

國立臺灣大學電機資訊學院資訊工程學系

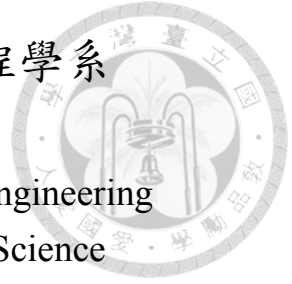
碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



公共議程新聞的自動化議題探勘

Mining Salient Issues from News Articles

on Public Agendas

倪嘉懋

Chia-Mau, Ni

指導教授：許永真博士

Advisor: Jane Yung-jen Hsu, Ph.D.

中華民國 102 年 7 月

July, 2013



國立臺灣大學碩士學位論文

口試委員會審定書

公共議程中的自動化議題探勘

Mining Salient Issues From News Articles On Public
Agendas

本論文係倪嘉懋君（學號 R00922033）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 102 年 7 月 31 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

許永真

（指導教授）

陳言平

劉昭麟

張嘉惠

蔡文瑜

許永真

系主任





誌謝

感謝我的父母在我的求學路上給我的支持，謝謝你們。

感謝許永真老師三年的教誨，包括專業領域上的指導，以及平日言教身教的啟發。許永真老師使我體悟知識的重要性，以及知識分子的社會責任。

感謝蔡宗翰老師在研究上的建議與討論，使我受益良多。

感謝怡亭學姊、麻立恆、蘇緯倫和 George 平日的討論與互相砥礪。

感謝 Commonsense Group 的夥伴與 iAgent 的各位。





摘要

公共議程 (Public Agenda) 係指一系列相關議題，引發公共討論並引起社會注意。在公共議程中，受關注的議題對於議程的發展方向有較高的影響力。然而，議題的顯著程度可能受到媒體的刻意操作，偏頗的媒體往往將違反其立場或利益的議題邊緣化。

本論文介紹一個透過分析新聞文章中的引述句以在公共議程中自動化探勘議題的方法。新聞中的引述句記錄了公眾人物與領域專家的意見，我們藉由將引述句依主題分群，以辨識出新聞文章中被大量辯論的議題。為了增進引述句的分群效果，我們提出「議題顯著詞」來代表每個引述句的主題。

我們收集了一份關於「核四停建」公共議程的新聞文章語料庫以驗證方法的效能，該公共議程是 2013 年台灣社會高度關注的議程之一。我們人工標記出新聞文章中的顯著議題，並將引述句依議題分類。我們使用這份語料庫來驗證引述句分群與議題探勘的效能。





Abstract

A public agenda is a set of issues or concerns that merit public attention. The issues that attract a lot of public attention are influential to the direction of the public agenda. However, the salience of issues might be purposely transferred by the media. Biased news organizations marginalize or filter issues that are against their positions or private interests.

In this thesis we propose a method to automatically mine salient issues from news articles. Quotations in news articles describe comments from public figure and domain experts. Our method for issue mining is based on quotation analysis. By clustering quotations according to their subjects, we identify issues that are widely debated on. We introduce issue significant terms to improve the performance of the method.

To evaluate the performance of issue mining, we compile a corpus of news articles about the public agenda on Lungmen Nuclear Power Plant. The public agenda is a focus for concern in Taiwan in 2013. We manually identify the ground truth of salient issues in the corpus and categorize quotations according to these issues. The performance of quotation clustering and issue mining is evaluated with the ground truth.





Contents

口試委員會審定書	iii
誌謝	v
摘要	vii
Abstract	ix
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
2 Background	3
2.1 Preliminaries	3
2.1.1 Public Agenda	3
2.1.2 Quotations in News Articles	5
2.2 Related Work	8
2.2.1 Topic Detection and Tracking	8
2.2.2 News Summarization	9
2.2.3 Mining Meaningful Targets from News	10
3 Issue Mining with Quotation Analysis	13
3.1 The Issue Mining Problem	13
3.2 Proposed Solution	14
3.3 Quotation Detection	15



3.3.1	Task Definition	15
3.3.2	Pattern-Based Quotation Detection	16
3.4	Issue Clustering	17
3.4.1	Preprocessing	18
3.4.2	Vector Space Model	19
3.4.3	Issue Significant Term Selection	20
3.4.4	Hierarchical Agglomerative Clustering	23
3.5	Issue Cluster Labeling	24
4	Evaluation	27
4.1	Data Preparation	27
4.1.1	Online News Articles	27
4.1.2	Salient Issue Annotation	28
4.1.3	Issue Category Annotation	29
4.2	Evaluate Issue Clustering	30
4.2.1	Evaluation Metrics	30
4.2.2	Compared Methods	32
4.2.3	Experiment Results	32
4.3	Evaluate Salient Issue Mining	34
4.3.1	Evaluation Criteria	35
4.3.2	Compared Methods	35
4.3.3	Experiment Results	36
5	Conclusion	39
5.1	Summary of Contribution	40
5.2	Future Work	40
	Bibliography	43



List of Figures

2.1	Elements in a Chinese news article	6
2.2	Comments gave by the opposition party is quoted separately in four paragraphs, and each paragraph contains a single issue.	8
3.1	System architecture of our solution	15
3.2	Two paragraphs from Chinese news articles. Multiple statements are quoted in these paragraphs. Each quotation might not contain complete semantic meaning alone.	16
3.3	The three components of issue clustering, including preprocessing, issue signifi- cant term (IST) selection and hierarchical agglomerative clustering.	18
3.4	An example of false clustering. Although Q-paragraphs 2 and 3 are talking about the same issue, Q-paragraphs 1 and 2 are more likely to be group together using VSM.	21
3.5	Terms with high SE in the public agenda on Lungmen Nuclear Power Plant	23
4.1	Distributions of top clusters generated by vector space model with all terms (VSM). . . .	33
4.2	Distributions of top clusters generated by vector space model with issue signifi- cant terms (ISTVSM).	34





List of Tables

2.1	Common sentence patterns of quotations in Chinese news articles.	7
3.1	Common reporting verbs in Chinese news articles	17
4.1	Corpus of articles is collected from 7 online news services in Taiwan.	28
4.2	Ground-truth list of salient issues and their available labels.	29
4.3	Issue categories and their sizes. Q-paragraphs are assigned to categories according to the issues they discuss.	30
4.4	Performance of issue clustering Q_{issue} and Q_{all}	32
4.5	Selected labels for annotators	36
4.6	Performances of methods on issue mining.	36





Chapter 1

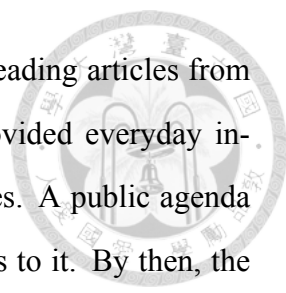
Introduction

A *public agenda* is a group of political controversies that commonly appear in public discussions and merit public attention. For instance, whether the construction should continue on Lungmen Nuclear Power Plant is a widely discussed public agenda in Taiwan in 2013. Each public agenda consists of several *issues*. In the power plant example, security detection on the power plant, disposal of nuclear waste, and current electricity storage are all issues on the public agenda. Given a public agenda, the goal of our work is to automatically mine issues from related news articles.

1.1 Motivation

Participants of public agendas may be individuals or groups. Their purposes are usually requesting the government to confront and solve some issues. These issues often have to compete with each other for visibility in the media. The more an issue is paid attention to by society, the more influence it has on the direction of a public agenda. Since trend of policies are often correlated to public agendas in a democratic regime, it is important that related issues are properly informed to the public.

However, surveys show that media are inherently biased [6] [24]. In the past decades, the ability of media to transfer the salience of issues in public agendas is recognized as *agenda-setting theory* [30]. Issues are likely to be filtered from the readers if they are against news organizations' interests.



A reader may obtain a relatively more balanced perspective by reading articles from different news providers. Nevertheless, the volume of articles provided everyday increases dramatically with the massive growth of online news services. A public agenda might have developed for certain duration before a user first exposes to it. By then, the newest articles are often lack of comprehensive information. On the other hand, there may be a considerable amount of retrospective articles that require to be organized before presenting to the readers. It is difficult for a user to efficiently understand issues in such a topic.

1.2 Objectives

The goal of our work is to mine issues from a corpus of news articles about the public agenda. The corpus can be constructed by collecting articles from multiple online news providers. By revealing issues in the corpus, issues that are blocked by some of the news providers could be exposed to readers. We assume that the subject of an issue can be expressed by a label. Thus, output of the problem is a list of labels such that each label represents an issue.



Chapter 2

Background

In this chapter, we introduce the background of our work. We specify relevant knowledge about journalism in Section 2.1. Related work on news understanding and delivery methods are introduced in Section 2.2.

2.1 Preliminaries

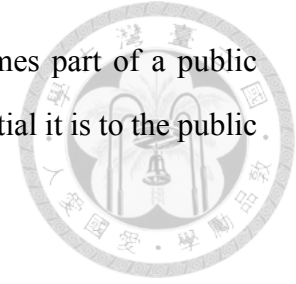
We introduce some relevant journalistic background in this section. First, we introduce the definition of a public agenda and the agenda-setting theory. Then, we describe the characteristics of quotations in news articles, which are utilized in our work.

2.1.1 Public Agenda

According to [16], there are two main steps in a policy agenda. A *public agenda*, also referred to as systemic agenda, is defined as a general set of political controversies that merit attention of the polity, while a *institutional agenda* is defined as a set of concerns or policies scheduled for active consideration by an institutional decision-making body [16]. If a public agenda gains enough attention from public discussion and the government, it may transform to a institutional agenda and cause actual actions from the decision-making groups.

Public agendas include several *issues*. An issue is an important problem for discussion or debate. Issues are raised by individuals or groups that recognize problems and request

for solutions. If an issue is widely known and discussed, it becomes part of a public agenda. The more public attention paid on an issue, the more influential it is to the public agenda.



