Techniques for intelligent novelty mining

Tsai, Flora S.

2010

Tsai, F. S. (2010). Techniques for intelligent novelty mining. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/46995>

<https://doi.org/10.32657/10356/46995>

Nanyang Technological University

Downloaded on 26 Apr 2023 15:59:37 SGT

2 *
244

TECHNIQUES FOR INTELLIGENT NOVELTY MINING

FLORA S. TSAI

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2010

Acknowledgement

I wish to thank my supervisor, Associate Professor Kap Luk Chan, for his guidance and support throughout the course of my work. His openness in allowing me the freedom to choose and expand the research topic is greatly appreciated. He was very patient with me throughout the course of my study, and his supervision and recommendations helped to significantly improve the contributions of the thesis.

I am also grateful for the support of the present and past management of my school for providing me the opportunity to further my studies while simultaneously maintaining a teaching career: Associate Professor Yap-Peng Tan, Professor Chan Hin Kam, Professor Alex Kot and Professor Meng Hwa Er.

Thanks also goes to numerous colleagues and friends for many helpful and insightful discussions, which have given me the motivation and confidence to attempt this endeavor. Finally, I am indebted to my family, who have given me continued care and support throughout the years, and without whom I would not be able to complete this endeavor at this point in my career.

Contents

Acknowledgement	i
Table of Contents	iii
Summary	vii
List of Figures	ix
List of Tables	xiv
1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Objectives	3
1.4 Major Contributions	4
1.5 Organization of Thesis	6
2 Literature Review	8
2.1 Introduction	8

<i>Contents</i>	<i>iv</i>
2.2 Novelty Mining Framework	9
2.2.1 Preprocessing	10
2.2.2 Categorization	14
2.3 Novelty Mining	19
2.3.1 Novelty Scoring	21
2.3.2 Novelty Decision Setting	22
2.3.3 Performance Evaluation	23
2.3.4 Sentence-Level Novelty Mining	25
2.3.5 Document-Level Novelty Mining	33
2.3.6 Summary of Previous Work	34
2.4 Chapter Summary	34
3 Metrics and Models for Novelty Scoring	39
3.1 Introduction	39
3.2 Novelty Metrics	42
3.2.1 Symmetric Metrics	42
3.2.2 Asymmetric Metrics	44
3.2.3 Factors Affecting Novelty Metric Selection	45
3.3 Blended Metrics for Novelty Measurement	48
3.4 Document-to-Sentence Model	49
3.5 Experiments and results	53
3.5.1 Novelty Datasets	53

<i>Contents</i>	<i>v</i>
3.5.2 Evaluation Measures	56
3.5.3 Metrics Comparison Under Various Situations	56
3.5.4 Blended Metrics vs. Individual Metrics	64
3.5.5 Evaluation of Metrics on Business Blogs	71
3.5.6 Effects of Weight Variation on Blended Metrics	73
3.5.7 Comparison of D-NM with D2S	74
3.5.8 Comparison of D-NM with D2S using Blended Metrics	76
3.5.9 Usage of D2S with Different Percentages of Novel Documents	76
3.6 Chapter Summary	81
4 A Learning Model for Adaptive Novelty Decision	83
4.1 Introduction	83
4.2 Adaptive Novelty Decision in Novelty Mining	85
4.2.1 Motivations	85
4.2.2 Learning-Based Adaptive Novelty Decision (LAND)	86
4.2.3 Implementation Issues	92
4.3 Experimental Study	93
4.3.1 Datasets	93
4.3.2 Results	93
4.3.3 Working Example	101
4.4 Chapter Summary	104

<i>Contents</i>	vi
5 Contextual Information Annotations for Novelty Feedback	115
5.1 Introduction	115
5.2 Definition of Context	117
5.3 Novelty Feedback	120
5.4 A Contextual Annotation-based Algorithm for Novelty Feedback	122
5.4.1 Basic Idea for CIAN	122
5.5 Experimental Study	126
5.5.1 Results	128
5.6 Chapter Summary	134
6 Conclusions and Future Research	138
6.1 Conclusions	138
6.2 Relation to Previous Work	142
6.3 Suggestions for Future Research	143
A Novelty Mining Application	147
A.1 Graphical User Interface	147
A.1.1 Categorization	149
A.1.2 Document-Level Novelty Mining	151
A.1.3 Sentence-Level Novelty Mining	156
B Author's Publications	158
Bibliography	163

Summary

Intelligent novelty mining addresses the domain-specific problem of mining novel information from text data with specific regard to the user context. Traditional novelty mining focuses on optimizing the retrieval of novel information with a predefined evaluation measure. This tends not be useful in realistic situations where a fixed unit of measure is unable to adapt to the performance requirements of different users. Intelligent novelty mining, on the other hand, aims to balance the technical significance and business concerns to create techniques that are useful in real-world scenarios. This thesis addressed the problem of intelligent novelty mining, which is a big challenge in the data mining and business community. The proposed techniques aim to leverage the use of novelty metrics, novelty decision, and novelty feedback to improve the results of mining new information from text data.

First, an original framework is proposed for measuring the novelty by a mixture of novelty metrics (BLEND) that can integrate the merits of different types of metrics and hence generalize better over different data. This new framework of novelty measurement is important because it is adaptable to different performance requirements and for data with different percentages of novel sentences, which was not previously addressed. The experimental results show that the blended metric not only enhances the overall performance of real-time novelty mining, but also makes the novelty mining more flexible to different performance requirements. Furthermore, a novel Document-To-Sentence (D2S) algorithm is proposed for document-level novelty mining, which segments a document into sentences, determines the novelty of each sentence, then computes the document-level novelty score based on a fixed decision point. This novel

method achieves significant improvement on the experimental document-level data.

Second, a new learning model for adaptive novelty decision (LAND) is proposed specifically for novelty mining. It is a score distribution-based novelty decision method, and models the score distributions of both novel and non-novel documents. The score distributions of both classes provide the global information of data, from which an optimization criterion can be constructed for searching the optimal novelty decision. The experimental results indicate that the algorithm not only generalizes well on both document-level and sentence-level novelty mining, but also performs robustly in realistic situations where only partial feedback is available.

Finally, an original method for learning contextual information in novelty feedback (CIAN) is proposed, which retrieves the novel information missed by the original novelty mining based on the annotation of named entities (such as number, time, location, person). This algorithm solves the important problem of novelty mining for highly similar sentences, and outperforms the original novelty mining in the range of recall from 0.4 to 1. This indicates that annotation-based novelty mining is important for mining the linguistic-level novelty in highly similar sentences.

List of Figures

2.1	Overview of a Novelty Mining Framework.	10
2.2	Preprocessing Steps	11
2.3	Novelty Mining Process.	20
2.4	Overview of Pattern-based Approach.	32
3.1	Novelty Mining Process of D2S	51
3.2	A New Model for Novelty Mining.	52
3.3	Performance of various novelty metrics for the TREC 2003 and 2004 Novelty Track data. Dashed lines denote asymmetric metrics and solid lines denote symmetric metrics.	58
3.4	Performance of various novelty metrics for the TREC 2003 and 2004 Novelty Track data. Dashed lines denote asymmetric metrics and solid lines denote symmetric metrics.	62
3.5	Blended metrics vs individual metrics for all 100 topics in high-precision (a) and high-recall (b) system in TREC 2003 and 2004 Novelty Track data	64

3.6	Blended metric M3 vs individual metrics for the topics with low (a-b), medium (c-d) and high (e-f) novelty ratio in TREC 2003 and 2004 Novelty Track data	66
3.7	Average best F scores (above) and the corresponding number of topics (below) in 10 different levels of ratios of novel sentences in TREC 2003 and TREC 2004 Novelty Track data.	68
3.8	Precision-Recall-F (PRF) curve	71
3.9	Redundancy-PRF curve	72
3.10	Effect of weight variation ($\alpha=0.1, 0.2$ to 0.9) on the performance of blended metrics on 100 topics from TREC 2003 and 2004 Novelty Track data.	73
3.11	R-PR curves for document-level NM on APWSJ: compare D-NM and D2S. The grey dashed lines show contours at intervals of 0.1 points of RF	75
3.12	R-PR curves for document-level NM on TIPSTER: compare D2S and D-NM using blended metrics.	77
3.13	PR curves for document-level NM on document-level TREC 2004: compare D2S and D-NM.	78
3.14	PR curves for document-level NM on document-level TREC 2003: compare D2S and D-NM.	79
3.15	R-PR curves for document-level NM on document-level TREC 2004: compare D2S and D-NM.	80
3.16	R-PR curves for document-level NM on document-level TREC 2003: compare D2S and D-NM.	80

4.1	Empirical and probability distribution approximation for TREC 2004 Novelty Track data topic N54.	88
4.2	Empirical and probability distribution approximation for TREC 2004 Novelty Track data topic N69.	89
4.3	Novelty mining combined with LAND	107
4.4	Sentence-level novelty mining on TREC 2004 Novelty Track data: <i>PR</i> curves of novelty mining with fixed decision point, simple decision and adaptive decision by LAND (tuned for F_β) with complete user's feedback.	108
4.5	Document-level novelty mining on APWSJ data: <i>R-PR</i> curves of novelty mining with fixed decision point, simple decision point and adaptive decision point by LAND (tuned for RF_β) with complete user's feedback.	108
4.6	Sentence-level novelty mining on TREC 2004 Novelty Track data: <i>PR</i> curves of LAND (tuned for F_β) with different percentages of the user's feedback.	109
4.7	Document-level novelty mining on APWSJ data: <i>R-PR</i> curves of LAND (tuned for RF_β) with different percentages of the user's feedback.	109
4.8	Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_β) with complete user's feedback on document-level TREC 2003 Novelty Track data (with PNS decision point 0.25).	110

<i>List of Figures</i>	xii
4.9 Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_β) with complete user's feedback on document-level TREC 2004 Novelty Track data (with PNS decision point 0.03).	111
4.10 Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_β) with complete user's feedback on document-level TREC 2004 Novelty Track data (with PNS decision point 0.5).	112
4.11 Sentence-level novelty mining results for TREC03 topic N39.	113
4.12 Decision adjustment to 0.6000 for Sentence 16 after running LAND.	113
4.13 Sentence-level novelty mining results for TREC03 topic N39 after running LAND.	114
5.1 Use Case of Novelty Feedback	122
5.2 Snapshot of GUI for additional contextual information	122
5.3 A contextual annotation-based novelty mining algorithm with the user feedback of "number" and "location".	123
5.4 Framework of the CIAN algorithm using one-to-all sentence comparison	126
5.5 CIAN using one-to-one vs. one-to-all, $\delta = 0.1$	129
5.6 CIAN using one-to-one vs. one-to-all, $\delta = 0.2$	130
5.7 CIAN using one-to-one vs. one-to-all, $\delta = 0.4$	131
5.8 Novelty feedback: "number"	132
5.9 Novelty feedback: "location"	133

<i>List of Figures</i>	xiii
5.10 Novelty feedback: “person”	134
5.11 Novelty feedback: “time”	135
5.12 Precision-Recall curve for the H1N1 swine flu topic.	136
5.13 Novelty results using contextual annotations of new number. . .	136
A.1 Main window of our novelty mining system.	149
A.2 Options for categorization.	150
A.3 Options for document-level novelty mining.	151
A.4 Options for sentence-level novelty mining.	157

List of Tables

2.1	Term Weighting Schemes of High Efficiency [87]	13
2.2	SVD Components within LSA Context [9]	19
2.3	Word patterns for five types of NE-questions (from [37])	36
2.4	Average novelty precision of ip-BAND and other baseline methods [37]	36
2.5	Measure for redundancy in [95]	37
2.6	Average performance of different redundancy measure with a simple decision algorithm, measured on 33 topics with AP News & Wall Street Journal dataset. Only <i>absolutely redundant</i> documents are treated as redundant	37
2.7	Summary for the previous work in novelty mining	38
3.1	Symmetric vs. Asymmetric Metrics	42
3.2	Explicit and Implicit Factors Of Novelty Metric Selection	46
3.3	Statistics of experimental data.	53
3.4	Statistics of Experimental Data	54
3.5	Summary of the novelty ratios	55

3.6	Top two novelty metrics of high-recall, high-precision and high- F -score novelty mining algorithm for the TREC 2003 and 2004 Novelty Track data. * denotes the asymmetric metrics.	59
3.7	Number of topics with low, medium and high novelty ratio in both sentence- and document-level data.	61
3.8	Top two novelty metrics for various performance requirements and novelty ratios for the TREC 2003 and 2004 Novelty Track data. * denotes the asymmetric metrics.	64
3.9	Average best F scores at different novelty ratios r_N . $\blacktriangle$ denotes the weak zones of novelty metrics.	67
3.10	Average recall and precision corresponding to the best F score (as reported in Table 3.9) at different novelty ratios r_N	69
3.11	Average number of retrieved sentences at different novelty ratios r_N	69
3.12	Percentage of Overlap on the Sentence Level	79
4.1	Contingency table in novelty mining.	89
4.2	Comparison of performance evaluated by F_β ($\beta = 0.2, 0.5, 0.8$) on TREC 2004 Novelty Track data.	96
4.3	Comparison of performance evaluated by F_β ($\beta = 0.2, 0.5, 0.8$) on APWSJ data.	97
4.4	Performance of LAND with different percentages of the user's feedback, $P_{\text{feedback}} = 100\%, 80\%, 50\%, 20\%$ and 10% , on TREC 2004 Novelty Track data and APWSJ data.	99

<i>List of Tables</i>	xvi
5.1 Statistics of TREC 2004 Novelty Track Data	127

Chapter 1

Introduction

1.1 Background

With the abundance of information and news articles released frequently, many of the articles may contain similar information. Thus, there is a need for identifying novel and relevant information out of a mass of incoming text documents. Novel information in this case refers to text that contain fresh content and *novelty mining*, or novelty detection, is the process of singling out novel information from a given set of text documents. Through this process, users can save time by reading only the novel information while the repeated information is filtered away. Novelty mining is an increasingly important problem because of the rising issue of information overload, where manual filtering of information by humans will eventually be replaced by automatic detection by computers. Thus, intelligent, automated, and adaptable techniques are critical to solve the important problem of novelty mining.

Novelty mining has been performed at three different levels: event level, sentence level and document level [36]. Research on novelty mining at the event

level arises from the Topic Detection and Tracking (TDT) research, which is concerned with online new event detection/first story detection [3, 11, 21, 60, 83, 84]. Previous works on the sentence and the document level novelty mining tend to apply some promising content-oriented techniques. Other studies utilize ontological knowledge, especially taxonomy, such as WordNet [5, 21, 28, 47, 95], synonym dictionary [21], HowNet [20], etc.

1.2 Motivation

Many studies related to novelty mining originated from the TREC Novelty Track, where 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 [58]. Recent studies suggested several ways to improve the performance of novelty mining by integrating various natural language processing (NLP) techniques, such as part-of-speech (POS) tagging, named entity recognition (NER), etc [37, 43, 91, 96]. These techniques facilitate the usage of additional contextual information, beyond the simple strategy using bag-of-words models in novelty mining.

Considering the diverse and changing scenario in the real world, this thesis proposes 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 proposed techniques are useful from both the technical and business perspectives.

1.3 Objectives

The objective of this thesis is to develop new and effective models for intelligent novelty mining using novelty metrics, novelty decision, and novelty feedback for detecting novel sentences and documents. The proposed models should improve the performance, predictive accuracy, and adaptability in novelty mining. In particular, three problems are targeted in this thesis. The first problem is to explore the models and metrics for novelty scoring. The objectives are to describe the different types of novelty metrics as well as the advantages and disadvantages of each. A thorough comparative study will be performed on different types of novelty metrics. A new framework of novelty measurement, a mixture of both types of novelty metrics, will be introduced. The goal of the blended metric is to integrate the merits of both metrics and hence generalize better over different topics. Furthermore, another novelty model is proposed which can make document-level novelty mining more effective by adopting the techniques for the sentence level. The second problem the thesis will address is of setting an adaptive decision point by utilizing the user's feedback over time. The goal is to develop a suitable adaptive novelty decision algorithm that can be tuned according to different performance requirements varying from high-precision to high-recall, by combining with different optimization criteria, which is very important for novelty mining. The final objective of the thesis will address 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 proposed algorithm will allow for novelty mining of highly similar sentences, using contextual information specified by the user.

In summary, intelligent novelty mining techniques that can extract novel information from text data with specific regard to the user context are extremely useful in real-world scenarios. This thesis analyzed the problem of intelligent novelty mining, which is a big challenge in the data mining and business community. The objective of the research work in this thesis aims to improve the performance of intelligent novelty mining by tackling the issues in novelty scoring, novelty decision, and novelty feedback.

1.4 Major Contributions

The major contributions of this research are summarized below.

1. Metrics and Models for Novelty Scoring

A new framework is proposed for measuring the novelty by a mixture of both symmetric and asymmetric novelty metrics. The goal of the blended metric is to integrate the merits of both metrics and hence generalize better over different topics. This new framework of novelty measurement is adaptable to different performance requirements and for data with different percentages of novel sentences/documents. Since it does not require any prior information from data, blended metrics are suitable for real-time novelty mining where little knowledge of the data is known beforehand. Furthermore, a new algorithm is proposed that integrates sentence-level

novelty mining with document-level novelty mining. The goal of the integrated technique is to improve the performance of detecting novel documents by utilizing sentence-level information. This original technique has never been proposed for novelty mining, and is superior to traditional document-level novelty mining if the novelty ratio (percentage of novel documents) is high. In addition, this algorithm shows a good capability to find redundant documents across different novel-ratio decision points, which is important for the adaptability of novelty mining.

2. Learning Models for Adaptive Novelty Decision

A novel learning model for adaptive novelty decision is proposed, which solves the problem of setting novelty decision points according to performance requirements which may vary from high precision to high recall. This problem is important to accommodate for the varying performance requirements of different users. The algorithm can be tuned automatically for different topics under different performance requirements. As a score distribution-based decision setting method, it models the score distributions of both novel and non-novel data. The score distributions of both classes provide the global information of data, from which an optimization criterion can be constructed for searching the optimal novelty decision value.

3. Learning Contextual Information for Novelty Feedback

Traditional novelty mining methods cannot detect novelty in a sentence that looks similar to its history sentences, but contains information that is useful in the user's current context, such as an analyst interested in price fluctuations of oil. An original method is proposed to accommodate the user's context in novelty mining by using the linguistic-level information,

i.e. named entities. This annotation-based novelty mining algorithm utilizes the acquired contextual information from novelty feedback, and is significantly useful for mining the linguistic-level novelty in highly similar sentences.

1.5 Organization of Thesis

This thesis is organized as follows:

- Chapter 1 begins with a general introduction that includes background, motivation, objectives, and contributions of the thesis.
- Chapter 2 reviews existing novelty mining techniques at both the sentence and document level.
- Chapter 3 describes metrics and models for novelty scoring. First, different novelty metrics are evaluated, then a framework is proposed for measuring the novelty by a mixture of novelty metrics, and the results are compared with other novelty metrics on experiments using real data. Another novelty model, the Document-to-Sentence technique is described, and experimental results of the proposed model are presented.
- Chapter 4 describes a new learning model for automatic novelty decision, and compares the results with those achieved using fixed and simple decision points.
- Chapter 5 proposes an annotation-based feedback developed specifically for novelty mining. The proposed algorithm is compared with the baseline in experimental studies with sentence-level data.

- Lastly, conclusions and recommendations for future work are presented in Chapter 6.
- All the techniques described in this thesis have been implemented in a real novelty mining application, the details of which are described in Appendix A.

Chapter 2

Literature Review

2.1 Introduction

Nowadays, a multitude of information such as news articles [22], blogs [13], and social networks [75] are readily available on the Internet. This leads to the problem of information overload [69] due to the large quantity of irrelevant and redundant information contained in these documents [32]. 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. Novelty mining can be used to solve many business problems, such as in corporate intelligence [73, 74], cyber security [68, 70, 66], and Web services [71, 85]. The pioneering work of novelty mining was proposed by Zhang et al. at the document level [95]. In their work, “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. In practice, however, it may be less meaningful to merely return novel documents for the end-user because there are still a lot of irrelevant, redundant information

in these documents [76]. Therefore, most of the later contributions have been made to sentence-level novelty mining [33, 34, 43, 77, 96, 97, 80]

Many studies related to novelty mining originated from the TREC Novelty Tracks [24, 58, 59]. The novelty track was introduced in the eleventh Text REtrieval Conference (TREC) in 2002 [24]. This track was designed to investigate systems' abilities in locating new and relevant information from a given document set which is categorized into topics. Other works on the sentence and the document level novelty mining tend to apply some promising content-oriented techniques [2, 3, 21, 36, 60, 83, 84].

Recent studies suggested several ways to improve the performance of novelty mining by integrating various natural language processing (NLP) techniques, such as part-of-speech (POS) tagging [4, 60], named entity recognition (NER) [7, 11, 26, 37, 91, 93, 96, 99], WordNet [5, 21, 28, 43, 47, 95], synonym dictionary [21, 43], HowNet [20], etc. These NLP techniques facilitate the usage of additional contextual information, beyond the simple strategy using bag-of-words models in novelty mining.

This chapter reviews techniques for novelty mining at both the sentence and document level. Section 2.2 gives the overview of a novelty mining framework, and describes the basic steps in preprocessing and categorization. Section 2.3 describes the process of novelty mining and reviews current approaches for novelty mining. Finally, this chapter is summarized in Section 2.4.

2.2 Novelty Mining Framework

A novelty mining framework includes three main parts, i.e. (i) preprocessing, (ii) categorization, and (iii) novelty mining, as shown in the Figure 2.1. Prepro-

Figure 2.1: Overview of a Novelty Mining Framework.

cessing is used to prepare a clean data matrix (term document/sentence matrix) which can be easily processed by a computer. Categorization classifies each incoming document or sentence (later, we only refer to documents without loss of generalization) into its relevant topic bin. Then, the novelty mining algorithm detects whether a new document is novel based on its history documents.

Before describing the details of novelty mining, the next sections briefly describe the steps in preprocessing and categorization.

2.2.1 Preprocessing

Before performing categorization and novelty mining, the documents must first be preprocessed.

The entire preprocessing steps can be seen in Figure 3.1.

Figure 2.2: Preprocessing Steps

Stop Words Removal

The first step for preprocessing is stripping all the formatting (i.e. capitalization and punctuation) and extraneous markup (i.e. dateline), from all documents, forming a term list from the documents. A stop words list is used as a filter to eliminate a large number of the commonly used English words, such as “the”, “to”, “have”, from the obtained term list, hence reducing much noise in the collection.

Word Stemming

After stop words removal, the remaining words are then stemmed, which further removes common endings from words. Stemming, by definition, is a process of reducing the inflected (or sometimes derived) words to their root forms. For example, words ‘opens’ and ‘opened’ become ‘open’ after being stemmed.

The Porter stemming algorithm [46] is used for word stemming. This al-

gorithm removes the commoner morphological and inflexional endings from the English words.

Term-Document Matrix (TDM) Generation

From the output of the preprocessing steps, a bag of words is obtained. The term-document matrix (TDM) converts the text into the numerical representation as the input for respective algorithms [9]. A TDM is generated by arranging the list of all content words along the vertical axis, and a similar list of all documents along the horizontal axis, in any order. The TDM is then filled in with the frequency of occurrence of each word in each corresponding document. Since each document only contains a portion of words of the whole content word list, the resultant matrix is very sparse with the existence of many zeros. Although this simple version of TDM can give acceptable results, the performances can be further improved by applying the term weighting schemes and document length normalization techniques.

