ABSTRACT

Title of dissertation: Microblogging Temporal Summarization: Filtering Important Twitter Updates for Breaking News

Tan Xu, Doctor of Philosophy, 2015

Dissertation directed by: Professor Douglas W. Oard
College of Information Studies

While news stories are an important traditional medium to broadcast and consume news, microblogging has recently emerged as a place where people can discuss, disseminate, collect or report information about news. However, the massive information in the microblogosphere makes it hard for readers to keep up with these real-time updates. This is especially a problem when it comes to breaking news, where people are more eager to know "what is happening". Therefore, this dissertation is intended as an exploratory effort to investigate computational methods to augment human effort when monitoring the development of breaking news on a given topic from a microblog stream by extractively summarizing the updates in a timely manner.

More specifically, given an interest in a topic, either entered as a query or presented as an initial news report, a microblog temporal summarization system is proposed to filter microblog posts from a stream with three primary concerns: topical relevance, novelty, and salience. Considering the relatively high arrival rate

of microblog streams, a cascade framework consisting of three stages is proposed to progressively reduce quantity of posts. For each step in the cascade, this dissertation studies methods that improve over current baselines.

In the relevance filtering stage, query and document expansion techniques are applied to mitigate sparsity and vocabulary mismatch issues. The use of word embedding as a basis for filtering is also explored, using unsupervised and supervised modeling to characterize lexical and semantic similarity. In the novelty filtering stage, several statistical ways of characterizing novelty are investigated and ensemble learning techniques are used to integrate results from these diverse techniques. These results are compared with a baseline clustering approach using both standard and delay-discounted measures. In the salience filtering stage, because of the real-time prediction requirement a method of learning verb-noun usage from past relevant news reports is used in conjunction with some standard measures for characterizing writing quality.

Following a Cranfield-like evaluation paradigm, this dissertation includes a series of experiments to evaluate the proposed methods for each step, and for the end-to-end system. New microblog novelty and salience judgments are created, building on existing relevance judgments from the TREC Microblog track. The results point to future research directions at the intersection of social media, computational journalism, information retrieval, automatic summarization, and machine learning.

MICROBLOGGING TEMPORAL SUMMARIZATION : FILTERING IMPORTANT TWITTER UPDATES FOR BREAKING NEWS

by

Tan Xu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Assistant Professor Vanessa Frias-Martinez
Professor Jimmy Lin
Principal Research Scientist Paul McNamee
Professor Douglas Oard, Chair/Advisor
Professor Philip Resnik

Acknowledgments

The road to a doctorate is long, challenging and requires exceptional effort. Luckily, many people helped me along this process. Here, I would like to thank those that make this road smoother.

First and foremost, I would especially like to thank my advisor, Professor Douglas Oard, for his fully support and contributions throughout my doctoral study and this dissertation work. Working with Doug was a blissful experience. Ive benefited so much from his expertise and knowledge as a great researcher, a mentor and a friend. Pursing a doctoral degree is challenging but his support carried me through many difficult moments. He gave me immeasurable advice and invaluable opportunities to work on many interesting projects. It has always been a pleasure to work with and learn from such an extraordinary individual.

Second, I am grateful to the committee members, Dr. Paul McNamee, Professor Jimmy Lin, Professor Philip Resnik, and Dr. Vanessa Frias-Martinez, for their valuable comments and suggestions. I appreciate all of their time and effort serving on my committee and willingness to improve the quality of this dissertation. I am also grateful to Professor Ira Chinoy for his valuable advice and guidance in the proposal stage, which is a critical part of this dissertation.

Third, I would like to greatly thank my earlier academic advisor, Professor Dagobert Soergel (now at University at Buffalo) who encouraged and supported me at the beginning two years of my Ph.D. studies. For all the iSchool and computer science department faculties, who taught me in courses and discussed research ideas

with me, words cannot express the gratitude.

Fourth, I would like to acknowledge financial support from the Human Language Technology Center of Excellence (HLTCOE). Thanks are due to Dr. James Mayfield, Dr. Paul McNamee, Dr. Veselin Stoyanov, Dr. Tim Finin, and Dr. Dawn Lawrie. Without their extraordinary theoretical ideas, computation and programming tutorial, this thesis would have been a distant dream. I would also like to thank the Office of Research Administration for providing me two year's of graduate assistantships. Thanks Vonnie Perkins and Zachary Friedman for all the support and helps.

Fifth, I would like to thank the 9 annotators who participated in this dissertation research. This dissertation would not have been possible or completed without their novelty and salience assessment. I appreciate all of them for volunteer their time and effort to help me finish this study. In addition, I would like to acknowledge the NIST TREC organizers for allowing me to use their Microblog track data.

My colleagues at the Computational Linguistic and Information Processing laboratory and Ph.D cohorts have enriched my graduate life in many ways and deserve a special mention. I'd like to express my gratitude to Dr. William Webber, Mossaab Bagdouri, Ning Gao, Dr. Jiaul Paik, Jyothi Vinjumur, Rashmi Sankepally, Dr. Zhongqiang Huang, Dr. Tamer Elsayed, Dr. Asad Sayeed, Dr. Yejun Wu and Yulu Wang, for their friendship and support.

I would also like to extend my thanks to the staff members in the iSchool and the Institute for Advanced Computer Studies at University of Maryland for their care during my graduate studies. They made the paperwork process much easier by

providing timely and continuous support.

I reserve a special thank you to my family for their unconditional support and encouragement throughout my dissertation work. I would like to recognize the invaluable support and patience of my wife, You Zheng.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all and thank God!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter introduces the research motivation and objective of the thesis, followed by an illustration of an example application scenario. Then, major concepts and terminology used in this thesis are defined. The chapter continues with an analysis of the research problem structure, along with a system framework design, and then highlights three concrete research questions. Finally, the chapter concludes by an overview of this thesis.

1.1 Motivation and Objective

Microblogging has become an important platform for global conversation. It is a new type of broadcast medium wherein the published content takes a form of a weblog but with shorter length. This new medium, not only enables an individual's voice to be heard globally, but also provides us a daily "fast food" for public topics. Microblog function as a sort of digest for readers. According to a 2012 Pew Research Center study, 19% of Americans consumed news or news headlines on "yesterday" social network updates.[1] This is particularly true for breaking news events, where news consumers are eager to be provided with quick, up to the minute updates. As a result, they actively seek new information.

[1] http://www.pewinternet.org/Reports/2012/Connected-viewers.aspx

However, in order to expand the role of microblog in the information delivery ecosystem, the obstacle of excessive quantity must be overcome. For example, on Twitter—one of the most visible microblogging platforms active today—there are over 58 millions tweets (a tweet is a microblog post sent using Twitter) published every day by around 550 million active registered users.[2] As a result, each Twitter search or query could potentially return hundreds of thousands results, far too many for any one person to interpret. Moreover, when incorporating live event monitoring, even more results are returned. In order to meet the information needs of populace, queries would need to be conducted repeatedly. According to the stop-searching theory provided by Bates's, such high numbers could lead to an early search termination based on greater needs of the user [17].

This thesis is intended as an exploratory effort to investigate computational methods seeking to augment human efforts when monitoring the development of breaking news events from microblog by extractively summarizing microblog posts in a timely manner. This thesis aims to recommend methods for querying topics and extracting valuable microblog posts in real time based on a sequence of timely ordered and rapid inputs. This is advantageous for two main reasons: (1) it relieves the searcher from having to perform repeated searches and therefore consolidates effort; and (2) it assists tweet publishers in directly disseminating their content to an interested audience.

Figure 1.1 illustrates this expected improvement in function. On January 24th,

[2]http://www.statisticbrain.com/twitter-statistics/

2011 at 13:32 (UTC),[3] a suicide bomber attacked the Domodedovo International Airport in Moscow. The bombing killed 37 people and injured 180. Within a two-hour window immediately following the incident (from 14:00 to 16:00), multiple tweets discussing the event were published. Note that Figure 1.1 displays only 1% of the sampled English language tweets that were manually selected according to topical relevance. This indicates that, in reality, there were many thousands of additional English language tweets not included in Figure 1.1. We must consider both that a Twitter search is less exacting than a human assessor, and that a human assessor would have needed to perform a manual search several times during the two-hour period. Only the tweets highlighted in yellow are necessary for the searcher to understand the major perspectives on the real time event. As a result, the new query function could save many hours of human labor.

The belief that microblog posts should be presented in a more concise manner has gained traction in recent months. Bandari et al. investigated predicting the major discussion topic trends in microblog [14]; Amer-Yahia et al. analyzed people's overall opinions or sentiments reflected from their microblog posts [8]; Shamma et al. attempted to use responsive tweets' volume over time as well as salient keywords among these tweets to identify shifts in topics of interest or momentary topics [169, 168, 170]; Finally, in 2014, Text Retrieval Conference (TREC) Microblog track organized a tweet timeline generation task [107]. Nevertheless, there is still a need for the consideration of future microblog posts. As seen in Figure 1.2 if we

[3]Coordinated Universal Time. For convenience, the rest of the paper uses UTC time by default unless otherwise noted.

14:05:21	At least two dead, dozens injured in blast at Moscow's Domodedovo airport - Interfax news agency #Russia
14:08:48	Explosion with casualties at Moscow airport: Source: www.forexlive.com --- Monday, January 24, 2011Perhaps a bit... http://bit.ly/eoSMXk
14:20:13	Blast at Moscow's Domodedovo airport was suicide attack - source citing initial probe #news
14:23:11	BBC News - Domodedovo airport: Blast rocks Moscow's main airport bbc.co.uk/news/world-eurâ€¦ #breakingnews #moscow
14:23:21	Explosion shakes Moscow's busiest airport \n (AP)\n http://bit.ly/dRgipc & http://dld.bz/Hket
14:27:15	MT @BBCBreaking: Russian media reporting a suicide bomber killed at least 10 people at Moscow's Domodedovo Airport
14:27:18	10 killed, 20 injured in Moscow airport blast. Russian media: Suicide bombing at arrivals hall of Domodedovo Airport.
14:28:27	Update: About 20 killed in Moscow airport suicide attack - Itar Tass
14:29:19	BBC News - Domodedovo blast: Explosion rocks Moscow's main airport http://www.bbc.co.uk/news/world-europe-12268662
14:40:50	[REUTERS]: Suicide bomber at Moscow airport kills 10 - report: MOSCOW (Reuters) - A suicide bomber at Moscow's ... http://reut.rs/hAFCzF
14:40:55	MOSCOW (BNO NEWS) -- Death toll from suicide attack at Moscow airport rises to 23, at least 130 injured.
14:41:10	Deadly suicide bombing hits Moscow airport: At least 10 people killed and 20 injured in suicide bomb blast at Mo... http://bit.ly/fz3ejE
14:41:36	Report: 10 dead in Moscow airport explosion: At least 10 people were killed and 20 injured in a suicide b... http://on.msnbc.com/flzVXg
14:43:02	Explosion rocks Moscow airport: MOSCOW - An explosion at Moscow - Domodedovo Airport has killed at least ten peo... http://bit.ly/g1dmI9
14:47:01	*UPDATE" Death toll at Moscos airport bombing climbs to 31, over 100 injured: reports
14:48:05	とりあえず友達は大丈夫みたい・・・ Deadly blast at Moscow's Domodedovo airport - RT http://bit.ly/hJFs3A
14:49:13	LATEST: At least 20 killed in explosion at Moscow's busiest airport http://huff.to/f3sSXw
14:49:18	Bad news. Early report is 23 dead, 130 injured (HT BizballMaury) RT @nytimes: NYT NEWS ALERT: Explosion Is Reported at a Moscow Airport
14:49:22	Blast at Moscow Airport.. http://news.yahoo.com/s/nm/20110124/ts_nm/us_russia_blast_airport
14:52:29	Oh my.. :(RT @washingtonpost: Explosion rocks Domodedovo, Moscow's busiest airport; at least 20 injured http://wapo.st/fnYC5H
14:53:40	Blast rocks Moscow's main airport: Moscow's Domodedovo airport - the busiest in the Russian capital - is h... http://tinyurl.com/4lr5kbo
14:54:23	Report: Explosion kills 23 at Moscow airport \n (AP)\n: AP - The Russian state RIA Novosti news agency says a... http://bit.ly/i3mgRZ
14:54:52	?!? ?!? ?!? ?!?.. RT @BBCBreaking Russian media reporting that at least 23 people were killed and 100 injured in Moscow airport bombing
14:56:20	Domodedovo blast: Explosion rocks Moscow's main airporthttp://armenion.com/?p=2134 #домодедово #domodedovo
14:57:34	Russian media now reporting that at least 31 people were killed and 130 injured in bombing at Moscow's Domodedovo airport, from AFP via BBC
14:59:37	Check out this article from FOX News. Report: At Least 10 Dead in Explosion in Moscow's Busiest Airport: http://fxn.ws/fPeWtK
15:00:04	Report: Explosion kills 23 at Moscow airport http://bit.ly/hwGiTm
15:02:08	From #AP: Bomb at #Moscow airport kills 31, injures 130. Story @ http://bit.ly/ga3Nnd
15:03:05	Security stepped up at Moscow airports after 31 people died in an apparent terrorist blast in #Domodedovo. Vnukovo and Sheremetyevo affected
15:04:03	At least 10 killed in apparent suicide bombing at Moscow airport. Sarah Palin can see the result of her hate spe... http://fk.cm/5902836
15:06:30	@cnnbrk: Blast that killed 31 at #Moscow airport was terrorist attack, Russian authorities say. http://on.cnn.com/f3zHcL"
15:08:04	deadly blast in #Moscow airport leave dozens dead - 31 killed and 130 wounded. Prayers for the victims
15:16:51	EWN International Explosion rocks Moscow's busiest airport: Moscow's busiest airport - Domodedovo has been rocke... http://bit.ly/hktYzw
15:17:06	Up to 31 killed in Moscow airport attack. Police seeking 3 men, according to report http://wapo.st/eA61BZ
15:18:58	At least 31 killed in Moscow airport suicide bomb (AFP): AFP - A suspected suicide bombing Monday killed at least 31 people and wound...
15:19:00	At least 31 killed in Moscow airport suicide bomb (AFP): AFP - A suspected suicide bombing Monday killed at least 31 people and wound...
15:20:21	Blast rocks Moscow's main airport - http://www.bbc.co.uk/news/world-europe-12268662
15:21:43	At least 31 dead in Moscow airport explosion http://bit.ly/gHDBVd (via @cnni)
15:22:53	Suicide bombing kills 31 people: At least one police officer was shot Monday in St. Petersburg, Fla., and the sh... http://bit.ly/e7yk8B
15:23:08	Terrible news from Moscow #Domodedovo aiport: at least 31 dead and 130 injured as a result of explosion. Russian media slow to report news.
15:24:26	ViewHeadlines.com - Deadly blast at Moscow airport viewheadlines.com/News/Article.aâ€¦ via @wibiya
15:25:33	Explosion at Moscow's Domodedovo airport kills at least 10 people http://tinyurl.com/6ypgpy6
15:25:53	Deadly Blast Strikes in MoscowÂ's Main Airport: ---Quote--- MOSCOW Â— An explosion rocked an international ... http://bit.ly/dQUNrP #tcot
15:26:26	Explosion Kills 31 At Moscow Airport: An explosion ripped through the international arrivals hall at Moscow's bu... http://bit.ly/hfLg12
15:28:44	Bomb kills passengers at Moscow airport http://wapo.st/gzLYBl
15:30:30	Live: Moscow airport explosion. http://news.bbc.co.uk/2/hi/europe/9372022.stm
15:31:57	Suicide attack at Moscow airport kills over 30 (WRAPUP 1) en.rian.ru/russia/2011012â€¦ via @ria_novosti
15:32:20	Fatal Blast Reported At Moscow Airport Terminal: Russian authorities Monday reported a fatal explosion at ... http://tinyurl.com/4eb838w
15:32:49	Explosion Kills 31 At Moscow Airport: An explosion ripped through the international arrivals hall at Moscow's... http://on.wesh.com/gPRKOW
15:42:21	At least 31 dead in Moscow airport explosion: An explosion attributed by Russia's Investigative Committee to... http://bit.ly/dKwJF7 CNN
15:43:41	Moscow Bombing :: Medvedev postpones his scheduled flight to Davos
15:47:03	Moscow airport bombing :S
15:48:37	Καμικάζι σε αεροδρόμιο: At least 30 people killed and 50 others injured in explosion at Moscow's Domodedovo international airportâ€
15:51:06	Graphic video from inside Domodedovo Moscow airport following terrorist blast - http://bit.ly/e7ac5V via @RodrigoEBR
15:51:13	At least 31 people were killed and more than 100 injured in a suicide bombing at Russia's biggest airport, Russi... http://bit.ly/eyk4Dm
15:51:56	Suicide bomber blamed for #Moscow airport attack that killed 31, Russian state TV reports. http://on.cnn.com/f3zHcL
15:52:53	Suicide bomber kills 31 and injured at least 100 at moscow airport.
15:54:31	Russian President: Apparent â€˜Terrorist Attackâ€™, witnesses say carried out by two suicide bombers, according to RIA: http://abcn.ws/eJlBnW
15:58:40	Deadly Blast at Moscow's Domodedovo Airport http://twurl.nl/kssyzu