However, not every issue can obtain a position in a public agenda. “Demands for change in the existing allocation of benefits and privileges in a community can be suffocated before they are even voiced; or kept covert; or killed before they gain access to the relevant decision-making arena. [5]” Advantaged groups in a political system might try to avoid issues against their interests being brought to public agendas.

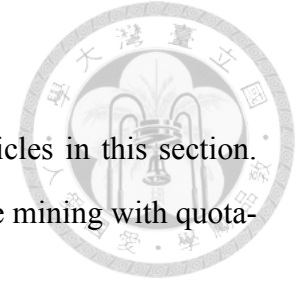
Mass media are the main channel where public opinions are spread. Therefore, mass media are often capable of transferring public attention among issues. *Agenda-setting theory* is an extensively researched media theory under social science. It describes the ability of media to change the salience of issues in a public agenda. The theory is raised by McCombs and Shaw in a study on the presidential election in 1968 [30]. They showed that citizens’ understanding of salient issues is highly correlated to the ones reported by the mass media.

The main assumption under agenda-setting theory is that media do not reflect but shape the reality. In addition, a number of studies show that the media is inherently biased [24] [6]. News organizations might limit journalists’ discretion on expressions against their positions or profit. Moreover, competition between news organizations could raise the restriction among journalists and enforce bias [6]. Eventually, readers have to choose among competing publications with opposing biases. Individuals’ knowledge on a public agenda is filtered and limited when biased media selectively report issues based on private interests.

Democracy demands well-informed electorates. Policies that affect public interests are often motivated by public agendas, while public agendas are based on individuals’ understanding to issues in the society. It is an important matter that issues related to a public agenda are properly shown to readers.

2.1.2 Quotations in News Articles

We show the function and properties of quotations in news articles in this section. These characteristics of quotation inspire us to solve the task of issue mining with quotation analysis.



Role of Quotations

News articles describe newsworthy events that happened in the real world. To ensure convenience of reading, journalists usually follow certain structure when writing news articles. According to Garret et al., news articles are constructed by materials from five main categories: *central occurrence*, *background*, *commentary*, *consequence*, and *follow-up* [19].

Quotation is the major kind of commentaries in daily news articles. It is an expression being restated as part of another. In news articles, public figures' comments or reactions to an event would be stated as quotation to let readers understand the article better.

Figure 2.1 shows an example of a typical Chinese news article. The first paragraph specifies the fact that a referendum motion was made and adopted, while the second paragraph gives more detail on the event. Then a comment made by a councilor is quoted in the third paragraph, which brings up a concern that the title of the referendum may be improper. The two following paragraphs support the article with some background of the event. At the end of the article, there is another quotation that describes the following procedure of the motion.

As we can see, quotations serve as explanations to events. Through the explanation of quotations, readers understand why a certain event is important, and what are the considerations and possible effects. In quotations, issues are often raised and discussed. Observations also show that issues are usually mentioned in quotation rather than other part of news articles.

Quotations play very important roles in public agendas. Commentaries made by public figures and domain experts are brought to readers via quotations in news articles. When quotations spread out through media, public discussion would be gradually established.

<p>環盟核四公投案宜縣審議會通過</p>	
<p>宜蘭縣環保聯盟發起「核四公投」提案，昨天獲宜蘭縣公投審議委員會通過，接下來將核對提案人身分，若符合規定，將送行政院審核。</p>	Event description
<p>宜蘭縣公投審議委員會 15 位委員有 9 人出席，佛光大學主任秘書許文傑擔任主委，不投票，另冬山鄉代會主席黃強呈提前離席，因此有 7 人投票，其中縣議員陳金麟棄權，此案以 6 票贊成、1 票棄權通過。</p>	Event description
<p>陳金麟強調，他不反對公投，但覺得題目太過於專業，民眾可能看不懂，如要停建就直接說停建，主題不明確，要人民怎麼行使公投權利？公投要編預算，勞師動眾要有效果，不能浪費公帑。</p>	Quotation
<p>「核四公投」由宜蘭縣環保聯盟前理事長張捷隆領銜提案，連署人數達 3525 人，超過門檻，宜蘭環盟上周將連署書送到縣政府，完成形式審查後，送縣公投審議委員會。</p>	Background
<p>宜蘭環盟提的公投主文為「宜蘭縣位處在台電公司核四廠緊急應變計畫區範圍（逃命圈）8 公里內，你是否同意台電公司核四廠進行裝填核燃料棒並試運轉？」</p>	Background
<p>宜蘭環盟理事長張曜顯說，提案將送交行政院審核，希望行政院尊重宜蘭人的公民自主權，讓此案通過。</p>	Quotation

Figure 2.1: Elements in a Chinese news article

Eventually, public agendas take shape from these public discussions. A study done by Gibson et al. shows the impact of selected quotations in news articles to readers' perspective [20], which is related to the agenda-setting theory we referred in Section 2.1.1.

Quotation Writing Rules

Several writing rules exist in journalism when quoting a speech or statement. These properties are useful when extracting and analyzing quotations.

In a quotation, the statement given by the original speaker, namely *reporting speech*, is introduced by an *reporting verb*, such as “say”, “claim”, and “point out” in English or “說”, “表示”, and “指出” in Chinese. There are two types of quotations:

- *Direct quotation*: The exact words of reporting speech are stated in the quotation,

Sentence Pattern	Quotation Type
$\langle word \rangle^+ \langle reporting_verb \rangle \langle comma \rangle \langle reporting_speech \rangle$	Indirect
$\langle word \rangle^+ \langle reporting_verb \rangle \langle reporting_speech \rangle$	Indirect
$\langle word \rangle^+ \langle colon \rangle [\langle quote \rangle] \langle reporting_speech \rangle [\langle quote \rangle]$	Direct

Table 2.1: Common sentence patterns of quotations in Chinese news articles.

indicated by quotation marks.

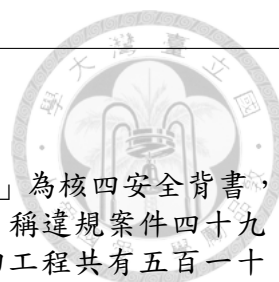
- *Indirect quotation*: A paraphrase of the reporting speech by the reporter. The quotation should not be placed in quotation marks.

In news articles, there are several common sentence patterns for direct and indirect quotations. The three most common patterns in Chinese news articles are listed in Table 2.1, in which each $\langle quote \rangle$ indicates a quotation mark of any kind. Among them, the first pattern, which is an indirect quotation with a comma-separated reporting verb and reporting speech, is the most frequently used pattern in Chinese news writing.

Quotations are usually short. According to *The News Manual* [23], short paragraphs are encouraged in news writing. Journalists have to deliver lots of ideas in a short news article. It keeps things simple for readers by serving short paragraphs with one or two sentences that each contains a simple topic. The writing style of short paragraphs applies to quotations, too. Speakers sometimes raise multiple issues in one speech. These issues are likely to be quoted in separated paragraphs when they are written by journalists.

Figure 2.2 shows an example of how quotations are separated into paragraphs. The news article is about the public agenda on Lungmen Nuclear Power Plant. A few days before the article, a statement was given by the government. The article specifies the response of the opposition party. The opposition party gives comments on four issues, which are security detection, electricity cost, nuclear waste disposal, and current electricity storage. Reporting speech is quoted in four separated paragraphs, and each paragraph contains a single issue.

The writing style of short paragraphs is useful in our work. It inspires us to cluster quotations according to their target issue.



核四安全？工程違規 512 件

〔記者李欣芳／台北報導〕經濟部日前公布「核能議題問答集」為核四安全背書，民進黨政策會撰寫的說帖質疑經濟部迴避許多核四工程缺失，稱違規案件四十九件、被台電列管十八項工安問題，但根據原能會的資料，核四工程共有五百一十二件違規事件。

對於經濟部稱「核四平均每度電價成本不到二元」，民進黨說帖中也引用美國官方數據反駁，指依「美國能源部二〇一一年的能源展望」的估算，核電廠成本換算台幣一度約三．三元。而台灣都是採用美國核電系統，再怎麼算，台灣的核電成本也不可能比美國便宜。

有關核廢料爭議，說帖指出，經部稱「我國已具備低階核廢處理的技術能力，並與國際同步發展高階核廢料的最終處置技術能力」，這種說法其實已承認且印證外界對台灣沒有能力處理核廢料（特別是高階核廢料）的質疑。事實上，台灣低階核廢料最終處置的場址遍尋不到，遑論高階核廢料的最終處置場。

民進黨說帖並指出，經濟部稱不蓋核四，二〇一八年後台灣會缺電，但早在一九九一年時，國民黨政府就說過，如果不蓋核四，台灣就會限電，可見這是國民黨政府慣用的恐嚇人民的手法。實際上，近年的幾次大規模限電與備用容量並無直接關係，從一九九七年以來，台灣的電源供應趨於穩定。

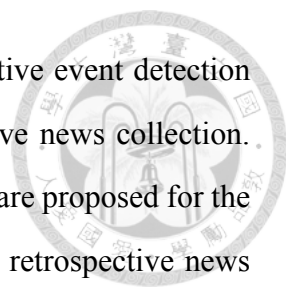
Figure 2.2: Comments gave by the opposition party is quoted separately in four paragraphs, and each paragraph contains a single issue.

2.2 Related Work

With the growth of information, there are a number of work and services for improving the effectiveness of news reading. News recommender systems [7, 8, 10, 12, 29] focus on serving interesting news according to reader preferences. Other approaches on enhancing news reading usually involve understanding content of news articles, including research field like topic detection and tracking and news summarization.

