelty mining output. For example, when the user does not want to miss any novel information, a high-recall system which only filters out the most redundant sentences is desired. On the other hand, when the user wants to read the sentences with the most significant novelty in a short time, a high-precision system which only retrieves highly novel sentences is preferred. It has been acknowledged that there is a trade-off between recall and precision; thus, users need to decide on which performance measure to optimize. The trade-off between recall and precision can be adjusted by the novelty threshold. As high-F score is set as the primary performance requirement in previous studies (Soboroff, 2004), based on the assumption that the user wants to keep a balance between high-precision and high-recall. This assumption may not work for the most of the users, but can be used if the user does not know in advance.

**Specific domain interest:** Last but not least, understanding the user's context is very helpful for the system to set the mining target. For example, a user who is interested in the financial news may be sensitive to different numbers, currency symbols, etc. To make use of the user's contextual information, we have to answer several questions, i.e. what contextual information can be acquired in novelty mining, how to make automatic annotation about the context and how to design the user interface for the user's context.

Tang *et al*. (2009) addressed the problem of accessing the user's context in novelty mining via novelty feedback, an important problem that has not been addressed by traditional novelty mining algorithms, which regard novelty as the opposite of similarity. The technique, contextual information annotation of named entities algorithm, was proposed to solve the issue of detecting novelty in highly similar sentences and was developed to interact with the user's novelty feedback. The algorithm is useful for users interested in detecting changes in number, location, person, or time and can be generalized for other named entities as well. The linguistic-level information, i.e., named entities, was used to accommodate the user's context. In the algorithm, the user is allowed to feedback his/her preferred significant information to the algorithm directly. The experiments on different data showed that the method can solve the important problem of mining the novel information in highly similar sentences at the lingui.

**Summary of intelligent novelty mining:** In relation to previous studies, intelligent novelty mining techniques are more suitable for real-world novelty mining. Other techniques cannot adapt to different users who have different definitions of novel information, whereas the adaptive techniques can. Intelligent techniques can also address the differences in users' performance requirements, whether they require a high-recall system which only filters out the most redundant sentences is desired, a high-precision system which only retrieves highly novel sentences, or a combination of both. In addition, the techniques make use of the user's contextual information in novelty mining, which was not addressed in previous work. In summary, the intelligent techniques described in this study are able to leverage the use of novelty metrics, novelty decision and novelty feedback to improve the results of mining new information from text data to overcome the challenge of information overload.

## CONCLUSIONS

This study reviewed intelligent novelty mining techniques for alleviating the information overload problem. Intelligent novelty mining addresses the domain-specific problem of mining novel information from text data with specific regard to the user context. Intelligent novelty mining aims to balance the technical significance and business concerns to create techniques that are useful in real-world scenarios. The techniques can leverage the use of novelty metrics, novelty decision and novelty feedback to improve the results of mining new information from text data.

## REFERENCES

Allan, J., C. Wade and A. Bolivar, 2003. Retrieval and novelty detection at the sentence level. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, July 28-Aug. 01, Toronto, Canada, pp: 314-321.

Chen, Y., F.S. Tsai and K.L. Chan, 2007. Blog search and mining in the business domain. Proceedings of the International Workshop on Domain Driven Data Mining, San Jose, California, Aug. 12, ACM, New York, USA., pp: 55-60.

Harman, D., 2002. Overview of the TREC 2002 novelty track. Proceedings of the 11th Text Retrieval Conference, (TREC'02), National Institute of Standards and Technology, Gaithersburg, pp: 46-55.

Hoffmann, H., 2007. Kernel PCA for novelty detection. Pattern Recogn., 40: 863-874.

Kwee, A.T. and F.S. Tsai, 2009. Mobile novelty mining. Int. J. Adv. Pervasive Ubiquitous Comput., 1: 43-68.

Kwee, A.T., F.S. Tsai and W. Tang, 2009. Sentence-Level Novelty Detection in English and Malay. In: Advances in Knowledge Discovery and Data Mining, Theeramunkong, T. *et al.* (Eds.). Springer, Berlin, Heidelberg, pp: 40-51.

Liang, H., F.S. Tsai and A.T. Kwee, 2009. Detecting novel business blogs. Proceedings of the 7th International Conference on Information, Communications and Signal Processing, Dec. 8-10, Macau, pp: 1-5.