Under the application of term weighting schemes, the values of a cell in the raw TDM can be transformed so that the frequency of each term in a document is recalculated. This transformation include two components: a global weighting and a local weighting [18]. Each element, a_{ij} , of a TDM can be expressed as

$$a_{ij} = l_{ij}g_in_{ij} \quad (2.1)$$

where l_{ij} is a local factor that measures the importance of term i in document j , g_i is a global factor that measures the importance of term i in the entire collection, and n_{ij} is a normalization factor.

Studies were done on examining the efficiency of various term weight-

ing schemes [14, 18, 30, 87]. Two global weighting and three local weighting schemes, which are found to give relatively high efficiency, are listed in Table 2.1. Symbol f_{ij} denotes term frequency, i.e., the number of times term i appears in document j . The probability of term i appearing in the document collection is defined as:

$$p_{ij} = \frac{f_{ij}}{\sum_k f_{ik}} \quad (2.2)$$

and the binary function is defined as:

$$b(f_{ij}) = \begin{cases} 1, & \text{if } f_{ij} \neq 0 \\ 0, & \text{if } f_{ij} = 0 \end{cases} \quad (2.3)$$

Table 2.1: Term Weighting Schemes of High Efficiency [87]

Symbol	Name	Type
	<i>Local term-weighting</i> (l_{ij})	
tf	Term frequency	f_{ij}
lo	Logarithmic	$\log_2(1 + f_{ij})$
al	Alternate log	$b(f_{ij})(1 + \log_2 f_{ij})$
	<i>Global term-weighting</i> (g_{ij})	
EN	Entropy	$1 + \sum_j (p_{ij} \log_2(p_{ij})) / \log_2 n$
IDF	Inverse document frequency	$\log_2(n / \sum_j b(f_{ij}))$

Normalization penalizes the term weights for a document in accordance with its length to fairly retrieve documents of all lengths [57]. The improved pivoted cosine normalization scheme, which reduces the gap between the relevance and the retrieval probability (retrieved using the normalization scheme), was proposed and implemented in a Matlab toolbox for term-document matrix generation [87], and it is denoted as:

$$(\sum_j (g_i l_{ij})^2)^{-1/2} \quad (2.4)$$

This formula replaces the cosine normalization and is widely used in the information retrieval domain.

Sentence Segmentation

For sentence-level novelty mining, the steps for generating the term-sentence matrix (TSM) is exactly the same as the term-document matrix; thus, the description is not repeated. However, before the generation of the TSM, the documents need to first be segmented into sentences. Sentence segmentation is the problem of dividing a string of written language into its component sentences. In English, using punctuation, particularly the full stop character is a reasonable approximation. However this problem is not trivial due to the use of the full stop character for abbreviations, which may or may not also terminate a sentence. For example, “Dr.” is not its own sentence in “Dr. Lee went to the shops in Main Street.” For sentence segmentation, Sentence-boundary [41] was found to have a lower error rate than Sentseg [29], and is therefore adopted as the segmenter in the experiments [79].

2.2.2 Categorization

The objective of categorization is purely to filter documents into topics that a user wishes to track for novel information [54]. Inputs for training the module would be topical bins containing relevant documents to that topic as well as user input in the form of keywords. Once categorization is completed, novelty mining would be carried out on some of the bins [98]. The rationale behind

categorization is that events belonging to the same topic often share a set of keywords. Those keywords are informative for discriminating relevant (on-topic) and non-relevant (off-topic) documents [84]. In text categorization, there are three major subtasks: (1) batch filtering, (2) routing, and (3) adaptive filtering. Batch filtering and routing are more traditional machine learning tasks which begin with a large sample of evaluation training documents; while adaptive filtering begins with only a topic statement (key words) and a very small number of relevant documents [15, 50]. Clearly, the adaptive filtering subtask is quite suitable to the requirements of categorization with an accumulating documents stream. Next, the details of adaptive filtering are described.

Adaptive filtering starts with a set of topics and a document stream, and a small number of examples of relevant documents. For each topic, it must create an initial filtering profile and then analyze the incoming documents and make a binary decision to retrieve or not to retrieve each document on a profile by profile basis. If a document is retrieved for a given profile and a relevance judgment exists for that document/topic pair, this judgment is provided and the information can be used to update the profile. This step is designed to simulate interactive user feedback. Relevance judgments cannot be used for documents which are not retrieved by the appropriate filtering profile or for any other topics. Documents must be processed in the specified time order; once a document has been selected or rejected for a specific topic it is no longer available for subsequent reconsideration for that topic. (The accumulating collection of processed documents may however be used for such information as term statistics or score distributions.) The final output is an unranked set of documents for each topic.

Rocchio Algorithm

A widely used profile updating method is the incremental Rocchio algorithm [51], which is a query modification process that has been extensively investigated in the literature and used in information retrieval and novelty mining [23, 51, 80, 98]. Relevance feedback improves the query with terms that are considered relevant to the information seeking. This is done iteratively and automatically by selecting a predefined set of the top retrieved sentences as relevant (pseudo-relevance feedback). The aim is to find the optimum query, which can maximize the similarity with relevant sentences while minimizing similarity with non-relevant sentences. If $R(\bar{R})$ is the set of relevant(non-relevant) sentences, then a query, Q_{opt} , can be defined such that:

$$Q_{opt} = \operatorname{argmax}_q (sim(q, R) - sim(q, \bar{R})) \quad (2.5)$$

where $sim(q, R)$ denotes the similarity of the query to the set of relevant sentences. Using cosine as the similarity measure:

$$\vec{Q}_{opt} = \frac{1}{|R|} \sum_{s_j \in R} \vec{s}_j - \frac{1}{|\bar{R}|} \sum_{s_j \in \bar{R}} \vec{s}_j \quad (2.6)$$

where the optimum query is a vector defined by the difference of the centroids of the relevant and non-relevant sentences. Then, the similarity will be calculated on each sentence with the constructed query of each topic. If the similarity between the sentence and the query is above the decision point, the sentence is considered as relevant to this topic. The definition of similarity is:

$$\cos(s_t, q_i) = \frac{\sum_{k=1}^n w_k(s_t)w_k(q_i)}{\|s_t\| \cdot \|q_i\|} \quad (2.7)$$

where $w_k(s)$ is the weight of k^{th} element in the sentence vector s and $w_k(q)$ is the weight of k^{th} element in the query vector q .

The Rocchio algorithm is popular for two reasons. First, it is computationally efficient for online learning. Secondly, compared to many other algorithms, it works well empirically, especially at the beginning stage of adaptive filtering where the number of training examples is very small.

Latent Semantic Analysis

Another method for categorization is latent semantic analysis (LSA), also known as Latent semantic indexing (LSI), is an automatic indexing and retrieval technique originally designed for improved detection of relevant documents on the basis of search queries [17]. The implicit higher-order structure of terms with documents is employed in the technique, hence the term “semantic structure”. It applies the singular value decomposition (SVD) method to project the document vectors into a lower dimensional subspace, measures the semantic similarity, reflects the major associative pattern in the indexing data, and ignores the less representative information.

LSA is a dimensionality reduction technique [69] addressing two issues from information retrieval, broadly named as synonymy and polysemy. The vector space analysis, singular-value decomposition (SVD), determines the correlation between the occurrence of terms and documents in the term by document matrix, and hence evaluates the “real” association of terms and documents. Also, some pattern of the occurrence of some words provides evidence of the likely

occurrence of other words. Hence, terms with similar meanings of the search query are emphasized and the synonymy problem is almost resolved. At the same time, the co-relational structure analysis on predicting the likely occurrence of other terms based on the pattern of the occurrence of the query tries to anticipate the searcher's desired knowledge, and hence it reduces the effects caused by the polysemous issue [14].

Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is commonly applied in the solution of unconstrained linear least squares problems, matrix rank estimation, and canonical correlation analysis [86]. The singular value decomposition of the $m \times n$ matrix A (where $m \geq n$, and $\text{rank}(A) = r$), denoted by $\text{SVD}(A)$, is defined as

$$A = U \Sigma V^T \quad (2.8)$$

where $U^T U = V^T V = I_n$ and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$, $\sigma_i > 0$, for $1 \leq i \leq r$, $\sigma_j = 0$, for $j \geq r+1$. In LSA, matrix A is the term-document matrix, representing the occurrence of each term in the documents of the corpus, denoted as $A = [a_{ij}]$. By applying the Eckart-Young theorem, A_k is an equivalent to the approximation of original matrix A , with k factors or k largest singular triplets, denoted as

$$A_k = \sum_{i=1}^k u_i \cdot \sigma_i \cdot v_i^T \quad (2.9)$$

Each of the SVD components is interpreted in meaning context within LSA (refer to Table 2.2). The choice of the value of k , which is the reduced

dimensional representation, is an interesting problem, and it is difficult to find the best solution. Indeed, the best value for k is usually determined through experiments. By examining the relation between performances and the number of dimensions, it is deduced that the performance of search is the best between 70 to 100 dimensions [17].

Table 2.2: SVD Components within LSA Context [9]

SVD Components	Interpretation within LSA
A_k	Best rank- k approximation to A
U	Term vector
Σ	Singular values
V	Document vectors
m	Number of terms
n	Number of documents
k	Number of factors
r	Rank of A

In short, the SVD method projects a typical large term-document matrix, which may have tens of thousands of dimensions, into a much smaller set of dimensions, preserving as much information as possible about the relative distances between the document vectors. During the collapse of the original term-document matrix into its reduced approximation, the “noise” information is lost, but the content words are superimposed on one another, revealing latent similarity in the document collection [86].

2.3 Novelty Mining

Once categorization has been performed, the novelty mining process detects novel information in relevant documents in a given topic. The major components in novelty mining are (i) novelty scoring, (ii) novelty decision making and (iii) performance evaluation, as shown in Figure 2.3. A *novelty score* is a calcu-

Figure 2.3: Novelty Mining Process.

lated 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” [39, 44, 64, 72, 81]. 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 determines the cutoff point for novelty. Finally, users can give *feedback* on the performance of the novelty mining results so as to improve the future predictions. As an adaptive filtering algorithm, novelty mining is one of the most challenging problems in information

retrieval.

A typical novelty mining algorithm can be described as follows. Given a specific topic, all the relevant documents are arranged in chronological order, i.e. $d_1, d_2, \dots, d_n$. The time sequence is determined by the creation date specified either in the file name or as an attribute of the document. Hence, the documents cannot be randomized. For each document d_t ($t = 1, \dots, n$), the degree of novelty of d_t is determined by a novelty metric, based on its history documents, i.e. d_1 to d_{t-1} . The novelty decision point determines whether a document is rated novel or redundant. Finally, the current document is pushed into the history document list.

2.3.1 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 [35]. 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. The following example uses the cosine distance metric.

Cosine similarity is a symmetric measure related to the angle between two vectors. If a document d is represented as a vector $d = [w_1(d), w_2(d), \dots, w_n(d)]^T$, the definition of cosine similarity is:

$$\cos(d_t, d_i) = \frac{\sum_{k=1}^n w_k(d_t)w_k(d_i)}{\|d_t\| \cdot \|d_i\|} \quad (2.10)$$

where $w_k(d)$ is the weight of k^{th} element in the document vector d .

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 (α) to finally decide whether the current document is redundant [96]. If the maximum redundancy score exceeds α , 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 [100]. The simple similarity method [95] and the overlap method [93] are based on the one-to-one comparison paradigm, while the all-to-one paradigm is adopted for the simple pool method [93] and the interpolated aggregate smoothing language model [5].

2.3.2 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 a *novelty decision point*. As an adaptive filtering algorithm, setting the decision point of novelty scores is one primary challenge in novelty mining [63]. If the novelty score of the document is above the novelty score decision point α , the document is considered as novel. The novelty score is defined as follows:

$$NoveltyScore(d_t) = \min_{1 \leq i \leq t-1} [1 - \cos(d_t, d_i)] \quad (2.11)$$

where d_i represents one of the history documents appearing before the incoming document d_t .

2.3.3 Performance Evaluation

The F-score is a popular evaluation *measure* (here we use *measure* instead of *metric* to avoid confusion with *novelty metric*) that is used for evaluating the results of novelty mining [97], 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.

The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$F\text{-score} = \frac{2 \times P \times R}{P + R} \quad (2.12)$$

where precision P , a measure of exactness, is defined as:

$$P = \frac{M}{S} \quad (2.13)$$

and recall R , a measure of completeness, is:

$$R = \frac{M}{A} \quad (2.14)$$

where M: # of matched novel documents between those returned by the system & ground truth, S: # of novel documents returned by the system, and A: # of novel documents by ground truth. The system-returned results are the results returned by the novelty mining algorithm.

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 [97]. Redundancy-precision and redundancy-recall are defined as:

$$RP = \frac{Mr}{Sr} \quad (2.15)$$

$$RR = \frac{Mr}{Ar} \quad (2.16)$$

where Mr: # of matched *redundant* documents between those returned by the system & ground truth, Sr: # of *redundant* documents returned by the system, and Ar: # of *redundant* documents by ground truth.

Redundancy F-score is:

$$RF = \frac{2 \times RP \times RR}{RP + RR} \quad (2.17)$$

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.

As the focus of this thesis is on novelty mining rather than preprocessing or categorization, the detailed methodologies for novelty mining are described next.

2.3.4 Sentence-Level Novelty Mining

Many studies related to sentence-level novelty mining, or novelty mining, originated from the TREC Novelty Tracks [24, 58, 59]. The novelty track was introduced in the eleventh Text REtrieval Conference (TREC) in 2002 [24]. 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* [64, 76, 97] is preferred in order to avoid confusion with novelty detection in outlier detection [25, 40, 45, 48, 56], 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 [58]. 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.

Task Definitions

The following four tasks allowed the participants to test their approaches to novelty mining. These four tasks were used both in 2003 and 2004 [58, 59].

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 thesis 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 (see Equation 2.12).

Review of Related Work

Dublin City University

At the most recent and final TREC Novelty Track in 2004, Dublin City University (DCU) [10] scored the highest among all participants, with the F-score of 0.622. DCU focused on comparisons of named entities such as people, organizations, locations, unspecified names, common nouns, cash amounts, and

verbs. Their algorithm also compared the words in a sentence against accumulated words in all previous sentences. Two of their formulas, ImportanceValue and UniqueHistory are described in the following sections.

ImportanceValue Technique

This formula observed the frequency of the entity from both within a sentence and the collection of sentences from which it originated. An entity with a high term frequency (tf) and a high inverse document frequency (idf) would be deemed as a valuable term in providing novel information.

$$ImportanceValue = \sum_{i=1}^n tf \sum_{i=1}^n idf \cdot \frac{1}{N} \quad (2.18)$$

UniqueHistory Technique

This technique determined the novelty score of a sentence by looking at the number of unique entities it contained. It will be compared to a list of accumulated unique entities for the topic. If the number of unique entities exceeded a particular decision point, the sentence was declared as novel.

University of IOWA

University of IOWA (UI) [19] also did well in TREC 2004. In 2003, they started with doing a count of new named entities and noun phrases in a sentence. In 2004, they improved the former algorithm into one which used synonyms in addition to the words for novelty comparisons, and also another one using word-sense disambiguation. The former embellishment expanded all noun phrases to their respective synonyms set, and subsequently used the synonyms for novelty comparisons.

Columbia University

Columbia University (CU) [53] managed to obtain the highest precision scores in TREC 2004. They focused on entity types of nouns, persons and organizations which were obtained by using a named entity recognizer. Next, using a classifier, they obtained weights for different classes of unique words. A sentence is declared novel if the summation of weights of the sentence is above a particular decision point. TREC Novelty Track 2004 Results For Task 2 of the novelty track, the minimum standard (TREC's benchmark) for the average F-score for 50 TREC- selected topics is 0.577, with an average precision of 0.42 and a maximum recall of 1. The average F-scores achieved ranged between 0.56 to about 0.62 [58].

Chinese Academy of Science

The Chinese Academy of Science (CAS) [27] achieved the highest F-score of 0.819 for Task 2 of the TREC 2003. For their best-scoring system, they adopted a static threshold and a value called 'New Information Degree'(NID) to present whether a sentence includes new information related to the former sentences. If the value of NID is big, this sentence is reserved, or it will be discarded. The equation for NID is given as follows:

$$NID_1 = \frac{\text{Sum of the 'idf' value of words appearing in both sentences}}{\text{Sum of the 'idf' value of words appearing in latter sentences}} \quad (2.19)$$

Novelty Measures

Allen et al. [5] performed sentence-level novelty mining using several commonly used novelty measures to show that the relevant analysis appeared to be the more difficult part of this task. The performance of novelty measures could be very sensitive to the presence of non-relevant sentences. In other words,

if non-relevant sentences have not been filtered out at first, they could be regarded as novel sentences in the novelty mining. Seven novelty measures have been compared with each other, including the vector space model measures: simple new word count, set difference, cosine similarity and the language model measures: interpolated aggregate smoothing, Dirichlet smoothing, “shrinkage” smoothing and sentence core mixture model.

This paper reviewed both situations: (i) all the relevant sentences were known, and (ii) only the relevant sentences were unknown but were given by the top 10% relevant sentences retrieved. In the first situation, all the novelty measures performed very well. The cosine similarity and language model measures worked better than the simple new word count and set difference. However, in the second situation, the performance degraded a lot. The performance of simple new word count and set difference would work better than others.

The paper also experimented extensively with parameter tuning for each model. They applied three different well-known retrieval models: the vector space model with tf-idf weighting (TFIDF), a language modeling approach with the Kullback-Leibler divergence (KLD) as scoring function, and a two stage smoothing language modeling approach (QM) as described by Zhai et al.[90]. Experimental results showed that when the novelty mining component is handed sentences that are all relevant, the language modeling and cosine distance similarity functions work best. In addition, the cosine metric is the only measure that degrades smoothly as recall and precision decrease.

Semantic Distance

Zhang et al. [91] computed the semantic distance between sentences based on WordNet. They constructed a binary classification problem (novel vs. non-

novel) for novelty mining and then collected 5 features to train a classifier including F1: the semantic distance from the sentence to the topic; F2: the distance from the sentence to the previous relevant context occurring before it; F3: word overlapping from sentence to the topic; F4: word overlapping with previous valid context; F5: is the sentence a head sentence. In the experimental study, they used the TREC2003 data for task 4. The experimental results showed that the classification using only F1 and F2 will obtain the best performance.

Overlap Measure

Zhao et al. [100] indicated that although similarity such as cosine similarity was proven to be effective experimentally [95], when one sentence’s meanings were covered by another, this relation is not necessarily symmetric, because sentences may contain multiple meanings. So, an asymmetric overlap measure should be used eventually. Their study performed task 2 of the TREC novelty track. Similarity metric (s0.4, similarity metric with decision point 0.4) and overlap metric (o0.7, overlap metric with decision point 0.7) were tested on the TREC Novelty 2004 data and TIPSTER data. The experimental results shows no significant difference between symmetric and asymmetric novelty metrics.

Pattern-based Approach

Li et al. [37] proposed a unified information-pattern-based approach for the sentence-level novelty mining, as shown in Figure 2.4. This approach consists of three steps: query analysis, relevant sentence retrieval and novelty sentence extraction. They defined novelty mining as finding the new answer to the question (i.e. query of a topic). Therefore, they generated one or more specific questions or general questions based on name entities. For example, the named entities are related to the questions about DATE (“when”), LOCATION (“where”), the PERSON (“who”), ORGANIZATION (“what/who”) and NUMBER (“how many”). Based on named entities extracted from TREC2002 data, they also generated the “answers” for these five questions manually. Therefore, finding the novel sentences for a specific topic (related to the specific questions) is to detect the new word patterns (for some answer types) in the coming sentences. The word patterns corresponding to different answer types are shown in the table 2.3.

They also noticed that in some cases, no specific question can be generated

Figure 2.4: Overview of Pattern-based Approach.

based on named entities, where they will evaluate a sentence by detecting both the new named entities and the new words in it. In the experimental study, they reported the final result for relevant sentence retrieval + novelty sentence detection, i.e. TREC novelty track task 1. The average precisions in top-5, 10, 15, 20 and 30 sentences in each topic of TREC 2002, TREC 2003 and TREC2004 were summarized in Table 2.4. The proposed method, ip-BAND, has been compared with the other four baselines and shown better precision in relevant and novel sentence retrieval.

Semantic Similarity

Aminul Islam et.al [26] presented a method for measuring the semantic similarity of texts using a corpus-based measure of semantic word similarity and a normalized and modified version of the Longest Common Subsequence

(LCS) string matching algorithm. They focused on computing the similarity between two sentences or two short paragraphs. The proposed method determined the similarity of two texts from semantic and syntactic information (in terms of common-word order) that they contain. They considered three similarity functions in order to derive a more generalized text similarity method. First, string similarity and semantic word similarity were calculated and then they used an optional common-word order similarity function to incorporate syntactic information. Finally, the text similarity was derived by combining string similarity, semantic similarity and common-word order similarity with normalization. Evaluation results on two different data sets showed that this method outperforms several competing methods, such as human similarity and Li et al.'s method [38].

2.3.5 Document-Level Novelty Mining

Zhang et al. [95] performed the document-level novelty detection based on the distance between the new document and the previously delivered documents (in history), i.e. document-document distance. This method is simple but robust. It simulates the way users view novelty because it is easiest for an assessor to identify a new document as a redundancy one with a specific previously seen document, and harder to identify it as redundant with a set of previously seen documents. Therefore, the redundancy of the new document $R(d_t|Hist(t))$ was set equal to the value of the maximally similar value in all $R(d_t|d_i)$:

$$R(d_t|Hist(t)) = \arg \max_{d_i \in Hist(t)} R(d_t|d_i) \quad (2.20)$$

where $Hist(t)$ denotes all the relevant documents delivered in time t . In their

paper, five novelty measures have been used to evaluate the similarity between d_t and d_i , i.e. $R(d_t|d_i)$, including set difference, geometric distance (cosine similarity), distributional similarity with Dirichlet smoothing, distributional similarity with shrinkage smoothing, and a mixture model.

In the experimental study, all the relevant documents have been assumed to be known. The experimental results on TIPSTER data showed that the cosine similarity metric was very effective and robust, as shown in the Table 2.6.

2.3.6 Summary of Previous Work

The previous work for both sentence and document-level novelty mining has been summarized in Table 2.7.

2.4 Chapter Summary

The major components in novelty mining are (i) novelty scoring, (ii) novelty decision and (iii) performance evaluation. Many of the previous studies in novelty mining optimized a single performance measure, such as the F score in the TREC Novelty Tracks [58, 59], 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). This assumption may not hold for many users, and therefore, the previous methods or some recommended parameters optimized based on F score may not be suitable for real-life applications. Therefore, the techniques proposed in this thesis are designed to perform optimally in a variety of performance requirements, and will focus on solving the problems in novelty scoring (which depend on the

selection of novelty metrics), novelty decision, and performance evaluation in novelty feedback.