2.2.1 Topic Detection and Tracking

Topic detection and tracking (TDT) [4, 46, 2, 9, 26, 37, 22] is an extensively studied research field. Its goal is to detect topic in a corpus and assign news articles to topic category. A topic is a set of articles that are highly related to a real-world event and its series of events [2]. TDT is composed of several subtasks, including retrospective



event detection, new event detection, and topic tracking. Retrospective event detection (RED) [46, 28] is the task of recognizing events from a retrospective news collection. Solutions by hierarchical clustering [46] or probabilistic models [28] are proposed for the task. Salient issue mining and RED both extract knowledge from a retrospective news corpus, except that RED focus on historical events.

There are several shortages of TDT which is identified by Feng et al. [17]:

- Topics are indivisible.
- Topics do not overlap.
- Topics are independent.

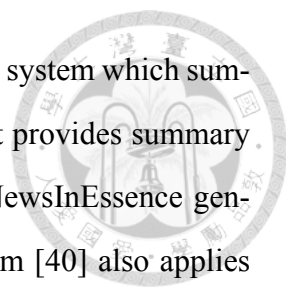
Topics are indivisible in TDT. With the massive growth of articles, it becomes not effective enough to serve readers with all the articles under a topic. The lack of inner and intra structure of topics inspires works that intend to obtain more fine-grained information from news. Incident threading [33, 17, 18] focus on extracting real-world incident from news articles and their relations among each other's. It provides a clearer view of the causal relations among incidents.

In TDT, topics are independent and do not overlaps. News issue construction [41] strives to fix the shortage of independence among topics by constructing “news issues” that composed of several topics.

We consider our work as providing rich information under public agendas, which is a specific type of topic. Informing readers with related salient issues is our concentration, in the view of it being a crucial utility in news reading.

2.2.2 News Summarization

The goal of news summarization [31, 21, 34, 40, 32, 25] is to automatically generate a summary of single or multiple articles. Summarization on one or more documents are often achieved by selecting significant sentences that are potentially capable of covering the major semantic meaning originally served. Goldstein et al. [21] provide a sentence selecting solution based on their statistical and linguistic features to achieve single news article



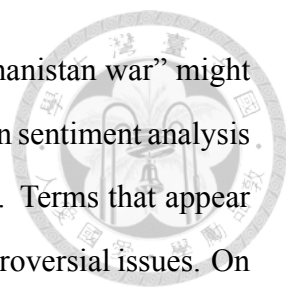
summarization, while McKeown et al. [31] present a natural language system which summarizes multiple news articles. NewsInEssence [34] is a service that provides summary on a topic. By applying TDT techniques to cluster news articles, NewsInEssence generates summaries over clusters according to their centers. CollabSum [40] also applies clustering to news articles, but focus on making use of mutual influences of documents within a cluster context. The work introduces a graph-ranking based algorithm for collaborative document summarizations within each cluster. Newsblaster [32, 25] is an online news summarizing service that summarizes news on a daily basis. TDT techniques are applied to daily new coming articles to generate several news clusters. The system automatically judges the type of a news cluster and applies various summarization strategies. Summaries provided by Newsblaster is evaluated and proved to help readers perform fact gathering [25].

There is also a number of work that concentrate on automatically generating timelines, which is a special kind of summary, for a topic. Allan et al. [3] brought up a temporal summarization on a topic. The solution is based on the assumption that a topics is composed of a series of events. Query based event extraction [14] extracts events from articles retrieved by a query. These events are summarized and put on a timeline to help readers understand a topic. Evolutionary timeline summarization and evolutionary trans-temporal summarization [44, 43] select significant sentences from articles related to a query or a topic. The selection of sentence takes concern on its quality on both articles within a short time range and the whole collection.

Timeline is an effective summarization for a topic. It represents a topic by events. However, events are only one aspect in a public agenda. To understand discussions and concerns raised under a public agenda, aspect of related issues is desired. We represent a public agenda with its salient issues in our work.

2.2.3 Mining Meaningful Targets from News

A work proposed by Choi et al. [15] strives to identify controversial issues and their subtopics. A controversial issue is a term that invokes conflicting sentiment or views



when the term is mentioned in articles. For example, the word “Afghanistan war” might be commonly involved in conflicting sentiment. An algorithm based on sentiment analysis is proposed to identify controversial issues or topics in news articles. Terms that appear frequently in both positive and negative context are considered as controversial issues. On the other hand, subtopic is an entity that is meaningfully associated and subordinated with a controversial issue. Significant terms or phrases are further extracted as the subtopics of an issue by evaluating their co-occurrence in articles.

The work is closely related to ours in the concept of mining issues from news. The main difference that separates our tasks is that we do not restrict our target issues to the controversial ones. Issues that are related to a public agenda might not always be controversial. Real-world problem that are difficult to be solved are usually mentioned in negative contexts. For example, nuclear waste disposal usually appears in negative contexts. Moreover, these issues are often marginalized and filtered from biased news organizations. We would like mine these issues and expose them to the readers.

Another task named news clustering [36] serves news as clusters of articles. It is a similar task as *search result clustering*. The proposed solution is based on finding labels that are useful for over-viewing a collection of news articles. The labels are mined from the collection first, and then news articles are assigned to related labels. Salient issue mining and news clustering both provides labels to overview the collection of news articles. The difference is that we focus only on issues.





Chapter 3

Issue Mining with Quotation Analysis

We introduce the issue mining problem and our proposed solution in this chapter. The goal of our work is to discover issues on a public agenda. In Section 3.1, we introduce the formal definition of the issue mining problem. The overview of the proposed solution is described in Section 3.2, while the details of components are specified separately afterward.

3.1 The Issue Mining Problem

A public agenda P consists of multiple issues. The issue mining problem takes as input a corpus of news articles A about P , and outputs a set of labels:

$$L = \{ l_1, l_2, \dots, l_\alpha \} \quad (3.1)$$

such that each label l_a represents an issue in A .

We assume that each issue can be represented by a label. A label is a keyword or phrase that describes the subject of an issue. For instance, for the issue of nuclear waste disposal, “nuclear waste” and “radioactive waste” are examples of suitable labels.

3.2 Proposed Solution

The main challenge of issue mining is that issues are often not the subject of news articles. A news article is usually triggered by an event rather than an issue. The subject of a news article is more likely to be an event or its follow-ups. Some issues are mentioned in commentaries frequently, but seldom become the subject of news articles. The issue of nuclear waste disposal is a typical example. Although the issue is constantly mentioned by public figures, few articles are written specifically on the issue. Identifying the issue from articles is a challenge.

We propose a solution based on quotation analysis. In Section 2.1.2 we discuss the characteristics of quotations, which can be summarized as follow:

1. Issues are often involved in quotations.
2. Several common sentence patterns are used for quotations.
3. Quotations are likely to be short and contain single issue.

These characteristics inspired us to solve the task by clustering quotations. By grouping quotations that are discussing the same issue together, we can understand the salient issues in the corpus.

Architecture of our solution is shown in Figure 3.1. The solution contains three main components, which are *quotation detection*, *issue clustering*, and *issue cluster labeling*.

In the first step, we apply quotation detection to articles in A . Quotation detection is based on common sentence patterns. We show a solution of quotation detection in Chinese in Section 3.3. After quotation detection, we obtain a set of text passages that each contains one or more quotations.

Issue clustering is applied on these passages. The goal of issue clustering is to group together quotations that discuss the same issue. Our approach is based on vector space model and hierarchical agglomerative clustering. To improve the performance of issue clustering, we introduce *issue significant terms*.

At last, a label is selected for each issue cluster. Significant terms and phrases are extracted as labels.

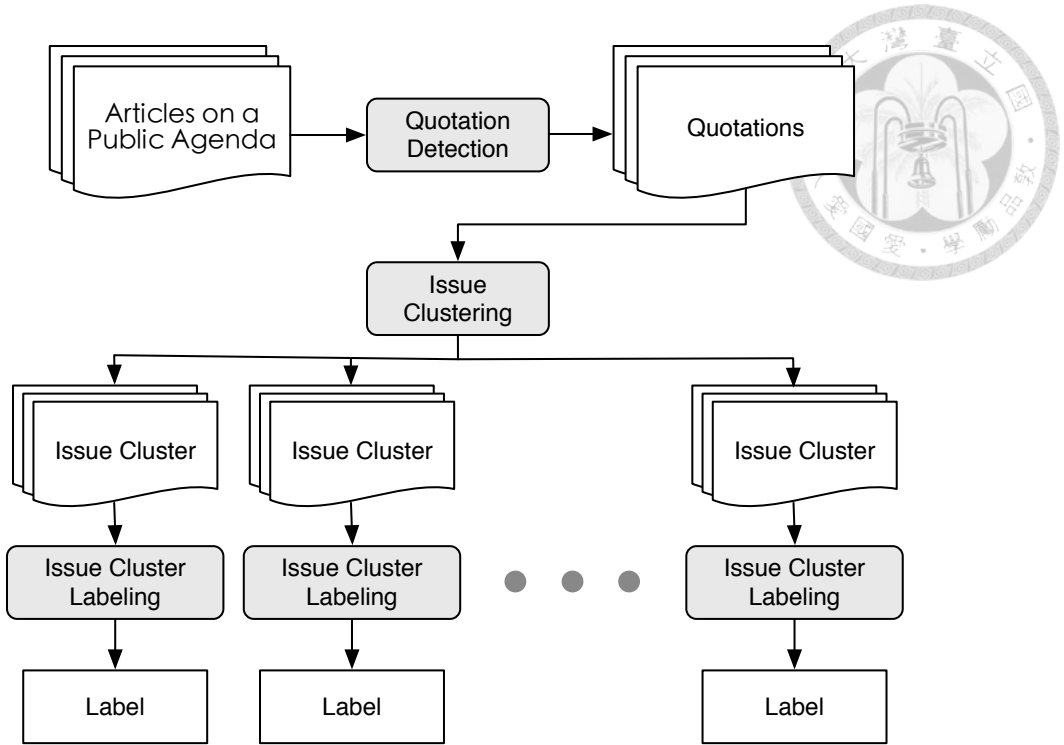


Figure 3.1: System architecture of our solution

3.3 Quotation Detection

In this section, we first describe the definition of quotation detection. Then a solution for Chinese quotation detection is demonstrated afterward.