Figure 1.1: Example Twitter temporal summary for 2011 Moscow airport bombing incident.

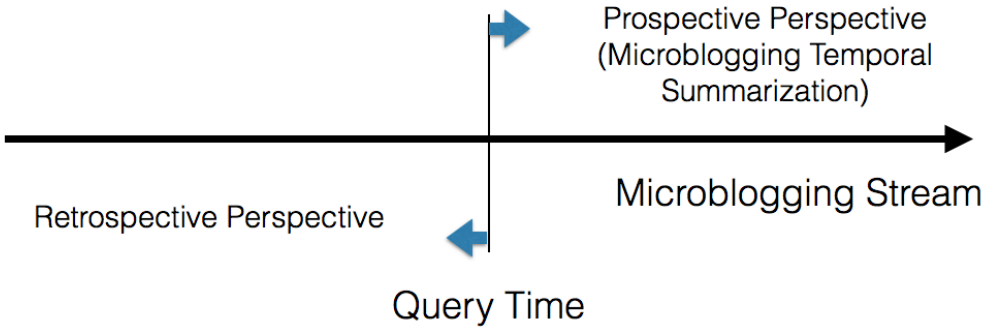


Figure 1.2: Timeline perspective of microblogging temporal summarization.

specify a query issue time, then most current methods focus on microblog posts published prior to this specific time point. This thesis aims to shed light on this prospective problem. The term "temporal summarization," first introduced in the TREC Temporal Summarization track [13] is used to denote the expected outcome of a timely ordered relevant and reliable microblogging update regarding a specific query topic.

1.2 Example Application: Tweets in Tomorrow's News

To evaluate and compare different microblog filtering methods in a mathematical framework, I choose to explore the problem of predicting whether a tweet's content would show up in a future news report. There are many factors that determine whether a tweet is newsworthy. In fact, in journalism, there have long been serious discussion regarding what constitutes "news values" [19, 65, 66, 143, 77, 148, 166]. It is necessary for journalists to have information resource from which they can derive a good news report. They must also be able to generate new insights or angles for the story. For journalism today, social media, particularly Twitter, functions as

one of the most important sources for this purpose [135]. According to a study conducted by Oriella PR Network, 54% of modern journalists find their news sources from tweets.[4] In order to cope with a rapidly changing news cycle, journalists use optimized tweet input to generate stories. If the content of a tweet content appears in a future news report, we can assume that the tweet could potentially be useful for a journalist and thus is worthy of a recommendation.

The following three specific subcases for tweet recommendations are of the most practical importance:

- **Recommending tweets that are topically relevant:** the input Twitter stream composed of tweets created for various reasons, i.e., daily chatter, conversation, sharing information, and reporting news [89]. However, according to Diakopoulos et al., tweets deserve further consideration of newsworthiness provided they are still on topic [52]. Therefore, a robust method for filtering out irrelevant tweets must be devised.

- **Recommending tweets that share novel information:** duplicated tweets reporting similar information are the greatest waste of resources. This problem is exacerbated by the ability to copy-and-paste or a more convenient "retweet" (a re-posting feature provided by Twitter for quick information sharing). Identifying tweets that share only new information can effectively reduce journalists' efforts by avoiding reading of previously discovered information.

- **Recommending tweets that sharing nontrivial information:** Balanc-

[4]http://www.oriellaprnetwork.com/sites/default/files/research/Brands2Life_ODJS_v4.pdf

ing the number a tweets to which a journalist is exposed can be more easily accomplished by eliminating trival tweet updates. This is, however, difficult, because it risks the loss of unusual (unique or unexpected) content. Exploring a cut-off boundary opens up the possibility of a potential attempt to further improve the application.

In addition to the benefit to journalists, who can then utilize the power of the human sensors web, a synthesized stream of texts also confront a number of other professions across multiple domains. For example, for years, there have been efforts underway to address the problem of topic tracking in news articles [5]; from academic publications or patents, many are interested in understanding the evolutionary pattern of certain technologies or research topics, so that they could identify and investigate a prospective future one [120]; and last but not least, in social media, many industries have expressed a growing interest in reputation control [10].

1.3 Structure of the Problem

In this thesis, I design a microblogging temporal summarization framework for this tweet recommendation application, as depicted in Figure 1.3. A continuous stream of chronologically ordered microblog posts according to their publication time, input at a high arrival rate (thousands per second in the case of Twitter). Given a query topic, a relevance filtering component filters out irrelevant posts; a novelty detection component detects posts containing new updates against past ones; and a salience detection component finally selects posts reporting important

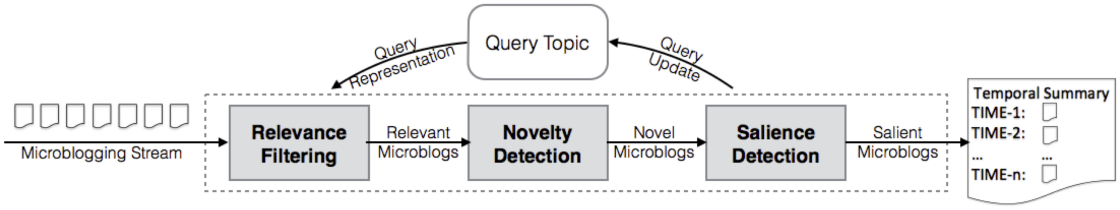


Figure 1.3: Query-focused microblogging temporal summarization framework.

content, and outputs this information to a temporal summary of the topic. This framework is designed to work in an online mode, thus the temporal summary is built up incrementally, in a simulation of the real-time tweet recommendation. Additionally, the query topic's internal representation is updated continuously during entire process according to what has been seen, itself a simulation for the evolution of our knowledge regarding the topic.

This framework is able to address the three aforementioned crucial problems in a cascade way.

1.3.1 The Relevance Filtering Problem

Relevance filtering and retrieval is an usual preprocessing step adopted by many related query-focused microblog mining studies for efficiency consideration, i.e., online reputation management [10, 9], tweet-based news exploration and information visualization [8, 52, 23], and social media opinion mining and sentiment analysis [145, 150]. In order to simplify the problem, these researchers generally begin with sets of documents that are "known" to be relevant. The assumption is that the process of finding relevant material has already produced relatively mature solu-

tions and thus can be studied separately. However, as concluded in the 4-year TREC Microblog track, isolating relevant tweets from irrelevant ones is still a relatively difficult problem [105, 107, 147, 176]. Thus, I examine this artificial assumption by first inspecting how to filter relevant microblog posts effectively. This is an especially important problem to solve considering the following step is novelty detection, which greatly depends on the system's ability to begin with relevant posts, Without this ability, irrelevant posts will be identified as containing new information.

To develop a robust relevance filtering component, the following questions need to be answered:

- How does query expansion utilizing local and Web resources affect microblog filtering effectiveness?
- What is the effect of adding word embedding to microblog filtering?
- Is supervised learning from labeled training data resulting in more effective filters than unsupervised models?

1.3.2 The Novelty Detection Problem

Reducing duplicated microblog posts in the relevant microblog stream is an additional step that can be taken in order minimize the users investment of effort. Therefore, recommending a microblog post to a user reports new information against past recommendations. However, as detailed in an early novelty detection work, the first difficulty of this task is to define what is meant by "new" [7]. Ma and Perkins further asserted that an accurate understanding of novelty is also required [115].

Although, designating a definition for novelty without a context is open-ended, within the scope of this study, I focus on textual cues to represent a microblog post's novelty based on the intuition that it is very likely that novel microblog posts use very different words.

More specifically, designing an effective novelty detection involves the following questions:

- What are the most potential features for representing a microblog post's novelty?
- Is the ensemble learning approach helpful for the novelty detection effectiveness?
- Can a batch mode approach (i.e. clustering-based) be an effective novelty detection method with the consideration of delayed prediction?

1.3.3 The Salience Detection Problem

For the final step, trivial updates regarding a query topic should be ignored. Salience detection is one method for accomplishing this goal, as in a typical automatic summarization task, salience is an important consideration when selecting content for the summary [117]. However, existing methods have been designed based on the computing of the center or grouping information of the source document(s), i.e., the graph-based centrality approach [61, 127, 156], the clustering approach [79, 185] and the recent topic modeling approach [40, 49, 73]. This thesis introduces the difficulty of seeing the source of microblog posts incrementally.

Therefore, similar to novelty detection, unless we delay the emission of the temporal summary, the topical salience cannot be computed with consideration of later selected microblog posts. As a solution, this thesis explores an alternative method that attempts to obtain topical salience from past related news reports. For example, if the query topic regards an earthquake, then we could learn what would be important for such a topic from past earthquake reports.

More specifically, the following questions must be addressed:

- How effective is using a microblog post's quality measurements in salience prediction?
- Can features extracted from past relevant news reports be helpful in deciding a microblog post's salience?

1.4 Contributions

Inspired by the story that Twitter first broke the news of Bin Laden's death,[5] and some unhelpful real-life Twitter search experiences, I attempt to show how computational methods - with its tremendous processing efficiency - can effectively improve manpower for the needs of news tracking in the microblogosphere. The contributions of this thesis include:

- Introducing a framework to systematically study the problem of microblogging temporal summarization and, thereby, find optimal solutions with respect to

[5] http://www.cnet.com/news/twitter-delivers-news-of-bin-ladens-death-first/

the three sub-problems of relevance filtering, novelty detection and salience detection.

- Following a Cranfield-like evaluation paradigm, a series of experiments are designed to evaluate the proposed methods for each sub-problem, and for the end-to-end system.

- Creating microblogging novelty and salience test data based on TREC Microblog evaluation.

- Demonstrating improved microblog relevance filtering by integrating query expansion, document expansion and word embedding to overcome the issues of data sparsity and language gap.

- Demonstrating improved microblog novelty detection by investigating various textual novelty measurements and the ensemble learning techniques.

- Demonstrating improved query-focused hierarchical clustering technique by predicting query optimal threshold.

- Demonstrating improved microblog salience detection by utilizing prior verb phrase usage propensity learned from past related news reports.

Ultimately, this research not only develops a running system using the proposed conceptual framework to predict the appearance of a tweet's content in a future news, but also explores particular methods and illustrates its effectiveness. The results could point to future research directions at the intersection of social

media, computational journalism, information retrieval, automatic summarization, and machine learning.

1.5 Overview of the Thesis

The remainder of this thesis is organized as follows: Chapter 2 reviews related previous research work on user information seeking behaviors, microblog-based news exploration, topic tracking and novelty detection, and detailed studies of automatic summarization and microblog retrieval and filtering; Chapters 3, 4, and 5 detail methods addressing the problem of microblog filtering, novelty detection, and salience detection, respectively, with experimental design and result analysis. Chapter 6 concludes the entire thesis and describes limitations and future research.

Chapter 2

Literature Review

This chapter reviews related research from the following perspectives: user information seeking behaviors in microblogosphere and the general microblog-based news exploration; a review of news topic tracking and novelty detection; and detailed studies of generic extractive summarization, update summarization and preivous works in temporal summarization. Finally, this chapter concludes by examining related research in microblog retrieval and filtering.

2.1 Microblogging Monitoring Behaviors

Microblog has become a primary channel by which people not only share information, but also search for information. Because the information in microblog grows fast, updates frequently, and covers a wide range of topics - as with the World Wide Web - it is necessary for interested users to effectively and efficiently find desired information. Therefore, search intentions similar to those available to traditional Web search can be found in the microblogosphere. They include: navigational (i.e., finding a particular user or group), and informational (i.e., finding information on a particular topic) [33]. However, according to Mishne and De Rijke's blog search behavior study, searchers' informational intent can be further divided into two classes: tracking references to named entities, and identifying blogs or posts that focus on

Type of Information	**Explanations**
Timely information	Breaking news, current events, real-time reporting, friends daily activities
Social information	People with specific interests, information or microblog posts of a specific user or group, and peoples' overall opinions on a particular topic
Topical information	Similar to traditional Web search, people also search on Twitter for information of specific interest

Table 2.1: Information people interested to seek from microblog. [178]

a specific topic [131]. By analyzing user queries issued to Twitter, Teevan et al. observed this phenomenon, where an extensive reuse of the same Twitter queries (56%) were issued more than once by the same user [178]. According to a qualitative analysis, the purpose of this repetition of search is to monitor topics over time. This was also confirmed by Teevan et al's user studies about which information users are most interested in obtaining via microblogs. As shown in Table 2.1, the two primary types of information are timely and topical [178].