Categories	Answer types	Word patterns
Name	Person	who, individual, person, people, participant, candidate, customer, victim, leader, member, player, name
	Organization	who, company, companies, organization, agency, agencies, name, participant
	Location	where, location, nation, country, countries, city, cities, town, area, region
Time	Date	when, date, time, which year, which month, which day
Number	Number	how many, how much, length, number, polls, death tolls, injuries, how long

Table 2.3: Word patterns for five types of NE-questions (from [37])

Dataset	Top sentences	B-NN	B-NW	B-NWT	B-MMR	ip-BAND
TREC02	5	0.1878	0.1959	0.2000	0.2204	0.2327
	10	0.1939	0.1918	0.2041	0.1980	0.2163
	15	0.1891	0.1946	0.1986	0.1946	0.2136
	20	0.1837	0.1867	0.1929	0.1776	0.2051
	30	0.1728	0.1762	0.1721	0.1653	0.1878
TREC03	5	0.4480	0.4680	0.4880	0.4600	0.5760
	10	0.4520	0.4820	0.5200	0.4880	0.5760
	15	0.4400	0.4920	0.5160	0.4907	0.5613
	20	0.4400	0.4930	0.5280	0.4700	0.5560
	30	0.4247	0.4773	0.5267	0.4747	0.5507
TREC04	5	0.2280	0.2360	0.2400	0.2320	0.2280
	10	0.2120	0.2100	0.2160	0.2120	0.2400
	15	0.2027	0.2120	0.2160	0.2040	0.2413
	20	0.1990	0.2090	0.2150	0.1990	0.2360
	30	0.1880	0.1973	0.2060	0.1913	0.2247

Table 2.4: Average novelty precision of ip-BAND and other baseline methods [37]

Measures	Definition for the redundancy $R(d_t d_i)$
Set Difference	$\ Set(d_t) \cap Set(d_i)\ $: the new words in the current document d_t which are not covered by previous seen document d_i
Cosine Distance	the cosine similarity between the d_t and d_i
LM: Shinkage Dirichlet Prior Mixture Model	Assume is the probabilistic language model for document d . The distribution similarity measure, Kullback-Leibler divergence, is a way to measure the redundancy of one document given another: $R(d_t d_i) = -KL(\theta_{d_t}, \theta_{d_i})$ $= -\sum_{w_i} P(w_i \theta_{d_t} \log \left(\frac{P(w_i \theta_{d_i})}{P(w_i \theta_{d_t})} \right))$

Table 2.5: Measure for redundancy in [95]

Measure	Recall	Precision	Mistake
Set Distance	0.36	0.29	28.1%
Cosine Distance	0.48	0.33	18.7%
LM: Shrinkage	0.375	0.45	21%
LM: Dirichlet Prior	0.375	0.45	21%
LM: Mixture Model	0.46	0.40	16.7%

Table 2.6: Average performance of different redundancy measure with a simple decision algorithm, measured on 33 topics with AP News & Wall Street Journal dataset. Only *absolutely redundant* documents are treated as redundant

Works	Level	Task	Start point	Data	Representation
[95] compare 5 novelty metrics	Document	Task 2	Given all relevant docs	TIPSTER	tf.idf or language model
[5] compare 7 novelty metrics	Sentence	Task 1 & 2	Given all sentences	TREC Novelty Track 02-03	tfidf or language model
[91] Semantic distance metric	Sentence	Task 4	Given relevant and partial novel sentences	TREC Novelty Track 03	tf.idf+WordNet
[100] similarity vs. overlap metric	Sentence/ Document	Task 1	Given all docs/sentences	TREC Novelty Track 02-04	tf.idf
[37] ip-BAND	Sentence	Task 1	Given all sentences	TREC Novelty Track 02-04	Named entities and length of sentences

Table 2.7: Summary for the previous work in novelty mining

Chapter 3

Metrics and Models for Novelty Scoring

3.1 Introduction

As presented in Chapter 2, the first major component in novelty mining is novelty scoring. To recap, 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 *metric* or distance function 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” [44, 52, 64, 72, 92, 81]. The cosine distance metric (see Eq. (2.10)) is a typical example. Using this measure, the novelty score is simply defined as 1 minus the cosine similarity score, as in Eq. (2.11). Since similarity-based metrics are symmetric since they do not take the ordering of sentences into account, they may cause some problems in terms of novelty scoring. For example, suppose there are two sentences s_1 and s_2 as

follows.

$$s_1 = \{A, B, C, D, E\}, \quad s_2 = \{A, B\}$$

where A, B, C, D, E are five different terms, and s_2 appears right after s_1 . Based on the similarity metrics, s_2 is likely to be predicted as “novel” due to the low similarity between s_1 and s_2 . While in fact, s_2 should be predicted as “non-novel”. This example also shows how important it is to consider the ordering of sentences in novelty scoring, because the so-called “novel” information totally depends on the information that has appeared in the history.

On the other hand, some asymmetric metrics have been proposed to measure the “novelty” of sentences by considering the ordering between sentences. The new word count metric [11], which counts the number of new words in the incoming sentence given its history sentences, is a typical example here. Zhao et al. [100] indicated that the asymmetric metrics are more suitable to the nature of novelty mining. However, symmetric metrics were shown to be empirically best across different corpora [95]. One question raised here is why the asymmetric metrics do not outperform the symmetric metrics for real-world data?

One possible reason is that the previous studies evaluated the performance of different novelty metrics globally, where the limitations of metrics in some situations may be canceled out by their strengths in other situations [78]. To the best of our knowledge, this problem has not been thoroughly investigated. This chapter attempts to detect the strengths and limitations of the different novelty metrics by a comparative study in a series of predefined situations.

A new framework of novelty measurement, i.e. a blended mixture of both types of novelty metrics (BLEND), is then proposed. The experimental result

shows the superior performance of the new framework of novelty measurement under different performance requirements and for data with different percentages of novel sentences. Since it does not require any prior information from data, the new framework of novelty measurement is suitable for real-time applications like novelty mining where no training data is available beforehand.

In addition to blended metrics, another model for novelty scoring is proposed for document-level novelty mining, which identifies novel documents from a given group of relevant documents. The new model is based on document-to-sentence (D2S) annotations, which combines sentence and document-level novelty mining. By using D2S, a document which shares one single sentence with each of the history documents could be correctly recognized as redundant, whereas general document-level methods may incorrectly label the document as novel. For example, here are three incoming simple documents in time sequence:

D1: Singapore is an island city-state located at the southern tip of the Malay Peninsula. It lies 137 kilometers north of the equator.

D2: Singapore is an island city-state located at the southern tip of the Malay Peninsula. The population of Singapore is approximately 4.86 million.

D3: Singapore lies 137 kilometers north of the equator. The population is approximately 4.86 million.

If we directly apply general document-level novelty mining (D-NM), *D3* is easier to be predicted as novel because it contains new information when comparing with *D1* and *D2* individually. However, if we segment these three documents into sentences, *D3* will be correctly predicted as redundant because all its sentences have appeared in the former sentences. This method can effectively help to detect this kind of text-reuse redundancy.

This chapter is organized as follows. Section 3.2 introduces several popular novelty metrics from both symmetric and asymmetric types, and analyzes the factors that may affect the performance these metrics. Section 3.3 introduces a new framework of novelty measurement, i.e. the mixture of both types of metrics, for novelty mining. Section 3.4 introduces another model for novelty scoring, namely Document-to-Sentence Annotations (D2S). Experiments on sentence-level and document-level data are presented in Section 3.5. Section 3.6 concludes the entire chapter.

3.2 Novelty Metrics

Novelty metrics can be divided into two types based on whether the ordering of sentences is taken into account, i.e. symmetric and asymmetric [100], as summarized in Table 3.1. In this section, we will introduce some popular metrics for each type separately.

Table 3.1: Symmetric vs. Asymmetric Metrics

	Symmetric metric	Asymmetric metric
Definitions	A metric M yields the same result regardless of the ordering of two sentences, i.e. $M(s_i, s_j) = M(s_j, s_i)$.	A metric M yields different results based on the different ordering of two sentences, i.e. $M(s_i, s_j) \neq M(s_j, s_i)$.
Typical metrics in novelty mining	Cosine similarity, Jaccard similarity	New word count, Overlap

3.2.1 Symmetric Metrics

A symmetric metric M yields the same result regardless of the ordering of two sentences, i.e. $M(s_i, s_j) = M(s_j, s_i)$. The cosine distance metric is one of the

most popular symmetric metrics in information retrieval [5]. To recap from Section 2.3.1, the cosine distance novelty metric first calculates the similarities between the current sentence s_t and each of its history sentences s_i ($1 \leq i \leq t-1$). Then, the novelty score is simply one minus the maximum of these cosine similarities, as shown in Eq. (3.1).

$$N_{cos}(s_t) = 1 - \max_{1 \leq i \leq t-1} cos(s_t, s_i) \quad (3.1)$$

$$cos(s_t, s_i) = \frac{\sum_{k=1}^n w_k(s_t) \cdot w_k(s_i)}{\|s_t\| \cdot \|s_i\|}$$

where $N_{cos}(s)$ denotes the novelty score of the sentence s based on cosine similarity and $w_k(s)$ is the weight of k^{th} element in the sentence weighted vector s . The term weighting function used is tf.isf (term frequency times inverse sentence frequency) as defined below.

$$w_k(s_i) = tf_{w_k, s_i} \log \left(\frac{n+1}{sf_{w_k} + 0.5} \right) \quad (3.2)$$

where tf_{w_k, s_i} is the frequency of the word w_k in the sentence s_i ; sf_{w_k} is the number of sentences, in which the word w_k appears in the collection; n is the number of sentences in the collection.

Another popular similarity-based metric in information retrieval is Jaccard similarity [31]. Jaccard similarity metric is used to measure the similarity between two sets, which is defined as the ratio of the intersection size and the union size of two sample sets. Similar to cosine similarity, Jaccard similarity is converted to a novelty score subtracting the maximum Jaccard similarity between s_t and s_i ($1 \leq i \leq t-1$) from one, as shown below.

$$\begin{aligned}
N_{Jaccard}(s_t) &= 1 - \max_{1 \leq i \leq t-1} Jaccard(s_t, s_i) \\
Jaccard(s_t, s_i) &= \frac{|s_t \cap s_i|}{|s_t \cup s_i|} \\
&= \frac{\sum_{k \in s_t \cap s_i} \min(w_k(s_t), w_k(s_i))}{\sum_{k \in s_t \cup s_i} \max(w_k(s_t), w_k(s_i))}
\end{aligned} \tag{3.3}$$

3.2.2 Asymmetric Metrics

An asymmetric metric M yields different results for different orderings of two sentences, i.e. $M(s_i, s_j) \neq M(s_j, s_i)$. A typical asymmetric metrics is the overlap metric. Overlap metrics have proven to be stable among different data collections [93] and to be more suitable for the nature of novelty because they take the ordering of sentences into account [100]. The overlap novelty metric is also calculated by comparing the current sentence s_t with any of its history sentences, s_i ($1 \leq i \leq t-1$), as defined in Eq. (3.4).

$$\begin{aligned}
N_{overlap}(s_t) &= 1 - \max_{1 \leq i \leq t-1} overlap(s_t | s_i) \\
overlap(s_t | s_i) &= \frac{|s_t \cap s_i|}{|s_t|} \\
&= \frac{\sum_{k \in s_t \cap s_i} \min(w_k(s_t), w_k(s_i))}{\sum_{k \in s_t} w_k(s_t)}
\end{aligned} \tag{3.4}$$

where $overlap(s_t | s_i)$ is the overlapping part of the sentence s_t by a history sentence s_i , $1 \leq i \leq t-1$. $w_k(s_i)$ is the weight of the word k in the sentence s_i , as defined in Eq. (3.2).

The new word count novelty metric is another popular asymmetric metric,

which was proposed for sentence-level novelty mining [5]. The idea of the new word count novelty metric is to assign the incoming sentence the number of the new words that have not appeared in its history sentences, as defined in Eq. (3.5).

$$newWord(s_t) = |W(s_t)| - |W(s_t) \cap [\cup_{i=1}^{t-1} W(s_i)]| \quad (3.5)$$

where $W(s_i)$ is the set of words in the sentence s_i . The values of the new word count novelty metric for an incoming sentence are non-negative integers such as 0, 1, 2, etc. To normalize the values of the novelty scores into the range of 0 to 1, the new word count novelty metric can be normalized by the total number of words in the incoming sentence s_t as below.

$$N_{newWord}(s_t) = 1 - \frac{|W(s_t) \cap [\cup_{i=1}^{t-1} W(s_i)]|}{|W(s_t)|} \quad (3.6)$$

where the denominator $|W(s_t)|$ is the word count of s_t . This normalized metric, $N_{newWord}$, has the range of values from 0 (i.e. no new word) to 1 (i.e. 100% new words).

3.2.3 Factors Affecting Novelty Metric Selection

For a variety of novelty metrics available, selecting a suitable metric is crucial for the final performance of novelty mining. In previous studies, many efforts have been put on creating some new metrics or comparing different metrics across different corpora, to find the empirically suitable metrics in a very general situation [1, 5, 95]. However, not much guidance has been provided for the selection of a suitable novelty metric in a *specific* situation. For example, what kind of

metrics could perform better for high-recall/high-precision requirements? Or which kind of metrics could perform better for the data with a small percentage of novel sentences? To answer these questions, we first need to explore the factors that may affect the performance of novelty metrics.

The differences between various novelty metrics are actually due to how they measure the “novelty” of the incoming sentence. Therefore, the performance of novelty metrics may be affected by the user’s requirement about his/her desired novel sentences explicitly, and the distributions of novelty based on the ground truth of the data implicitly, as summarized in Table 3.2. We describe both the explicit and implicit factors in this section, in order to benefit the different situations that occur in the real-world applications of novelty mining.

Table 3.2: Explicit and Implicit Factors Of Novelty Metric Selection

Explicit factor	Performance requirements, e.g. high-recall, high-precision, or high- F -score.
Implicit factor	Distributions of novelty based on the truth of the data, e.g. high, medium or low novelty ratios in a specific topic.

Explicit Factor: Performance Requirements

In novelty mining, performance requirements are optimized through maximizing the corresponding evaluation measures (refer to Section 2.3.3). Two measures, recall and precision, best gauge the overall effectiveness of information retrieval performance. In the context of sentence-level novelty mining, these measurements are used to assess the likelihood of missing novel sentences and retrieving actually non-novel sentences for the user, respectively. It has been acknowledged that there is a trade-off between recall and precision, i.e. high-recall and

high-precision cannot be achieved simultaneously.

Although F -score was used as the primary performance requirement in previous studies [58], in practical applications, the user may require high recall or high precision rather than high F . For example, when users do not want to miss any novel information, a high-recall algorithm which only filters out the most redundant sentences is desired. On the other hand, when users want to read the sentences with the most significant novelty in a very short time, a high-precision algorithm which only retrieves highly novel sentences is preferred. In high-recall/precision cases, novelty metrics that are good at detecting the redundancy/novelty of sentences may be a better choice. The type of novelty metrics that will favor high-recall or high-precision situations will be explored in the experimental study (see Section 3.5.3).

Implicit Factor: the Distributions of Novelty Based on the Truth of the Data

The distribution of novelty based on the truth of the data can be described by the prior probability $P(novel)$, i.e. the probability that the next incoming sentence is novel. This prior probability reflects our prior knowledge of how likely we are to get a novel sentence before it actually appears.

For the topics containing many highly similar sentences, the novelty metric is expected to be sensitive to novel information in the highly similar sentences. On the other hand, for the topics containing many dissimilar sentences, we hope the novelty metric is good at detecting the redundant information from these dissimilar sentences. How the performance of the novelty metrics are influenced by the topics with different percentage of novel sentences will also be tested in the experimental study (see Section 3.5.3).

3.3 Blended Metrics for Novelty Measurement

Because the strengths of symmetric metrics and asymmetric metrics can complement each other, a new framework is proposed for measuring the novelty by a mixture of both types of novelty metrics. The goal of the blended mixed metric (BLEND) is to integrate their merits and hence generalize better over different topics. From a classification point of view, constructing a blended metric that combines the measurements from multiple metrics is equivalent to constructing a classifier using multiple features rather than only one feature. Since the characteristics of the two metrics differ as discussed in the previous section, some samples that overlap in the one dimensional feature space could be separable in the two dimensional feature space. Moreover, the decision boundary in the two dimensional feature space which is defined by the blending strategy and decision point can be more flexible than that in the one dimensional feature space [65].

Two major issues for constructing a blended metric are (i) the scaling problem that ensures different component metrics are comparable and consistent and (ii) the combining strategy that defines the way of fusing the outputs from different component metrics. As the novelty scores from all novelty metrics ranges from 0 (i.e. redundant) to 1 (i.e. totally novel), they are both comparable and consistent.

For the combining strategy, a new framework of measuring through combining two types of metrics is formulated, as shown in Eq. (3.7).

$$M(s_t) = \alpha N_{sym}(s_t) + (1 - \alpha) N_{asym}(s_t) \quad (3.7)$$

where α is the combining parameter ranging from 0 to 1. The bigger the α , the

heavier the weight for the symmetric metrics. Without losing any generalization, equal weights of 50% ($\alpha = 0.5$) are chosen for both symmetric and asymmetric metrics and compare the performance of the blended metrics with those of the individual metrics. In Section 3.5.6, the effects of weight variation will be discussed in detail.

Four different blended metrics can be generated from this new framework using Eq. (3.7), i.e. M_1 to M_4 , as follows.

$$M_1(s_t) = \alpha N_{cos}(s_t) + (1 - \alpha) N_{newWord}(s_t) \quad (3.8)$$

$$M_2(s_t) = \alpha N_{cos}(s_t) + (1 - \alpha) N_{overlap}(s_t) \quad (3.9)$$

$$M_3(s_t) = \alpha N_{Jaccard}(s_t) + (1 - \alpha) N_{newWord}(s_t) \quad (3.10)$$

$$M_4(s_t) = \alpha N_{Jaccard}(s_t) + (1 - \alpha) N_{overlap}(s_t) \quad (3.11)$$

The experimental results for the blended metrics are shown in Sections 3.5.4 and 3.5.5.

3.4 Document-to-Sentence Model

This section proposes a new model for novelty scoring which is based on document-to-sentence (D2S) annotations. By using D2S, a document which shares one single sentence with each of the history documents could be correctly recognized as redundant, whereas general document-level methods may incorrectly label the document as novel.

Sentence-level annotations have been discussed in some other fields, for text reuse on the Web [8]. However, the D2S model emphasizes on the novelty mining issue and investigates how to use the sentence-level information to predict the

novelty of a document.

The basic idea of the model D2S is to segment the document into sentences and perform novelty mining at sentence level. It is assumed that all the sentences from one document are relevant to the topic of the document. This assumption is the prerequisite for sentence-level novelty mining in D2S.

The novelty of the sentences in one document can reflect the novelty of the document. It simulates the way users view novelty. It is easier for an assessor to identify a new document as a redundant one with several specific previously seen sentences. Then, the novelty of a new document was set to the proportion of its novel sentences which are judged through comparing the similarity with all the history sentences by our algorithm. Therefore, we define $NovelRate_s$ to measure the novelty of a document. In other words, $NovelRate_s$ is the ratio of novel sentences of a given document as predicted by the system (novelty mining algorithm). If the $NovelRate_s$ of one document is larger than a predefined decision point, we judge it to be a novel document. The definition of $NovelRate_s$ is:

$$NovelRate_s = \frac{num(novel\_sentences\ retrieved\ by\ system)}{num(all\_sentences)} \quad (3.12)$$

where $num(all\_sentences)$ is the number of sentences in one document and $num(novel\_sentences\ retrieved\ by\ system)$ is the number of novel sentences which are predicted by the sentence-level NM algorithm for the document.

To better understand the usage of D2S, later we generate another two document-level datasets based on TREC 2004 and 2003 Novelty Track. We use

Figure 3.1: Novelty Mining Process of D2S

$NovelRate_a$ to measure the percentage of real novel sentences for a document by TREC assessors. $NovelRate_a$ is defined as:

$$NovelRate_a = \frac{num(novel\_sentences \text{ labeled by assessor})}{num(all\_sentences)} \quad (3.13)$$

where $num(all\_sentences)$ is the number of sentences in one document and $num(novel\_sentences \text{ labeled by assessor})$ is the number of novel sentences which are labeled by TREC assessors for the document.

Generally speaking, $NovelRate_s$ is our proposed discriminating decision point for predicting the novelty of each document on the retrieved results. And $NovelRate_a$ is a measurement to evaluate the percentage of novel sentences for a document on assessors' label results.

The novelty mining process using D2S can be seen in Fig. 3.1.

Figure 3.2: A New Model for Novelty Mining.

First, we segment the documents into sentences and apply sentence-level NM on these sentences. Then based on the prediction of each sentence, we can calculate $NovelRate_s$ for each document. If the $NovelRate_s$ of one document is larger than the decision point, it will be predicted as a novel document.

This new algorithm can improve code reuse in a novelty mining algorithm because we only need to focus on the development of the sentence-level modules. Moreover, it can facilitate the maintenance and improve the NM performance, which will be tested in the following experiments. Fig. 3.2 shows the new framework of NM with the D2S algorithm. This new framework can lower the granularity of similarity computation from documents to sentences. When a document arrives, it is first segmented into sentences. Then these sentences are preprocessed and categorized. Finally, sentence-level novelty mining is performed on these sentences, the $NovelRate_s$ is calculated, and the novelty prediction is performed. For the novelty mining module, one decision function is added, which is to differentiate between document-level NM and sentence-level NM.

3.5 Experiments and results

3.5.1 Novelty 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, i.e. TREC 2002 [24], 2003 [59], 2004 [58] Novelty Track data. In this experimental study, the TREC 2002 Novelty Track data is 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 3.3.

Table 3.3: Statistics of experimental data.

data	topics	# relevant sentences	# novel sentences	average novelty ratio
TREC2003	N1-N50	15557	10226	65.7%
TREC2004	N51-N100	8343	3454	41.1%

Table 3.4: Statistics of Experimental Data

Dataset	Novel	Non-novel
APWSJ	10839(91.10%)	1057(8.90%)

APWSJ Document-Level

For document-level novelty mining, the APWSJ novelty data [95] 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. Similar to a previous study [100], we 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*. In our experiments, we adopt the strict definition used in [95] 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 details of the APWSJ data are summarized in Table 3.4.

TREC 2003/4 Document-Level

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 [79, 64]. 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 $NovelRate_a$ for each document, which is the actual percentage of novel sentences in that document. If we set a low $NovelRate_a$ 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 $NovelRate_a$ decision points, we can observe performances on datasets with different percentages of novel documents.

BizBlogs Document-Level

The novelty BizBlogs dataset [39], a novelty annotated version of the BizBlogs07 [14] dataset of business blogs, was also used in our experiments. The dataset contains 1269 blog posts in four categories: product, company, marketing and finance [14]. 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 [14].

Table 3.5 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 [67].

Table 3.5: Summary of the novelty ratios

Category	# Blogs	# Novel blogs	Percentage of novel blogs
1(product)	392	379	96.68
2(company)	264	251	95.08
3(marketing)	267	260	97.38
4(finance)	346	321	92.77
sum	1269	1211 average =	95.48

3.5.2 Evaluation Measures

Refer to Section 2.3.3 for the definitions of the evaluation measures used for novelty mining. For the evaluation of sentence-level novelty mining, the novel sentences selected manually by the NIST assessor who created the topic were considered as the truth data. The experiments use the evaluation measures proposed for the TREC Novelty Track [58], i.e. set-based recall, precision and F -score averaged across all 100 topics or subsets of topics in the TREC 2003/2004 Novelty Track data (see Equations 2.12-2.14).

Because there is a large percentage of novel documents in APWSJ, redundancy-precision (RP), redundancy-recall (RR) and redundancy F-Score (RF)(see Equations 2.15-2.17 were used to evaluate the performance of document-level NM on APWSJ.

Precision-recall-F (PRF) and redundancy-precision-recall-F ($R-PRF$) curves were adopted as the standard graphs for novelty mining [97]. The larger the area under the PRF curve/ $R-PRF$ curve, the better the algorithm. In the graphs, we show the standard F/RF contours [97] are shown, which indicate the F/RF Score values when setting precision and recall from 0 to 1 with a step of 0.1. These contours can facilitate us to compare redundancy F Scores/ F Scores in $R-PR$ curves/ PR curves.