3.3.1 Task Definition

Paragraphs in news articles may contain more than one quotation. We show two examples in Figure 3.2. Each paragraph contains multiple quotations and some short background about the scenario. If we extract each quotation separately, the extracted quotation would lack complete semantic meanings. On the other hand, paragraphs in news articles tend to be short and hold a single subject. Even if multiple statements are quoted in a paragraph, they are likely to be discussing the same issue.

We define a *Q-paragraph* as a paragraph that contains one or more quotations and their contextual information. The goal of quotation detection is to extract a set of *Q-paragraphs* Q from a news article corpus A :

$$Q = \{ q_1, q_2, \dots, q_\beta \} \quad (3.2)$$

Example 1

台電過去對外宣稱，核四試運轉測試中已完成百分之七十三，可準備裝填燃料棒。對此林宗堯表示他「非常不同意」。他說，台電在預算壓力、工期壓力下被逼急了，現在說要裝燃料棒草草交差，「台電現在的做法很難確保核四廠未來四十年能安全運轉」。張家祝也回應他說：「政府一定會解決，會嚴謹處理。」

Example 2

對於行政院長江宜樺說「公投門檻不宜太低」，蔡英文反嗆：「這是一個政治學者講的話嗎？」她與民進黨主席蘇貞昌都主張修正公投法，降低公投過關門檻。蘇貞昌並批評國民黨想用公投護航核四興建，一直站在「反動的一方」，是阻礙台灣進步的反動力量。他呼籲所有反核的朋友，「大家要有為建立非核家園一戰的準備和決心」。

Figure 3.2: Two paragraphs from Chinese news articles. Multiple statements are quoted in these paragraphs. Each quotation might not contain complete semantic meaning alone.

where a q_i is a Q-paragraph extracted from some news article in A . Quotation detection can be reduced to the task of deciding whether a quoted statement exists in a given paragraph.

3.3.2 Pattern-Based Quotation Detection

Quotations can be detected by matching common sentence patterns. We demonstrate a pattern-based solution for Chinese quotation detection. Common sentence patterns for quotations in Chinese news articles are listed in Table 2.1. The usage of reporting verb is also quite limited in Chinese news writing. We manually identify seven common reporting verbs in news articles. They are listed in Table 3.1.

For each paragraph in articles from A , we check whether one of the three patterns exists in the sentences. If such pattern exists, the paragraph is added to Q . The first quotation pattern covered most of the quotations in news articles.

A common false alarm occurs in the current approach. There are two common patterns for indirect quotations. The first one contains a comma that separates reporting verb and reporting speech, while the other do not. False alarms often occur with the later pattern. The reason is that the reporting verbs are also often used for expressing emotions. Sentences that contain terms such as “表示滿意” or “表示遺憾” are detected as quotations, but these sentences merely describe the emotions of characters. To avoid these false

表示	呼籲	說	指出
強調	認為	喊話	

Table 3.1: Common reporting verbs in Chinese news articles



alarms, we set a threshold on length of the reporting speech. If the reporting speech is no longer than 5 Chinese words, we do not detect the sentence as a quotation.

We validate our pattern-based quotation detection with 30 Chinese articles. The task can be reduced to a binary classification problem, where the algorithm judges whether a paragraph is a Q-paragraph. Given a ground truth, we can evaluate the extracted result with its precision and recall. Among 142 annotated Q-paragraphs, 119 are successfully extracted. Only 2 false positive Q-paragraphs are selected by our approach. In quotation detection, the precision of extraction is more important than its recall. Since we make use of characteristics of Q-paragraphs in later steps, it is important that the precision of retrieved paragraphs is high. The experiment shows a 98% of precision and 84% of recall.

3.4 Issue Clustering

The purpose of issue clustering is to group together Q-paragraphs that are discussing the same issue. Given a Q-paragraph set Q , the output of issue clustering is a cluster set:

$$\begin{aligned}
 C &= \{c_1, c_2, \dots, c_\gamma\} & (3.3) \\
 s.t. \quad &\bigcap_{a=1}^{\gamma} c_a = Q \\
 &c_a \neq \phi, \forall a \\
 &c_a \cap c_b = \phi, \forall a, b \text{ where } a \neq b
 \end{aligned}$$

Under the assumption that each group of Q-paragraphs should have similar semantic appearance, we adopted the well-known *vector space model* (VSM) [35]. Furthermore, *issue significant terms* (IST) are introduced in our work to capture the characteristics of Q-paragraphs. Each Q-paragraph is described as an IST vector. Eventually, we apply hierarchical agglomerative clustering on these vectors. The three components of issue

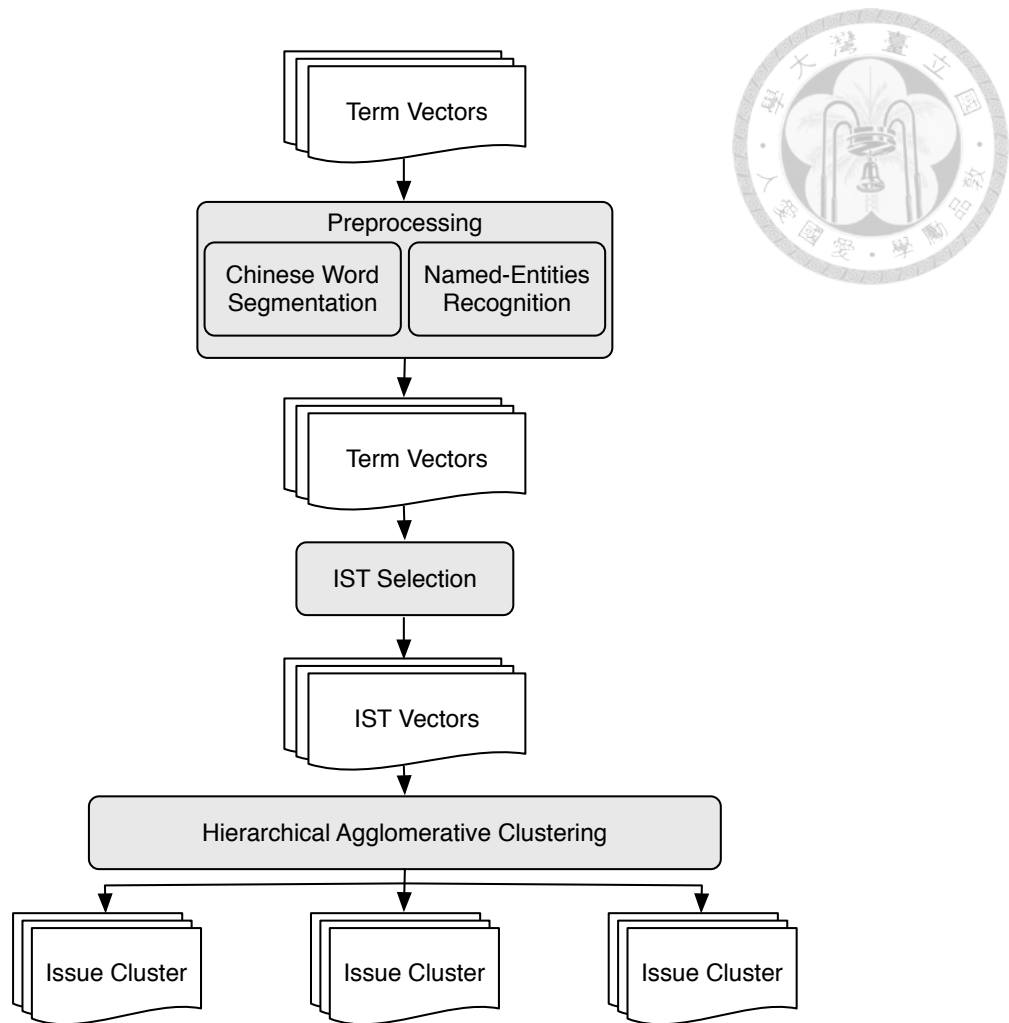


Figure 3.3: The three components of issue clustering, including preprocessing, issue significant term (IST) selection and hierarchical agglomerative clustering.

clustering is shown in figure 3.3.

3.4.1 Preprocessing

There are two preprocessing steps. First, we apply word segmentation to tokenize Chinese sentences. Then, named entities are recognized as rich information.

Word Segmentation for Chinese Text

The major distinction of Chinese text processing is the segmentation of Chinese words. We use *Stanford Word Segmenter* [11] to bridge over the gap. Sentences are first tokenized to words before further analyzed.



Named Entity Recognition

To obtain better knowledge on sentences, we also include a named entity recognizer in our solution. *Mencius* [38] [42] [39] is a Chinese named entity recognizer that recognizes names of people, organizations, and locations. Characters in a public agenda often give comments on various issues. Since our goal is to group together Q-paragraphs that discuss the same issue, these terms are not good materials for clustering Q-paragraphs. Named entity recognition helps us separate these terms from others.

3.4.2 Vector Space Model

After word segmentation, each segmented token is treated as a *term*. We can index all the terms that appear in Q as:

$$T = \{ t_1, t_2, t_3, \dots, t_\theta \} \quad (3.4)$$

In vector space models, Q-paragraphs are mapped to a vector space for calculating similarities. We represent each Q-paragraph q_i by a θ dimension *term vector* \vec{v}_{q_i} , in which every dimension measures a term:

$$\vec{v}_{q_i} = (w_{i,1}, w_{i,2}, \dots, w_{i,\theta}) \quad (3.5)$$

where $w_{i,j}$ is the measurement for importance of term t_j in q_i .