When using an information system, users normally have a specific information need. The user's objective is to satisfy this information need, and the role of the information system is to satisfy that information need with minimum expenditure of effort by the user. With regard to the microblogging monitoring need, the expression of this need from the user, the presentation of the results system returns, and

the interaction between the user and the system may take various forms. For example, Naaman et al. observed that an information user may "follow" other users for their latest updates, or they may even ask direct question regarding updates in the form of their own microblog posts [134]. However, conducting a search is one of the most popular and straightforward strategies utilized for returning on-topic updates. Applying Wilson's nested model of information seeking [196], an information users microblogging seeking behavior is a subset of their microblogging behavior, particularly as concerns the variety of methods the user employs to discover and gain access to microblog posts. Additionally, a user's microbloging search behavior is a subset of their microblogging seeking behavior, particularly concerning the interactions between users and computer-based search systems for microblog posts [200]. Wilson classified this monitoring behavior as "ongoing search," which refers to a continuous search that is carried out to update or expand one's current framework of knowledge, ideas, beliefs, or values [195].

Now that we consider microblogging monitoring behavior as a series of interconnected search on a single, problem-based theme - as pointed by ODay and Jeffries - one of the fundamental issues every searcher faces is to determine when to stop searching [146]. This can be done either by terminating the search completely or by starting a new search. In their work, ODay and Jeffries defined four triggers that can lead to a new search, as well as describing three stopping circumstances. These can be summarized by stating that users will stop searching when there are no additional compelling triggers for further search or when they have done an appropriate amount of searching for the task [146]. Bates takes a different approach,

using a cost-benefit analysis, to characterize the decision to stop searching with a similar conclusion [17]. The underlying assumption is that searchers will make a decision that maximizes expected utility: if stopping yields higher expected utility than continuing, the searcher will cease his search.

However, monitoring microblog by conducting continuing search results with limited support influences the decision about when to stop a search. For every search, searchers have to expand effort to (1) browse numerous returning posts which contain much duplicated content and (2) identify and digest valuable posts that match the searcher's news topic monitoring intent. Thus, in this situation, every continuing search will cause an earlier stop, not because the user is satisfied with the search results, but because of the increased cost of further searching. For this reason, designing a system to automatically identify topical updates from microblog streams and summarize important results minimizes each monitoring searcher's effort and improves the overall capability to satisfy his/her information need.

2.2 Microblog-based News Exploration

Microblog has become an increasingly critical platform for distributing both global and local news. In recognition of the fact that microblog is a excellent source for news stories, professional journalists have begun using it to drive new insights and angles [135]. According to a study conducted by Oriella PR Network, 54% of modern journalists find their news sources from microblog.[1] Therefore, efforts

[1]http://www.oriellaprnetwork.com/sites/default/files/research/Brands2Life_ODJS_v4.pdf

are underway to automatically reveal and analyze the structure and dynamics of microblog posts covering news events, such as atomic facts and relations between named entities.

Shamma et al. explored how Twitter's usage pattern can reflects the structure of live media [169, 168, 170]. More specifically, the researchers focused on using information of responsive tweets' volume over time - including salient keywords among these tweets - to detect topic-of-interest shifts in order to help identify a news topic's thematic segments or momentary topics. Diakopoulos et al. extended this idea in their Vox Civitas system by utilizing more powerful filtering and visual mechanisms, which enabled an interactive exploration of a news topic's responsive tweets for journalistic purposes [52]. Similar systems that visually analyze microblog posts over a particular news topic include: TwitInfo, which provides a timeline-based analytic dashboard for real-time Twitter feeds that can help users detect sub-topics and explore the query topic further via geolocation and sentiment analysis [23]; MAQSA provides a similar dashboard display as well as extra displays that list relevant news articles and extracted entities and sub-topics [8]; and Dork et al. presented a Visual BackChannel for visualizing topics discussed on microblog posts, where the primary view is a Topic Stream, an interactive stacked graph that visualizes live-changing textual data as depicted in Figure 2.1 [54].

In the figure, **left**: the development of a query topic over time intervals t_i is represented as a stream-like shape defined by two cubic Bezier curves whose control points, $p_{ia,b,c}$ and $q_{ia,b,c}$, define the changing widths w_i of the topic stream; **middle**: the chosen color scale ranges between a rich, bright green for newer topics and a

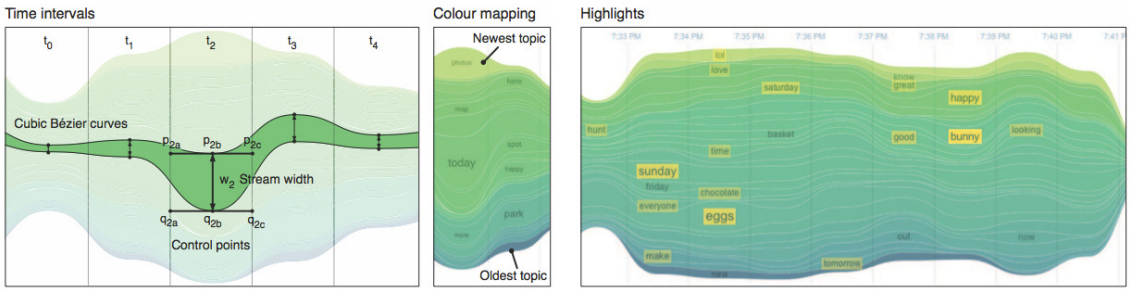


Figure 2.1: Visual BackChannel topic stream [54].

matte, dark blue for older topics; and **right**: the current activity in the backchannel is highlighted with a yellow-colored background [54]. Note that given an news topic, the newest detected "sub-topic" is placed at the top of the stream and the most recently mentioned old "sub-topic" terms are highlighted by a temporarily yellow-colored background in the stream.

In addition to explore microblogs covering news, another group of studies that more closely relates to this thesis research attempts to highlight microblog posts worth reading. Criteria that has been tried includes topic relevance, author's reputation, microblog post's popularity, and recency [136, 149, 183].

2.3 Topic Detection and Tracking

With the exponential growth of information on the Internet, monitoring news from a stream of multi-source newswire articles has been explored by the Topic Detection and Tracking (TDT) works. The main purpose of the TDT is to organize news articles according to the event topics reported [5]. More concretely, it includes: (1) *story segmentation* task, which breaks text into cohesive story segments; (2)

first story detection task, which recognizes the onset of a new topic; (3) *clustering detection* task, which groups together news stories that discuss the same topic; (4) *tracking* task, which traces the development of a news topic; and (5) *story link detection* task, which decides whether two randomly selected news stories cover the same topic [5].

The sub-task most closely related to this thesis is the *tracking* task. This is similar because topic tracking also attempts to process streaming news reports and begins with a given news topic of interest that represents a few news sample reports. A wide range of statistical and machine learning techniques can be used as a solution. For example, some techniques apply classical vector-space and probabilistic models to calculate similarity between a news topic and an incoming news report [142, 151]. Others apply a variety of language modeling approaches for this same purpose [90, 112, 201, 110, 95, 137, 100, 101, 189]. With regard to the binary classification decision for each incoming news report - regardless of the fact that most frequently adopted and straightforward approach is to threshold its similarity score with a target news topic - there are other more sophisticated supervised machine learning approaches, such as k-Nearest Neighbor (kNN) [37, 202], Decision Tree induction [37], the Boosting method [164], Neural Networks [158], and Support Vector Machines (SVM) [69, 209]. Finally, Yang et al. found that combining outputs of alternative methods for producing a joint topic tracking decision can improve the performance of any single method [203].

Updating news topics from evolving data streams is another research subject addressed by both TDT and this thesis [87, 121, 124]. Related works also apply

TDT technologies to news monitoring tasks in a microblog stream. Petrovic et al. examined the first story detection problem in Twitter streams [154, 155]; Lin et al. researched a general topic tracking problem, also in Twitter streams [106].

2.4 New Information Detection

Given the "information overload" of email, Web content, and social media, detecting relatively new information for users has become problematic in recent years. The first mention of this problem occurred in the TDT work in 1999 during a summer Novelty Detection workshop at John Hopkins University's Center for Speech and Language Processing [7]. In that workshop, the New Information Detection (NID) task was defined as identifying the onset of new information within a topic by flagging the first sentence that contained a mention of the topic. Despite the fact that this was the first trial that raised the issue, continued research on temporal summarization did not make significant progress due to an unclear definition of "new". Without this clear definition, 80% of the sentences in news articles were found to contain some new information [6].

Research from the the TREC Novelty Track conducted from 2002-2004 continued this line of questioning. However, the goal shifted to exploring methods that reduced the amount of redundant material that was shown to a searcher. The task was defined as: given a topic and a chronologically ordered set of topically relevant documents segmented into sentences, return sentences that are both relevant and new from what have been previously seen [175]. Because the setup of the evalua-

tion, most participating groups adopted a similar two-step system framework, which addressed identifying relevant sentences and novel sentences separately. Moreover, according to an overview of this track, there was no dramatic difference in the range of approaches, where relevant sentences were selected by measuring similarity to the topic, and novel sentences were selected by measuring dissimilarity to past sentences [174, 173].

Considering the feature space, Soboroff and Harman categorized the assortment of all novelty metrics that were used into two broad categories: statistical metrics and semantic metrics [175]. The former includes metrics from traditional retrieval models such as vector space or language modeling, query expansion techniques, and document sentence term expansion with dictionaries or corpus-based information. The latter category includes metrics relying on deep natural language processing (NLP), such as named entity recognition and alignment, use of verbs and verb phrases, and in one case, use of ontology to conceptually expand topics. Tsai et al. approached from a different perspective and categorized novelty metrics into two different types based on whether the ordering of two comparing sentences was taken into account when computing features [180]. Since they argued that the nature of novelty detection required to consider the ordering of seeing sentences, they distinguished that asymmetric metrics yield different results for different ordering of the same two sentences - such as word-overlapping count - from symmetric metrics, which yield the same results regardless of the ordering of two sentences - such as cosine similarity [212].

Regarding the novelty decision, many participants employed a threshold-based

approach, which determined whether a sentence was novel based on whether the sentence's novelty metrics generated a score that exceeded a pre-determined threshold [1, 165, 25]. This threshold could be determined using either 2003 track data, or in an ad-hoc manner. Teams also used Support Vector Machines (SVM) to learn a classifier in order to make a binary decision about whether a sentence was novel or not [91, 179]. This approach was also adopted by some later studies of novel sentence detection [70].

In addition to conducting novelty detection at the sentence-level, similar research was conducted at the document level and the topic level. For example, Tsai and Zhang attempted to identify on-topic documents that contained novel information by composing judgments from sentences [181]. Zhang et al. performed adaptive information filtering that learned and identified relevant and novel documents [211]; and the TDT first story detection task targeted detecting novel topics from newswire streams.

2.5 Extractive Multi-Document Summarization

Automatic text summarization is another way researchers attempt to solve the "information overload" problem caused by the continually increasing amount of textual information. For generic summarization tasks, there are two categories of approach: extractive summarization and abstractive summarization. The former extractively selects sentences or text units from the original text to construct a summary, while the latter tends to build a summary from the original text seman-

tically to simulate the method by which humans summarize. Even though abstractive summarization is the ideal way to automatically build a summary, extractive summarization is a more practical approach, because: (1) extracted sentences can directly form a valid yet readable summary; (2) it can be evaluated automatically by comparing sentences that one method selects to the set that is known to be in a good summary; and (3) the resulting summary can be further processed and used as a basis for constructing abstractive summarization. Therefore, inspired by the extractive summarization approach, this thesis attempts to formulate a temporal summary by extracting microblog posts.

In previous works, sentence graph-based methods are widely used for extractive multi-document summarization. In this type of graph, each vertex represents a sentence (or a textual unit), and each edge represents a relationship between two sentences, such as various lexical similarities. Then, certain graph-based ranking methods are used to select sentences. For example, sentences may be selected by using the centroid of a sentence cluster [79, 157, 139]; by using various voting techniques [61, 127]; or through a stationary distribution after Markov random walks over the entire graph [60, 185, 68].

Additionally, another group of term graph-based extractive summarization methods exists. In this type of graph, each vertex denotes a term, which could be a named entity or a verb. The edges then represent relationship between two terms, which could be either co-occurrence information or semantic dependency relations. In fact, many studies on the principle of utilizing information about terms to produce coherent and semantically relevant summaries have been performed. For example,

there are works focusing on using term frequency related features or alternative term scoring functions to determine a sentence's importance [140, 48]. Others attempts a purely data-driven approach to learn term weights for sentences [16, 49, 40]. The hypothesis behind these approaches is that they would be able to provide finer text representation and thus, could be favorable to sentence compression that was targeted to include more informative contents in a fixed-length summary. Nevertheless, these advantages relied on appropriately defining, selecting and scoring terms.

Another way of viewing existing summarization methods is to considering their learning techniques. Many of the above mentioned works employed unsupervised machine learning techniques, which are based on the rationale that terms and sentences can reinforce each other to determine their salience. On the other hand, Kupiec et al. performed one of the earlier studies that applied supervised machine learning techniques to document summarization [97]. They attempted to classify a sentence either as an in-summary or a non-summary sentence. A later work by Conroy and O'leary also addresses this approach [47]. Amini and Gallinari explored using unlabeled data to improve the classification performance [11]. This could also be called a semi-supervised machine learning approach. Metzler and Kanungo used a learning-to-rank technique for sentence ranking and selection [126]. In addition to these above approaches, a more recent trend of automatic summarization depends on purely data-driven methods, which have shown remarkable improvements [16, 49, 40].

Moreover, following some previous work in summarization, this thesis utilizes filtering and novelty detection as summarization preprocessing steps. Summariza-

tion preprocessing normally attempts to prune irrelevant sentences in advance by using less complex computational methods to improve system efficiency. For example, Zhang et al. used a series of content filters to capture changed information over the historical document collection for the update summarization task [207]. There are also summarizaton postprocessing works that attempt to refine sentences into the final summary to improve system effectiveness. For example, Lin et al. applied a modified Maximal Marginal Relevance (MMR) algorithm to incrementally re-rank sentences to put into a final summary [108].

2.6 Update Summarization

At the onset of summarization studies, the primary research focused on summarization techniques perform in a batch mode, such as those given a collection of documents (or a single document) [117]. However, in realizing the value of information that was consistent with previously available textual resources but contained new content, summarizing the "update" has become one of the focuses of recent automatic summarization research. As a leading effort of evaluating auto-summarization systems, DUC [2] introduced Update Summarization task in 2007. The task seeks to generate a 100-word multi-document summaries of a set of newswire articles under the assumption that the user has already read a set of

[2]The NIST (National Institute of Standards and Technology) initiated the DUC (Document Understanding Conference) series in 2000 to evaluate automatic text summarization. It enables participating groups to compare their system with each other and provides manual evaluation of their summaries.

earlier articles. Its goal is to inform readers of novel information about a particular topic [3]. DUC moved to TAC in 2008 with new summarization evaluation tracks, and the Update Summarization task was kept in TAC until 2011.