3.5.3 Metrics Comparison Under Various Situations

In this section, various symmetric and asymmetric metrics are compared and analyzed under different conditions. As we described in Section 3.2.3, the performance of novelty metrics will be influenced by both the explicit factor, i.e. the user's performance requirements and the implicit factor, i.e. the prior prob-

ability of novel sentences in the processed data.

Seven metrics including cosine, Jaccard, overlap metrics combined with tf and tf.isf weighting functions, new word count metrics are compared in different situations. The reason to include tf term weighting function is to ensure that the different performance in each situation mainly depends on different types of novelty metrics rather than the different term weighting functions.

Performance of Metrics under Different Performance Requirements

A novelty mining algorithm needs to retrieve the novel information for the user according to his/her performance requirements. When users do not want to miss any novel information, a high-recall algorithm which only filters out the most redundancy is desired. When users want to read the sentences with the most significant novelty, a high-precision algorithm which only retrieves highly novel sentences is preferred. There is a tradeoff between precision and recall, which can be adjusted by the novelty decision point explicitly. A high decision point will only let sentences with high novelty scores be passed to user, and this novelty mining algorithm will show the high-precision property. On the other hand, a low decision point will only eliminate very redundant sentences, and the novelty mining algorithm will show the high-recall property.

In this empirical study, we adjusted the novelty decision point in a range from 0.05 to 0.95 with a step length equal to 0.05, from which 19 precision-recall pairs (also the corresponding F -scores) are obtained. The increasing novelty decision points create a series of novelty mining algorithms from high-recall to high-precision artificially. For each novelty decision point, the precision, recall and F -scores of 100 topics in the TREC 2003 and 2004 Novelty Track data are averaged over topics and reported in graphs and tables.

For demonstrating how various metrics perform in high-precision and high-recall cases, we use Precision-F score (PF) curve and Recall-F score (RF) curve, where the x-axis is the precision or recall and the y-axis is the F score. Figure 3.3 shows the performance of seven different novelty metrics in 2 different ways, including *F*-score-Precision (FP) curves (Fig.3.3(a)) and *F*-score-Recall (RF) curves (Fig.3.3(b)). In these curves, the high-precision/recall region can be zoomed in to the level which can clearly show the difference between various novelty metrics. For example, in the range of precision between 0.68 and 0.74, the asymmetric metrics clearly outperform the symmetric metrics.

The top two novelty metrics of high-recall, high-precision and high-*F*-score novelty mining algorithm for the TREC 2003 and 2004 Novelty Track data are summarized in Table 3.6. From this experiment, we can find that symmetric metrics will outperform asymmetric metrics in high-recall algorithms and underperform asymmetric metrics in high-precision algorithms. But neither type of

Figure 3.3: Performance of various novelty metrics for the TREC 2003 and 2004 Novelty Track data. Dashed lines denote asymmetric metrics and solid lines denote symmetric metrics.

metrics can perform well in both high-recall and high-precision novelty mining algorithms.

Table 3.6: Top two novelty metrics of high-recall, high-precision and high- F -score novelty mining algorithm for the TREC 2003 and 2004 Novelty Track data. * denotes the asymmetric metrics.

	rank 1	rank 2
high-precision	new word*	overlap*
high-recall	Jaccard	cos
high- F -score	Jaccard	cos

This result can be explained by the characteristics of metrics. The symmetric metrics were initially designed to measure the similarity (redundancy) between sentences [95]. In novelty mining, the precondition of using symmetric metrics for the novelty measure is to assume that novelty is the opposite of redundancy. In the high-recall algorithm, since only the most redundant sentences will be detected and filtered out, the symmetric metrics which are suitable to measure the redundancy can perform very well. On the other hand, asymmetric metrics are more suitable to the nature of novelty mining [100] and hence perform better when detecting highly novel information. Therefore, the asymmetric metrics can perform best in high-precision algorithms.

To our best knowledge, there is no previous work discussing the novelty metric selection problem. This empirical study is valuable for selecting a suitable metric for the novelty mining algorithm based on the user’s performance requirements.

Performance of Metrics on Data with Different Prior Probabilities

In the previous section, we have investigated how the performance of novelty metrics is affected by the algorithm performance requirements. In this section,

we continue to investigate the effects of the prior probability of data itself on the novelty metrics. Since novelty mining is a real-time process in which no training data is available, the prior probability of the processed data is usually unknown beforehand. However, it is still important to understand how the different types of metrics are affected by the prior probability of data, because the principles provide guidelines for improvement in novelty measurement.

Given a sentence stream s_1 to s_n , we can define whether the sentence s_i ($i = 1$ to n) is novel as a variable Y . The prior probability $P(Y)$ is a description of the probability that the next incoming sentence is novel before actually seeing it. Using the maximum likelihood estimation, the prior probability of novelty can be estimated by the ratio of novel sentences, denoted as novelty ratio r_N .

$$r_N = \frac{\text{\#novel sentences}}{\text{\#relevant sentences}} \quad (3.14)$$

The goal of this experiment is to investigate the performance of various metrics in the data with different prior probabilities. Therefore, we calculate the prior probability of each topic and divide all the 100 topics in the TREC 2003 and 2004 Novelty Track data (see Section 3.5.1) into different categories corresponding to different levels of novelty ratio r_N . Then, the same seven metrics are tested and compared in each category separately. Notice that the novelty ratios which are assumed to be known is only used for stratifying the topics and will never be used in the novelty mining process. Therefore, the novelty mining results can still simulate the performance of novelty metrics in real applications.

For ease of understanding, novelty ratios ranging from 0 to 1 are divided into 3 levels, i.e. low ($r_N \leq 0.33$), medium ($0.33 < r_N \leq 0.66$) and high ($r_N > 0.66$) levels. The number of topics with low, medium and high novelty

Table 3.7: Number of topics with low, medium and high novelty ratio in both sentence- and document-level data.

data	novelty ratio		
	low $r_N \leq 0.33$	medium $0.33 < r_N \leq 0.66$	high $r_N > 0.66$
TREC2003	1	23	26
TREC2004	17	32	1
total	18	55	27

ratio in each data is summarized in Table 3.7.

In each category, we repeat the experiments from the previous section. The novelty decision point is varied from 0.05 to 0.95 with an equal step of 0.05. For each decision point, the precision, recall and F -scores for the topics in each category are averaged and reported in the graphs and tables.

The experimental results of various metrics for the topics with low, medium and high novelty ratios in the TREC 2003 and 2004 Novelty Track data are shown in three rows of Figure 3.4, respectively. We summarized the experimental results in this figure as follows.

- (i) The same types of novelty metrics perform similarly in each category of topics. See the comparison between asymmetric metrics (dashed lines) and symmetric metrics (solid lines).
- (ii) Asymmetric metrics (i.e. overlap and new word count metrics) perform better for the topics with low novelty ratio, as shown in Figure 3.4(a-b).
- (iii) Symmetric novelty metrics (i.e. cosine and Jaccard similarities) perform better for the topics with high novelty ratio, as shown in Figure 3.4(e-f).
- (iv) Both types of metrics perform similarly for the topics with medium novelty ratio, as shown in Figure 3.4(c-d).

3.5 Experiments and results

62

(a) topics with low r_N , high-precision (b) topics with low r_N , high-recall (c) topics with medium r_N high-precision (d) topics with medium r_N , high-recall (e) topics with high r_N , high-precision (f) topics with high r_N , high-recall

Figure 3.4: Performance of various novelty metrics for the TREC 2003 and 2004 Novelty Track data. Dashed lines denote asymmetric metrics and solid lines denote symmetric metrics.

Moreover, to also consider the effects from different requirements, from high-precision to high-recall, the top two novelty metrics of various performance requirements and novelty ratios for the TREC 2003 and 2004 Novelty Track data are summarized in Table 3.8, where * denotes the asymmetric metrics. From Figure 3.4 and Table 3.8, we can conclude as follows.

- (i) As observed from Figure 3.4, asymmetric metrics such as new word count and overlap metrics always win in high-precision case (see Figure 3.4(a), (c), (e)). In the high-recall case, asymmetric metrics will outperform symmetric metrics when the novelty ratio is low (see Figure 3.4(b)), and underperform symmetric metrics when the novelty ratio is high (Figure 3.4(f)). This experimental result shows the clear complementary behavior of different types of metrics in various situations.
- (ii) Jaccard and cosine similarities achieve the globally highest F -score as shown in the lower-right corner of Table 3.8, which is consistent with the empirically best performance of similarity-based novelty metrics reported in the previous papers [95, 5]. However, by testing in different situations, the cosine similarities as well as Jaccard will not consistently perform best for novelty mining locally, especially when the high-precision novelty mining algorithm is desired.
- (iii) The new word count metric cannot get a very high recall. As new word count is the number of new words in the incoming sentence that does not appear in any history sentence, some novel sentences with no new words will never be retrieved. Therefore, the new word count metric is not recommended for high-recall novelty mining algorithms.

The advantage of this experiment is to understand the strengths and limita-

Table 3.8: Top two novelty metrics for various performance requirements and novelty ratios for the TREC 2003 and 2004 Novelty Track data. * denotes the asymmetric metrics.

	novelty ratio			
	low	medium	high	total
high-precision	new word* overlap*	new word* overlap*	overlap* new word*	new word* overlap*
high-recall	new word* Jaccard	overlap* Jaccard	Jaccard cos	Jaccard cos
high- <i>F</i> -score	new word* overlap*	overlap* Jaccard	Jaccard cos	Jaccard cos

tions of novelty metrics in different situations. The experimental results indicate that symmetric and asymmetric novelty metrics cannot perform well in various situations simultaneously, while their strengths may complement each other. In the next section, we will show the results of the blended metrics.

3.5.4 Blended Metrics vs. Individual Metrics

Figure 3.5: Blended metrics vs individual metrics for all 100 topics in high-precision (a) and high-recall (b) system in TREC 2003 and 2004 Novelty Track data

The series of blended metrics from M_1 to M_4 have been tested on all 100 topics from TREC 2003 and 2004 Novelty Track data and compared with 4 individual metrics, cosine, Jaccard, overlap and new word count metrics, with the experimental results shown in Figure 3.5. The tf.isf term weighting function is used for both individual and blended metrics. The F scores of various metrics are shown in both the high-precision region (Figure 3.5(a)) and the high-recall region (Figure 3.5(b)). Since the highest F scores appear in the high-recall region, the global view is omitted. From Figure 3.5, within the individual metrics, asymmetric and symmetric metrics perform well in high-recall and high-precision cases respectively, but not both. On the other hand, the blended metrics can perform well in both cases.

To analyze the comparison results more clearly, each blended metric can be compared with its component individual metrics separately. Next, the comparison is shown of M_3 and its component individual metrics, i.e. Jaccard similarity metric and new word count metric, as the comparison results for the other three blended metrics are quite similar.

From Figure 3.5, the strengths and limitations can clearly be observed of two individual metrics in the high-recall and high-precision case, respectively. The blended metric M_3 can perform well in both cases. Although the improvement of F score from the blended metric is not significant, it effectively avoids the significant performance drop of using only one individual metric in either high-recall or high-precision cases.

Figure 3.6 also compare the performance of the blended metrics with the those of the individual metrics for topics with (a) low, (b) medium, and (c) high novelty ratios, respectively. Again, the superior performance can be observed from the blended metrics in both high-recall and high-precision cases.

Figure 3.6: Blended metric M3 vs individual metrics for the topics with low (a-b), medium (c-d) and high (e-f) novelty ratio in TREC 2003 and 2004 Novelty Track data

Table 3.9: Average best F scores at different novelty ratios r_N . $\blacktriangle$ denotes the weak zones of novelty metrics.

data	metrics	best F score			
		r_N			overall
		low	medium	high	
TREC2003	cos	N.A.	0.7209 $\blacktriangle$	0.9006	0.8323
	nwr	N.A.	0.7214	0.8716 $\blacktriangle$	0.8145
	blended	N.A.	0.7267	0.9002	0.8343
TREC2004	cos	0.4854 $\blacktriangle$	0.6551 $\blacktriangle$	0.8253	0.6416
	nwr	0.5246	0.6615	0.8131 $\blacktriangle$	0.6496
	blended	0.5242	0.6657	0.8263	0.6538

Table 3.9 shows the average best F score at different novelty ratios r_N for TREC 2003 and 2004 Novelty Track data, respectively. From Table 3.9, the first observation is that the cosine similarity and new word ratio metrics have different “weak zones”, as denoted by $\blacktriangle$. The cosine similarity metric performs worst when the ratio of novel sentences is smaller than 0.3, while the new word ratio metric performs worst when the ratio of novel sentences is larger than 0.6. This observation shows that the “weak zones” of two single novelty metrics do not overlap.

Secondly, we can observe that the blended metric performs generally better than both cosine similarity and new word ratio metrics. The blended metric obtains the highest F scores, 0.8343 and 0.6538 in TREC 2003 and 2004 Novelty Track data, respectively. The second highest F scores are 0.8323 (cosine similarity in TREC 2003) and 0.6496 (new word ratio novelty metric in TREC 2004). These results prove that different types of single metrics which have their own pitfalls can benefit from fusing the information from each other. This is especially beneficial for real-time detection of novel information, where prior novel rates cannot be estimated.

Fig. 3.7 integrates the performance of three novelty metrics over all 100

Figure 3.7: Average best F scores (above) and the corresponding number of topics (below) in 10 different levels of ratios of novel sentences in TREC 2003 and TREC 2004 Novelty Track data.

topics in TREC 2003 and TREC 2004 Novelty Track data, at 10 different levels of ratios of novel sentences. From this figure, it is clear that the F score of all three metrics are highly correlated with r_N . It can also be observed that cosine similarity performs worst when the ratio of novel sentences is smaller than 0.5. On the contrary, new word ratio performs worst when the ratio of novel sentences is greater than 0.5. The blended metric performs well at all levels of r_N . These results also prove that the “weak zones” for cosine similarity and new word ratio do not overlap, and that the blended metric can utilize the merits of different types of novelty metrics.

Table 3.10 shows the recall and precision for these three metrics in different levels of ratios of novel sentences. One significant phenomenon is that the best

Table 3.10: Average recall and precision corresponding to the best F score (as reported in Table 3.9) at different novelty ratios r_N .

data	metrics	recall				precision			
		r_N			overall	r_N			overall
		low	medium	high		low	medium	high	
TREC2003	cos	N.A.	0.9365	0.9806	0.9638	N.A.	0.5907	0.8378	0.7439
	nwr	N.A.	0.9137	0.9010	0.9058	N.A.	0.6021	0.8507	0.7562
	blended	N.A.	0.9332	0.9750	0.9591	N.A.	0.6033	0.8411	0.7507
TREC2004	cos	0.7980	0.9225	0.9815	0.9047	0.3704	0.5159	0.7175	0.5110
	nwr	0.6869	0.8668	0.9122	0.8327	0.4328	0.5465	0.7389	0.5446
	blended	0.6965	0.8935	0.9566	0.8578	0.4328	0.5392	0.7332	0.5391

Table 3.11: Average number of retrieved sentences at different novelty ratios r_N .

data	metrics	number of retrieved sentences			
		r_N			overall
		low	medium	high	
TREC2003	cos	N.A.	200.84	298.35	261.30
	nwr	N.A.	189.53	259.71	233.04
	blended	N.A.	198.53	294.87	258.26
TREC2004	cos	91.20	129.56	137.83	122.88
	nwr	61.18	117.18	124.17	105.70
	blended	62.27	119.67	131.17	108.42

F score of cosine similarity corresponds to high recall and low precision. Also, the cosine similarity metric retrieves more sentences than the other two metrics especially when the novelty ratio is low, as shown in Table 3.11. This implies that the cosine similarity metric may lead in the situation where high recall is required, such as the case where missing out important information is especially detrimental. On the other hand, the new word ratio metric may outperform in the situation where high precision is required.

For a real-time novel sentence mining system that needs to satisfy different users’ requirements, the flexibility of the system to adapt to different requirements is especially critical. For example, when users do not want to miss any novel information, a high-recall system which may retrieve more sentences is de-

sired. When users want to read the most novel sentences first, a high-precision system is preferred. It is well known that there is a tradeoff between recall and precision, which can be adjusted by a predefined decision for novelty scores. However, the decision can only adjust recall-precision tradeoff externally. While the intrinsic flexibility of novelty metric decides the adjustable ranges of recall and precision. For example, Section 3.5.3 showed that cosine similarity is prone to retrieve more sentences, and hence may result in a system exhibiting high recall and low precision, which may not satisfy a precision-oriented requirement well.

For the precision, the range of values for the precision of the cosine similarity metric is much narrower than those of new word ratio and blended metrics. For the recall, the range of values of the new word ratio is smaller than those of cosine similarity and blended metrics. The maximum recall obtained by the new word ratio is around 0.93, while the maximum recall obtained by the other two metrics can approach 0.99. The blended metric has a wider range of values for both recall and precision. In addition, the F-Scores achieved by the blended metric can surpass the highest F-Scores obtained in the TREC Novelty Tracks. Furthermore, this new framework of novelty measurement has practical significance for the real-time novelty mining algorithm where no training data is available beforehand. By using this framework which is the mixture of two types of metrics, the real-time novelty mining performs well under different requirements and user's preference without requiring any prior information from data.

Figure 3.8: Precision-Recall-F (PRF) curve

3.5.5 Evaluation of Metrics on Business Blogs

We used the novelty BizBlogs dataset [39], a novelty annotated version of the BizBlogs07 [14] dataset of business blogs, which contains 1269 blog posts in four categories: product, company, marketing and finance. Refer to Section 3.5.1 for more details on this dataset.

To see how well our algorithm detects novel blogs, we calculated the precision (P) and recall (R) for different decision points. The Precision-Recall-F (PRF) curve [97] for three different metrics (cosine, blended, and new word count) is shown in Fig. 3.8. In the graph, the F score is plotted according to its P and R components. The grey dashed lines show contours at intervals of 0.1 points of F score.

Figure 3.9: Redundancy-PRF curve

From the results, we can observe that our novelty mining algorithm can detect novel blogs with extremely high precision, recall, and F-score. Cosine and blended metrics performed better overall than the new word count metric.

To see how well our algorithm detects redundant (or non-novel) blogs, we calculated the redundancy-precision (RP) and redundancy-recall (RR) for different decision points.

The R - PRF curve is shown in Fig. 3.9. In the graph, the RF score is plotted according to its RP and RR components. The grey dashed lines show contours at intervals of 0.1 points of RF score. From this graph, we observe that the decision point that gives the optimal RF -score is 0.65, which happens to be the same optimal decision point that maximizes the F -score.

3.5.6 Effects of Weight Variation on Blended Metrics

Figure 3.10: Effect of weight variation ($\alpha=0.1, 0.2$ to 0.9) on the performance of blended metrics on 100 topics from TREC 2003 and 2004 Novelty Track data.

To test the performance of the blended metrics in the most general case, equal weights of 50%, i.e. $\alpha = 0.5$ are used for the pair of individual metrics. To better understand the effects of weight variation on blended metrics, various weights were tried where α is set at a range from 0.1 to 0.9 with an equal step length 0.1.

The upper left graph in Figure 3.10 clearly showed that using new word count metrics helps to increase the average F score of the cosine similarity metric when assigning similar weights for the pair of individual metrics. When a low weight smaller than 0.3 is assigned to cosine similarity, the average F score will

decrease gradually. The lower left graph in Figure 3.10 showed that when equal or higher weight is assigned to Jaccard similarity metric, integrating the new word count metric is very helpful to improve the average F score. Again, the average F score will decrease gradually in terms of a weight smaller than 0.3.

By investigating the effects of weight variation on blended metrics, two useful observations can be summarized.

- (i) An equal weight assigned to cosine similarity and a higher weight assigned to Jaccard similarity can obtain the largest improvement using the blended metrics.
- (ii) Although the new word count metric obtained the lowest macro-averaged F score, integrating the new word count metric will be more helpful for the blended metrics. This may be because the new word count metric employs one-to-multiple sentence comparison. The other asymmetric metric, overlap metric, uses one-to-one sentence comparison, which is the same with the symmetric metrics cosine and Jaccard similarities. The more metrics used in the new framework, the better the performance.

3.5.7 Comparison of D-NM with D2S

For sentence segmentation, we utilized Sentence-boundary (see Section 2.2.1), and the novelty metrics used were cosine, new word count, and blended metrics.

We compared the performance of original document-level NM (D-NM) with that of D2S. D-NM is based on calculating the cosine similarity between documents, where the term frequency of each word in the document represents one feature in the document vector. For D-NM, we choose a series of novelty score decision points, varying from 0.05 to 0.65 with the step of 0.1. The greater the

Figure 3.11: R-PR curves for document-level NM on APWSJ: compare D-NM and D2S. The grey dashed lines show contours at intervals of 0.1 points of RF

novelty score, the higher probability the document is novel. For D2S, we set a fixed novelty score decision point $nt = 0.05$ on the sentence-level NM. Then, given different $NovelRate_s$ decision points, varying from 0.4 to 1.0 with a step of 0.1, we show the R - PR curves in Fig. 3.11.

From Fig. 3.11, we notice D2S has a higher value of RR and RP than D-NM on several $NovelRate_s$ decision points. Using a $NovelRate_s$ decision point of 0.95, the best RP and RR pairs are higher than the cosine benchmark reported by [95]. This indicates that D2S is effective in detecting the novel documents for APWSJ. Using a $NovelRate_s$ decision point of 1.0 means that the document is redundant if all the sentences are redundant. Thus, RR cannot be equal to 1 unless all sentences in the document are redundant. The NM performance is

improved when we segment the document first and apply sentence-level NM to judge the novelty of a document based on $NovelRate_s$. However, how is the performance of D2S affected when using the blended mixed metrics? In the following discussion, we analyze this situation.

3.5.8 Comparison of D-NM with D2S using Blended Metrics

The previous experiments compared the results using the cosine distance metric. The following experiments also analyzed the usage of blended metrics for document-level novelty mining and D2S. As seen from Fig. 3.12, D2S shows a good capability to find redundant documents using the blended metric if the goal is optimizing redundancy F-Score (RF) and redundancy-precision (RP) (as defined in Equations 2.17 and 2.15, respectively).

However, does D2S always show a better performance than D-NM? In the following discussion, we will answer this question and report the applicability of D2S.

3.5.9 Usage of D2S with Different Percentages of Novel Documents

Through setting several $NovelRate_a$ decision points, we can obtain datasets with different percentages of novel documents. Then, we compare the performance of D2S with D-NM when setting different $NovelRate_a$ decision points on document-level TREC 2004 and 2003. On D-NM, novelty decision points were selected from 0.05 to 0.95, with an equal step of 0.1. On D2S, the novelty

Figure 3.12: R-PR curves for document-level NM on TIPSTER: compare D2S and D-NM using blended metrics.

decision point for sentence-level NM was 0.45. The $NovelRate_s$ decision points for D2S were varying from 0.1 to 1.0 with a step of 0.1.

Fig. 3.13 shows the PR curve between D2S and D-NM when setting two $NovelRate_a$ decision points, 0.1 and 0.9 on the document-level TREC 2004. The reason that we used PR curves instead of $R-PR$ curves to evaluate document-level TREC data is because there are not so many novel documents as in AP-WSJ. Therefore, for document-level TREC data, it is fair to use PR curve to evaluate the document-level NM performance.

When setting $NovelRate_a$ decision point=0.1, we obtain a dataset with a high percentage of novel documents, where most of the documents are labeled to be novel. From this figure, we can observe that D2S shows a good perfor-

Figure 3.13: PR curves for document-level NM on document-level TREC 2004: compare D2S and D-NM.

mance when the percentage of novel documents in one dataset is high. It has been proved in the above experiments on APWSJ that we can achieve a better performance using D2S because APWSJ is a dataset with a high percentage of novel documents. However, D2S is not suitable for the dataset with a low percentage of novel documents. Moreover, we notice that the D-NM achieves a better performance when recall is low. Therefore, for low recall cases, it is better to use D-NM.