The product of term frequency and inverse document frequency (TF-IDF) is a widely used term weighting method [35]. Since Q-paragraphs are usually short, we use *raw frequency*, which is the number of times that a term t_j appears in a Q-paragraph q_i , as the frequency measurement $TF(t_j, q_i)$. Inverse document frequency measures whether a term is common or rare. The more frequent a term appears in a corpus of documents D , the less likely the term has a specialized semantic meaning. We use *logarithmically scaled*

frequency for IDF measurement:

$$\text{IDF}(t_j, D) = \log \frac{(|D|)}{|\{d_i : d_i \in D, d_i \text{ contains } t_j\}|} \quad (3.6)$$



We weight the importance of a term by the product of its term frequency and inverse document frequency. The effect of stopwords and general terms that lack of distinct semantic meanings are greatly reduced.

We could treat Q as a document set and obtain IDF measurements for terms. However, Q contains only Q-paragraphs on a specified public agenda. Agenda-related terms would appear frequently while holding distinct semantic meanings. If we use the inverse document frequency in Q , the importance of these terms would be substantially underestimated.

We obtain another set of IDF measurements for terms via an external corpus A_E , which is composed of general news articles:

$$\text{EIDF}_{t_j} = \text{IDF}(t_j, A_E) \quad (3.7)$$

EIDF_{t_j} measures the rarity of a term t_j in general news articles. Semantic meanings of agenda-related terms tempt to be distinct in general news articles, leading to high EIDFs. The final weight of t_j is decided by the mixture of the two weighting methods:

$$\text{MIDF}_{t_j} = \text{IDF}(t_j, Q) \cdot \text{EIDF}_{t_j} \quad (3.8)$$

3.4.3 Issue Significant Term Selection

There are several obstacles when clustering Q-paragraphs. We propose a term selection process to improve the performance of issue clustering.

Obstacles for Issue Clustering

There are three major challenges when applying vector space model to issue clustering:

- Q-paragraphs are usually very short, term frequency often fails to show the subject

of a Q-paragraph.



- Quotations are comments made by characters. The selection of terms may be very different when talking about the same issue.
- Important public figures may give comments on various issues. Named entities further confused the process of issue clustering.

Q-paragraph 1

崔愷欣表示，雖然國民黨宣稱核四興建後，經評鑑安全才會決定運轉，但目前台電與原能會的品管失靈，也沒有國際組織為核電廠安全背書的前例。

Q-paragraph 2

根據經濟部昨向國民黨立院黨團簡報資料指出，有關高階核廢料（用過核燃料）最終處置場，台電規畫於二〇二八年選定場址，預計二〇五五年完成，並將尋求國際合作處理或處置機會。

Q-paragraph 3

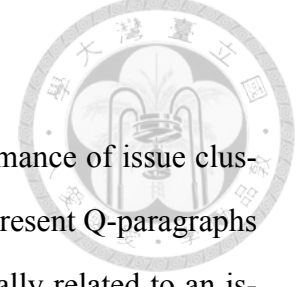
地球公民基金會花東辦公室主任蔡中岳說，核電廠集中北部，但核廢料卻都往東部丟，要解決核廢料的最根本辦法，就是「核電歸零」。今年東台灣的遊行將在台東新生公園集結，而原住民團體也將在 228 當天施放狼煙，預告進一步的行動。

Figure 3.4: An example of false clustering. Although Q-paragraphs 2 and 3 are talking about the same issue, Q-paragraphs 1 and 2 are more likely to be group together using VSM.

In Figure 3.4 we show an example of false clustering. All three Q-paragraphs are related to the public agenda on whether the government should continue constructing Lungmen Nuclear Power Plant. Q-paragraph 1 give comments on the security detection issue, while Q-paragraphs 2 and 3 talk about nuclear waste disposal.

In issue clustering, we would like to group Q-paragraphs 2 and 3 together. However, Q-paragraphs 2 and 3 do not share many overlapping term beside the term “核廢料”. The cosine similarity of their term vectors is relatively low. Since Q-paragraphs 1 and 2 share several terms with high IDF like “國民黨”, “臺電”, and “國際”, they are more likely to be grouped together while clustering.

Selecting Issue Significant Terms with Significance Estimation



We propose a procedure of term selection to improve the performance of issue clustering. Instead of using term set T that includes all the terms, we represent Q-paragraphs with a subset of T . The subset only includes terms that are potentially related to an issue. These terms are referred to as *issue significant terms* (IST). For instance, in the issue of nuclear waste disposal, “waste”, “disposal”, and “fuel” would be ISTs. We name this subset T_{IST} .

Recognizing ISTs before recognizing issues is a challenge. There are some properties of ISTs that we use for extraction: First, an IST usually contains a specific semantic meaning. Its IDF in the external corpus is likely to be high. Second, an IST tend to be significant in A . If a term is widely used in an issue, its document frequency would not be too low. A significant term in A has a higher possibility to be an IST.

To measure the significance of a term in a corpus D , we adopt and modify an extensively used procedure in news keyword extraction [27] [41]. For every document d_i in D , a few terms are selected as key terms. The set of key terms for d_i is denoted as K_{d_i} . Various criteria can be applied for selecting K_{d_i} . We evaluate the significance of a term in D by the frequency of the term being selected as key term. The *significance estimation* (SE) of a term is formulated as follow:

$$SE(t_j, D) = \frac{|\{d_i : d_i \in D, t_j \in K_{d_i}\}|}{|D|} \quad (3.9)$$

We treat the set of Q-paragraphs Q as the corpus. A term with high SE in Q has a higher possibility to be an IST.

In our work, we select 10 terms with the highest weight in $v_{q_i}^{\vec{}}$ as key terms for q_i . By doing so, we filtered most of the stopwords and general terms that do not contain much semantic meanings.

We extract T_{IST} from the original term set T using NER and the SE measurement. First, all named entities are excluded. Since characters often give comments on different issues, named entities are not good materials for clustering Q-paragraphs. Then, the terms

that are generally used in all issues should also be filtered. For example, the term “Lungmen” would not be an IST when dealing with the public agenda on Lungmen Nuclear Power Plant. It is a term directly about the public agenda rather than a specified issue. We name these term as *agenda terms*. Observations show that if a term appears in more than one third of the Q-paragraphs, it is likely to be an agenda term. These terms are also excluded. Finally, we select only the top 20 percent of the remaining terms according to their SE in Q . The selected terms formed a new term set T_{IST} , which defines a reduced vector space in our model. Some selected terms are shown in Figure 3.5, the numbers before terms indicate their SE rankings in T_{IST} .

1 公投	2 停建	3 預算	4 反核	5 追加
6 黨團	7 立委	8 核四廠	9 遊行	10 續建
11 計劃	12 非核	13 核能	14 提案	15 朝野
16 退回	17 家園	18 核能廠	19 會期	20 能源
21 核電	22 安全	23 運轉	24 核災	25 違憲
26 公投案	27 門檻	28 電價	29 年金	30 核廢料

Figure 3.5: Terms with high SE in the public agenda on Lungmen Nuclear Power Plant

3.4.4 Hierarchical Agglomerative Clustering

After representing Q-paragraphs with IST vectors, we can apply clustering on them. Hierarchical agglomerative clustering (HAC) is a bottom-up clustering algorithm. Each Q-paragraph is treated as a cluster at the beginning. The two most similar clusters are merged in every iteration. The merging process continues until all Q-paragraphs are merged to one big cluster or a stopping criterion is reached.

There are different linkage criteria that decide which two clusters should be merged in every iteration. We use average linkage clustering in our work. The similarity between two clusters c_a, c_b is defined as:

$$\text{SIM}(c_a, c_b) = \frac{1}{|c_a||c_b|} \sum_{q_s \in c_a} \sum_{q_t \in c_b} \frac{v_{q_s}^{\vec{}} \cdot v_{q_t}^{\vec{}}}{|v_{q_s}^{\vec{}}||v_{q_t}^{\vec{}}|} \quad (3.10)$$

In every iteration, the two most similar clusters are merged.

HAC generates a possible cluster set in every iteration. For β Q-paragraphs, $\beta - 1$

cluster sets are generated through the clustering process. The selection of final cluster set includes three criteria:


- The algorithm do not select cluster sets in which the 10 largest clusters cover less than 70 percent of the Q-paragraphs.
- The algorithm do not select cluster sets in which the largest cluster covers more than half of the Q-paragraphs.
- Among the cluster sets that qualify to the first two criteria, elbow criterion is applied on one-way ANOVA F-test for the final choice [1]. F-test is the ratio of the between-cluster variance to the total variance. It measures the explained variance of a cluster set. As the number of clusters decreases in the merging process of HAC, value of F-test gradually falls. At some point, the merging of two clusters causes the value of F-test to decrease dramatically. According to elbow criterion, we select the cluster set before the merging occur.

3.5 Issue Cluster Labeling

After issue clustering, every issue cluster is composed of Q-paragraphs that discuss the same issue. To let user understand the issue in a issue cluster c_a , a label l_a is given to represent the issue. The most significant term in c_a is a possible choice of l_a . We also extract phrases from c_a as label candidates, since phrases are usually more complete in semantic meanings than words. l_a is selected by comparing the SE measurement of terms and phrases in c_a .