Most works addressing the DUC/TAC Update Summarization task were aimed at the extraction aspects of summary generation. The key challenge is to select sentences that are biased to the given topic and that also contain evolving content. To this end, various sentence selecting methods have been attempted. Hickl et al. constructed certain knowledge representations from a cluster of documents, and then selected sentences that could add new facts of the current knowledge into the update summary [82]. This was found to be the best system at the DUC 2007 update summarization track. Witte et al. used a heuristic fuzzy coreference cluster graph to select new sentences [22]. Although this method requires manually tuning the sentence ranking mechanism, it can be generalized to apply to any summerization tasks. He et al. employed manifold-ranking frame for sorting sentences, and used an iterative feedback mechanism to model the dynamically evolving topics [163].

Because traditional summarization techniques are designed in an offline fashion, they can not easily handle newly added content easily. When encountering a new piece of information, the entire graph needs to be recalculated. Therefore, Wang and Li proposed using a COBWEB algorithm to incrementally update a current hierarchical sentence clustering tree [187]. The sentences in the final summary are then selected from each node of the hierarchical tree from top to a user specified layer. This method was tested on real-world disaster management data and TAC

[3]http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html

benchmark data, and outperforms classical clustering and sentence-graph methods. Meanwhile, another group of summarization strategies avoided this problem by considering each sentence's selection decision independently. Kupiec et al. performed one of the earlier works that applied supervised machine learning techniques to classify each sentence either as an in-summary or a non-summary sentence [97]. Later work by Conroy and O'Leary also used this method [47]. Hickl et al. maintained certain knowledge representations from a historical documents, and selected new sentences by checking whether they could add new facts to the current knowledge base [82]

There were also update summarization strategies that relied on post-processing to reduce redundant sentences in the summary. For example, Lin et al. applied a modified Maximal Marginal Relevance (MMR) algorithm to select sentences by incremental sentence re-ranking [108]. Following this idea, Boudin et al. introduced a scalable MMR algorithm [27].

2.7 Microblog Retrieval

As determined by the TREC Novelty track, the performance of novel sentence detection is very sensitive to the presence of non-relevant sentences, and isolating relevant sentences appeared to be a more difficult task [175]. Given a collection of microblog posts, the Information Retrieval (IR) literature offers many approaches to conduct retrieval based on textual queries. For example, Yan Li et al. used vector space models and query expansion techniques based on term semantic similarity and

the co-occurrence for the TREC 2011 Microblog Track [103]; Bandyopadhyay et al. explored the Google search API to expand query terms [15]; and systems developed in the work of [55, 122, 57, 44], adopted probabilities models, such as - Okapi BM25 and language models - as their baseline standard IR system for microblog retreival.

However, many researchers acknowledged that standard IR techniques, which mainly use term frequency, document length and inverse document frequency, and are unlikely to perform optimally, due to the short document length and vocabulary mismatch [15, 63]. To overcome these problems, several strategies were examined. Naveed et al. modified vector space model, which removed the document length normalization to avoid the term sparsity issue caused by the short length of tweets [138]. Metzler and Cai made use of the Markov random field model (full dependence variant) for text scoring, Latent Concept Expansion for pseudo-relevance feedback, and linear learning-to-rank to combine evidence of a variety of features (text score and twitter-specific features, such as whether a tweet contains a hashtag, out-of-vocabulary term percentage, language identification, etc.) [125]. Both learning-to-rank and pseudo-relevance feedback approaches were proven to be effective with the TREC 2011 Microblog Track data.

Another IR strategy employed by many researchers was to rank results by considering document quality. Similar to Web search, the document equal quality assumption does not hold true for microblog posts. Thus, distinguishing informative or authoritative content from less useful content is important when ranking retrieved microblog posts. Bendersky et al. showed that in Web searches, quality-biased ranking using content-based features can improve the retrieval performance [20].

Following this idea, Massoudi et al. incorporated quality indicators (e.g. emoticons, post length, shouting, capitalization, hyperlinks, reposts, followers, and recency) into a generative language model for searching microblog posts given a topic of interest [122]. Choi et al. also examined the role of document quality in microblog retrieval [44]. The research illustrated that a quality model learned from retweet behavior (a user quoting or forwarding other users content on Twitter) can improve the baseline retrieval model. Naveed et al. further learned a likelihood of retweet prediction function of tweets from a set of high and low-level content-based features of tweets [71]. The low-level features comprise the words contained in a tweet, (the tweet being a direct message), including the presence of URLs, hashtags, usernames, emoticons, and question and exclamation marks as well as terms with a strong positive or negative connotation. The high-level features are formed by associating tweets to topics and by determining the sentiments of a tweet. By using this interestingness as static quality measure, another work filtered and re-ranked tweets retrieved from an IR system during a one week period, as described in [3, 138].

In addition to content features, some researchers argue that the social status strongly correlates with the likelihood that a microblog post is interesting (as indicated by retweet ratio in Twitter) and thus, is more relevant to a given topic of interest. Nagmoti et al. not only used the relative length and the presence or a URL in the tweet, but also considered social network properties of the authors (e.g. the number of followers and followees) to rank microblog posts [136]. Hong et al. used machine learning techniques to predict the retweet possibility by combining features from both tweet content and the author's social graph [85]. Duan et

al. used a learning to rank approach to combine standard IR ranking (e.g. Okapi BM25, cosine similarity, and document length) with a number of content features (e.g. the presence of URL and URL count, retweet count, contained hashtag frequency, whether the tweet was a reply tweet, and ratio of out-of-vocabulary words) and authority features (e.g. number of followers, number of times a user is referred, number of joined lists, and popularity based on retweet relations) [55].

2.8 Summary

In this chapter, we reviewed major studies of search and topic tracking behaviors in microblogoshpere and microblog-focused news exploration. We also reviewed major approaches used in topic detection and tracking, novel information detection, extractive multi-document summarization, update summarization and microblog retrieval. This review helps us to highlight several gaps between what has been done and what needs to be done. In the remain thesis, additional related works will be reviewed in greater detail as the study progresses.

Chapter 3

Microblogging Relevance Filtering

This chapter addresses the problem of microblogging relevance filtering. It begins with a problem introduction, following by detailed methods. Then, an experiment is designed based on TREC Microblog real-time filtering evaluation. Lastly, this chapter concludes with results and error analysis.

3.1 Introduction

Information filtering has always been of great interest of people since the spread of information overload in the second half of the 20th century. Early attempts to devise an automated way to efficiently reduce unwanted information can be traced back to the early 1960s selective dissemination of information (SDI) systems in which the purpose was to help librarians route journal articles to readers according to their fixed profiles [29]. After the birth of the Internet, the demand for information filtering let to an increase in specialized filtering systems, such as news filtering, spam email filtering, and event filtering, to name just a few. This thesis focuses on microblog filtering.

Generally, a filtering system processes a stream of documents and for each document a binary decision is made regarding whether the document should be sent to the user [64]. There are two primary strategies to make such a decision. The

content-based filtering approach relies on a user query or profile to match relevant documents, like the SDI system. A modified version, adaptive filtering, that allows the system to refine or update original user profiles or to query incrementally with newly available knowledge, such as to personalized information, relevance feedback or pseudo relevance feedback (PRF) [4, 133]. The collaborative filtering approach (aka. social filtering) tries to filter documents based on similarities between tastes of different users [30, 171]. This thesis explores only on the content-based filtering approach, with consideration given to adaptive filtering.

Although content-based filtering has been studied for many years, microblogging data brings new challenges, most specifically, the enlarged language gap between a query and relevant documents. Because microblogs are mostly composed of informal, subjective, and innovative language, and are short in length, given a query, we could expect a data sparsity issue when using traditional methods, such as vector space model, to capture a similar microblog post. For example, if the query is about President Obama, then we might miss tweets using words "barackobama", "teamobama", "presobama", "thanksobama", "nobama", "pro-obama", etc. Although there have been linguistic theories that try to explain this language variability phenomenon [99, 190], and dedicated studies to understand certain word variation patterns [39, 72], it is hard to to design a generalized solution.

One popular solution to tackle this issue is query expansion. The basics behind query expansion is that by adding some related words to the original query, the chance of adding overlapping words with a relevant document is increased. Albakour et al. employed Incremental Rocchio to expand original queries from pseudo-relevant

tweets judged by a trained logistic regression classifier [2]. Lin et al. applied a uni-gram language modeling approach to model query from the most recent small set of on-topic tweets smoothed with a large background tweet corpus (comprising tweets published during a period of 1-month prior to the query) [106]. Although both showed the feasibility of this approach on Twitter data, and demonstrated some effectiveness, words are still compared by their surface forms, thus suffering from homonymy and synonymy problems. For example, "obama" and "presobama" remain completely different words.

Because this is not a unique problem in microblogs, many solutions have been developed. The first type of approach is to utilize various information sources for word sense disambiguation, such as manually crafted rules [92], dictionaries [102, 96], knowledge base [83], and a second language [34]. Voorhees was the first to use WordNet for the ambiguous nouns in information retrieval [184]. However, according to Schütze and Pedersen, all the above methods share two problems: (1) failure in choosing the correct word sense and (2) lack of coverage [167]. Many words in microblogs may not be covered by these generic lexical resources; and it may still be difficult to disambiguate senses thanks to the short context. Therefore, a coarse-grained but consistent disambiguation approach is proposed. Kim et al. used the root sense in the WordNet hierarchy to disambiguate nouns instead of accurate disambiguation [93]. Alternatively, a recent approach is the uses distributed representations, a so-called embedding of words. The idea is to map a word in the vocabulary to vectors of real numbers in a low dimensional space ("continuous space"). Similar words should be mapped to nearby points in the space. This mapping can

be learned with [51, 84, 24] or without [21, 128, 153] supervision. Because word embedding provides a way to capture higher-level dependency between words - either syntactic or semantic - when using it as the underlying document representation, it has been shown to boost the performance of information retrieval [45], collaborative filtering [94], and other natural language processing tasks, such as syntactic parsing and sentiment analysis [116].

Therefore, in this chapter, I exploit a mixture approach that combines these two solutions: query expansion and word embedding, for the task of microblog filtering. More specifically, the following three research questions are addressed:

- How does query expansion utilizing local and Web resources affect microblog filtering effectiveness?
- What is the effect of adding word embedding on microblog filtering?
- Does supervised learning from labeled training data result in more effective filters than unsupervised models?

3.2 Preliminaries

Let us use the following formalism to describe the microblog filtering problem. A query q_0 explicitly expresses user's search interest in text. The query may take the form of a few keywords, or a paragraph of short description. Additionally, q_0 needs to indicate the starting time t_0 of the filtering, which, by default, is the query issue time. Then, for a sequence of temporally ordered tweets $\mathcal{D}$ entered

into the system, a binary decision $y \in \{0, 1\}$ needs to be made regarding whether or not to return a tweet $d_i \in \mathcal{D}$ to a user who provided a query q_0. For both q_0 and d_i, a bag-of-words (BOW) representation is used as $\{w_1 \cdots w_n\}$. Since we are also interested in representing a word with word embedding, each word w is then represented by a vector of size e: $w \to \vec{E_w} = [E_w^1 \cdots E_w^e]$, thus transforming $\vec{q_0}$ and $\vec{d_i}$ into $\{w_1 \cdots w_n\} \to \{\vec{E_{w_1}} \cdots \vec{E_{w_n}}\}$. This is named the bag-of-embedded-words (BOEW) representation [45].

BOEW stems from BOW, which is a simple but commonly used representation of text disregarding grammar and word order but maintaining multiplicity. As an example, consider the following query and two tweets:

q_0: "Moscow airport bombing"

d_1: "At least 31 dead in Moscow airport explosion."

d_2: "Breaking News: 10 killed in Baghdad bombing"

A dictionary can be constructed (case-insensitive, ignoring numbers, and removing stop-words) as shown in Table 3.1, with each words's document frequency ($DF(w, \mathcal{D})$, the count of documents a corresponding word occurs), and inverse document frequency ($IDF(w, \mathcal{D}) = \log \frac{N}{DF(w, \mathcal{D})}$, where N is the number of documents in $\mathcal{D}$). In total, we find that there are 10 distinct words after removing stop-words and numbers. Using the indexes of these words in the dictionary, we can represent the query and tweets by a 10-dimension vector as:

$\vec{q_0}$: [1, 0, 0, 1, 0, 0, 0, 0, 1, 0]

Index	Word	DF	IDF
1	airport	2	0.1760
2	baghdad	1	0.4771
3	breaking	1	0.4771
4	bombing	1	0.4771
5	dead	1	0.4771
6	explosion	1	0.4771
7	killed	1	0.4771
8	least	1	0.4771
9	moscow	2	0.1760
10	news	1	0.4771

Table 3.1: Example bag-of-words model dictionary.

$\vec{d_1}$: [1, 0, 0, 0, 1, 1, 0, 1 ,1, 0]

$\vec{d_2}$: [0, 1, 1, 1, 0, 0, 1, 0, 0, 1]

where each entry of the vector is the occurrence count of the corresponding entry indexed in the dictionary. This representation is referred to as using term frequency (TF) as the BOW model term weighting scheme. An alternative common term weighting scheme is the TF-IDF, which is calculated for a term w using Equation 3.1. TF-IDF reduces term weights if a word appears frequently in the corpus, which suggests a word has a relatively general meaning.

$$TF \cdot IDF(w, d, \mathcal{D}) = TF(w, d) \times IDF(w, \mathcal{D}) \tag{3.1}$$

Thus, the above query and tweets can be represented as:

$\vec{q_0}$: [0.1760, 0, 0, 0.4771, 0, 0, 0, 0, 0.1760, 0]

$\vec{d_1}$: [0.1760, 0, 0, 0, 0.4771, 0.4771, 0, 0.4771 ,0.1760, 0]

$\vec{d_2}$: [0, 0.4771, 0.4771, 0.4771, 0, 0, 0.4771, 0, 0, 0.4771]

At this point, we can already apply the vector space model to measure the similarities between each tweet and the query by their cosine-distance, as defined by Equation 3.2. Now, suppose we know a 25-dimension embedding for each word in the dictionary $\hat{\mathbf{E}} \in \mathbb{R}^{10\times25}$, as shown in Figure 3.1, as a heat map reflecting the embedding values of each word. In Section 3.4, I introduce various ways to create word embeddings. Then we can have a BOEW representation (a 10×25 matrix) for the query and tweets, with each word $\vec{E_w} = TermWeight(w) \times \hat{\mathbf{E}}_w$. However, when calculating similarities between q_0 and d_i, we must also calculate similarities between each pair of word embeddings first, and then aggregate them on the query and document level. It obviously increases computational complexity, especially considering that in a real application, the dictionary may be composed of hundreds of thousands of words, or perhaps over a million when dealing with tweets.

$$cosine\_similarity(q_0, d_i) = cos(\theta) = \frac{\vec{q_0} \cdot \vec{d_i}}{\parallel \vec{q_0} \parallel \parallel \vec{d_i} \parallel} \tag{3.2}$$

Therefore, inspired by the success of various aggregation methods used in computer vision, we attempt to combine individual local word embeddings and transform