Furthermore, we performed the same experiments on document-level TREC 2003 Novelty Track. We can get same conclusions as those reported on TREC 2004 (See Fig. 3.14). Furthermore, we notice that D2S is better at finding novel information when there are many overlaps between texts, such as news sets like APWSJ, TREC 2004, because usually later text will reuse some sentences from the former text. Table 3.12 shows the percentage of same sentences in each

Figure 3.14: PR curves for document-level NM on document-level TREC 2003: compare D2S and D-NM.

Table 3.12: Percentage of Overlap on the Sentence Level

Dataset	Percentage of Overlap Sentences
APWSJ	12.99%
Document-level TREC 2004	13.96%
Document-level TREC 2003	13.72%

dataset after document segmentation. We learn that for these three datasets, an average of more than 10% of the sentences are the same. Therefore, it is recommended to use D2S when there are some overlap sentences between text.

Moreover, in order to check the consistency between R - PR and PR curve, we also show the R - PR curve on the document-level TREC 2004 and TREC 2003 Novelty Track. As seen from Fig. 3.15 and Fig. 3.16, it is learnt that D2S shows a good capability to find redundant documents no matter whether the percentage of novel documents in one dataset is high or not.

In conclusion, we find that it is good to use D2S when the dataset has a

Figure 3.15: R-PR curves for document-level NM on document-level TREC 2004: compare D2S and D-NM.

Figure 3.16: R-PR curves for document-level NM on document-level TREC 2003: compare D2S and D-NM.

high percentage of novel documents. In addition, D2S is more appropriate to be utilized in the dataset where there are many overlaps between texts, such as news. Finally, D2S is better at finding redundant documents than retrieving novel documents.

3.6 Chapter Summary

In this chapter, 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). The performance of novelty metrics was influenced by both explicit factors, i.e. the performance requirements, and implicit factors, i.e. the novelty ratios of data. Considering the effects from these two factors, complementary behavior was observed in the symmetric and asymmetric novelty metrics.

Then, a new framework of novelty measurement, a mixture of both types of novelty metrics, was introduced. The experimental results showed the superior performance of this new framework under different performance requirements and for data with different prior probabilities. The blended metric also achieved the highest F-scores as compared to the results from the TREC 2003 and 2004 Novelty Tracks. 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-time application such as novelty mining algorithms where no training data is available beforehand.

Another model for novelty scoring was proposed, which felicitously applied document-to-sentence annotations (D2S) for document-level novelty mining. This model can make document-level NM more effective by adopting the techniques for the sentence-level. Experiments on APWSJ show that the D2S method can greatly improve the document-level NM performance by achieving higher redundancy-precision and redundancy-recall scores as compared to the benchmark cosine method. Through applying experiments on document-level TREC 2004 and TREC 2003 datasets, we also find that it is good to use D2S when the percentage of novel documents is high.

Chapter 4

A Learning Model for Adaptive Novelty Decision

4.1 Introduction

As presented in Chapter 2, the second major component in novelty mining is novelty decision. A novelty *decision* point or value is defined as the boundary value that determines the cutoff point or threshold for novelty. One primary challenge in novelty mining is how to set the decision value of novelty scores adaptively, under different performance requirements varying from high-precision to high-recall. Since novelty mining is a real-time technique where little or no training information is available, the decision point cannot be pre-defined with confidence. However, it is possible to acquire some user feedback as time goes by, from which the information should be maximally utilized to meet the user's preferences.

Few studies have focused on adaptive novelty decision in terms of novelty mining area. Zhang et al. [95] proposed a simple novelty decision algorithm,

which increases the decision point a little bit if a novel document is retrieved as a non-novel one based on the user's feedback. Clearly it is a weak algorithm because it can only increase the decision point. This chapter addresses the problem of setting an adaptive decision point to meet different requirements.

This problem is related to the score distribution-based novelty decision algorithms that have been proposed for relevant document or sentence retrieval [6, 49, 88, 94]. Some ideas of score distribution modeling can be borrowed from the relevance retrieval area; however, the novelty score in novelty mining has its distinctive characteristics. The score distributions can be approximated by Gaussian distributions (detailed in Section 4.2.2). On the other hand, in relevance retrieval, the scores of non-relevant documents follow exponential distributions [6, 94]. This also implies that most of the non-relevant documents do not overlap much with relevant ones.

The proposed method, learning-based adaptive novelty decision (LAND) algorithm is a generic algorithm, which can be tuned according to different performance requirements, by employing different optimization criteria. For example, the F_β score (Eq. 4.6), which is the weighted harmonic average of precision and recall, where β controls the tradeoff between them.

Novelty mining combined with LAND has been tested on both document-level and sentence-level data and compared to novelty mining using various fixed decision points [63] and simple novelty decision [95]. The experimental results show that good performance of LAND can be obtained at both levels. In order to test LAND on a more practical level, different percentages of user's feedback were also used.

This chapter is organized as follows. Section 4.2 first analyzes factors that may affect the novelty decision in novelty mining, and then introduces

a learning-based adaptive novelty decision (LAND) algorithm. Section 4.3 tests LAND with both complete and partial feedback. Section 4.4 concludes the whole chapter.

4.2 Adaptive Novelty Decision in Novelty Mining

4.2.1 Motivations

There are several reasons motivating us to design an adaptive novelty decision algorithm real-time novelty mining. First of all, there is little or no training information in the initial stages. Therefore, a fixed decision point is difficult to predefine. The training information that is necessary for novelty decision includes the statistics of data and the users' reading habits. For example, a topic with 90% novel documents needs a relatively low decision point for novelty scores to retrieve most of the documents.

On the other hand, different users may have different definitions of "novel" information. For example, a user may regard a document with 50% novel information as a novel document while another user may only regard a document with 80% novel information as a novel document. The decision point of novelty scores should be higher for the user with a stricter definition for the "novel" document. As novelty mining is an accumulating technique, more training information will be available for novelty decision, based on the user's feedback given over time. The adaptive novelty decision algorithm is able to utilize these available information and makes novelty mining customized to the user's needs.

Satisfying different performance requirements is another important motiva-

tion to employ an adaptive novelty decision algorithm in novelty mining. When the users do not want to miss any novel information, a high-recall technique which only filters out the most redundant documents is desired. When users want to read the documents with the most significant novelty, a high-precision technique which only retrieves very novel documents is preferred. Therefore, the decision point should be tuned according to different performance requirements.

Next, the proposed method, Learning-Based Adaptive Novelty Decision (LAND), is introduced and described.

4.2.2 Learning-Based Adaptive Novelty Decision (LAND)

LAND is a score distribution-based novelty decision method. Therefore, the first important issue here is to investigate score distributions of both novel and non-novel documents. The score distributions of both classes provide the global information of data, from which an optimization criterion can be constructed for searching the optimal decision point. Next, these two issues will be introduced separately.

Novelty Score Distributions

Assume there are n training sentences, i.e. $s_1, s_2, \dots, s_n$, each of which belongs to either a novel class c_1 or a non-novel class c_0 . For any sentences s_i , $i = 1, 2, \dots, n$, the novelty score, i.e. x_i , can be estimated by some novelty metric such as the blended mixed metric as defined in Eq. (3.7).

To find the empirical novelty score distributions of data, some training datasets are needed. Here, topics N54 and N69 are used from TREC 2004 Novelty Track data [58], and all the novel and non-novel sentences are assumed

to be known beforehand. The following steps are processed on both training datasets separately.

step1: calculating the novelty scores x_i of each sentence s_i , $i = 1, 2, \dots, n$ using Eq. (3.1), where the history sentence list includes all the history novel sentences.

step2: the scores of each class are divided into several equal width bins with the bin width equals to $\frac{\max(scores) - \min(scores)}{\text{no. of bins}}$, where the no. of bins for any class c_k equals to the maximum integer smaller than $\frac{n_k}{10}$. Then, the number of sentences falling in l^{th} bin of k^{th} class can be obtained, denoted as $n_{k,l}$, where, $l = 1, 2, \dots, \text{no. of bins}$ and $k \in \{0, 1\}$.

step3: to obtain the empirical distributions of novelty scores, the number of sentences in each bin is normalized as follows.

$$p_e(x|c_k) = \frac{\text{no. of bins}}{n_k} \times n_{k,l} \quad (4.1)$$

where n_k and $n_{k,l}$ are the total number of sentences in the class c_k and the number of sentences falling in the l^{th} bin of class c_k , respectively. The empirical distributions of novelty scores for topic N54 and N69 are as shown in Figure 4.1 and Figure 4.2, respectively.

The Gaussian distribution (also called normal distribution) in the random variable X with mean μ and variance σ^2 has the probability density function ($p(x)$) as below.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} \quad (4.2)$$

Figure 4.1: Empirical and probability distribution approximation for TREC 2004 Novelty Track data topic N54.

Assuming that both novelty scores of the novel class and non-novel class follow the Gaussian distributions, for any class c_k , $k \in \{0, 1\}$, the maximum likelihood estimations for the Gaussian probability density function $p(x|c_k) \sim G(\mu_k, \sigma_k^2)$ is:

$$\mu_k = \frac{1}{n_k} \sum_{i \in c_k} x_i \quad (4.3)$$

$$\sigma_k^2 = \frac{1}{n_k} \sum_{i \in c_k} (x_i - \mu)^2 \quad (4.4)$$

The Gaussian probability density function is plotted by the dashed lines, as shown in Figure 4.1 and Figure 4.2. From the figures, the novelty scores can be fitted well using Gaussian distributions.

Optimization Criterion

Assume that there is an incoming sentence stream $s_1, s_2, \dots, s_n$, of which n_1 are novel. After filtering the sentence stream by novelty mining with a decision point θ , any sentence can be classified to one of four classes as shown in Table 4.1.

Figure 4.2: Empirical and probability distribution approximation for TREC 2004 Novelty Track data topic N69.

Table 4.1: Contingency table in novelty mining.

	novel	non-novel
retrieved	R_1	R_0
non-retrieved	N_1	N_0
total	n_1	n_0

The precision and recall can thus be written as:

$$\text{precision} = \frac{R_1}{R_1 + R_0} \quad \text{recall} = \frac{R_1}{n_1} \quad (4.5)$$

The F score (F1 score) is a special case of F_β score, i.e. the weighted harmonic average of precision and recall:

$$F_\beta = \frac{1}{\frac{\beta}{\text{precision}} + \frac{1-\beta}{\text{recall}}} \quad (4.6)$$

where the β is the parameter to control the weights of precision and recall.

The number of sentences in each class of the contingency table, R_1 , R_0 , N_1 and N_0 , are the functions of θ and can be approximated by the probability distributions of the novel and non-novel classes. For example, given a decision point θ , the estimation of $R_1(\theta)$ is proportional to the probability of novel

sentences with novelty scores greater than the θ , as below.

$$R_1(\theta) = n_1 \cdot P(x > \theta|c_1) = n_1 \cdot \int_{\theta}^{+\infty} p(x|c_1)dx \quad (4.7)$$

Similarly, other functions can be obtained:

$$R_0(\theta) = n_0 \cdot P(x > \theta|c_0) \quad (4.8)$$

$$N_1(\theta) = n_1 \cdot P(x < \theta|c_1)$$

$$N_0(\theta) = n_0 \cdot P(x < \theta|c_0)$$

Substituting Eq. (4.7) and Eq. (4.8) into Eq. (4.5), the precision and recall can be rewritten as the functions of the decision point θ , as follows.

$$\text{precision}(\theta) = \frac{P_{c_1}P(x > \theta|c_1)}{P_{c_1}P(x > \theta|c_1) + P_{c_0}P(x > \theta|c_0)} \quad (4.9)$$

$$\text{recall}(\theta) = P(x > \theta|c_1) \quad (4.10)$$

where P_{c_1} and P_{c_0} are the prior probabilities of the novel and non-novel classes and can be estimated as follows.

$$P_{c_1} = n_1/n \quad P_{c_0} = n_0/n \quad (4.11)$$

After obtaining the precision and recall as the functions of θ , the optimization criterion can be constructed for searching the best decision point. Substituting Eq. (4.9) and Eq. (4.10) into Eq. (4.6), the criterion $F_{\beta}(\theta)$ is obtained, whose

maximum value corresponds to the best decision point θ^* .

$$\begin{aligned}\theta^* &= \arg \max_{\theta} F_{\beta}(\theta) = \arg \max_{\theta} \frac{1}{\frac{\beta}{\text{precision}(\theta)} + \frac{1-\beta}{\text{recall}(\theta)}} \\ &= \arg \max_{\theta} \frac{P(x > \theta | c_1)}{\beta[P(x > \theta | c_1) + \frac{P_{c_0}}{P_{c_1}}P(x > \theta | c_0)] + (1-\beta)}\end{aligned}\quad (4.12)$$

LAND is a generic method that can be tuned automatically according to different performance requirements, by employing the different optimization criteria. The alternative evaluation measures, i.e. redundancy-precision (RP) and redundancy-recall (RR) are usually used when the ratio of novel documents is very high [95].

$$RP = \frac{N_0}{N_1 + N_0} \quad RR = \frac{N_0}{n_0} \quad (4.13)$$

The corresponding redundancy- F_{β} score, i.e. RF_{β} , is defined as:

$$RF_{\beta} = \frac{1}{\frac{\beta}{RP} + \frac{1-\beta}{RR}} \quad (4.14)$$

Similarly, the best decision point can be obtained that maximizes the RF_{β} as given in Eq. (4.15).

$$\begin{aligned}\theta^* &= \arg \max_{\theta} RF_{\beta}(\theta) \\ &= \arg \max_{\theta} \frac{1}{\frac{\beta}{RP(\theta)} + \frac{1-\beta}{RR(\theta)}} \\ &= \arg \max_{\theta} \frac{P(x < \theta | c_0)}{\beta[P(x < \theta | c_0) + \frac{P_{c_1}}{P_{c_0}}P(x < \theta | c_1)] + (1-\beta)}\end{aligned}\quad (4.15)$$

4.2.3 Implementation Issues

There are several implementation issues in LAND. The first issue is how LAND is combined with real-time novelty mining. Figure 4.3 shows the flowchart of novelty mining combined with LAND. After the method predicts the i^{th} sentence s_i , $i = 1, 2$, etc., it checks whether there is any new user's feedback for the current sentence or any history sentence. If there is any new feedback available, the current decision point will be updated by LAND. Finally, the method will use this newly updated decision point to predict the next incoming sentence.

The second implementation issue is how many feedback scores are necessary for the probability estimation in the initial stage of novelty mining. From the experiments with TREC2004 data, the minimum number of feedback scores n_{min} of both novel and non-novel sentences should not be less than 4. A smaller n_{min} will degrade the accuracy of probability estimation and lead to an unreliable probability model while a big n_{min} will not start the adaptive novelty decision until the method accumulates enough feedback. In the experimental study, $n_{min} = 4$ is set for both novel and non-novel sentences.

When the number of feedback scores does not meet the requirement of n_{min} , an initial decision point is necessary. Therefore, setting the initial decision point is another implementation issue. Due to the characteristics of novelty mining, there are more novel sentences in the early stage of accumulating sentences. Therefore, the initial decision point should be slightly lower, such that most of the sentences can be retrieved. As the accumulating sentences increases, the possible feedback also increases to trigger LAND. In the experiments, the initial decision point is set to $\theta_0 = 0.3$ for novelty scores.

4.3 Experimental Study

4.3.1 Datasets

Three public datasets, TREC 2004 Novelty Track data [58], TREC 2003 Novelty Track data [59], and APWSJ novelty mining data [95], are used in the experiments. Refer to Section 3.5.1 for the details of the data.

4.3.2 Results

When novelty mining adopts a fixed decision point, feedback is not considered and the whole novelty mining process is unsupervised. When novelty mining adopts an adaptive novelty decision algorithm, the decision point of novelty mining will be updated automatically once there is any new feedback from the user. In the experiments, this process is simulated by assuming the user's feedback for the current document, if any, is returned immediately after the prediction is made. The new feedback, as well as the novelty score given by the method, will be input into the adaptive novelty decision algorithm. Finally, the new decision point output by this algorithm will replace the current one and be used for future incoming documents. Notice that when no feedback is received, the method will not change the value of the decision point (fixed at the default initial value).

In this experimental study, LAND is tested by assuming (i) the complete and (ii) partial user's feedback. The first case is an ideal case. However, the 100% user's feedback cannot be guaranteed in real-world applications since feedback is very time-consuming, thus users do not like to give feedback. Therefore, LAND is also tested on the same datasets with partial feedback from random

sampling. This experiment simulates the performance of LAND in practice.

In this experimental study, the precision-recall (PR) curve is adopted to demonstrate the performance of different sentence-level novelty mining algorithms. In information retrieval, PR curves are usually used to compare algorithms, where the algorithm with larger area under the curve is regarded as a better algorithm [16]. For each graph, the F score is plotted according to its precision and recall components. The lines show contours at intervals of 0.1 points of F score. For example, Figure 4.4 illustrates PR curves of two methods along with the contours of the F score in the PR space. Given any point on the PR curve, the corresponding region of F score can be seen. Similarly, the redundancy- PR ($R-PR$) curve is adopted to demonstrate the performance of different document-level novelty mining algorithms. The contours of RF score is also plotted according to its RP and RR components. Using this graph, the show precision, recall and F score can be shown in one graph.

Case 1: Complete Novelty Feedback

In this experiment, LAND is compared with two benchmark novelty decision methods: (i) fixed decision points (ii) Zhang's simple novelty decision algorithm as described in [95]. The simple novelty decision algorithm for learning novelty decision point can be described as follows.

- Initialize redundancy decision point to a value that is so low that only very redundant documents are considered redundant; and
- For each document delivered as novel, ask the user if the document is novel. If the document is redundant and if $N(s_t) < N(s_i)$ for all s_i in the history list,

then the decision point $\theta = N(s_t)$
 else the decision point $\theta = \theta + \frac{N(s_t) - \theta}{10}$

The initial decision point is set for the simple novelty decision as 0.05, where the other values like 0.1, 0.2 will not largely influence the final result of this algorithm. One problem of this algorithm is that it only increases the decision point for novelty scores. Moreover, this algorithm cannot adjust the tradeoff between recall and precision based on performance requirements.

For novelty mining using fixed decision points, as shown in Figure 4.4, the corresponding *PR* curve (black line) is plotted by varying the fixed decision point from 0.05 to 0.95 with an equal step length 0.05. For each decision point, precision and recall for each topic are calculated and the average precision and recall over 50 topics are reported in the figure. For novelty mining with LAND, the *PR* curve is plotted by varying the parameter β of the optimization criterion F_β score from 0.2 to 0.9. Again, for each value β , precision and recall for each topic are calculated and the averaged precision and recall over 50 topics are reported in the figure.

From this figure, novelty mining with LAND outperforms novelty mining with fixed decision points. The precision and recall obtained by novelty mining with LAND will not fall in the regions of the extreme values, in which the F score can be very low. In practice, the users usually require a high-recall technique with an acceptable precision value, or a high-precision technique with an acceptable recall value. The extremely high recall with a extremely low precision makes little sense because this technique just marks almost all documents as novel. On the other side, the extremely high precision with an extremely low recall means the method only marks very few documents as novel, which is also less meaningful.

Moreover, since there is no prior information available for the user to choose a suitable fixed decision point, the method with a predefined decision point can hardly lead a suitable tradeoff between precision and recall, and hence can hardly obtain a good F score. On the contrary, LAND will optimize the F score automatically, based on the user's feedback overtime.

Although both the fixed decision point and LAND's parameter β control the tradeoff between precision and recall, they play different roles in novelty mining. For the fixed decision point, the quantitative relationship to precision-recall tradeoff is uncertain. Therefore, setting a fixed decision point may not accommodate promising results under different requirements. On the contrary, the parameter β in LAND reflects weights of precision and recall (β is the weight of precision while $1 - \beta$ is the weight for recall) in the optimization process, therefore, can be set directly to meet different performance requirements.

For the APWSJ data, redundancy-precision (RP) and redundancy-recall (RR) as defined in Eq. (2.15 and 2.16) were used, since the number of novel documents is far greater than the number of non-novel documents [95]. LAND in this case is tuned for RF_β as defined in Eq. (2.17), accordingly. Figure 4.5 shows the redundancy-PR ($R-PR$) curves of novelty mining with fixed decision point, simple novelty decision [95] and adaptive decision point by LAND. From this figure, similar experimental results can be observed with the first dataset.

Table 4.2: Comparison of performance evaluated by F_β ($\beta = 0.2, 0.5, 0.8$) on TREC 2004 Novelty Track data.

	Performance of novelty mining employing		
	adaptive decision by LAND (β)	best fixed decision (θ)	simple novelty decision
$F_{0.2}$	0.7706 (0.2)	0.7758 (0.15)	0.6469
$F_{0.5}$	0.6155 (0.5)	0.6126 (0.45)	0.5913
$F_{0.8}$	0.5396 (0.8)	0.5281 (0.60)	0.5525

Table 4.3: Comparison of performance evaluated by F_β ($\beta = 0.2, 0.5, 0.8$) on APWSJ data.

	Performance of novelty mining employing		
	adaptive decision by LAND (β)	best fixed decision (θ)	simple novelty decision
$RF_{0.2}$	0.4191 (0.2)	0.4416 (0.35)	0.3689
$RF_{0.5}$	0.3985 (0.5)	0.3955 (0.25)	0.2817
$RF_{0.8}$	0.4201 (0.8)	0.4499 (0.20)	0.2483

In order to investigate the abilities of all three algorithms to meet different performance requirements, they are evaluated on F_β score and RF_β according to different performance requirements, i.e. $\beta = 0.2, 0.5$ and 0.8 , as shown in Table 4.2 and Table 4.3. For novelty mining employing the LAND algorithm, the parameter β is set to $0.2, 0.5$ and 0.8 accordingly. For novelty mining employing the fixed decision point, the highest F_β scores are reported in the tables. From Table 4.2, by comparing with best fixed decision point, LAND can obtain a similar or slightly better results for TREC 2004 Novelty Track data. The best fixed decision points for $F_{0.2}$, $F_{0.5}$ and $F_{0.8}$ are $0.15, 0.45$ and 0.60 , respectively. By checking the PR curves in Figure 4.4, the corresponding region can also be found of the best fixed decision points is covered by the PR curve of LAND. This implies that LAND is effective to search the best decision point in novelty mining, under different performance requirements. Table 4.3 shows the comparison of three algorithms for APWSJ data using RF_β score. For this data, the performance of LAND is also close to the best fixed decision points. Again, the PR curve of LAND covers the corresponding region of the best fixed decision points in Figure 4.5.

Since the simple novelty decision algorithm [95] that is unable to adjust the tradeoff between precision and recall, the performance of this algorithm is much inferior to that of LAND under various performance requirements on

both sentence and document-level data as shown in Figure 4.4 and Figure 4.5, respectively.

Case 2: Partial Novelty Feedback

In the previous section, LAND has been tested by assuming that user feedbacks are provided for all documents. However, in the real world application, since assessing documents is time-consuming, the 100% user's feedback cannot be guaranteed.

The more practical case is to assume feedbacks from the user are available for partial documents. In order to simulate this situation, different percentages of documents by random sampling are selected. The selected documents are given the user's feedbacks, while those unselected documents are not given any feedback and will not be used to update the decision point.