Markou, M. and S. Singh, 2003. Novelty detection: A review-part 1: Statistical approaches. Signal Process., 83: 2481-2497.

Ng, K.W., F.S. Tsai, L. Chen and K.C. Goh, 2007. Novelty detection for text documents using named entity recognition. Proceedings of the 6th International Conference on Information, Communications and Signal Processing, Dec. 10-13, Singapore, pp: 1-5.

Ong, C.L., A.T. Kwee and F.S. Tsai, 2009. Database optimization for novelty detection. Proceedings of the 7th International Conference on Information, Communications and Signal Processing, Dec. 8-10, Macau, pp: 1-5.

Soboroff, I. and D. Harman, 2003. Overview of the TREC 2003 novelty track. Proceedings of the 12th Text Retrieval Conference, (TREC'03), National Institute of Standards and Technology. Gaithersburg, MD, pp: 38-53.

Soboroff, I., 2004. Overview of the TREC 2004 novelty track. Proceedings of the 13th Text Retrieval Conference, (TREC'04), National Institute of Standards and Technology, Gaithersburg, pp: 1-16.

Tang, W. and F.S. Tsai, 2009. Threshold setting and performance monitoring for novel text mining. Proceedings in Applied Mathematics 3 Society for Industrial and Applied Mathematics-9th SIAM International Conference on Data Mining, Jan. 04-06, New York, pp: 1310-1319.

Tang, W., A.T. Kwee and F.S. Tsai, 2009. Accessing contextual information for interactive novelty detection. Proceedings of the European Conference on Information Retrieval (ECIR) Workshop on Contextual Information Access, Seeking and RetrievalEvaluation. http://www.irit.fr/CIRSE09/000_ECIR%202009%20Workshop_proceedings.pdf.

Tang, W., F.S. Tsai and L. Chen, 2010. Blended metrics for novel sentence mining. Expert Syst. Appl., 37: 5172-5177.

Tsai, F.S. and K.L. Chan, 2007. Detecting Cyber Security Threats in Weblogs Using Probabilistic Models. In: Intelligence and Security Informatics, Yang, C.C. *et al.* (Eds.). Vol. 4430. Springer, Berlin, Heidelberg, ISBN: 978-3-540-71548-1, pp: 46-57.

Tsai, F.S., W. Han, J. Xu and H.C. Chua, 2009. Design and development of a mobile peer-to-peer social networking application. Expert Syst. Appl., 36: 11077-11087.

Tsai, F.S. and K.L. Chan, 2010. Redundancy and novelty mining in the business blogosphere. The Learning Organization, Vol. 17.

Tsai, F.S. and Y. Zhang, 2010. D2S: Document to sentence framework for novelty detection. Knowl. Inform. Syst.

Tsai, F.S., M. Etoh, X. Xie, W.C. Lee and Q. Yang, 2010a. Introduction to mobile information retrieval. IEEE Intell. Syst., 25: 11-15.

Tsai, F.S., A.T. Kwee, W. Tang and K.L. Chan, 2010b. Adaptable services for novelty mining. Int. J. Syst. Service-Oriented Eng., Vol. 1.

Tsai, F.S., W. Tang and K.L. Chan, 2010c. Evaluation of metrics for sentence-level novelty mining. Inform. Sci., 180: 2359-2374.

Yee, K.Y., A.W. Tiong, F.S. Tsai and R. Kanagasabai, 2009. OntoMobiLe: A generic ontology-centric service-oriented architecture for mobile learning. Proceedings of the 10th International Conference on Mobile Data Management (MDM) Workshop on Mobile Media Retrieval (MMR), May 18-20, Taiwan, pp: 631-636.

Zhang, Y., J. Callan and T. Minka, 2002. Novelty and redundancy detection in adaptive filtering. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR'02), Tampere, Finland, pp: 81-88.

Zhang, Y. and F.S. Tsai, 2009a. Chinese novelty mining. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, (EMNLP'09), Singapore, pp: 1561-1570.

Zhang, Y. and F.S. Tsai, 2009b. Combining named entities and tags for novel sentence detection. Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval, (ESAIR'09), Barcelona, Spain, pp: 30-34.

Zhang, Y., A.T. Kwee and F.S. Tsai, 2010. Multilingual sentence categorization and novelty mining. Inform. Process. Manage., 10.1016/j.ipm.2010.02.003