	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	e14	e15	e16	e17	e18	e19	e20	e21	e22	e23	e24	e25
airport	-2.09	0.00	0.85	-1.46	-0.84	-0.80	0.44	0.37	1.40	0.41	-0.10	0.78	-2.84	0.58	0.70	-0.98	0.20	0.37	-1.04	-0.28	-0.16	-1.45	0.43	-1.13	-0.25
baghdad	-0.49	0.46	0.31	0.57	-1.54	-0.66	-0.20	-1.04	1.33	1.46	0.17	-0.28	-1.06	0.37	0.36	-0.42	0.07	0.08	1.09	-0.17	0.36	0.14	0.50	-2.25	0.69
breaking	0.40	0.70	0.41	-0.05	-0.42	-0.22	1.36	0.16	0.01	-0.09	0.60	0.54	-3.28	0.12	0.85	-0.21	-0.31	0.51	-0.33	-0.71	-0.66	0.35	0.15	-1.20	-0.24
bombing	0.14	0.51	0.45	0.25	-0.46	-1.28	0.79	-1.23	0.12	1.06	0.90	0.63	-1.82	1.28	1.12	0.44	0.49	0.58	0.27	-0.73	0.55	0.07	-0.16	-2.72	0.33
dead	0.48	-0.07	0.81	0.19	-0.15	0.10	1.51	-0.06	-0.26	0.19	-0.23	0.29	-3.81	0.10	0.46	0.28	0.07	0.35	-0.82	-0.39	0.68	0.37	0.82	-0.87	0.83
explosion	-0.37	-0.25	0.36	0.11	0.06	-0.12	1.08	-0.90	0.84	0.62	0.86	0.25	-1.95	1.32	1.19	0.62	0.43	1.14	0.17	-0.32	0.29	-0.56	0.86	-1.77	1.03
killed	-0.11	0.77	1.02	0.27	-0.55	-0.78	0.91	-0.38	-0.47	0.95	0.69	1.36	-3.66	0.79	0.22	-0.07	0.65	0.41	0.33	-0.97	1.13	1.07	0.32	-2.14	0.74
least	0.49	0.27	0.28	-0.45	-0.64	0.22	1.39	-0.55	-0.60	0.38	-0.01	0.34	-4.33	0.74	0.38	0.10	0.88	-0.66	0.12	-0.94	0.11	0.44	-0.15	0.02	-0.87
moscow	-0.80	0.20	-0.13	-0.89	-1.65	-0.67	0.36	-0.40	0.94	-0.03	0.29	0.32	-2.06	0.54	0.41	-0.74	-0.06	0.06	0.11	0.15	-0.39	-0.79	-0.21	-1.85	0.85
news	0.98	0.18	-0.22	-1.09	0.08	-0.89	1.11	0.41	0.70	0.07	0.71	0.37	-3.73	0.31	0.26	-1.12	-0.49	0.18	-0.51	-0.18	-0.13	-1.14	-0.81	-1.68	-0.02

Figure 3.1: Example heat map of 25-dimension word embeddings. More **green** denotes larger value; more **yellow** denotes smaller value.

them into a global vector representation with a fixed-length. When processing images, the method must also handle local descriptors, where each is characterized by low-level properties such as color, texture or shape. In order to avoid intensive computation, various aggregation methods are proposed. A simple solution is to take the average of local word embeddings. Thus, by $\vec{q_0} \cdot \hat{\mathbf{E}}$ and $\vec{d_i} \cdot \hat{\mathbf{E}}$, we will have 25-dimension vector representation for both query and tweet in the semantic space, and can apply Equation 3.2 to calculate the similarity. Note that, other more sophisticated methods do exist for this purpose of measuring distance between two text document represented by BOEW. For example, Fisher kernel based aggregation has been proven to be more effective in text retrieval and clustering tasks [45, 208]. In a recent work, Kusner et al. proposed a Word Mover's Distance, which instead of using the aggregation approach, accumulates words' travel distance (cost) from one document to the point cloud of the other document [98]. To address the problem of tractability, word centroid distance and nearest neighbor search methods were investigated.

$\vec{q_0}$: [-0.44 0.28 0.34 -0.29 -0.65 -0.87 0.51 -0.59 0.47 0.57 0.46 0.49 -1.73

0.81 0.73 -0.10 0.26 0.35 -0.03 -0.37 0.16 -0.36 -0.04 -1.82 0.26]

$\vec{d_1}$: [-0.22 0.01 0.82 -0.48 -0.78 -0.17 2.04 -0.73 0.40 0.63 0.33 0.61 -5.68 1.23 1.16 0.17 0.68 0.47 -0.41 -0.81 0.42 -0.28 0.77 -1.78 0.58]

$\vec{d_2}$: [0.44 1.25 0.94 -0.02 -1.38 -1.83 1.89 -0.99 0.80 1.64 1.46 1.25 -6.46 1.37 1.34 -0.66 0.20 0.84 0.40 -1.32 0.60 0.23 -0.00 -4.77 0.71]

3.3 Query Expansion

As documented by Belkin and Croft, the essence of information filtering is closely related to information retrieval [18]. When working with microblog ad-hoc retrieval in the TREC Microblog track, studies have shown substantial, consistent and significant improvements in retrieval effectiveness from the use of query expansion. Therefore, the following three expansion techniques were tried for this filtering task.

- **Initial query expansion based on Web search.** Web search results have been reported to offer a useful basis for microblog retrieval query expansion [58, 59, 199]. Thus a Google Custom Search Engine (GSE)[1] is used to find related Web pages indexed before the query issue time, that can be used to expand the original query q_0. More specifically, Rocchio's relevance feedback algorithm is used for the expansion as defined in Equation 3.3 [162].

[1]https://www.google.com/cse

$$\vec{q_{exp}} = \vec{q_0} + a \cdot \frac{1}{|\mathcal{D}_r|} \cdot \sum_{d_i \in \mathcal{D}_r} \vec{d_i} - b \cdot \frac{1}{|\mathcal{D}_{nr}|} \cdot \sum_{d_i \in \mathcal{D}_{nr}} \vec{d_i} \qquad (3.3)$$
$$s.t.\ a \in [0,1], b = 0$$

where, $\mathcal{D}_r$ are set of relevant Web pages, $\mathcal{D}_{nr}$ are a set of irrelevant Web pages, and a, b are parameters. However, when conducting query expansion from Web search, there is no relevance judgment available, therefore PRF is used, which assumes the top k searched Web pages are relevant given the query q_0. Thus, we set $b = 0$, and constrain a to the space $\{a \in [0,1]\}$. Figure 3.2 uses an uninterpolated precision-recall plot to show the improvement that GSE query expansion (black dotted, tuned $a = 0.2, k = 30$) can achieve over a baseline (black solid), in which only original query terms are used. These results are the average over 10 training topics from the TREC 2012 Microblog real-time filtering task.

- **Incremental query expansion based on self-training.** Because the task of information filtering spans a period of time, the initial query can possibly be improved over time. This is the case with adaptive filtering [4]. Since this thesis treats the filtering decision for each microblog post as a binary decision, without any explicit relevance judgments, a self-trained approach can be applied to utilize past decisions. This approach has been shown to improve (average) filtering effectiveness in various tasks. An incremental query expansion technique achieved one of the best results in news filtering in the TREC Filtering track [161]. Following the work of Albakour et al., who applied

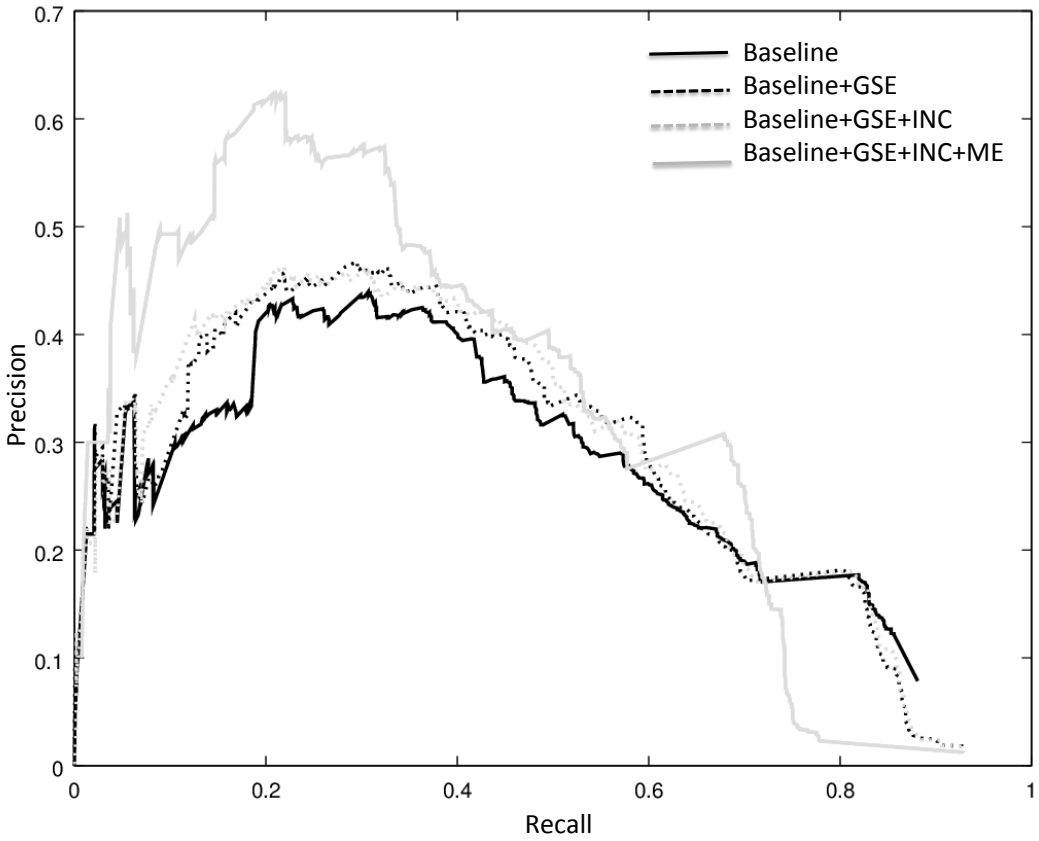


Figure 3.2: Uninterpolated precision recall trade-off of expansion methods on TREC 2012 Microblog Filtering training topics.

the same technique to microblog filtering, Equation 3.4 is used to update the query vector at the time t from microblog posts that the system considers relevant [2].

$$\vec{q_t} = \vec{q_{exp}} + c \cdot \frac{1}{|\mathcal{S}_t|} \cdot \sum_{d_i \in \mathcal{S}_t} \vec{d_i} + \sigma c \cdot \frac{1}{|\mathcal{L}_t|} \cdot \sum_{d_i \in \mathcal{L}_t} \vec{d_i} \tag{3.4}$$

where $\mathcal{S}_t$ refers to the most recent s relevant microblog posts judged by the system at time t; $c \in [0, 1]$ is a parameter similar to a in Equation 3.3 to control the contribution of terms from $\mathcal{S}_t$; $\mathcal{L}_t$ refers to all l older system judged

relevant microblog posts from a longer period of time since the filtering start time t_0 until the earliest microblog publication time in $\mathcal{S}_t$; and $\sigma \in [0, 1]$ is a decay factor that further discounts the contribution of terms from $\mathcal{L}_t$, in order to simulate the drifting interest of the user. Lin et al. also considered this topic drift issue by applying a history retention technique to only use the most recent tweets for query expansion [106]. Figure 3.2(grey dotted) shows the effectiveness of this technique (tuned $threshold = 0.2, s = 5, c = 0.1$, and $\sigma = 0.5$) on the training topics.

- **URL expansion.** For each microblog post that is under examination, we can follow any embedded Web links and parse the content text from the raw HTML.[2] If this extracted text is judged by the system to be relevant to the query (implemented as having a relatively high cosine similarity with a threshold of 0.3), then we can use it to expand the original microblog post. Here, we include the requirement that the page used for expansion must be relevant because in initial experiments it was found that in many cases links had been made to Web pages that contained non-relevant (or principally non-text) content. As shown in Figure 3.2(grey solid), this technique boosts the filtering effectiveness for microblog posts with relatively high cosine similarity with a query, while it begins to introduce more noise than useful information when this similarity decreases.

[2]We use Goose to extract the text; see https://github.com/GravityLabs/goose

3.4 Word Embedding

Two commonly used approaches to map a word to a semantic vector space include: (1) global matrix factorization methods and (2) local context window methods [153]. This first approach begins with the idea of using word co-occurrence statistics gathered directly from the corpus of interest to represent words [167]. This approach is furthered by applying dimension reduction on the word co-occurrence matrix in order to capture a low dimensional of "latent concepts" to represent words, which is still an implicit measurement of co-occurrence. Some well known methods that follow this approach are Latent Semantic Indexing (LSI) [51], probabilistic LSI (pLSI) [84], and Latent Dirichlet Allocation (LDA) [24]. The second approach is to use Neural Network to learn from local context window. A seminal work in this field is by Bengio et al., where a 3-layer Neural Network was trained with observed n-grams [21]. This work has later been refined by [46, 132, 129, 128]. Among these, Mikolov et al's skip-gram model proposes a simple single-layer architecture instead of the full neural network for efficiency, and is capable of phrase representation in addition to word representation.

However, according to Pennington et al., both of these two approaches suffer drawbacks [153]. The first approach performs poorly on the word analogy task, suggesting a sub-optimal word embedding. And the second approach poorly utilizes statistics of the corpus as it is trained on separate local context instead of the global co-occurrence matrix. Therefore, they propose a global log-bilinear regression model with a weighted least square method to train on global word co-occurrence statistics,

which they named GloVe (for Global Vectors) model. The cost function is defined in Equation 3.5.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(E_{w_i}^T \tilde{E_{w_j}} + b_i + \tilde{b_j} - \log X_{ij})^2 \tag{3.5}$$

where X_{ij} is the entry of the word co-occurrence and matrix $\mathbf{X}$ tabulates the number of times word j occurs in the context of word i. $f(x)$ is defined in Equation 3.6, with $x_{max} = 100$ and $\alpha = 0.75$.

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \tag{3.6}$$

E_{w_i} is the word embedding (vector) for word i. $\tilde{E_{w_j}}$ is a word vector for a context word j from a separate word embedding $\tilde{\mathbf{E}_\mathbf{w}}$. When $\mathbf{X}$ is symmetric, $\mathbf{E_w}$ and $\tilde{\mathbf{E}_\mathbf{w}}$ are equivalent and only differ as a result of random initialization. Otherwise, they are two word embeddings produced by the model and should perform equivalently. As the final word embeddings, Pennington et al. chose to sum $\mathbf{E_w} + \tilde{\mathbf{E}_\mathbf{w}}$, which boosts the word embedding performance of various tasks. b_i and $\tilde{b_j}$ are the bias introduced for E_{w_i} and $\tilde{E_{w_j}}$ respectively.