For each topics, four different percentages of documents are selected, i.e. $P_{feedback} = 80\%$, 50% , 20% and 10% , being given feedbacks from the user. For each percentage, the algorithms are repeated 100 times and each time, a documents subset that will be given feedbacks is generated by random sampling. The experimental results reported in graphs and tables are the average values over these 100 runs.

Figure 4.6 and Figure 4.7 show the PR curves and $R-PR$ curves of novelty mining with LAND on TREC 2004 Novelty Track data and APWSJ data, respectively. From these figures, the performance of LAND algorithm on TREC 2004 Novelty Track data has no significant reduction with the decreasing percentage of user feedback; while the performance of LAND on APWSJ data is slightly reduced with the decreasing percentage of user feedback. These results

show that precision and recall of LAND are not necessary lower for partial feedback. Another observation is that the adjustable regions of precision and recall shrink with the decreasing percentage of user feedback. In the extreme case that there is no feedback at all, the PR curves will concentrate on one point, where the corresponding precision and recall are obtained by novelty mining with the initial decision point θ_0 .

Table 4.4: Performance of LAND with different percentages of the user's feedback, $P_{\text{feedback}} = 100\%$, 80% , 50% , 20% and 10% , on TREC 2004 Novelty Track data and APWSJ data.

P_{feedback}	F score/ RF score	
	TREC2004 data	APWSJ data
100%	0.6155	0.3985
80%	0.6139 ± 0.0017	0.3948 ± 0.0049
50%	0.6146 ± 0.0024	0.3899 ± 0.0043
20%	0.6152 ± 0.0022	0.3841 ± 0.0039
10%	0.6121 ± 0.0018	0.3837 ± 0.0026

Table 4.4 shows the F score and RF score for TREC 2004 Novelty Track data and APWSJ data, respectively. From this table, the reduction of the F score for TREC 2004 Novelty Track data is insignificant. For APWSJ data, the RF score decreases slightly with the decreasing percentage of user feedbacks. This experimental result shows that the LAND algorithm can work robustly when the percentage of user feedback decreases.

In the following sections, we test LAND by assuming complete feedback for document-level novelty mining data with low, medium, and high novelty ratios (refer to Section 3.5.1). This will provide us some guidelines on using LAND in practice.

High Novelty Ratio

To construct a document-level novelty mining data with high novelty ratio, we choose TREC 2003 Novelty Track data because the ground truth novelty ratio at sentence level is naturally high (65.73%). As set the PNS decision point to 0.25, the document-level novelty ratio is 79.20%, i.e. 79.20% incoming documents are novel. In this case, we found the result of LAND is not as good as the best result of the fixed decision point, as shown in Figure 4.8.

Medium Novelty Ratio (30%-75%)

We construct a document-level novelty mining data with medium novelty ratio by using TREC 2004 Novelty Track data because the ground truth novelty ratio at sentence level is 41.40%. As set the PNS decision point to 0.03, the document-level novelty ratio is 47.73%, i.e. 47.73% incoming documents are novel. In this case, the result of LAND is comparable with the best result of the fixed decision point, as shown in Figure 4.9.

Low Novelty Ratio (30%)

We also construct a document-level novelty mining data with low novelty ratio by using TREC 2004 Novelty Track data. As we set the PNS decision point to 0.5, the document-level novelty ratio is 27.71%, i.e. 27.71% incoming documents are novel. In this case, the LAND algorithm outperforms the best result of the fixed decision point, as shown in Figure 4.10.

Discussion

Although both the fixed decision point and the LAND's parameter β control the tradeoff between precision and recall, they play different roles in novelty mining. The fixed decision point cannot reflect the tradeoff between precision and recall directly. Due to different data may have different characteristics and different metrics may output different values of novelty scores, the fixed decision point can hardly be predefined with confidence. On the contrary, the parameter β in LAND reflects the weights of precision and recall directly (β is the weight of precision while $1 - \beta$ is the weight for recall), and hence can be set based on the performance requirement directly.

From our experimental results on document-level novelty mining data with low, medium and high novelty ratios, we find LAND is extremely useful for the data with low novelty ratios, useful for the data with medium novelty ratios, while work not as good as the best fixed decision point for the data with a novelty ratio higher than 75%. Therefore, LAND is not recommended for the topic with high novelty ratio. In this case, setting a lower fixed decision point to let most of the document to be “novel” could be a better choice.

4.3.3 Working Example

To illustrate the usage of LAND in practice, we show a working example of using LAND for sentence-level novelty mining. Consider the following sentences from topic N39 in TREC 2003:

1: CLUES POINT TO PHILIPPINE STUDENT AS VIRUS AUTHOR By
JOHN MARKOFF c.2000 N.Y. Times News Service

- 2: Law enforcement officials and computer security investigators focused on the Philippines Friday in their search for the author of a software program that convulsed the world's computer networks.
- 3: Investigators in both Asia and the United States said clues appeared to point to a college student in his early 20s using a Philippine Internet service provider.
- 4: The rogue program, borne as an attachment to an e-mail with the subject line "I Love You," surfaced in Asia on Wednesday.
- 5: It moved from there to Europe and the United States on Thursday, clogging or disabling corporate e-mail systems and destroying data on personal computers.
- 6: Although the spread of the infection appeared to slow Friday, at least eight variations of the original program had been identified by antivirus firms.
- 7: Once it is launched, the "I Love You" program, among other things, tries to fetch an additional program from a Philippine Web site enabling it to steal passwords from the victim's computer.
- 8: American security experts said they had found evidence that a person using the "spyder" alias found in the "I Love You" program had written two versions of a password-stealing program found in recent months.
- 9: "Our theory is that he had written this program twice and was looking for a way to get broader distribution for it," said Peter S. Tippett of ICSA.net, a computer security firm based in Reston, Va.

- 10: At the same time, Fredrik Bjorck, a Swedish computer security researcher who last year helped identify the author of a similar program called Melissa, told Swedish television that he had identified the perpetrator of the latest attack as a German exchange student named Mikael.
- 11: He said that Mikael was in his 20s and that he had used Internet service providers in the Philippines to spread his programs.
- 12: Bjorck said Mikael had published information on how to get rid of the “I Love You” program.
- 13: He did not identify Mikael’s location.
- 14: The ICISA.net researchers said they had disassembled one of the four components of the “I Love You” program and had discovered that its instructions closely matched two similar programs that they had captured last fall and in January.
- 15: Once a computer was infected, the program was set up to fetch the password-stealing component from a Philippine Web site.
- 16: After it was installed in the computer it was programmed to relay the stolen passwords to an e-mail account also in the Philippines.
- 17: But after the “I Love You” outbreak was detected on Wednesday, the company running the Philippine Web site, Sky Internet, quickly removed the password program from its system.
- 18: Computer investigators said that both the “I Love You” program and the password-stealing modules discovered earlier had references to Amable Mendoza Aguiluz Computer College, which they said had seven campuses

in the Philippines.

When sentence-level novelty mining is used with a fixed decision point of 0.55, Sentences 9, 11 and 15 are determined as non-novel, as shown in Figure 4.11 ('X' is Non-Novel, Check is Novel). If we provide feedback as shown and process the sentences again using the LAND option, then the decision point will be automatically adjusted. We set the feedback of "1" for "Novel" for sentences 2-5, and "0" for "Non-Novel" for sentences 8, 9, 11, and 15, as shown in Figure 4.11. In this scenario, after running LAND based on the user feedback, the decision point was automatically adjusted to 0.60 for sentences 16 and beyond, as shown in Figure 4.12. In this figure, Sentence 16 was compared with the most similar sentence, in this case Sentence 15, and because the novelty score of 0.5980 is below the decision point value, Sentence 16 is now rated as "Novel". As seen in Figure 4.13, the resulting novelty rating changed for the sentences 16, 17, and 18, from "Novel" to "Non-Novel", based on the decision point adjustment from the user feedback. This example shows how LAND works for a real-life scenario. For more details of the implemented novelty mining applications, refer to Appendix A.

4.4 Chapter Summary

This chapter addressed the problem of setting an adaptive decision point by utilizing the user's feedback over time. First, the score distributions of both novel and non-novel classes are analyzed and found to both follow the Gaussian distributions. This facilitated the score modeling in novelty mining. Then, a Learning-based Adaptive Novelty Decision (LAND) algorithm was proposed.

LAND is a generic method, which can be tuned according to different performance requirements varying from high-precision to high-recall, by combining with different optimization criteria. In this chapter, the most commonly used performance evaluation measure in novelty mining, F_β score, was employed as the optimization criterion. F_β score is the weighted harmonic average of precision and recall, where β and $(1-\beta)$ are weights for precision and recall, respectively. Thus, the performance requirement, which is related to the tradeoff between precision and recall, can be satisfied by LAND with an explicitly set parameter β .

In the experimental study, novelty mining employing LAND has been tested on both document-level and sentence-level novelty mining data. With complete user feedback, the experimental results showed the promising performance of LAND for novelty decision for real-time novelty mining. In order to test LAND on a more practical level, the algorithm was tested with partial feedback. With a decreasing percentage of user feedback, the F score of LAND on TREC 2004 Novelty Track data had no significant reduction, while the RF score of LAND on APWSJ data was slightly reduced. These experimental results indicate that LAND can work robustly with partial feedback from the user. LAND 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 LAND 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). The experimental results show LAND is very effective in finding the best decision point in the novelty mining algorithm. Moreover, LAND is able to meet the different performance requirements by

setting the weights of precision and recall externally.

Our findings also indicate that LAND is extremely effective for data with a low novelty ratio, useful for data with a medium novelty ratio, and not as effective for data with a high novelty ratio. This is consistent with our previous findings that the novelty ratio, or percentage of novel documents/sentences, of the data is one of the most important factors that determine which novelty metric and novelty decision point to select.

Further enhancements to the algorithms presented in this chapter include applying other learning algorithms for analysis of data input and description of output. The algorithm can automatically learn a function from training data, such that the function can sort documents according to their degrees of relevance, preference, or novelty as defined in a specific application. Examples of possible learning techniques that can be used include conventional approaches such as Least Square Retrieval Function, pointwise approaches such as Constraint Ordinal Regression, pairwise approaches such as RankBoost, and listwise approaches such as RankCosine.

Figure 4.3: Novelty mining combined with LAND

Figure 4.4: Sentence-level novelty mining on TREC 2004 Novelty Track data: PR curves of novelty mining with fixed decision point, simple decision and adaptive decision by LAND (tuned for F_β) with complete user’s feedback.

Figure 4.5: Document-level novelty mining on APWSJ data: R - PR curves of novelty mining with fixed decision point, simple decision point and adaptive decision point by LAND (tuned for RF_β) with complete user’s feedback.

Figure 4.6: Sentence-level novelty mining on TREC 2004 Novelty Track data: PR curves of LAND (tuned for F_β) with different percentages of the user's feedback.

Figure 4.7: Document-level novelty mining on APWSJ data: $R-PR$ curves of LAND (tuned for RF_β) with different percentages of the user's feedback.

Figure 4.8: Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_{β}) with complete user’s feedback on document-level TREC 2003 Novelty Track data (with PNS decision point 0.25).

Figure 4.9: Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_{β}) with complete user's feedback on document-level TREC 2004 Novelty Track data (with PNS decision point 0.03).

Figure 4.10: Precision-recall curves of novelty mining with fixed decision vs adaptive decision by LAND (tuning for F_β) with complete user’s feedback on document-level TREC 2004 Novelty Track data (with PNS decision point 0.5).

Figure 4.11: Sentence-level novelty mining results for TREC03 topic N39.

Figure 4.12: Decision adjustment to 0.6000 for Sentence 16 after running LAND.

Figure 4.13: Sentence-level novelty mining results for TREC03 topic N39 after running LAND.

Chapter 5

Contextual Information Annotations for Novelty Feedback

5.1 Introduction

As presented in Chapter 2, the third major component in novelty mining is performance evaluation. Evaluating the performance of predefined data such as the TREC Novelty Track data is simple as it uses the novelty judgments rated by assessors. However in a real-world scenario where different users have different definitions of novelty, adapting the definition of novelty to the context of different users is not a trivial matter. Traditional novelty mining techniques cannot detect novelty in a sentence that looks similar to its history sentences, but contains information that is useful in the user's current context, such as an analyst interested in price fluctuations of oil. Novelty mining for highly similar sentences was never addressed in previous studies.

This chapter attempts to bridge a gap between the explicit user's preference and the advance of contextual information, via novelty feedback. By studying the target users, named entities are found to provide a good level of description for novel information. Accordingly, use cases of novelty feedback and different usage scenarios are analyzed. Contextual information annotation of named entities (CIAN) is then proposed to integrate contextual information from the user's novelty feedback.

The main contributions of this chapter are as follows:

- 1) We 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.
- 2) The original technique, contextual information annotation of named entities (CIAN) algorithm, is proposed to solve the issue of detecting novelty in highly similar sentences, and was developed to interact with the user's novelty feedback.
- 3) This is the first study to use the linguistic-level information, named entities of number, time, location, and person, to accommodate the user's context in novelty mining through feedback of his/her preferred information.
- 4) The CIAN algorithm has been successfully implemented and tested on real data, with results indicating that the method can solve the important problem of mining the novel information in highly similar sentences at the linguistic level.

This chapter is organized as follows. Section 5.2 defines the concept of

context for novelty mining. Section 5.3 analyzes the user context in the novelty mining algorithm and the corresponding use case and scenario for novelty feedback. Section 5.4 introduces a contextual information annotation of named entities (CIAN) algorithm to utilize the acquired contextual information from novelty feedback. The experimental study is presented in Section 5.5. Some concluding remarks are given in Section 5.6.

5.2 Definition of Context

Recently, there has been increased interests in exploiting contextual information for data mining. This has led to research in the use of context in adaptive, interactive, personalized or collaborative systems, the design of explicit and implicit feedback techniques, the investigation of relevance, the application of a notion of context to problems like advertising or mobile search. 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. In data pre-processing, enriching the data with additional contextual information can lead to the generation of more meaningful patterns. Also, 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.

In addition, novelty mining can be applied in a number of different domains, such as bioinformatics, business intelligence, finance data analysis, location based services, network data analysis, and social link analysis. As a consequence, context-aware novelty mining can benefit research efforts from

different fields.

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 [58, 59], 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). This assumption may not hold for many users, and therefore, the previous methods or some recommended parameters optimized based on F score may not be suitable for real-life enterprise applications.

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

1. **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.

2. **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 [58], this 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). This assumption may not work for the most of the users, but can be used if the user has no idea about his/her desired system performance.
3. **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.

5.3 Novelty Feedback

In previous studies, contextual information accessed through relevance feedback has been successfully utilized for query expansion in relevant information retrieval [42, 51, 55]. In this case, user may either provide the relevance feedback for documents/passages [51, 82, 89] where the useful terms are extracted, or select the key terms directly [61].

In novelty mining, however, accessing the context information through user feedback has not been explored too much [62]. One possible reason is that it is much more difficult to describe the novel information using key words. In relevant document/passage retrieval, if “football” is selected as a key word by the user, documents/passages with “football” are more likely to be retrieved by a search engine. But in novelty mining, the user cannot always use “football” as novel term because any term becomes somewhat non-novel after its first appearance.

The target users are interested in two kinds of novel sentences. One is the highly novel sentences that have never been seen before. Detecting this kind of novel sentence is relatively easy and the bag-of-words models can perform well. The other kind of novel sentences may look similar to its history sentences but also be desired by the user because of the significant novel information given by one or two words. Novelty mining for highly similar sentences has never been addressed in previous studies. For example, Tom has defined a topic as “Oil price”. For this topic, Tom may care more about different numbers. Given a pair of sequential sentences s_1 and s_2 , s_2 contains novel information just because it consists of a new number, i.e. 48.

Example 1:

s_1 : Now the oil price is US\$50/barrel.

s_2 : Now the oil price is US\$48/barrel.

Another example from the user Annie, who has created a topic as “Schedule of meetings” and is interested in the time and date of different meetings, as shown in Example 2 below. In this case, s_2 is also novel due to the new date.

Example 2:

s_1 : The meeting will start at 10am, 2 April.

s_2 : The meeting will be at 10am, 2 May.

From these two examples, it is observed that the linguistic-level information, such as number, time, location and person, reflect the user’s context and potentially interest the user. That is because these linguistic entities show very natural aspects of significant information. Therefore, the individual or combinations of different linguistic entities can be used to describe the user’s contextual information and feedback.

The use case scenarios of novelty feedback are shown in Figure 5.1. Two scenarios are described for the users who have different levels of insights about their topics. In the scenario A, the user is assumed to be quite clear about their topics. In the scenario A, the user is assumed to be quite clear about his/her target novel information, such as Tom and Annie. Therefore, the contextual information can be feedback by directly selecting from the GUI as shown in Figure 5.2. In scenario B, the user is assumed to be unclear about his/her preferences until actually seeing some documents/sentences. In this case, the contextual information will be extracted automatically from the novel text feedback by the user.

Use case: Novelty feedback in Novelty Mining
Actor: A user of the system
Preconditions: The user has defined a topic

Scenario A:

- 1. The user selects “novelty feedback” on the GUI.
- 2. The GUI provides four categories of preferred novel information.
- 3. The user ticks one or more categories, for example, “number” and “location”.
- 4. The user inserts a set of documents.
- 5. The system will highlight the novel sentences and also bold the new information (such as number and location), based on the novelty feedback from the user.

Scenario B:

- 1. The user selects a piece of text in a document, string X, and feedback as desired novel information.
- 2. The system annotates the string X automatically, and then updates the activated categories accordingly.
- 3. The system will highlight the novel sentences and also bold the activated categories, based on the novelty feedback from the user.

Figure 5.1: Use Case of Novelty Feedback

Figure 5.2: Snapshot of GUI for additional contextual information

5.4 A Contextual Annotation-based Algorithm for Novelty Feedback

5.4.1 Basic Idea for CIAN

After extracting the contextual information from the user’s feedback, the next task is detecting the novel sentences by considering the contextual information. From the previous section, the novel sentences desired by the user can be broadly

Figure 5.3: A contextual annotation-based novelty mining algorithm with the user feedback of “number” and “location”.

divided into high novelty (based on bag-of-words) and low novelty (need to be detected at linguistic level). Therefore, the basic idea of the algorithm, contextual information annotation of named entities (CIAN), is to construct a cascading framework: the original novelty mining algorithm detects the high novelty at the first layer and in the second layer, the CIAN algorithm detects the low novelty based on linguistic information, as shown in Figure 5.3. In this algorithm, any new incoming sentence s_t will be predicted in 4 steps, i.e.

- step 1: Record the corresponding context information specified by the user, if any.
- step 2: Run the original novelty mining on s_t . Go to step 3 if the novelty mining prediction is “non-novel”, step 4 otherwise.
- step 3: Run CIAN on s_t . Go to step 4 if the CIAN prediction is “novel”.
- step 4: Predict s_t as novel.

The named entities in sentences are identified by a commercial named entity extractor [43]. Four types of entities are extracted including “number”, “loca-

tion”, “time” and “person”. Unlike the other three types of named entities, numbers themselves are not that meaningful. Therefore, the first word after the number is also accessed by the CIAN algorithm using part-of-speech (POS) tagging [12]. Only the number that appears in conjunction with a noun (e.g. player, mile, etc.) or cardinal number (e.g. million, billion, etc.) is recorded as a *number pattern*. For example, “20 people”, “3 million”, “45 people” will be recorded and compared in CIAN, instead of numbers themselves like “20”, “3” and “45”. Other named entities such as time, location, person show enough information by themselves and no additional information will be appended. For example, “Singapore”, “London” and “Liverpool” clearly show three different locations and some sentence is prone to be novel due to some new location that has never appeared before.

Contextual Rules in CIAN

CIAN compares the incoming sentence with its history sentences, and makes the final prediction based on a set of rules related to each kind of named entities, i.e. Number/Location/Time/Person (N/L/T/P). There are two strategies in sentence comparison: one-to-one and one-to-all. Novelty mining with one-to-one strategy compares the incoming sentence s_t with its history sentences one by one, and predicts the novelty of s_t based on its nearest neighbor sentence. Novelty mining with one-to-all strategy predicts the novelty of s_t by comparing it to the combination of all the history sentences. Two sets of rules are implemented for CIAN as below.

- *Rule-A using one-to-one strategy:*

the nearest sentence pair s_1 and s_2 have different N/L/T/P

$\Rightarrow s_2$ is novel

- *Rule-B using one-to-all strategy:*

satisfy rule-A AND s_2 has different N/L/T/P compared with the history list of N/L/T/P $\Rightarrow s_2$ is novel

Rule-B using the one-to-all strategy always performs better for detecting linguistic-level novelty in the CIAN algorithm. This indicates that one-to-all sentence comparison may be more effective for detecting the linguistic-level novelty in highly similar sentences. For example, suppose s_1 and s_2 in Example 2 are history sentences, one-to-all strategy combines the history sentences together: “the monthly meeting will be at 10am, 2 April, 2 May”, which is a proper summarization of history. However, for detecting the highly novel sentence using bag-of-words models, one-to-all strategy performs badly. That is because combining the highly novel sentences may not be reasonable. For example, two highly novel history sentences: “Tom will visit Singapore”, “Annie will visit England” will be combined to “Tom Annie will visit Singapore England”, which is quite misleading. On the other hand, the one-to-all comparison strategy conducts a tighter criterion for novel sentence retrieval, which is more suitable for highly similar sentences. Figure 5.4 shows how the CIAN algorithm works by using one-to-all sentence comparison strategy.

Searching Pool of CIAN

CIAN attempts to retrieve the novel information hidden in the highly similar sentences which are missed by the original NM algorithm, as shown in Figure 5.3. To construct a searching pool of CIAN, a simple way is to search all the sentences predicted as “non-novel” by the original NM. However, this searching pool is fixed and unable to adjust the trade-off between precision and recall for different users’ needs.

Figure 5.4: Framework of the CIAN algorithm using one-to-all sentence comparison

Instead of a fixed searching pool for CIAN, a soft searching pool is defined, by introducing the width of the searching pool, i.e. δ . The sentences with a novelty score from $(\theta - \delta)$ to θ (θ is the novelty decision point in the original NM) are selected into the pool. With the increasing value of δ , a larger pool of sentences are formed and then processed by the CIAN algorithm. From the physical point of view, the width of the searching pool reflects how much the user does not want to miss any novel information. Different parameter settings in CIAN will be tested in Section 5.5.

5.5 Experimental Study

TREC 2004 Novelty Track data was used as the experimental data (see Section 3.5.1, while F score, precision (P), and recall (R) were used as the evaluation

measures (see Section 2.3.3.

The proposed novelty mining algorithm, CIAN, may only affect the predictions of sentences that involve the named entities such as number, person, time and location, based on the user’s feedback. In order to evaluate the effectiveness of CIAN, a local evaluation method is proposed, i.e. still running the original NM and CIAN on all sentences, but evaluate the precision, recall and F score in the sentence subsets involving the related types of named entities. For example, for the CIAN algorithm with novelty feedback “number”, let M_n be the number of matched sentences in the subset of sentences involving “number”, A_n be the number of sentences that involve “number” selected by the assessor, S_n be the number of sentences involving “number” selected by the system. The corresponding P_n , R_n are modified to M_n/A_n and M_n/S_n , respectively. The corresponding F_n score is defined similarly as Eq. (2.12). The CIAN algorithm with the novelty feedback having two, three or four types of named entities can be evaluated in the similar way, except the different sizes of sentences subset. The numbers of sentences in the sentence subsets involving different types of named entities in TREC 2004 Novelty Track data are summarized in Table 5.1.