A phrase p is a sequence of words $c_1c_2\dots c_\lambda$, which contains a specified semantic meaning. Not any sequence of words can be considered as a phrase. A sequence of words that lack complete syntactic and semantic structure is merely a string. We adopt the solution brought up by Wang et al. to identify significant phrases [41]. A significant phrase p must satisfy the following criteria:

1. The number of documents that contains p is more than a threshold TH_D .

- 
2. p contains at least one of the significant terms in the collection. We select the top 20 percent of terms with highest SE as significant terms.
 3. p is a max-duplicated string in the issue cluster [45]. Its document frequency is larger than all other strings that treat p as a substring.
 4. The mutual information of its two longest substring should be high [13]. The two longest substrings of p are $ss_1 = c_1c_2\dots c_{\lambda-1}$ and $ss_2 = c_2c_3\dots c_\lambda$. The *substring mutual information* of p is defined as:

$$\text{SMI} = \frac{\text{Pr}(p)}{\text{Pr}(ss_1) + \text{Pr}(ss_2) - \text{Pr}(p)} \quad (3.11)$$

where the probability of a string is measured by its document frequency. SMI of a phrase p should be higher than a threshold TH_{SMI} .

We follow the parameter settings in the original work. All the strings that match the criteria are treated as significant phrases in a issue cluster.

After phrases are extracted, we adopt Equation 3.9 to measure the significance of a phrase. Since phrases are already judged as significant strings, we add all phrases that appears in q_i to K_{q_i} . The SE of a phrase is equivalent to the document frequency of the phrase.

The term or phrase with the highest SE measurement in c_a is selected as the label l_a . For a cluster set C , we output labels of the clusters that are large enough. Duplicate labels are removed from the final label set L .





Chapter 4

Evaluation

In this chapter we evaluate the performance our solution. The evaluation is consist of two parts. First, we evaluate the performance of issue clustering. Second, we evaluate the ability of different approaches to mine salient issues from articles.

4.1 Data Preparation

We compile a dataset on the public agenda of whether the construction of Lungmen Nuclear Power Plant should continue. The dataset is composed of three parts. First, a set of 221 articles related to the public agenda is collected from online news providers. 720 Q-paragraphs are extracted from these articles. Second, 6 salient issues and their available labels are identified. Last, a ground-truth issue category set is built for the extracted Q-paragraphs. Q-paragraphs are categorized according to the issues they discuss.

4.1.1 Online News Articles

News articles are collected from 7 major online news services in Taiwan as listed in table 4.1. A webpage crawler automatically collects 41174 news articles from these online news sites. The corpus of articles covers various domains.

In our work, we focus on the debate on whether the construction should continue Lungmen nuclear power plant. It is one of the most salient public agendas in Taiwan in 2013.

News service	Number of articles
中央通訊社	4,948
苦勞網	16
中時電子報	3,451
聯合新聞網	7,948
自由時報電子報	6,081
蘋果日報	2,382
ETtoday 東森新聞雲	16,348



Table 4.1: Corpus of articles is collected from 7 online news services in Taiwan.

A set of articles on the public agenda is manually identified. The article set contains 211 articles published in the duration from March to May, 2013.

We apply the method described in Section 3.3 to extract Q-paragraphs from articles. 720 Q-paragraphs are extracted from the 211 articles on the public agenda.

4.1.2 Salient Issue Annotation

We manually identify a ground-truth list of salient issues for experiments. 5 annotators are recruited for the task.

Before the annotators identify issues, we make sure that the annotators understand the task and the public agenda. First, the definition of public agenda and issues are explained to each annotator. Then, to ensure the annotators have enough knowledge on the public agenda, they are asked to read a number of related articles. Annotators are asked to browse through the titles of the articles and select at least 30 articles to read.

After annotators gain understanding to the agenda, they are asked to point out related salient issues. We use the method proposed by Wang et al. to generate a list of keywords and keyphrases for the public agenda [41]. These keywords and keyphrases are provided to the annotators as label candidates. Since named entities would not be an issue, they are manually removed. Given a list of label candidates, annotators are asked to point out all available labels for each issue they recognize.

Finally, we take an agreement on the annotated labels by the following procedure: First, all labels that are selected by more than half of the annotators are included to the ground truth. In this step, 13 labels are included. Then, we apply HAC to group together

Issue ID	Available Issue Labels
1	公投, 公投案
2	核廢料, 燃料棒, 放置燃料棒
3	核災
4	安全檢測
5	違憲
6	電價, 電價上漲, 電力, 用電需求, 備用容量率



Table 4.2: Ground-truth list of salient issues and their available labels.

labels of the same issue.

We use average linkage clustering on the task. Initially, each label is an issue itself. Pairwise similarity of labels is decided by annotators that select both of the labels. We calculate the ratio of annotators that consider two labels as the same issue. The ratio is treated as the pairwise similarity of two labels. Merging process continues until the highest average linkage between issues is lower than 0.5. The agreement of salient issues is shown in table 4.2 .

4.1.3 Issue Category Annotation

To evaluate the performance of issue clustering, we obtain a issue category set for evaluation benchmark.

$$C^* = \{ c_1^*, c_2^*, c_3^*, c_4^*, c_5^*, c_6^* \} \quad (4.1)$$

where c_a^* includes Q-paragraphs that are discussing the a 'th issue in Table 4.2. Not every Q-paragraph is assigned to a issue category. Q-paragraphs that are not related to any salient issue and Q-paragraphs that are discussing multiple salient issues are excluded from C^* . A Q-paragraph is not assigned to more than one issue category.

15 annotators participate in the task. For a given Q-paragraph, annotators are asked to either assign the Q-paragraph to a issue category in C^* or mark the Q-paragraph as “None” (Q-paragraph not relevant to all issues) or “Multiple” (Q-paragraph mentions multiple issues). Each Q-paragraph is annotated by 3 annotators. Ground-truth category of a Q-paragraph is decided by majority rule.

The result is shown in Table 4.3. In the 720 Q-paragraphs, 329 are assigned to an issue

category	c_1^*	c_2^*	c_3^*	c_4^*	c_5^*	c_6^*
size	109	16	16	74	72	42

Table 4.3: Issue categories and their sizes. Q-paragraphs are assigned to categories according to the issues they discuss.



category, 340 are marked as “None”, and 44 are marked as “Multiple”.

There are 7 Q-paragraphs that the 3 annotators all judge it differently. Situations are similar on these Q-paragraphs. Each Q-paragraph mentions several issues. Annotators assign the Q-paragraph to different issue categories or mark the Q-paragraph as “Multiple”, leading to conflict results. These Q-paragraphs are not assigned to any category.

We build two test sets for issue clustering. The first set Q_{issue} consists of the 329 Q-paragraphs that are included in one of the issue categories. The other set Q_{all} includes all 720 Q-paragraphs.

4.2 Evaluate Issue Clustering

In this section, we evaluate the performance of issue clustering. We compare the behavior of HAC using term vectors and IST vectors.

We run experiments on two test sets, namely Q_{issue} and Q_{all} . In Q_{issue} , every Q-paragraph belongs to one of the issue categories. We evaluate the ability of HAC to accomplish the clustering task. The other test set Q_{all} includes all 720 Q-paragraphs that are extracted from news articles. In our architecture of issue mining, Q_{all} is the set of Q-paragraphs that is passed to the issue clustering component. The difficulty of the task is higher.

4.2.1 Evaluation Metrics

The performance of a cluster set C is evaluated with a ground-truth issue category set C^* . Two metrics are used in the experiments.



Normalized Mutual Information

The main measure we use to evaluate how well a data is clustered is by computing *normalized mutual information* (NMI):

$$\text{NMI} = \frac{I(C; C^*)}{[H(C) + H(C^*)]/2} \quad (4.2)$$

where I is mutual information:

$$I(C; C^*) = \sum_{c_i \in C} \sum_{c_j^* \in C^*} Pr(c_i \cap c_j^*) \log \frac{Pr(c_i \cap c_j^*)}{Pr(c_i)Pr(c_j^*)} \quad (4.3)$$

and H is entropy:

$$H(C) = - \sum_{c_i \in C} Pr(c_i) \log Pr(c_i) \quad (4.4)$$

NMI measures how much statistical information does the cluster set C share with ground-truth categories C^* . Its range falls in $[0, 1]$. The greater NMI is, the closer is C to the ground truth C^* .

Clustering Q_{all} is more difficult than clustering Q_{issue} . To examine the different of the two tasks, we evaluate the NMI of clustering results of Q_{all} with two ways. In the first way, NMI is evaluated on all Q-paragraphs. In the second way, all Q-paragraphs that are not included by Q_{issue} are ignored when evaluating NMI. We refer to the later as *filtered NMI*.

Purity

Purity is another common measurement on clustering. To compute purity, each cluster is first assigned to the category that is most frequent in the cluster. Purity measures the ratio of data being put to the cluster with the correct category:

$$\text{purity}(C, C^*) = \frac{1}{\beta} \sum_{c_i \in C} \max_{c_j^* \in C^*} |c_i \cap c_j^*| \quad (4.5)$$

	Q_{issue}		Q_{all}			
	NMI	Purity	NMI	Filtered NMI	Purity	Total Purity
VSM (TF-IDF)	0.408	0.620	0.238	0.315	0.632	0.222
VSM (TF-MIDF)	0.455	0.741	0.235	0.305	0.690	0.222
ISTVSM (TF-IDF)	0.455	0.723	0.272	0.328	0.574	0.270
ISTVSM (TF-MIDF)	0.454	0.726	0.278	0.363	0.711	0.518

Table 4.4: Performance of issue clustering Q_{issue} and Q_{all} .

where β is the number of Q-paragraphs with issue categories. It is easy to see that *purity* = 1.0 when HAC starts, because each cluster contains only one Q-paragraph. As we merge clusters together, purity decreases gradually. A low purity shows that many Q-paragraphs with categories are not put in the correct clusters. Issue-related terms would not be significant enough, leading to a bad performance of issue mining.