This model has been shown to produce word embeddings with a meaningful linear substructure by its state-of-the-art performance of word analogy task on a set of 19,544 questions like " a is to b as c is to _ ? ", with an accuracy of 75%. As illustrated in Figure 3.4,[3] the difference between various pairs of word embeddings for which the underlying concepts in similar ways are reflected by sim-

[3]http://www-nlp.stanford.edu/projects/glove/

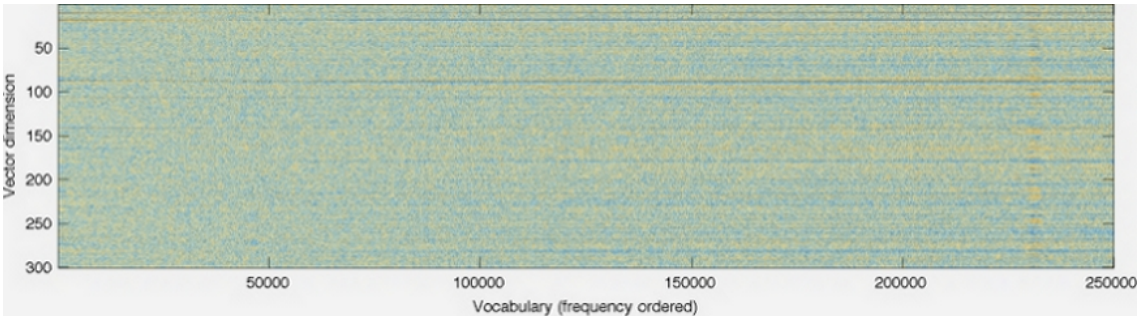


Figure 3.3: Example GloVe word embeddings. More **green** denotes larger value; more **yellow** denotes smaller value.

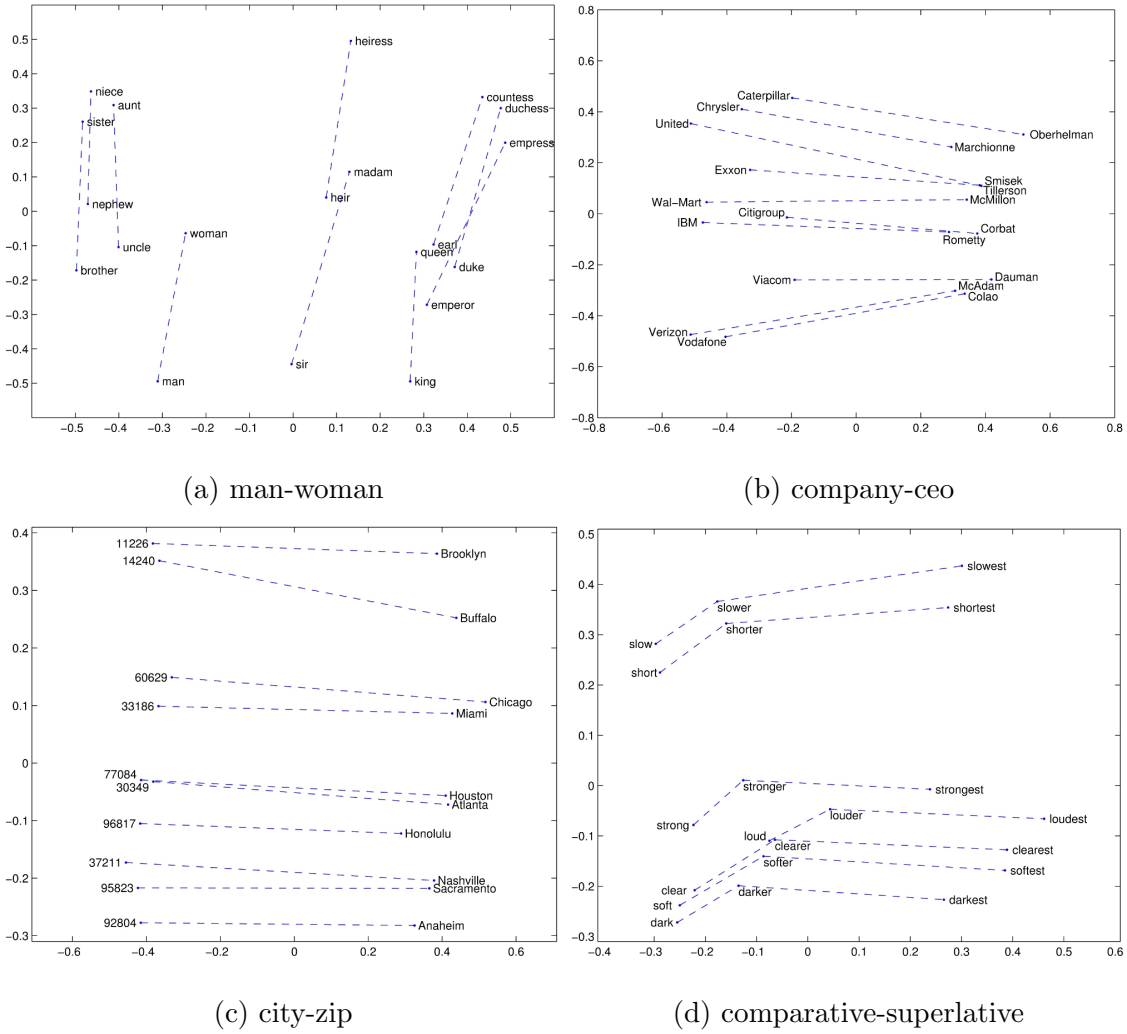
ilar differences in the word embeddings, i.e., man vs woman, is similar to king vs queen or brother vs sister. The model has demonstrated better results on several word similarity and named entity-recognition evaluations than other current competitive word embeddings, such as the Hellinger PCA[4] and word2vec.[5] As an example, by cosine similarity the top 20 most relevant words to "obama" are "barack", "romney", "president", "clinton", "biden", "potus"(President of the Untied States), "mitt"(Mitt Romney), "gop", "republicans", "hillary", "bush", "democrats", "debate", "republican", "obamas", "americans", "bill", "obamacare", and "says". Figure 3.3 illustrates the heat map for a sample of this produced word embeddings.

In this thesis, I am primarily interested in how much a word embedding representation can improve filtering effectiveness, I directly applied GloVe word embeddings trained on 27 billion tweets with dimension $e = \{25, 50, 100, 200\}$[6]. By averaging these word embeddings over all words in the query or tweets to produce the representations on which cosine similarity is then computed, Figure 3.5 shows

[4]http://lebret.ch/words/

[5]https://code.google.com/p/word2vec/

[6]http://www-nlp.stanford.edu/projects/glove/



(a) man-woman

(b) company-ceo

(c) city-zip

(d) comparative-superlative

Figure 3.4: Example linear substructures captured by the GloVe word embeddings.

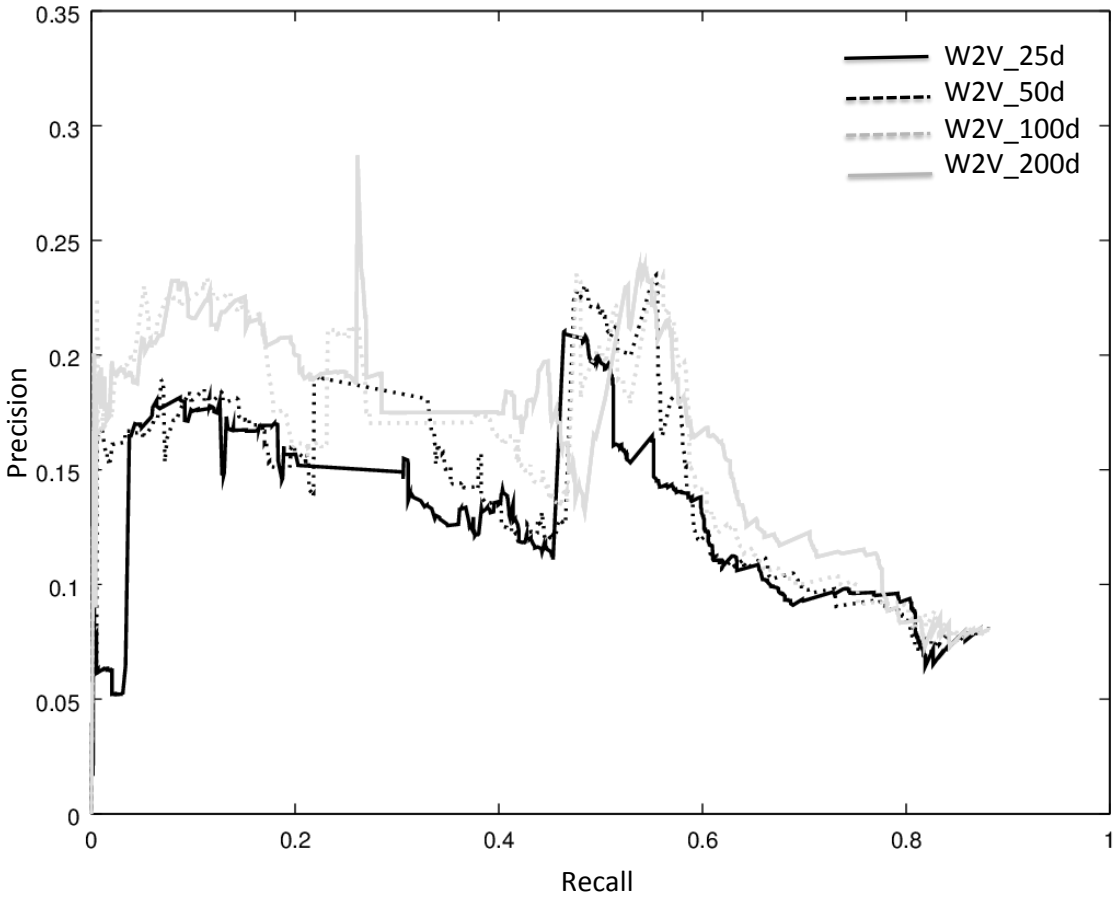


Figure 3.5: Uninterpolated precision recall trade-off of normalized BOEW on TREC 2012 Microblog Filtering training topics.

their precision-recall trade-off on the 10 training topics. As we can see from the results, better results can be achieved with higher dimensional word embeddings, which we would expect to capture more subtle words meanings. We can also see lower precision for low recall values than methods using the surface form of words because the BOEW model is designed as a recall-enhancing technique, which is not as discriminative as surface forms. We see higher precision for high recall values because surface forms of words do not work as well.

3.5 Joint Microblog Filtering

So far, given a query, I can filter a microblog stream according to the surface form of its words, and according to each words' semantic meaning. Each approach captures a complementary evidence from overlapping words and overlapping meanings. In this section, I create a filter that leverages both sources of evidence, together with additional sources of evidence. I first propose an unsupervised sigmoid combination, followed by a supervised logistic regression model.

3.5.1 Unsupervised Joint Microblog Filtering

The goal is to measure the similarity between a query and a microblog post in a latent space that could consider their similarity measured in the two observed spaces: BOW (word) space and BOEW (semantic) space.

$$Sim(q,d) = F(Sim_{BOW}(q,d), Sim_{BOEW}(q,d)) \tag{3.7}$$

The number of possibilities for F is vast. For example, similar to Rocchio algorithm in query expansion, a convex linear combination with parameters $\vec{\theta}$ can be used. However, instead of a fixed proportion, I argue that the weight given to the semantic space should decrease when the word space could give a stronger signal. By analogy, when manually judging a microblog post's relevance to a query, we only think deeply about its meaning when the words used are not obvious enough. With this insight, one class of functions that we could use is defined in Equation 3.8.

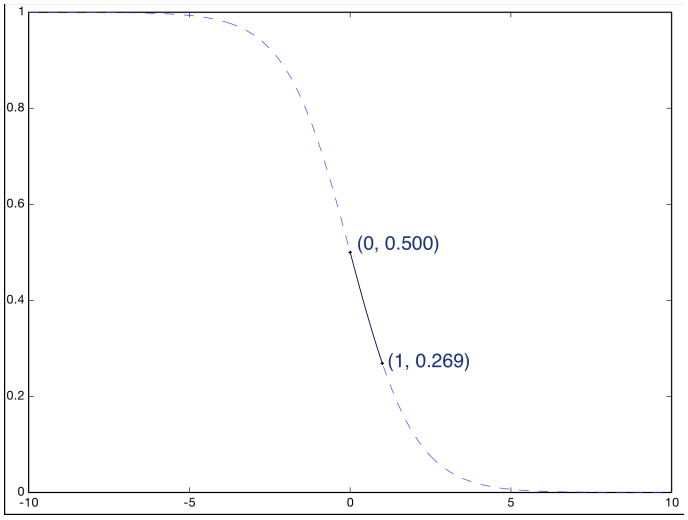


Figure 3.6: Sigmoid function $x \in [0, 1]$.

$$Sim(q, d) = Sim_{BOW}(q, d) + f(Sim_{BOW}(q, d))Sim_{BOEW}(q, d) \tag{3.8}$$

Of course a large number of functions f satisfy these properties, but the sigmoid function defined by Equation 3.9, and depicted in Figure 3.6 works well, where $x \in [0, 1]$.

$$\begin{gathered} f(x) = \frac{1}{1 + e^{(-(-x))}} = \frac{1}{1 + e^{x}} \\ s.t.\ x \in [0, 1] \end{gathered} \tag{3.9}$$

With this sigmoid combination, when $Sim_{BOW}(q, d) = 1$ (i.e., using exactly the same words in a microblog post as the query), a contribution from the semantic space is discounted by 0.269, and when $Sim_{BOW}(q, d) = 0$ (i.e., no overlapping words between a microblog and the query at all), the contribution from the semantic space is discounted by 0.5. Thus we never completely trust the semantic space, but rely on it more when the lexical space is demonstrably weak. The uninterpolated precision-