Table 5.1: Statistics of TREC 2004 Novelty Track Data

	# Rel.	# Novel	N/R
All sentences	8343	3454	41.4%
The sentences involving ...			
“number”	2558	1081	42.26%
“location”	2428	1029	42.38%
“person”	1826	765	41.08%
“time”	1710	711	41.58%

In this experimental study, the precision-recall (PR) curve is adopted to demonstrate the performance of different algorithms. For each graph, the F score is plotted according to its precision and recall components. The lines

show contours at intervals of 0.1 points of F score. For example, Figure 5.5 illustrates PR curves of two methods along with the contours of the F score in PR space. The grey dashed lines show contours at intervals of 0.1 points of F score. Given any point on the PR curve, the corresponding region of F score can be observed.

To draw the PR curve varying from high-recall to high-precision in PR space, the novelty decision point, θ , in the original NM is varied from 0.15 to 0.80 with an equal step length of 0.05. CIAN with three different widths of the searching pool, i.e. $\delta = 0.1, 0.2$ and 0.4 are tested and reported in graphs.

5.5.1 Results

CIAN with Different Sentence Comparison Strategies

In this experiment, the CIAN algorithm with two strategies, i.e. one-to-one (using rule-A) and one-to-all (using rule-B) sentence comparison, have been tested on the TREC 2004 Novelty Track data [58].

Figures 5.5 - 5.7 clearly shows that CIAN with one-to-all sentence comparison outperforms that with one-to-one sentence comparison using three different width of searching pool. The performance of the CIAN with one-to-one sentence comparison is even lower for a larger searching pool, as shown in Figure 5.7, while the performance of the CIAN with one-to-all sentence comparison consistently performs well in these situations. This result indicates that the CIAN with one-to-all sentence comparison is able to retrieve novel information from the highly similar sentences at a relatively stable level of precision. In the following experiment, CIAN is performed with one-to-all sentence comparison.

Figure 5.5: CIAN using one-to-one vs. one-to-all, $\delta = 0.1$

CIAN Vs. the Original NM

In this experiment, the CIAN algorithm using one-to-all comparison strategy is compared with the original NM algorithm.

Figures 5.8 - 5.11 show the experimental result of CIAN using one of four kinds of named entities, separately. From these figures, it is observed that:

1. With the increasing value of δ , the width of the searching pool, more sentences as missed by the original NM are processed in CIAN. For different types of named entities, CIAN with the parameter setting of $\delta = 0.2$ constantly performs better than $\delta = 0.1$ and $\delta = 0.4$. This result indicates that a suitable searching pool in the CIAN algorithm may improve the performance of CIAN. Note that the width of the searching pool $\delta = 0.2$

Figure 5.6: CIAN using one-to-one vs. one-to-all, $\delta = 0.2$

is better than other parameter settings in terms of the assessor’s judgments of novelty. In real-world applications, the parameter setting with $\delta = 0.2$ can be treated as the benchmark while the algorithm may increase δ accordingly if the high-recall is the main concern of the user.

- 2. For each fixed decision point θ , the recall of CIAN is higher than that of the original NM with little reduction in precision. CIAN with “person” and “location” outperforms the original NM while CIAN with “number” and “time” have no improvement. This may be due to the different contexts of assessors and the real users who prefer some specific types of novel information. Therefore, an independent assessor manually accessed the highly similar sentences retrieved by CIAN, and found that CIAN is helpful for mining the linguistic-level novelty in highly similar sentences.

Figure 5.7: CIAN using one-to-one vs. one-to-all, $\delta = 0.4$

H1N1 Swine Flu

In this experimental study, we manually gathered 100 news articles related to the H1N1 Swine Flu over the period from April 21 - May 12, 2009. The articles were arranged in a time sequence and assessed for novelty by an undergraduate student using a time window of 10 documents. Out of the 100 articles, 73 were judged as novel. The results for document-level novelty mining using the cosine distance metric and binary weighting are shown in Figure 5.12. Similar to our previous dataset, the graph indicates an extremely high precision across most ranges of recall, resulting in the highest F-Score of 0.917 at a decision point of 0.45, which is higher than the benchmark of 0.844.

For sentence-level novelty mining, an analyst who is interested in monitor-

Figure 5.8: Novelty feedback: “number”

ing the spread of the virus for national security can apply the CIAN algorithm for contextual annotations of named entities. For example, consider the following two sentences:

Sentence 1: Canada’s public health agency confirmed 36 new cases across the country Wednesday.

Sentence 2: Meanwhile, Canada’s public health agency confirmed 46 new cases across the country Monday.

Most sentence-level algorithms which count frequency or overlap of terms would rate *Sentence 2* as non-novel compared to *Sentence 1* because most of

Figure 5.9: Novelty feedback: “location”

the terms overlap. However, if an analyst is specifically interested in monitoring the number of new cases of the virus, the information contained in the number in “46 new cases” is considered as important and should be flagged as novel. In our implemented application (See Appendix A), the CIAN algorithm can be selected in the Sentence-Level dialog by checking “Yes” for the Contextual Annotation NM option. By selecting this option, the user can monitor for new number, time, location, or person information.

For a worldwide health pandemic such as the H1N1 swine flu, the CIAN algorithm is useful in monitoring the number of new cases and the new locations

Figure 5.10: Novelty feedback: “person”

(cities and countries) in the global spread of the virus. Figure 5.13 shows that using the CIAN number option, our system will correctly detect *Sentence 2* as novel even though the novelty score is below the decision point. This is because the number “46” is novel compared to “36” new cases from *Sentence 1*. This option is important in this situation where the user has specific requirements of novelty.

5.6 Chapter Summary

This chapter addressed the important problem of accessing the user’s context in novelty mining (NM) via novelty feedback which have rarely been discussed in previous studies. One possible reason is that it is very difficult for the user

Figure 5.11: Novelty feedback: “time”

to describe the novel information by some words, as people do in the relevant information retrieval, because a word may become non-novel after its first occurrence. The linguistic-level information, i.e. named entities, have been used to accommodate the user’s context in NM algorithms. The user is allowed to feedback his/her preferred significant information to the algorithm directly. For example, if the user is interested in monitoring new numbers, the “Number” option can be selected. If the user is interested in monitoring new locations, the “Location” option can be selected. Likewise, “Person” and “Time” can be selected if the user wishes to monitor changes in person or location. Accordingly,

Figure 5.12: Precision-Recall curve for the H1N1 swine flu topic.

Figure 5.13: Novelty results using contextual annotations of new number.

---

a contextual information annotation of named entities (CIAN) algorithm has been developed to interact with the user's novelty feedback. The experiment on TREC 2004 Novelty Track and Swine Flu data showed that CIAN was significantly useful for mining the novel information in highly similar sentences at the linguistic level.

Chapter 6

Conclusions and Future Research

6.1 Conclusions

Intelligent novelty mining is an important problem that has heretofore been little-studied in a systematic manner. One important challenge of intelligent novelty mining is the problem of how to detect novel information from text data with specific regard to the user context and performance requirements. Another important challenge is how to leverage the use of novelty metrics, novelty decision, and novelty feedback to improve the results of mining new information from text data. This thesis introduces the challenges and finds answers for two questions: where those challenges occur; and what are the best solutions for different types of realistic scenarios.

The central foundation of the thesis is that intelligent novelty mining is a domain-driven problem, one that can benefit from analysis of both technical requirements and business concerns. Although previous studies have analyzed

text documents and sentences for novelty based on optimizing fixed measures such as precision or F Score, none have yet to provide a unified approach to intelligent novelty mining based on the adapting to different performance requirements. Our proposed original techniques aim to leverage the adaptability of the algorithms to accommodate the changing conditions for real-world situations, which has rarely been explored for novelty mining.

Metrics and Models for Novelty Scoring

The first part of this thesis describes the different types of novelty metrics as well as the advantages and disadvantages of each. 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). The performance of novelty metrics was influenced by both explicit factors, i.e. the performance requirements, and implicit factors, i.e. the novelty ratios of data. Considering the effects from these two factors, complementary behavior was observed in the symmetric and asymmetric novelty metrics.

Then, a new framework of novelty measurement, a mixture of both types of novelty metrics, was introduced. The goal of the blended metric is to integrate the merits of both metrics and hence generalize better over different topics. The experimental results showed the superior performance of this new framework under different performance requirements and for data with different percentages of novel sentences. 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-time intelligent novelty mining. The experimental results show that the blended metric not only enhances the overall performance of real-time novelty mining, but also makes the algorithm more flexible to different users' requirements, an essential requirement for intelligent novelty mining. The blended metric also achieved the highest F-scores compared to the best F-scores reported in the TREC 2003 and 2004 Novelty Tracks.

Furthermore, another novelty model, document-to-sentence (D2S), was proposed for document-level novelty mining. This model can make document-level NM 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 NM 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, we also find that it is significantly better to use D2S when the percentage of novel documents is high.

New Learning Model for Adaptive Novelty Decision

The second part of the thesis 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 original algorithm, Learning Adaptive Novelty Decision algorithm for novelty mining (LAND), was proposed. It is a generic method, 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 employing LAND has been tested on both document-level and sentence-level novelty mining data. With complete user feedback, the experimental results showed the promising performance of LAND for novelty decision making for a real-time novelty mining algorithm. In order to test LAND on a more practical level, the algorithm was tested with partial feedback. The experimental results indicate that LAND can work robustly with partial feedback from the user. LAND 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 LAND 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). The experimental results show that LAND is very effective in finding the best decision point in the novelty mining algorithm. Therefore, the original algorithm LAND 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.

Contextual Annotation-based Novelty Mining

The third part of the thesis 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 original technique, contextual information annotation of named entities (CIAN) 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. CIAN 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, have been used to accommodate the user's context. In the CIAN algorithm, the user is allowed to feedback his/her preferred significant information to the algorithm directly. The experiments on different data showed that the CIAN method can solve the important problem of mining the novel information in highly similar sentences at the linguistic level.

6.2 Relation to Previous Work

In relation to previous studies, this thesis developed intelligent techniques more suitable for real-world novelty mining. Other techniques cannot adapt to different users who have different definitions of "novel" information, whereas our adaptive techniques can. Our 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, our techniques make use of the user's contextual information in novelty mining, which was not addressed in previous work. In summary, the intelligent techniques proposed in this thesis 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 for practical novelty mining, which has not been studied before.

6.3 Suggestions for Future Research

The following are suggestions for future research:

- **How to improve results on novelty mining with imperfect categorization**

In this thesis, novelty mining is performed on a set of relevant sentences or documents. This assumes that the categorization process perfectly categorizes the relevant sentences or documents. However, in the case where categorization is not perfect, the novelty mining algorithms need to be tuned accordingly. The performance measures (such as precision or F score) that need to be optimized, the evaluation strategies, the categorization ranking, and the ideal decision point are problems for further research.

- **Novelty mining on imbalanced data**

The quality and size of the training document set also affects the final results of novelty mining. A good training document set should have enough number of documents to ensure the good probability estimation and the document sizes in both novel and non-novel classes are balanced to ensure the good novelty mining result. Novel documents and non-novel documents in the training data set (labeled by the user) under some specific topic could be highly unbalanced since the users can only feedback for the documents delivered, which are top novel ones. In this case, the statistical measure from the non-novel documents may be statistically insignificant due to the limited number of non-novel documents available. This problem also requires further research.

- **Detecting changes in writing style**

Novelty mining is not necessarily fixed to detecting changes in content such as new words, new sentences, or new documents. It can also be applied to detect changes in writing style, such as those written by different authors. This task is closely related to authorship identification, which is a task of predicting the most likely author of a text given a predefined set of candidate authors and a number of text samples per author of undisputed authorship. Further studies on authorship identification, such as evaluation of stylometric features used in writing and Natural Language Processing (NLP) techniques, can be useful for detecting changes in writing style.

- **Detecting changes in writing structure**

In addition to writing style, novelty mining can be applied to detect changes in writing structure. Different types of text have different structures. Therefore, the writing structure of a document can provide us a hint to categorize it to its type. Some main types of documents include email, news, essays, reports, letters, etc. In addition, within the same type of document there can be different structures such as tables, figures, lines, words, footnotes, fonts, etc. Techniques for detecting changes in writing structure, which include text as well as image processing, is the subject of further research.

- **Multilingual novelty mining**

This thesis focused on novelty mining of English documents and sentences. In theory, our novelty mining algorithms can be applied to documents and sentences in languages other than English. However, problems of prepro-

cessing and possible adjustments in the algorithms for scoring, decision point, and novelty feedback need to be further studied.

Appendices

Appendix A

Novelty Mining Application

All the techniques described in this thesis have been implemented in a real novelty mining application. The details of the application are described in this Appendix.

A.1 Graphical User Interface

The Graphical User Interface (GUI) of our novelty mining system is shown in Figure A.1. The system is designed to accommodate different users with various requirements, and can be customized according to an individual user's preference.

The features of the main window are described as follows:

1. Select Database.

This shows the list of existing databases for the current user. The user can

also add new databases by clicking the Add button or remove unwanted databases by clicking the Remove button.

2. Topics.

This section shows the list of available topics given the database. In this section, the user can also add a new topic by clicking the Add button, modify an existing topic by pressing the Edit button, or even delete unwanted topic by clicking the Remove button.

3. Test documents.

In this section, user can input the testing documents by clicking on the Import button. The list box shows the list of documents that have not been categorized or undergone novelty mining, including documents are that just imported using Import button.

4. Process Documents.

The user can select the process (categorization only, novelty mining only, or both). Also, in this section, users can select the specific topics.

5. Processed Panel.

This section is divided into two tabs. One is Processed Docs and the other is Existing Docs. Processed Docs tab shows the list of the documents being processed during the current session. Existing Docs shows the list of document that were processed on previous sessions. This panel also shows topic id and topic title in the Topic Input box.

6. Viewing Panel.

This section also divided into two tabs. The first tab is the sentences view. In this view, all sentences are listed together with the novelty score and

Figure A.1: Main window of our novelty mining system.

prediction. The other tab is the document view. This shows the content of selected document.

A.1.1 Categorization

Figure A.2 shows the options for performing document-level novelty mining.

Users can set properties for the categorization process. The properties include:

1. Initial Threshold. This threshold is compared with each document’s score to determine the relevancy of given document to the specified topic. The

Figure A.2: Options for categorization.

document is predicted to be relevant to the given topic if the score is above this threshold.

2. Alpha, Beta, Gamma. These Parameters are used in the query expansion using Rocchio algorithm as follows:

$$q = \alpha \cdot q_{initial} + \beta \cdot \frac{\sum_{X_i \in R} X_i}{|R|} - \gamma \cdot \frac{\sum_{X_i \in NR} X_i}{|NR|} \quad (A.1)$$

where q is the new query updated by introducing training samples and $q_{initial}$ is the initial query learnt from topic information (including title, description and narratives). R is the set of training samples labeled as relevant and NR is the set of training samples labeled as irrelevant. Parameters α (Alpha), β (Beta), and γ (Gamma) control the weights between initial query, relevant training samples and non-relevant training samples.

3. Recommended Setting. This is the setting that is proven to be the best according to empirical results. Users can choose the desired setting according to their individual requirements.

Figure A.3: Options for document-level novelty mining.

A.1.2 Document-Level Novelty Mining

Figure A.3 shows the options for performing document-level novelty mining. This option allows users to specify the properties for document-level novelty mining. The properties in this tab are described in the following sections.

Metrics

Users can select the metric as a basic measurement for detecting novel documents. There are basically two types of novelty metrics, i.e. symmetric and asymmetric [100]. Symmetric metric is a metric that yields the same result regardless the ordering of two sentences. The most common symmetric metric is the cosine similarity metric, which calculates the similarity between two sentences. Then, the novelty score can be obtained by one minus the cosine similarity score.

Another type of novelty metric is the asymmetric metric. This kind of metric takes into the account the ordering of sentences. A popular asymmetric metric is “new word count”. This metric counts the number of new words

appearing in the current sentence based on history sentences [5]. This novelty metric was proposed for sentence-level novelty mining [5].

As investigated by Tang et al. [76], the strengths of symmetric metrics and asymmetric metrics may complement each other. Therefore, they also proposed a new framework for measuring the novelty by combining both types of novelty metrics. The objective of combining these metrics is to integrate their merits and thus this blended mixed metric can perform better in the general situation.

For the combining strategy, they [76] formulated a new framework of measuring through integrating both types of metrics linearly, one of each type, and concluded that this new type of novelty metric shows the superior performance under different performance requirements and for the data with different percentage of novelty ratios. Moreover, this blended mixed metric is suitable for real-time application for it does not require prior information about the data [76].

The three different metrics implemented in our system are as follows:

1. Cosine similarity.

This metric compares the similarity between current documents and previous documents in the history set. The less similar the current document is, the more novel the current document.

2. New Words Count.

This metric compares the number of the new words between the current documents and previous documents. The more new words found in the current document, the more novel the document is.

3. Mixed Metric.

This metric combines the cosine similarity metric and the new word count

metric. The portion between these two metrics is given by the value α , based on following equation:

$$MMScore = \alpha(1 - CSScore) + (1 - \alpha)NWCScore \quad (A.2)$$

where $MMScore$, $CSScore$, and $NWCScore$ denote as blended mixed metric score, cosine similarity score, and new word count score, respectively.

Term Weighting Function

This term weighting function (twf) is used only when user chooses to use cosine similarity metric. The options of this twf are

1. TF (Term Frequency).

This twf calculates the number of specific term appears in the document. The larger the number the heavier the weight of this term.

2. TF.IDF (Term Frequency/Inverse Document Frequency).

This twf not only calculates the number of specific term appears in the document, but also calculates the importance of that term across all documents (it means the number of documents that contain this term).

3. Binary.

This twf only calculates the existence of the word. In other word, this twf gives the same weight for each term regardless the frequency of the term.

Additional Features

This feature is the additional features that can be chosen. Two features that can be chosen are Name Entity and Term Extractor.

Threshold

Novelty threshold is used as a boundary value to predict whether the incoming document is novel. The question is how the user sets the optimum threshold so that the system can give optimal performance. Two types of threshold setting are available in our system, i.e. fixed threshold and adaptive threshold setting.

As the name suggests, on the fixed threshold setting, the same threshold is used to predict all incoming documents. One question may arise i.e. how to set the right threshold for all the incoming documents. It is nearly impossible to fix a threshold for all documents, since there is little or even no training information in the initial stage of novelty mining.

Therefore, the adaptive threshold setting is introduced as the other alternative in setting the threshold. In this threshold setting, the threshold value is adapted over time as more and more documents come in. We used the adaptive threshold setting algorithm proposed by Tang et al. [63] because it can not only generalize well on both document-level and sentence-level novelty mining but also perform robustly on a more practical level where only partial feedback is available [63]. Another merit of the algorithm is that it can satisfy the diverse users' requirements varying from high-precision to high-recall and also it can suit the diverse users' requirements of "novelty". By utilizing different optimization criteria, the algorithm is able to be tuned according to different performance requirements.

Process Order

This option represents the sequence of processing documents. Created Date Process Order means the documents are processed based on the created date of the documents, whereas Filename Process Order processes the document based on the filename (on descending order).

History Set

The documents that have already been processed are put into the history documents. This set is compared with the next incoming documents. This section gives user the option which documents are needed to be put inside the history set. All means put all already processed documents inside the history set. System Novel means only put the documents that are predicted as novel documents by the ND process.

Document-to-Sentence Technique

Document-to-Sentence (D2S) is a technique for detecting novelty of documents using sentence-level information [79]. In this technique, sentence level information was used to predict the novelty of a document, and experimental results showed that the D2S technique outperforms standard document-level novelty mining [79].

To use D2S, two parameters, Novel Rate and Sentence Threshold, need to be provided. Novel Rate is a threshold for deciding whether the documents are novel. The equation for Novel Rate as follows:

$$NovelRate = \frac{numberofnovelsentences}{numberofsentences} \quad (A.3)$$

Sentence threshold is the threshold used for sentence-level novelty mining.

Sentence Length

This option is used for segmenting documents into sentences. Maximum represents the maximum number of words in a sentence. If the real sentence contains more than Maximum words; then, it cuts the sentence and regards these Maximum words as a full sentence. Minimum represents the minimum number of words to be considered a full sentence. If the number of words in current sentence is less than Minimum; then, this sentence is combined the previous one.

Recommended Setting

The built in setting that believes to be the best setting based on conducted experiments.

A.1.3 Sentence-Level Novelty Mining

Figure A.4 shows the options for performing sentence-level novelty mining.

These options are basically the same as those on the document-level novelty mining (see Section A.1.2), with the addition of heuristic annotation novelty mining option [62], which uses high-level structures of words, i.e. named entities, to accommodate the user's context in the novelty mining system. Four type of entities were given, i.e. number, time, location, and person. This system uses two-layer architecture in detecting novel sentences. The first layer is the original

Figure A.4: Options for sentence-level novelty mining.

novelty mining algorithm. The second layer is using the heuristic annotation novelty mining algorithm. Therefore, every sentences predicted as “non-novel” in the first layer goes to the second layer. In this layer, the sentence is evaluated based on entities that the user chooses i.e. number/time/location/person. Thus, this annotation-based novelty mining is helpful for mining the linguistic-level novelty in highly similar sentences [62].

Appendix B

Author's Publications

1. **Flora S. Tsai**, W. Tang and K.L. Chan, "Evaluation of Metrics for Sentence-Level Novelty Mining", *Information Sciences*, 180(12):2359-2374, 2010.
2. **Flora S. Tsai** and K.L. Chan, "Redundancy and novelty mining in the business blogosphere", *The Learning Organization*, 17(3), 2010.
3. **Flora S. Tsai**, A.T. Kwee, W. Tang and K.L. Chan, "Adaptable Services for Novelty Mining", *International Journal of Systems and Service-Oriented Engineering*, 1(2), 2010.
4. Y. Chen, **Flora S. Tsai**, and K.L. Chan, "Machine Learning Techniques for Business Blog Search and Mining", *Expert Systems With Applications*, 35(3):581-590, 2008. Citations: 8
5. **Flora S. Tsai** and K.L. Chan, "An Intelligent System for Sentence Retrieval and Novelty Mining", *International Journal of Knowledge Engineering and Data Mining*, 1(1), 2010.

6. **Flora S. Tsai** and Y. Zhang , “D2S: Document-to-Sentence Framework for Novelty Detection”, *Knowledge and Information Systems*, 2010 (pending major revision).
7. W. Tang, **Flora S. Tsai**, and L. Chen, “Blended metrics for novel sentence mining”, *Expert Systems With Applications*, 37(7):5172-5177, 2010.
8. A.T. Kwee and **Flora S. Tsai**, “Mobile Novelty Mining”, *International Journal of Advanced Pervasive and Ubiquitous Computing*, 1(4):43-68, 2009.
9. Y. Zhang, **Flora S. Tsai**, and A.T. Kwee, “Multilingual Sentence Categorization and Novelty Mining”, *Information Processing and Management*, 46(5), 2010.
10. **Flora S. Tsai**, W. Han, J. Xu and H.C. Chua, “Design and development of a mobile peer-to-peer social networking application”, *Expert Systems With Applications*, 36(8):11077-11087, 2009. Citations: 2
11. W. Tang and **Flora S. Tsai**, “Adaptive Threshold Setting for Novelty Mining”, In M. Berry and J. Kogan (Eds.): *Text Mining: Applications and Theory*, Wiley, pp. 129-148, 2010.
12. A.T. Kwee, **Flora S. Tsai**, and W. Tang, “Sentence-Level Novelty Detection in English and Malay”, In T. Theeramunkong et al. (Eds.): *Lecture Notes in Computer Science (LNCS)*, Advances in Knowledge Discovery and Data Mining, 5476, pp. 40-51, Springer Berlin / Heidelberg, 2009. Citation: 1
13. **Flora S. Tsai** and K.L. Chan, “Blog Data Mining for Cyber Security Threats” (invited), In L. Cao, P.S. Yu, C. Zhang, H. Zhang (Eds.): *Data*

Mining for Business Applications, Springer, 2009.