When computing purity, all Q-paragraphs that are not in any issue category are ignored. However, observations show that in some clusters, Q-paragraphs with issue categories might be overwhelmed by Q-paragraphs without category. We calculate *total purity* by assigning all Q-paragraphs without categories to a new category c_0^* . If a cluster contains a lot of Q-paragraphs without category, the cluster might be assigned to c_0^* .

4.2.2 Compared Methods

We use HAC as our clustering method. Performances of different types of term vectors are compared.

For choice of vector space, we use the vector space generated by all terms or the one generated by ISTs. We refer to the vector space model that involves all terms as **VSM**, and the model that generates vector space with ISTs as **ISTVSM**.

Choice of term weighting methods includes TF-IDF and TF-MIDF. We use the corpus of all online news articles as the external corpus A_E . Each vector space is combined with either of the term weighting methods to create term vectors.

4.2.3 Experiment Results

The experiment results on the two test sets are shown in Table 4.4 .

The results show that in Q_{issue} , term weighting is more important than vector space. However, results in Q_{all} show that vector space is critical when dealing with all kinds of Q-paragraphs.

We observe that **VSM (TF-IDF)** does not perform well with Q_{issue} . Issue-related terms are frequently used in Q_{issue} . By weighting terms with IDF in Q_{issue} , the importance of issue-related terms are highly underestimated. Advantage of TF-MIDF is that it adjusts TF-IDF with an external corpus. Issue-related terms would be weighted highly in term vectors. On the other hand, there is no significant difference between TF-IDF and TF-MIDF with **ISTVSM**. Since we only use ISTs in term vectors, terms with low document frequency are already filtered. The clustering result is similar using different weight of terms.

In experiments on Q_{all} , the test set includes Q-paragraphs that mention none or multiple issues. IST selection became very crucial in the situation. We take a closer look on the clustering result of **VSM** and **ISTVSM** using TF-MIDF on Q_{all} . The distribution of the top 6 clusters generated by both methods are shown in Figure 4.1 and Figure 4.2 .

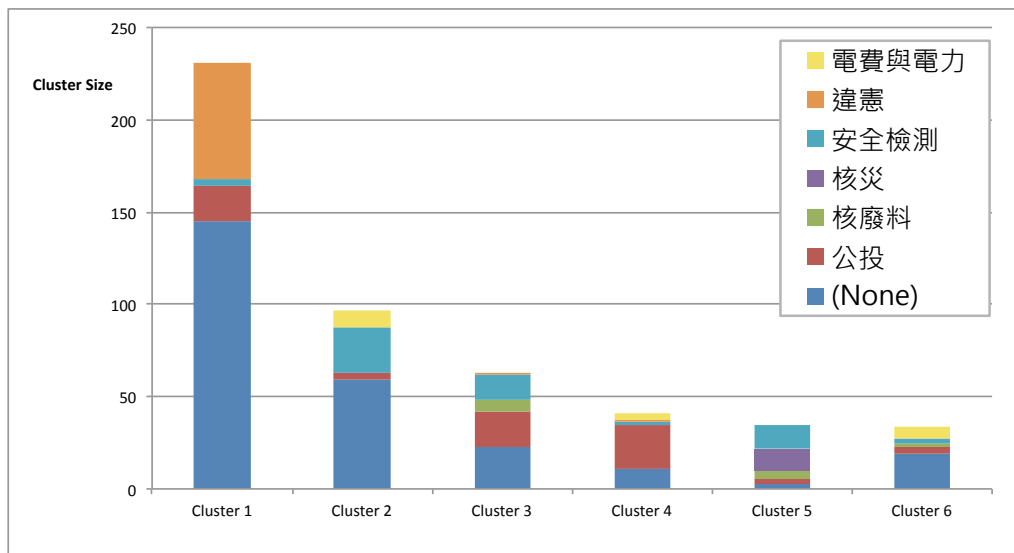


Figure 4.1: Distributions of top clusters generated by vector space model with all terms (**VSM**).

Both methods built a large cluster that is dominated by Q-paragraphs without issue categories. The content of the cluster is mostly about daily procedures of political agendas. Significant terms in the cluster are terms that are related to political procedures.

Issue No. 1 is about referendum, and issue No.4 is about judicial review of ceasing

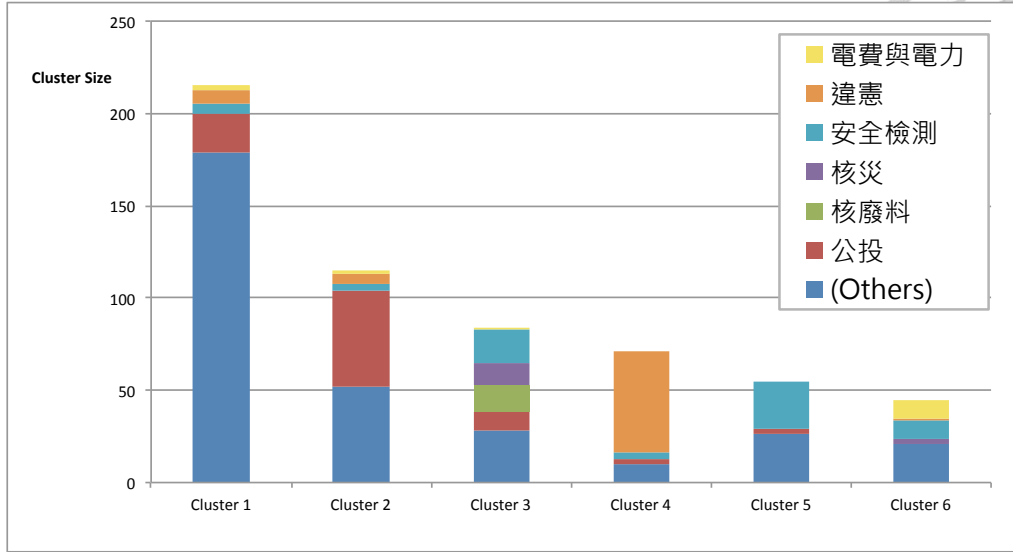


Figure 4.2: Distributions of top clusters generated by vector space model with issue significant terms (ISTVSM).

construction. Since these two issues are highly related to political procedures, they share a lot of common terms with these Q-paragraphs without issue category. In **VSM**, Q-paragraphs about these issues are easily merged into the largest cluster. These two issues are successfully distinguished in **ISTVSM**. By describing Q-paragraphs with only ISTs, a considerable amount of terms related to political procedures are filtered. The term “違憲”, which means violation of the constitution, is emphasized in IST vector space model and became very decisive in hierarchical clustering.

Excluding named entities from term vectors is also crucial. **VSM** tends to build clusters that focus on named entities. In Figure 4.1, the second largest cluster focus on Taiwan Power Company and the minister of economic affairs, while the third largest cluster focus on the mayor of Taipei City. As a consequence, both of the clusters contain several categories of Q-paragraphs from these public figures, which leads to a low purity.

4.3 Evaluate Salient Issue Mining

In this section, we show the issue lists generated by different issue mining methods and compare their performance. An issue list L is evaluated on two major aspects. First, we evaluate the amount of salient issues successfully revealed by the issue list. Second,

the list should provide as less false issues as possible.



4.3.1 Evaluation Criteria

In an issue list L , each issue is represented by a label l_a . We evaluate L by the issues it successfully revealed. For every salient issue in the ground-truth list, we check whether one of its available labels exists in L . If such label exists, we considered the issue successfully discovered by L .

We also evaluate whether an issue list is precise. The system should provide as less false issue as possible. Precision of a L is defined as the ratio of its labels that exist in the ground-truth list.

4.3.2 Compared Methods

We implemented two methods, which are **Q-VSM** and **Q-ISTVSM**, that are based on issue clustering. Q-paragraphs are first extracted by quotation detection which we introduce in Section 3.3. Then, we apply issue clustering to the extracted quotations. Vector space is generated by all terms in **Q-VSM**, while vector space is generated by ISTs in **Q-ISTVSM**. Both methods use TF-MIDF as term weighting method and HAC as clustering method.

To evaluate issue mining based on issue clustering, we build an algorithm, which is referred to as **S-VSM**, that mine issues with article clustering. For a given corpus, every news article is mapped to a term vector. We apply HAC on the term vectors to group together articles that have similar semantic appearance.

Three methods share the same procedure after the clustering step. First, we remove clusters that are not large enough. Clusters that do not contain more than 5 percent of Q-paragraphs/articles are removed. Then, a label is selected for each cluster using the techniques which we described in Section 3.5.

Issue lists generated by the three methods are shown in table 4.5.

S-VSM	Q-VSM	Q-ISTVSM
經濟委員會	經濟委員會	臺電預算
公投	臺電	公投
國民黨團	公投	核廢料
郝龍斌	核災	違憲
替代能源	非核家園	安全檢測
反核	黨團	說帖
	發電	

Table 4.5: Issue lists generated by issue mining methods.

	Precision	Issue Discovered
S-VSM	0.167	1
Q-VSM	0.286	2
Q-ISTVSM	0.667	4

Table 4.6: Performances of methods on issue mining.

4.3.3 Experiment Results

The performances of issue mining methods are shown in table 4.6.

The main obstacle of **S-VSM** on the task is that news articles are usually triggered by events instead of issues. Some issues are often mentioned in commentary paragraphs, but seldom become the subject of a article. A large amount of news articles are related to political procedures. Issues that are not related to political procedures, such as nuclear waste disposal and security detection, are easily overwhelmed. The only issue that **S-VSM** successfully found is about referendum, which is a very salient issue in the agenda.

The differences between **Q-VSM** and **Q-ISTVSM** are explained in Section 4.2.3. In **Q-VSM**, issue clusters often focus on named entities. With IST selection, **Q-ISTVSM** build clusters that focus on issues. Two extra issues are discovered by **Q-ISTVSM**. Precision of **Q-ISTVSM** is also higher.