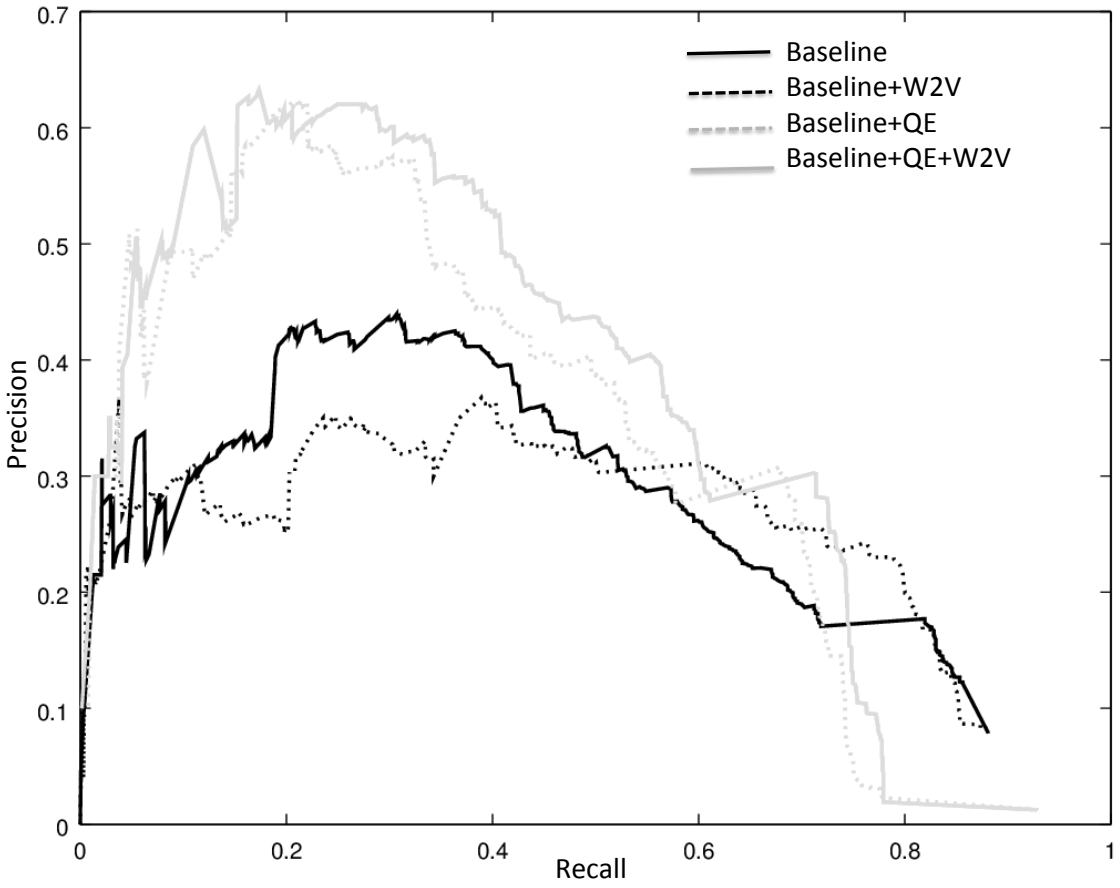


Figure 3.7: Uninterpolated precision recall trade-off of unsupervised combinations of filtering on TREC 2012 Microblog Filtering training topics.

recall trade-off resulting from the sigmoid combination is shown in Figure 3.7. We see improvement in precision over the baseline (black solid) for higher recall values (i.e., for the most difficult relevant microblogs to detect) from using word embedding (black dotted). As discussed in section 3.3, by using fully automatic expansion of the query representation (grey dotted), we are able to achieve improvement in precision for lower recall values (i.e., for the more easily detected relevant microblogs). When combining these methods, an improvement in precision across the full range of recall can be observed (grey solid).

3.5.2 Supervised Joint Microblog Filtering

As one of supervised machine learning approaches, Logistic Regression (LR) has been successfully applied to various information filtering tasks, which can be considered a binary classification problem given a document and a query (relevant or irrelevant) [2, 42, 210]. The formulation is that we could estimate the probability that a document is relevant to a query (an unobserved variable $y = 1$; if the document is irrelevant, then $y = 0$) given a vector of features (observed variables $\vec{x} \in \mathbb{R}^n$) using a logistic (sigmoid) function parameterized by $\vec{\theta}$, as defined in Equation 3.10

$$\begin{aligned} P(y = 1|\vec{x}; \vec{\theta}) &= g(\vec{\theta}^T \vec{x}) \\ g(z) &= \frac{1}{1 + e^{-z})} \end{aligned} \tag{3.10}$$

where the $\vec{x}$ could be coarsely-tuned but high dimensional features (i.e., words used in a document) or fine-tuned low dimensional features (i.e., the aforementioned cosine distance between word vectors or semantic vectors). In order to estimate parameters $\vec{\theta}$, from some training labeled instances (size of m), a gradient descent can be applied, with the L2-regularized cost function $J(\vec{\theta})$ and partial derivatives of parameters defined as 3.11

$$\begin{aligned} J(\vec{\theta}) &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log g(\vec{\theta}^T \vec{x}^{(i)}) + (1 - y^{(i)}) \log(1 - g(\vec{\theta}^T \vec{x}^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2 \\ \frac{\partial}{\partial \theta_0} J(\vec{\theta}) &= \frac{1}{m} \sum_{i=1}^{m} (g(\vec{\theta}^T \vec{x}^{(i)}) - y^{(i)}) x_0^{(i)} \\ \frac{\partial}{\partial \theta_j} J(\vec{\theta}) &= \frac{1}{m} \sum_{i=1}^{m} (g(\vec{\theta}^T \vec{x}^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j \end{aligned} \tag{3.11}$$

where θ_0 is the introduced bias with $x_0 = 1$, and $j \in n$.

In addition to $\vec{\theta}$, there are 3 hyper-parameters that need to be tuned: regularization parameter λ, number of gradient descent iteration, and higher-order polynomial degree to the features (where $degree = 1$ is the simple linear combination, and the above unsupervised sigmoid combination belongs to $degree = 2$). I apply a 10-fold cross validation on the 10 training topics, using each topic to find the optimized parameters and make an average for the final parameters. In Figure 3.9, each sub-figure shows the precision-recall trade-off validated on each topic while training on the other 9 topics with the optimized parameters (blue), compared with the results of the sigmoid combination (black). Figure 3.8 shows the decision boundary from the final parameters ($\lambda = 0.1, iter = 100, degree = 4$) on all training data.

One additional benefit of applying a supervised filterer is its incorporation of more relevance measurements in a straightforward way as new features. Given a query and a document, there could be multiple methods to measure their relevance. In addition to cosine distance, I also investigate Okapi-BM25 [160] and Kullback-Leibler divergence (KL divergence, the language modeling approach) [205, 204]. For the Okapi-BM25, the similarity function is defined by Equation 3.12, where the parameter average document length is set to $avgdl = 28$, and from training data, tuned $k = 0.1$ and $b = 0.2$.

$$BM25(q, d) = \sum_{w \in q} IDF(q_w, \mathcal{D}) \cdot \frac{TF(q_w, d) \cdot (k + 1)}{TF(q_w, d) + k \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \tag{3.12}$$

For the KL divergence using Dirichlet prior smoothing, the similarity is defined

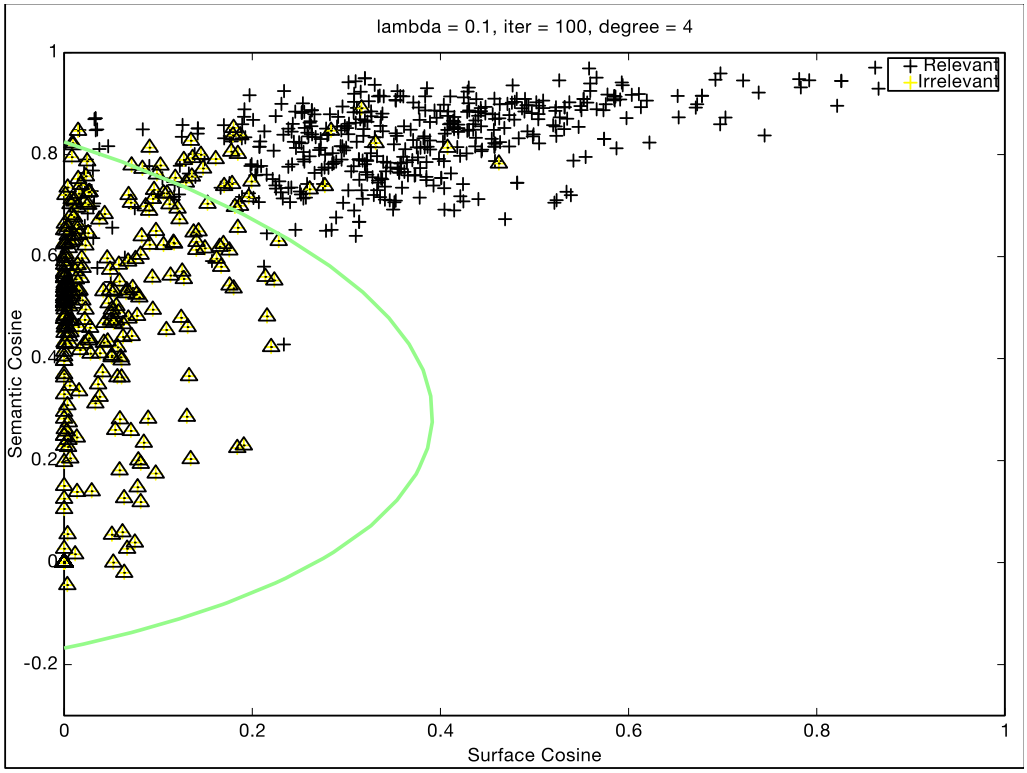


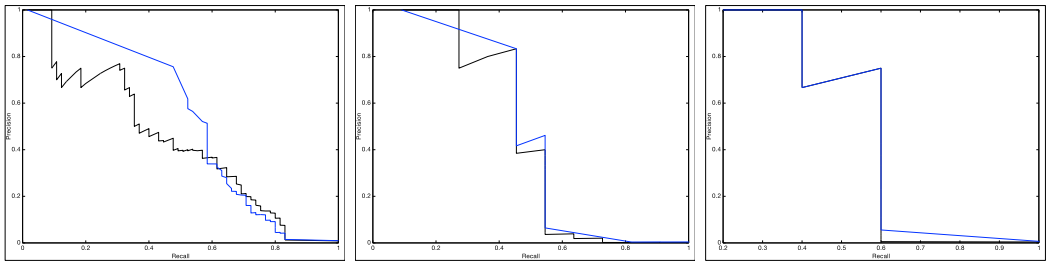
Figure 3.8: L2-regularized 4-degree polynomial logistic regression decision boundary on training tweets.

by Equation 3.13, where $p(q_w|\hat{q})$ is estimated from a query with the maximum likelihood, $p(q_w|\hat{\mathcal{D}})$ is estimated from a background corpus, and parameter μ is set to 1200.

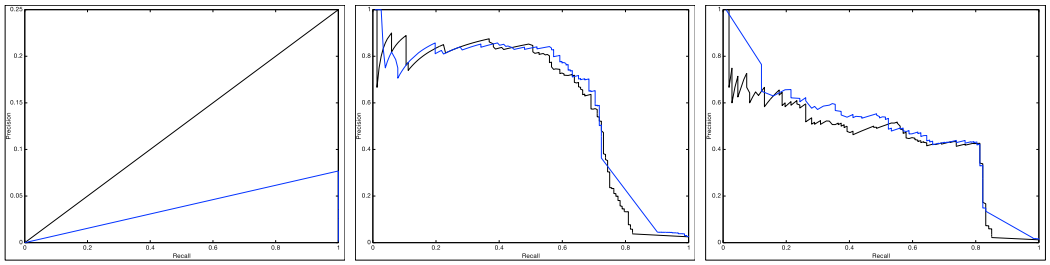
$$KL(q,d) = \sum_{w \in q} p(q_w|\hat{q}) \log(1 + \frac{TF(q_w,d)}{\mu \cdot p(q_w|\hat{\mathcal{D}})}) + \log \frac{\mu}{\mu + |d|} \tag{3.13}$$

3.6 Evaluation

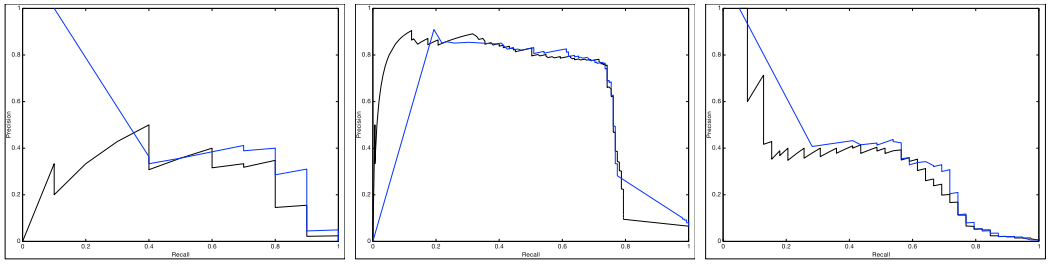
TREC 2012 Microblog filtering evaluation topics and relevance judgments are used to evaluate the effectiveness of the proposed microblog filtering system with



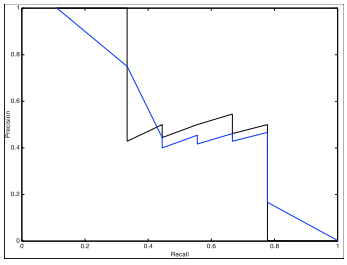
(a) topic 1 (b) topic 6 (c) topic 11



(d) topic 16 (e) topic 21 (f) topic 26



(g) topic 31 (h) topic 36 (i) topic 41



(j) topic 46

Figure 3.9: 10-fold cross validation precision recall trade-off on TREC 2012 Microblog Filtering training topics.

```
<top>
<num> Number: MB036 </num>
<title> Moscow airport bombing </title>
<querytime> Mon Jan 24 14:05:21 +0000 2011  </querytime>
<querytweettime> 29540259654012928 </querytweettime>
<querynewesttweet> 29674954899333120 </querynewesttweet>
</top>
```

Figure 3.10: Example query of topic MB036 in TREC 2012 Microblog Filtering track

the aforementioned methods.[7] In this section, I describe the evaluation setup and analyze the results.

3.6.1 Evaluation Setup

In total, there are 10 training topics and 36 test topics targeting a Twitter corpus containing 16 million tweets sampled over a period of two weeks (January 24th to February 8th, 2011). An example topic is given as follows, where the *query* field represents an user's information need with a few keywords at a specific time given by the *querytime* field. The *querytweettime* and *querynewesttweet* give the start and ending timestamp for the query in terms of the chronologically nearest tweet ID within the corpus.

On average, around 100 out of 1,000 manually judged tweets were manually marked by TREC assessors as relevant to a query topic. The evaluation metrics used in TREC are $F_{\beta=0.5}$ as defined by Equation 3.14, which is a precision-biased function of precision and recall when parameterized by $\beta = 0.5$; and $T11SU$ as

[7] https://sites.google.com/site/microblogtrack/2012-guidelines

defined by Equation 3.15, which is a utility measure for which a value of 1/3 can be achieved by removing everything (zero effort, that does not require the user to read anything) [176]. In this study, although I am more interested in $F_{\beta=1}$, with balanced combination of precision and recall, I also report these TREC measures to facility comparison to TREC submissions.

$$F_\beta = \frac{(1+\beta^2)Precision \cdot Recall}{\beta^2 Precision + Recall} \tag{3.14}$$

$$\begin{aligned} T11U &= 2 \times |\text{relevant retrieved}| - |\text{irrelevant retrieved}| \\ MaxU &= 2 \times |totalrelevant| \\ MinU &= -0.5 \\ NormU &= T11U/MaxU \\ T11SU &= \frac{max(NormU, MinU) - MinU}{1 - MinU} \end{aligned} \tag{3.15}$$

3.6.2 Result Analysis

Table 3.2 provides the evaluation results for each of the proposed methods with filtering threshold optimized on $F_{\beta=1}$ from training data. According to a two-tailed paired t-test, stars indicate statistically significant better $F_{\beta=1}$ than the previous row in the table.