14. **Flora S. Tsai**, Y. Chen, and K.L. Chan, "Probabilistic Latent Semantic Analysis for Search and Mining of Corporate Blogs", In C. Soares, Y. Peng, J. Meng, T. Washio, Z-H Zhou (Eds.): *Frontiers in Artificial Intelligence and Applications: Applications of Data Mining in E-Business and Finance*, 177, pp. 63 - 73, IOS Press, 2008. Citations: 2
15. **Flora S. Tsai**, Y. Chen, and K.L. Chan, "Probabilistic Techniques for Corporate Blog Mining", In T. Washio et al (Eds.): *Lecture Notes in Computer Science (LNCS)*, 4819, pp. 35 - 44, Springer-Verlag, 2007. Citation: 1
16. **Flora S. Tsai** and K.L. Chan, "Detecting Cyber Security Threats in Weblogs Using Probabilistic Models", In C.C. Yang et al (Eds.): *Lecture Notes in Computer Science (LNCS)*, 4430, pp. 46 - 57, Springer Berlin / Heidelberg, 2007. Citations: 2
17. W. Tang and **Flora S. Tsai**, "Threshold Setting and Performance Monitoring for Novel Text Mining", *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics 3*, pp. 1310-1319, 2009.
18. W. Tang, A.T. Kwee, and **Flora S. Tsai**, "Accessing Contextual Information for Interactive Novelty Detection", *European Conference on Information Retrieval (ECIR) Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, 2009.
19. H. Liang, **Flora S. Tsai**, and A.T. Kwee, "Detecting novel business blogs", *IEEE Seventh International Conference on Information, Communications and Signal Processing (ICICS)*, 2009.

20. C. Ong, A.T. Kwee, and **Flora S. Tsai**, "Database optimization for novelty detection", *IEEE Seventh International Conference on Information, Communications and Signal Processing (ICICS)*, 2009.
21. Y. Zhang and **Flora S. Tsai**, "Chinese Novelty Mining", *EMNLP: Empirical Methods for Natural Language Processing*, pp. 1561–1570, 2009.
22. Y. Zhang and **Flora S. Tsai**, "Combining Named Entities and Tags for Novel Sentence Detection", *ACM International Conference on Web Search and Data Mining (WSDM) Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR)*, 2009. Citation: 1
23. K.W. Ng, **Flora S. Tsai**, K.C. Goh, and L. Chen, "Novelty Detection for Text Documents Using Named Entity Recognition", *IEEE Sixth International Conference on Information, Communications and Signal Processing (ICICS)*, 2007. Citations: 3
24. Y. Chen, **Flora S. Tsai**, and K.L. Chan, "Blog Search and Mining in the Business Domain", *ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM)*, 2007. Citations: 7
25. K.Y. Yee, A.W. Tiong, **Flora S. Tsai**, and R. Kanagasabai, "OntoMo-biLe: A Generic Ontology-centric Service-Oriented Architecture for Mobile Learning", *IEEE Mobile Data Management (MDM) Workshop on Mobile Media Retrieval (MMR)*, pp. 631-636, 2009.
26. T. Chen, K. Wu, K.-H. Yap, Z. Li, and **Flora S. Tsai**, "A Survey on Mobile Landmark Recognition for Information Retrieval", *IEEE Mobile Data Management (MDM) Workshop on Mobile Media Retrieval (MMR)*, pp. 625-630, 2009.

27. K. Wu, K.-H. Yap, K.L. Chan, L.-P. Chau, **Flora S. Tsai**, and L. Guan, "Mobile Media Annotation and Sharing: Current Challenges and Future Trends" (invited paper), *ACM SIGIR Workshop on Mobile Information Retrieval (MobIR'08)*, 2008.
28. K.L. Chan, X. Xie, W.C. Lee, M. Etoh, and **Flora S. Tsai** (Eds), *Proceedings of the ACM SIGIR 2008 Workshop on Mobile Information Retrieval*, 2008.
29. **Flora S. Tsai** and K.L. Chan, "Dimensionality Reduction Techniques for Data Exploration", *IEEE Sixth International Conference on Information, Communications and Signal Processing (ICICS)*, 2007.
30. **Flora S. Tsai**, S.L. Leong, W.Y. Chin, L. Chen, K.L. Chan, "MedMine: Text Mining of Protein-Protein Interactions", *Asia Pacific Bioinformatics Conference*, Jan. 2005.
31. **Flora S. Tsai**, M. Singh, S. Hariharan, C.K. Chan, P.R. Kolatkar, K.L. Chan, "An Integrative Approach for Determining and Visualizing Protein Interactions", *Asia Pacific Bioinformatics Conference*, Jan. 2005.
32. **Flora S. Tsai**, P. Wasan, F.C. Chan, S. Ng, K.L. Chan, C.W. Lee, "BindAPP: Binding Affinity Prediction Program", *Asia Pacific Bioinformatics Conference*, Jan. 2005.
33. **Flora S. Tsai**, Y. Wu, and K. L. Chan, "Nonlinear Dimensionality Reduction Techniques and Their Applications", *EEE Research Bulletin*, 2004.

Bibliography

- [1] N. Abdul-Jaleel, J. Allan, W. Croft, F. Diaz, L. Larkey, X. Li, M. Smucker, C. Wade, UMass at TREC 2004: Novelty and HARD, in: Proceedings of TREC 2004 - the 13th Text Retrieval Conference, 2004, pp. 1–13.
- [2] J. Allan, V. Lavrenko, H. Jin, First story detection in TDT is hard, in: CIKM 2000, McLean VA, USA, 2000, pp. 374–381.
- [3] J. Allan, R. Papka, V. Lavrenko, On-line new event detection and tracking, in: ACM SIGIR 1998: the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 37–45.
- [4] J. Allan, R. Papka, V. Lavrenko, On-line new event detection and tracking, in: SIGIR 1998: Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 37–45.
- [5] J. Allan, C. Wade, A. Bolivar, Retrieval and novelty detection at the sentence level, in: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 314–321.

- [6] A. Arampatzis, J. Beney, C. H. A. Koster, T. P. Weide, KUN on the TREC-9 filtering track: Incrementality, decay, and threshold optimization for adaptive filtering systems, in: TREC 9 - the 9th Text Retrieval Conference, 2000, pp. 589–600.
- [7] L. D. Baker, T. Hofmann, A. K. McCallum, Y. Yang, A hierarchical probabilistic model for novelty detection in text, in: Inequalities Technical Report, 1999, pp. 1–7.
- [8] M. Bendersky, W. Croft, Finding text reuse on the web, in: WSDM 2009, Barcelona, Spain, 2009, pp. 262–271.
- [9] M. W. Berry, S. T. Dumais, G. W. O'Brien, Using linear algebra for intelligent information retrieval, *SIAM Rev.* 37 (4) (1995) 573–595.
- [10] S. Blott, O. Boydell, F. Camous, P. Ferguson, G. Gaughan, C. Gurrin, G. J. Jones, N. Murphy, N. E. O'Connor, A. F. Smeaton, B. Smyth, P. Wilkins, Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004, in: Proceedings of TREC 2004 - the 13th Text Retrieval Conference, 2004, pp. 15–19.
- [11] T. Brants, F. Chen, A. Farahat, A system for new event detection, in: SIGIR 2003, Toronto, Canada, 2003, pp. 330–337.
- [12] E. Charniak, Statistical techniques for natural language parsing, *AI Magazine* 18 (4) (1997) 33–44.
- [13] Y. Chen, F. S. Tsai, K. L. Chan, Blog search and mining in the business domain, in: DDDM '07: Proceedings of the 2007 international workshop on Domain driven data mining, ACM, New York, NY, USA, 2007, pp. 55–60.

-
- [14] Y. Chen, F. S. Tsai, K. L. Chan, Machine learning techniques for business blog search and mining, *Expert Syst. Appl.* 35 (3) (2008) 581–590.
  - [15] K. Collins-Thompson, P. Ogilvie, Y. Zhang, J. Callan, Information filtering, novelty detection, and named-page finding, in: *TREC 2002 - the 11th Text Retrieval Conference*, 2002, pp. 1–12.
  - [16] J. Davis, M. Coadrich, The relationship between precision-recall and ROC curves, in: *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
  - [17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* 41 (6) (1990) 391–407.
  - [18] S. Dumais, Improving the retrieval of information from external sources, *Behavior Research Methods, Instruments and Computers* 23 (2) (1991) 229–236.
  - [19] D. Eichmann, P. Srinivasan, Novel results and some answers, in: *Proceedings of TREC 2002 - the 11th Text Retrieval Conference*, 2002, pp. 1–7.
  - [20] D. Eichmann, Y. Zhang, S. Bradshaw, X. Y. Qiu, L. Zhou, P. Srinivasan, A. K. Sehgal, H. Wong, Novelty, question answering and genomics: the University of Iowa response, in: *Proceedings of TREC 2004 - the 13th Text Retrieval Conference*, 2004, pp. 1–10.
  - [21] M. Franz, A. Ittycheriah, J. McCarley, T. Ward, First story detection: combining similarity and novelty based approach, in: *Topic Detection and Tracking Workshop*, 2001, pp. 1–11.

-
- [22] E. Gabrilovich, S. Dumais, E. Horvitz, Newsjunkie: Providing personalized newsfeeds via analysis of information novelty, in: *Proceeding of the 13th International Conference on World Wide Web (WWW 2004)*, 2004, pp. 482–490.
- [23] A. Gkanogiannis, T. Kalamoukis, An algorithm for text categorization, in: *SIGIR 2008*, Singapore, 2008, pp. 869–870.
- [24] D. Harman, Overview of the TREC 2002 Novelty Track, in: *Proceedings of TREC 2002 - the 11th Text Retrieval Conference*, 2002, pp. 46–55.
- [25] H. Hoffmann, Kernel PCA for novelty detection, *Pattern Recogn.* 40 (3) (2007) 863–874.
- [26] A. Islam, D. Inkpen, Semantic text similarity using corpus-based word similarity and string similarity, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2 (2) (2008) 1–25.
- [27] Q. Jin, J. Zhao, B. Xu, NLPR at TREC 2003: Novelty and Robust, in: *Proceedings of TREC 2003 - the 12th Text Retrieval Conference*, 2003, pp. 126–137.
- [28] H. Kazawa, T. Hirao, H. Isozaki, E. Maeda, A machine learning approach for QA and novelty tracks: NTT system description, in: *TREC-10*, 2003, pp. 686–697.
- [29] A. Kojiro, S. Shimizu, *Sentence Segmenter* (2006).
URL <http://www.eng.ritsumei.ac.jp/asao/resources/sentseg/>
- [30] T. G. Kolda, *Limited-memory matrix methods with applications*, Ph.D. Dissertation, University of Maryland at College Park, pp. 1–102, 1997.

-
- [31] I. N. Kouris, C. H. Makris, A. K. Tsakalidis, Using information retrieval techniques for supporting data mining, *Data & Knowledge Engineering* 52 (3) (2005) 353–383.
  - [32] A. T. Kwee, F. S. Tsai, Mobile novelty mining, *International Journal of Advanced Pervasive and Ubiquitous Computing* 1 (4) (2009) 43–68.
  - [33] A. T. Kwee, F. S. Tsai, W. Tang, Sentence-level novelty detection in English and Malay, *Lecture Notes in Computer Science (LNCS)* 5476 (2009) 40–51.
  - [34] K. L. Kwok, P. Deng, N. Dinstl, M. Chan, TREC2002 web, novelty and filtering track experiments using PIRCS, in: *Proceedings of TREC11*, 2002, pp. 520–528.
  - [35] L. Lee, Measures of distributional similarity, in: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 25–32.
  - [36] X. Li, W. B. Croft, Novelty detection based on sentence level patterns, in: *CIKM 2005*, 2005, pp. 744–751.
  - [37] X. Li, W. B. Croft, An information-pattern-based approach to novelty detection, *Information Processing and Management: an International Journal* 44 (3) (2008) 1159–1188.
  - [38] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, K. Crockett, Sentence similarity based on semantic nets and corpus statistics, *IEEE Transactions on Knowledge and Data Engineering* 18 (2008) 1138–1149.

- [39] H. Liang, F. S. Tsai, A. T. Kwee, Detecting novel business blogs, in: Proceedings of the 7th IEEE International Conference on Information, Communications and Signal Processing, 2009, pp. 1–5.
- [40] M. Markou, S. Singh, Novelty detection: a review—part 1: statistical approaches, *Signal Process.* 83 (12) (2003) 2481–2497.
- [41] M. Munoz, R. Nagarajan, Sentence splitter (2001).
URL <http://12r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS>
- [42] Y. Nemeth, B. Shapira, M. Taeib-Maimon, Evaluation of the real and perceived value of automatic and interactive query expansion, in: ACM SIGIR 2004, 2004, pp. 526–527.
- [43] K. W. Ng, F. S. Tsai, K. C. Goh, L. Chen, Novelty detection for text documents using named entity recognition, in: Proceedings of the 6th IEEE International Conference on Information, Communications and Signal Processing, 2007, pp. 1–5.
- [44] C. L. Ong, A. T. Kwee, F. S. Tsai, Database optimization for novelty detection, in: Proceedings of the 7th IEEE International Conference on Information, Communications and Signal Processing, 2009, pp. 1–5.
- [45] P. Perner, Concepts for novelty detection and handling based on a case-based reasoning process scheme, *Eng. Appl. Artif. Intell.* 22 (1) (2009) 86–91.
- [46] M. Porter, An algorithm for suffix stripping, *Readings in information retrieval* (1997) 313–316.

- [47] H. Qi, J. Otterbacher, A. Winkel, D. R. Radev, The University of Michigan at TREC2002: Question answering and novelty tracks, in: TREC2002, 2002, pp. 1–12.
- [48] S. Roberts, Novelty detection using extreme value statistics, *IEE Proceedings on Vision, Image & Signal Processing* 146 (3) (1999) 124–129.
- [49] S. Robertson, Threshold setting and performance optimization in adaptive filtering, *Information Retrieval* 5 (2-3) (2002) 239 – 256.
- [50] S. Robertson, I. Soboroff, The TREC 2002 Filtering Track report, in: *Proceedings of TREC 2002 - the 11th Text Retrieval Conference*, 2002, pp. 1–13.
- [51] J. Rocchio, Relevance feedback in information retrieval, in: *The SMART retrieval system: experiments in automatic document processing*, 1971, pp. 313 – 323.
- [52] L. Ru, L. Zhao, M. Zhang, S. Ma, Improved feature selection and redundancy computing–THUIR at TREC 2004 Novelty Track, in: *Proceedings of TREC 2004 - the 13th Text Retrieval Conference*, 2004, pp. 1–5.
- [53] B. Schiffman, K. McKeown, Columbia university in the Novelty Track at TREC 2004, in: *Proceedings of TREC 2004 - the 13th Text Retrieval Conference*, 2004, pp. 1–8.
- [54] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys* 34 (1) (2002) 1–47.
- [55] X. Shen, B. Tan, C. Zhai, Implicit user modeling for personalized search, in: *ACM SIGIR 2005*, 2005, pp. 824–831.

-
- [56] S. Singh, M. Markou, An approach to novelty detection applied to the classification of image regions, *IEEE Trans. on Knowl. and Data Eng.* 16 (4) (2004) 396–407.
- [57] A. Singhal, C. Buckley, M. Mitra, Pivoted document length normalization, in: *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, 1996, pp. 21–29.
- [58] I. Soboroff, Overview of the TREC 2004 Novelty Track, in: *Proceedings of TREC 2004 - the 13th Text Retrieval Conference*, 2004, pp. 1–16.
- [59] I. Soboroff, D. Harman, Overview of the TREC 2003 Novelty Track, in: *Proceedings of TREC 2003 - the 12th Text Retrieval Conference*, 2003, pp. 38–53.
- [60] N. Stokes, J. Carthy, First story detection using a composite document representation, in: *HLT 2001*, 2001, pp. 134–141.
- [61] B. Tan, A. Velivelli, H. Fang, C. X. Zhai, Term feedback for information retrieval with language models, in: *ACM SIGIR 2007*, Amsterdam, The Netherlands, 2007, pp. 263 – 270.
- [62] W. Tang, A. T. Kwee, F. S. Tsai, Accessing contextual information for interactive novelty detection, in: *European Conference on Information Retrieval (ECIR) Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, 2009, pp. 1–4.
- [63] W. Tang, F. S. Tsai, Threshold setting and performance monitoring for novel text mining, in: *Society for Industrial and Applied Mathematics*

- 9th SIAM International Conference on Data Mining, Proceedings in Applied Mathematics 3, 2009, pp. 1310–1319.
- [64] W. Tang, F. S. Tsai, Adaptive threshold setting for novelty mining, in: Text Mining: Applications and Theory, Wiley, 2010, pp. 129–148.
- [65] W. Tang, F. S. Tsai, L. Chen, Blended metrics for novel sentence mining, Expert Syst. Appl. 37 (7) (2010) 5172–5177.
- [66] F. S. Tsai, Network intrusion detection using association rules, International Journal of Recent Trends in Engineering 2 (2) (2009) 202–204.
- [67] F. S. Tsai, Database optimization for novelty mining of business blogs, Nanyang Technological University Technical Report (2010) 1–31.
- [68] F. S. Tsai, K. L. Chan, Detecting cyber security threats in weblogs using probabilistic models, Intelligence and Security Informatics 4430 (2007) 46–57.
- [69] F. S. Tsai, K. L. Chan, Dimensionality reduction techniques for data exploration, in: IEEE Sixth International Conference on Information, Communications and Signal Processing, 2007, pp. 1–5.
- [70] F. S. Tsai, K. L. Chan, Blog data mining for cyber security threats, in: Data Mining for Business Applications, Springer, 2009, pp. 169–182.
- [71] F. S. Tsai, K. L. Chan, An intelligent system for sentence retrieval and novelty mining, International Journal of Knowledge Engineering and Data Mining (2010) 1–19.
- [72] F. S. Tsai, K. L. Chan, Redundancy and novelty mining in the business blogosphere, The Learning Organization 17 (2010) 1–11.

-
- [73] F. S. Tsai, Y. Chen, K. L. Chan, Probabilistic techniques for corporate blog mining, *Lecture Notes in Computer Science (LNCS)* 4819 (2007) 35–44.
  - [74] F. S. Tsai, Y. Chen, K. L. Chan, Probabilistic latent semantic analysis for search and mining of corporate blogs, *Frontiers in Artificial Intelligence and Applications: Applications of Data Mining in E-Business and Finance* 177 (2008) 63–73.
  - [75] F. S. Tsai, W. Han, J. Xu, H. C. Chua, Design and Development of a Mobile Peer-to-Peer Social Networking Application, *Expert Syst. Appl.* 36 (8) (2009) 11077 – 11087.
  - [76] F. S. Tsai, A. T. Kwee, W. Tang, K. L. Chan, Adaptable services for novelty mining, *International Journal of Systems and Service-Oriented Engineering* 1 (2) (2010) 1–16.
  - [77] F. S. Tsai, W. Tang, Context-aware novelty mining, in: *Nanyang Technological University Technical Report*, 2010, pp. 1–12.
  - [78] F. S. Tsai, W. Tang, K. L. Chan, Evaluation of metrics for sentence-level novelty mining, *Information Sciences* 180 (12) (2010) 2359–2374.
  - [79] F. S. Tsai, Y. Zhang, D2S: Document-to-sentence framework for novelty detection, *Nanyang Technological University Technical Report* (2010) 1–12.
  - [80] F. S. Tsai, Y. Zhang, Novel sentence retrieval and mining, *Nanyang Technological University Technical Report* (2010) 1–30.
  - [81] F. S. Tsai, Y. Zhang, A. T. Kwee, W. Tang, Multilingual novelty detection, *Nanyang Technological University Technical Report* (2010) 1–19.

-
- [82] J. Xu, W. B. Croft, Query expansion using local and global document analysis, in: *ACM SIGIR 1996*, 1996, pp. 4 – 11.
- [83] Y. Yang, T. Pierce, J. Carbonell, A study on retrospective and on-line event detection, *ACM Press*, 1998, pp. 28–36.
- [84] Y. Yang, J. Zhang, J. Carbonell, C. Jin, Topic-conditioned novelty detection, in: *SIGKDD 2002*, 2002, pp. 688 – 693.
- [85] K. Y. Yee, A. W. Tiong, F. S. Tsai, R. Kanagasabai, OntoMobiLe: A Generic Ontology-centric Service-Oriented Architecture for Mobile Learning, in: *Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, Workshop on Mobile Media Retrieval*, IEEE, 2009, pp. 631–636.
- [86] C. Yu, J. Cuadrado, M. Ceglowski, J. S. Payne, Patterns in unstructured data: Discovery, aggregation, and visualization (2002).
URL <http://www.hirank.com/semantic-indexing-project/lsi>
- [87] D. Zeimpekis, E. Gallopoulos, TMG: A MATLAB Toolbox for generating term-document matrices from text collections, in: *Grouping Multidimensional Data*, MIT Press, Cambridge, MA, 2006, pp. 187–210.
- [88] C. Zhai, P. Jansen, E. Stoica, N. Grot, D. A. Evans, Threshold calibration in CLARIT adaptive filtering, in: *Proceedings of Seventh Text Retrieval Conference TREC-7*, 1999, pp. 149–156.
- [89] C. Zhai, J. Lafferty, Model-based feedback in the language modeling approach to information retrieval, in: *Proceedings of the 10th international conference on information and knowledge management*, 2001, pp. 403 – 410.

- [90] C. Zhai, J. Lafferty, Two-stage language models for information retrieval, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, 2002, pp. 49–56.
- [91] H.-P. Zhang, J. Sun, B. Wang, S. Bai, Computation on sentence semantic distance for novelty detection, *Journal of Computer Science and Technology* 20 (3) (2005) 331–337.
- [92] H.-P. Zhang, H.-B. Xu, S. Bai, B. Wang, X.-Q. Cheng, Experiments in TREC 2004 Novelty Track at CAS-ICT, in: Proceedings of TREC 2004 - the 13th Text Retrieval Conference, 2004, pp. 1–6.
- [93] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, Expansion-based technologies in finding relevant and new information: THU TREC2002 novelty track experiments, in: Proceedings of TREC 2002 - the 11th Text Retrieval Conference, 2002, pp. 1–5.
- [94] Y. Zhang, J. Callan, Maximum likelihood estimation for filtering thresholds, in: ACM SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 294–302.
- [95] Y. Zhang, J. Callan, T. Minka, Novelty and redundancy detection in adaptive filtering, in: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, pp. 81–88.
- [96] Y. Zhang, F. S. Tsai, Combining named entities and tags for novel sentence detection, in: ESAIR 2009: Proceedings of the WSDM Workshop

- on Exploiting Semantic Annotations in Information Retrieval, 2009, pp. 30–34.
- [97] Y. Zhang, F. S. Tsai, Chinese novelty mining, in: EMNLP 2009: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2009, pp. 1561–1570.
- [98] Y. Zhang, F. S. Tsai, A. T. Kwee, Multilingual sentence categorization and novelty mining, *Information Processing and Management: an International Journal* (2010) 1–19.
- [99] L. Zhao, M. Zheng, S. Ma, The empirical impact of the nature of novelty detection, in: AIRS 2005, 2005, pp. 497–502.
- [100] L. Zhao, M. Zheng, S. Ma, The nature of novelty detection, *Information Retrieval* 9 (2006) 527–541.