Issue No. 6 is an issue that all methods fail to mine. It is the issue about electricity price and capacity. Comparing to other issues, it is an issue that covers diverse concepts. Electricity price, electricity capacity, and cost of power production are all included in this issue. Term usage is very different in this issue. By adopting vector space model, we inherit the assumption that all terms are independent. Two Q-paragraphs talking about “電力需求” and “供電” respectively may be considered irrelevant, even though the two

terms are highly related in human concept.







Chapter 5

Conclusion

Public agenda is one of the most important means for democratic participation. However, individuals face the threat that related issues might be understated or filtered by biased media. In a democracy polity, it is important that related issues of a public agenda are properly informed to electorates. We strive to extract salient issues from various news providers automatically.

This work proposed a solution for mining salient issues on public agendas based on quotation analysis. We showed the characteristics of quotations in news that benefits issue mining. A pattern-based quotation detector is introduced in our work to efficiently retrieve Q-paragraph from articles.

The goal of issue clustering is to group Q-paragraphs that discuss same issue together. We introduce a term weighting method, which is referred to as TF-MIDF, to conquer the problem that issue-related terms are underestimated with TF-IDF. A term selection procedure, namely IST selection, is proposed to retrieve terms that are potentially related to an issue.

A dataset of 211 news articles and 720 Q-paragraphs about a public agenda is compiled for evaluating our system. 6 salient issues are identified from the corpus, and Q-paragraphs are categorized according to these issues. We evaluate the behavior of issue clustering when using different types of term vector. We also compare the issue lists generated by methods that involve issue clustering and a method that does not.

5.1 Summary of Contribution



We summarize the contribution of our work as follows:

- We identify the importance and definition of salient issue mining on public agendas. A solution based on quotation analysis is proposed in our work.
- The challenges of clustering Q-paragraphs are revealed in this thesis. We introduce issue significant terms to cluster Q-paragraphs based on the issue they discuss. Experiments show that it outperforms the traditional vector space model on the task.
- A dataset that includes 211 news articles and 720 Q-paragraphs is compiled in our work, which can be used in future research. 6 salient issues are identified from the corpus, and Q-paragraphs are categorized according to these issues.

5.2 Future Work

There are several limitations in our work that requires future improvements. We identify them as follow:

- Table 4.3 shows the distribution of the 720 Q-paragraphs we annotated. According to our experiments, if we focus on only the 329 Q-paragraphs that belongs to a category, our methods could achieve better performance on clustering. Dealing with all 720 Q-paragraphs together raises the difficulty of clustering dramatically. The robustness and effectiveness of the issue mining solution could both increases if we build a filter that judges whether a Q-paragraph is likely to contain an issue. Future work on classifying Q-paragraphs is desired.
- The detection of public agendas from a general corpus of news articles is not included in our research scope. Techniques in the research field of topic detection and tracking (TDT) are able to detect topics and group articles by their topics [2]. In TDT, a topic is an event or activity, along with all directly related events. Generally speaking, we can treat public agendas as topics and apply TDT techniques to detect

agenda. Further research is required to automatically detect public agendas from news articles.

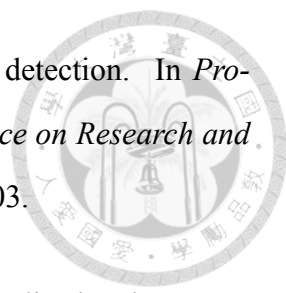


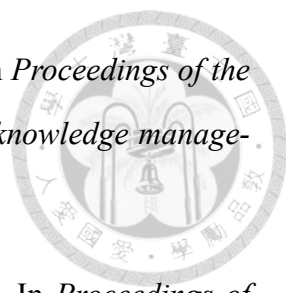


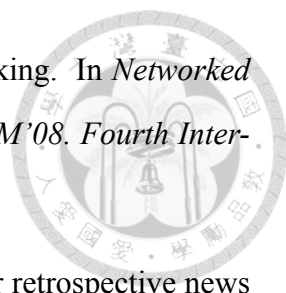


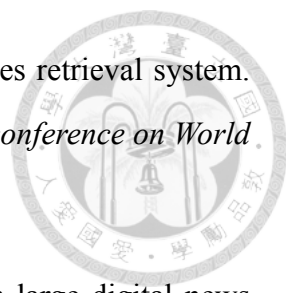
Bibliography

- [1] M. S. Aldenderfer and R. K. Blashfield. Cluster analysis: Quantitative applications in the social sciences. *Beverly Hills: Sage Publication*, 1984.
- [2] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.
- [3] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18. ACM, 2001.
- [4] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, 1998.
- [5] P. Bachrach and M. S. Baratz. *Power and poverty: Theory and practice*. Oxford University Press, 1970.
- [6] D. P. Baron. Persistent media bias. *Journal of Public Economics*, 90(1):1–36, 2006.
- [7] D. Billsus and M. J. Pazzani. A hybrid user model for news story classification. *COURSES AND LECTURES-INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES*, pages 99–108, 1999.
- [8] D. Billsus and M. J. Pazzani. A personal news agent that talks, learns and explains. In *Proceedings of the third annual conference on Autonomous Agents*, pages 268–275. ACM, 1999.

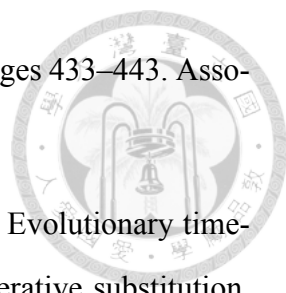
- 
- [9] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 330–337. ACM, 2003.
- [10] I. Cantador, A. Bellogín, and P. Castells. Ontology-based personalised and context-aware recommendations of news items. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 562–565. IEEE, 2008.
- [11] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232. Association for Computational Linguistics, 2008.
- [12] J.-H. Chiang and Y.-C. Chen. An intelligent news recommender agent for filtering and categorizing large volumes of text corpus. *International Journal of Intelligent Systems*, 19(3):201–216, 2004.
- [13] L.-F. Chien. Pat-tree-based keyword extraction for chinese information retrieval. In *ACM SIGIR Forum*, volume 31, pages 50–58. ACM, 1997.
- [14] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 425–432. ACM, 2004.
- [15] Y. Choi, Y. Jung, and S.-H. Myaeng. Identifying controversial issues and their sub-topics in news articles. In *Intelligence and Security Informatics*, pages 140–153. Springer, 2010.
- [16] R. W. Cobb and C. D. Elder. The politics of agenda-building: An alternative perspective for modern democratic theory. *The Journal of Politics*, 33(4):892–915, 1971.

- 
- [17] A. Feng and J. Allan. Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 821–830. ACM, 2007.
- [18] A. Feng and J. Allan. Incident threading for news passages. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1307–1316. ACM, 2009.
- [19] P. Garrett and A. Bell. Media and discourse: A critical overview. *Approaches to media discourse*. Oxford: Blackwell, pages 1–20, 1998.
- [20] R. Gibson and D. Zillmann. The impact of quotation in news reports on issue perception. *Journalism & Mass Communication Quarterly*, 70(4):793–800, 1993.
- [21] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128. ACM, 1999.
- [22] Q. He, K. Chang, and E.-P. Lim. Using burstiness to improve clustering of topics in news streams. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 493–498. IEEE, 2007.
- [23] P. Henshall and D. Ingram. The news manual (three volumes), 1992.
- [24] E. S. Herman and N. Chomsky. *Manufacturing consent: The political economy of the mass media*. Bodley Head, 2008.
- [25] J. Hirschberg, K. McKeown, R. Passonneau, D. K. Elson, and A. Nenkova. Do summaries help? a task-based evaluation of multi-document summarization. 2005.
- [26] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM, 2004.

- 
- [27] S. Lee and H.-j. Kim. News keyword extraction for topic tracking. In *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, volume 2, pages 554–559. IEEE, 2008.
- [28] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113. ACM, 2005.
- [29] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.
- [30] M. E. McCombs and D. L. Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.
- [31] K. McKeown and D. R. Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 1995.
- [32] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia’s newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, pages 280–285. Morgan Kaufmann Publishers Inc., 2002.
- [33] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 446–453. ACM, 2004.
- [34] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn. Newsinessence: summarizing online news topics. *Communications of the ACM*, 48(10):95–98, 2005.
- [35] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

- 
- [36] H. Toda and R. Kataoka. A clustering method for news articles retrieval system. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 988–989. ACM, 2005.
- [37] D. Trieschnigg and W. Kraaij. Hierarchical topic detection in large digital news archives. In *Proceedings of the 5th Dutch Belgian Information Retrieval workshop*, pages 55–62, 2005.
- [38] R. T.-H. Tsai. Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures. *Expert Systems with Applications*, 37(5):3553–3560, 2010.
- [39] T.-H. Tsai, S.-H. Wu, C.-W. Lee, C.-W. Shih, and W.-L. Hsu. Mencius: A chinese named entity recognizer using the maximum entropy-based hybrid model. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), 2004.
- [40] X. Wan and J. Yang. Collabsum: exploiting multiple document clustering for collaborative single document summarizations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 143–150. ACM, 2007.
- [41] C. Wang, M. Zhang, L. Ru, and S. Ma. An automatic online news topic keyphrase extraction system. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 214–219. IEEE, 2008.
- [42] C.-W. Wu, R. T.-H. Tsai, and W.-L. Hsu. Semi-joint labeling for chinese named entity recognition. In *Information Retrieval Technology*, pages 107–116. Springer, 2008.
- [43] R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Confer-*

ence on Empirical Methods in Natural Language Processing, pages 433–443. Association for Computational Linguistics, 2011.

- 
- [44] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary time-line summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 745–754. ACM, 2011.
- [45] W. Yang. Chinese keyword extraction based on max-duplicated strings of the documents. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–440. ACM, 2002.
- [46] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.