There are a few observations to be drawn from the results. First of all, we can see on average a 2.89% improvement of $F_{\beta=1}$ in absolute from tweet expansion (ME) and 2.13% improvement in absolute from the BOEW (W2V) when using the sigmoid combination. This suggests the effectiveness of these two complementary

	Precision	Recall	$F_{\beta=1}$	$F_{\beta=0.5}$	T11SU
Baseline	0.2074	0.2738	0.2080	0.2019	0.1931
Baseline+GSE	0.1654	0.3902	0.2085	0.1780	0.1196
Baseline+GSE+INC	0.2123	0.2801	0.2115	0.2064	0.2007
Baseline+GSE+INC+ME	0.2469	0.3233	0.2404*	0.2364	0.2379
Baseline+GSE+INC+ME+W2V	0.2445	0.3950	0.2617*	0.2465	0.2276
LR(Baseline+GSE+INC+ME, W2V)	0.3168	0.3897	0.2905*	0.2987	0.3018
LR(Baseline+GSE+INC+ME, W2V, BM25, KL)	0.4161	0.3863	0.3071*	0.3334	0.3508

Table 3.2: Effectiveness of filtering methods on TREC 2012 Microblog Filtering evaluation.

expansion methods on the microblog filtering task. However, we did not see statistically significant improvement from the query expansion using Web search (GSE) and only moderate improvement from the incremental query expansion using self-training (INC). One one hand, it could be argued that because of the limited amount of training data (10 topics), it is hard to optimize the parameters (i.e., the filtering threshold). On the other hand, when looking at the expansion source documents (i.e., the searched Web pages and pseudo-relevant tweets), comparing with the ad-hoc microblog retrieval task's expansion, we see that a larger number of irrelevant documents were used. This may be because some topics start immediately after the real-world events happened, when there was not yet enough information available on the Web about the topic, and incremental query expansion only works when we have a relatively strong start. An example top 10 Google search results are listed

in Table 3.3 for the topic of "Moscow airport bombing". Note that all these results are indexed before January 23rd, 2011 (one day before the query start time, and the incident happened on Januray 24th, 2011 at 13:32 UTC), and none of them are relevant to the topic, although they are on related past events, such as bombing incidents happened in Moscow before. So, in this case, even though the GSE expansion helps in increasing the recall from 0.974 to 1.000, it decreases precision from 0.325 to 0.065, and thus decreases $F_{\beta=1}$ from 0.488 to 0.122.

From the evaluation results, we can also observe the effectiveness of supervised machine learning technology for the filtering task. We see a statistically significant improvement from using L2-regularized 4-degree polynomial logistic regression to combine surface cosine distance and the semantic cosine distance between a query and a tweet. In addition, we see statistically significant improvement when introducing more features in a straightforward way.

When selecting a threshold that could optimize $F_{\beta=0.5}$ on training data, $F_{\beta=0.5}$ and $T11SU$ measurements of the proposed methods are shown in Figure 3.11 (marked as $\triangle$), comparing with 60 submissions of TREC 2012 Microblog track (marked as $+$). The scatterplot is shown with a vertical line at the utility point of zero effort.

From this figure, we see that 3 out of the 7 proposed methods produce results that can exceed the zero efforts. The best method, which combines several BOW-based features and semantic features using a supervised approach, achieves second best $T11SU$ and fifth best $F_{\beta=0.5}$ for submissions that exceed the zero efforts. I reviewed these TREC works that outperformed my system [12, 76, 104, 206], and noticed that one dramatic difference between my system and theirs was the use of

Rank	Text
1	The 2010 Moscow Metro bombings were suicide bombings carried out by two ..... off from Domodedovo International Airport, previous Moscow metro bombings, ...
2	Mar 29, 2010 ... A commuter wounded in the bombing at the Park Kultury subway station in Moscow, shortly after the blast on Monday morning. More Photos ...
3	Mar 29, 2010 ... The attack has struck fear into Muscovites and refocused attention on the ... that took off from Moscow airport, bombed the Moscow metro twice, ...
4	About Category:Terrorist incidents in Moscow and related categories This category's scope includes pages on ... D. Domodedovo International Airport bombing ...
5	For 2007 bombing, see 2007 Nevsky Express bombing. ... the Russian cities of Moscow and Saint Petersburg causing derailment near the town of Bologoye, ...
6	The 1977 Moscow bombings were a series of three bombings in Moscow ... At the Tashkent Airport, a KGB officer noticed a woman carrying a bag similar to the ...
7	Airport Attacks in Rome and Vienna, December 27, 1985: Four gunmen .... Attack on U.S. Embassy in Moscow, September 13, 1995: A rocket-propelled grenade ...
8	Dec 24, 1991 ... A powerful bomb exploded today in the path of a bus carrying Soviet Jewish emigrants to the Budapest airport for their flight to Israel. ... continue to provide transit to the Jewish emigrantstraveling from Moscow to Tel Aviv.
9	Terrorist attacks and suicide bombings in Russia ... Domodedovo Airport ... The August 2004 Moscow metro bombing took place in the morning on August 31, ...
10	Aug 24, 2004 ... Both planes took off from the same Moscow airport within minutes of ... for numerous bombings and other attacks in Russia in recent years, ...

Table 3.3: Top 10 Google search results for topic MB036 "Moscow airport bombing" indexed before Jan. 23rd, 2011.

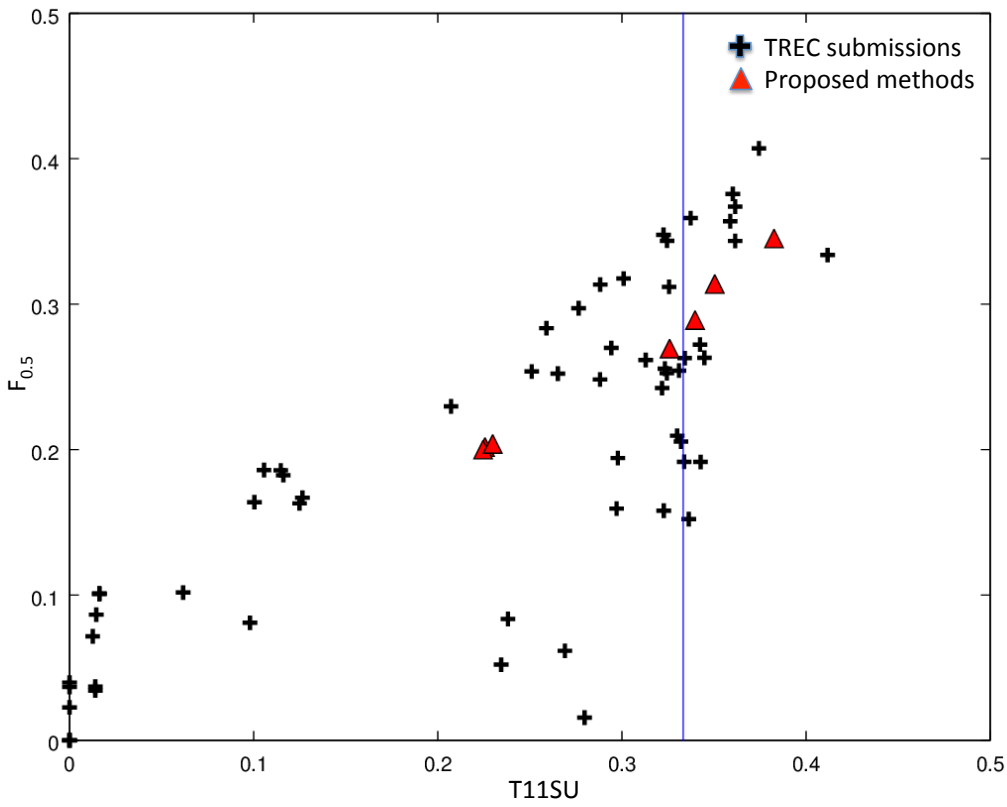


Figure 3.11: Effectiveness comparison with TREC 2012 Microblog Filtering submissions.

future information and additional manual annotation other than the training data. When designing this microblog filtering system, I was very careful to avoid using any of these two types of information. For example, I avoided calculating the IDF and background language model from targeting TREC tweet corpus. Instead, I used another around 1 billion English tweets to form a 15% sample of all tweets sent between 5/25/2009 and 10/17/2010 for the purpose [145]. In addition, because on average only 1,000 tweets were assessed by TREC, which were pooled from the participating runs's filtering results, there are tweets that are in fact relevant, retained by my system, however received no judgment from TREC, and are consid-

ered as irrelevant in this evaluation. For example, Figure 3.12 lists all the identified tweets that are evaluated as irrelevant to the topic of "Moscow airport bombing" according to TREC's relevance judgment. After reviewing these tweets, we can tell that they are actually all relevant to the topic. For this topic, the system got a $recall = 1.000, precision = 0.065$, and $F_{\beta=0.5} = 0.080$. However, as studied by Buckley and Voorhees, it is still reliable to compare the relative effectiveness of different systems with incomplete relative judgment because the document reduction into the judgment pool is systematically unbiased [35].

3.7 Conclusion

In this chapter, I investigated various expansion techniques and similarity measurements for the problem of microblog filtering. In addition to using the traditional BOW model to represent a query and a microblog, I also derived the BOEW model, which maps text from its surface form into semantic space. This leads to a useful degree of mitigation for the data sparsity issue. The power of this mapping was experimentally demonstrated using unsupervised and supervised machine learning techniques, combined with other BOW-based similarity measures. Finally, I demonstrated the state-of-the-art effectiveness of the proposed microblog filtering system by using TREC 2012 Microblog Filtering evaluation.

01/24/11 09:23:31 EST 29544831210496001 0 Two reported dead in Moscow airport blast - World news - Europe - msnbc.com http://t.co/eixMesL via @msnbc
01/24/11 09:40:22 EST 29549070322245632 0 Explosion shakes Moscow's busiest airport - Yahoo! News: http://yhoo.it/i3VlFe via @addthis
01/24/11 09:41:37 EST 29549388292431872 0 BBC News - Domodedovo blast: Explosion rocks Moscow's main airport http://www.bbc.co.uk/news/world-europe-12268662
01/24/11 09:42:59 EST 29549730866401280 0 cnn europe with telefone interview on air from Moscow Domodedovo (DME) airport after explosion
01/24/11 09:44:26 EST 29550097238855680 0 Watching the news and there's already been 6 homicides in T.O, a suicide bomber in Moscow airport and its currently -21 outside...
01/24/11 09:54:20 EST 29552588470882306 0 Suicide Bomber at Moscow's Domodedovo Airport Kills at Least 10: A suicide bomber detonated in the international... http://bit.ly/g49TWU
01/24/11 09:56:03 EST 29553017262964736 0 News Alert: Deadly Blast Strikes in Moscow's Main Airport http://nblo.gs/doB2A
01/24/11 10:00:13 EST 29554067818680321 0 Moscow airport explosion kills 31 #news
01/24/11 10:07:29 EST 29555897399246848 0 @delusionalGod Wow dude. Awful timing. There's just been a bombblast at Moscow airport.
01/24/11 10:08:26 EST 29556135979655168 0 BBC News - Domodedovo blast: Explosion rocks Moscow's main airport www.bbc.co.uk/news/world-europe-12268662 - more on #bbcnews channel
01/24/11 10:09:56 EST 29556512460382209 0 Bomb in Moscow airport...I hope everyone is okay. I'm not letting my mom leave.
01/24/11 10:20:09 EST 29559083182850048 0 Deadly Blast Strikes in Moscow's Main Airport: There were conflicting reports on the number of dead and injured... http://nyti.ms/eXnJrA
01/24/11 10:30:33 EST 29561700906704896 0 Hey people, the Moscow airport was bombed.
01/24/11 10:32:34 EST 29562207364714496 0 A woman jumped off a 23rd story window in Buenos Aires and survived. 31 people went to the airport in Moscow and died. C'est la mort.
01/24/11 10:34:24 EST 29562669941923840 0 Hmm blast rocks Moscow airport...guess I can't fly into Zurich tonight right? Raaa #iloveny
01/24/11 10:36:42 EST 29563248529379328 0 Pleased to hear that @hunternigel is at the other airport in Moscow.
01/24/11 10:45:16 EST 29565404418408448 0 BREAKING: dozens killed in Moscow airport blast (update) - http://tinyurl.com/4hopfp2 - solarp
01/24/11 10:48:49 EST 29566296760778752 0 I love reading about airport explosions before heading to the airport. Sure it was in Moscow but STILL. Not cool!!!
01/24/11 10:53:19 EST 29567432465055744 0 @NuuChiiiii explosion in an airport in Moscow cuz of a suicide-bomber --
01/24/11 10:56:36 EST 29568258826829824 0 i wonder if someone's gonna cancel my flight tomorrow since there was an explosion in the OTHER airport (still in Moscow) today.
01/24/11 11:04:09 EST 29570157340790785 0 At least 31 killed in Moscow airport explosion http://usat.me?139849
01/24/11 11:07:09 EST 29570912642662401 0 Omg! suicide bomb @ moscow
01/24/11 11:15:31 EST 29573016115478528 0 Bombing reported in Moscow DME airport; prayers to all the families.
01/24/11 11:32:20 EST 29577249497817088 0 (AP) - Airport spokeswoman increases death toll to 35 in Moscow airport explosion.
01/24/11 11:36:39 EST 29578335377952768 0 Bombing tragedy in Moscow at the airport: http://bbc.in/fYM4Ai
01/24/11 11:40:18 EST 29579254492565504 0 Tue, 25 Jan 2011 00:35:42 +0800 Blast kills 31 at Moscow airport http://t.rthk.hk/37wj4
01/24/11 11:43:54 EST 29580158918721537 0 Bombing in moscow? dont fuck with the #russians
01/24/11 11:47:32 EST 29581076179451905 0 Thank God for #Euronews which has a more international focus on the news than #bbc #sky #cnn - eg, Moscow bomb blast.
01/24/11 11:59:24 EST 29584060695183360 0 Pray for the dead and injured in Moscow Airport terror attack. Pics posted on Twitter taken inside the airport are horrific. #Domodedovo
01/24/11 12:11:01 EST 29586984024412161 0 Wh spokesman says Obama briefed on Moscow airport attack, offers condolences to victims, says US expresses "solidarity" w/ Russia
01/24/11 12:13:20 EST 29587566013456384 0 Explosion Reported in Moscow's Busiest Airport Prompts CHF Buying: Source: www.ibtimes.com --- Monday, January 2... http://bit.ly/h7o6XV
01/24/11 12:18:46 EST 29588934317056001 0 Explosion rips through Moscow's busiest airport, killing 31 http://bit.ly/fiTtG4
01/24/11 12:19:08 EST 29589027917144064 0 Moscow Airport Blasts: FSB False Flag, Part Of Anti-Putin Campaign Or Real Separatist Attack? http://bit.ly/haPFmD
01/24/11 12:26:13 EST 29590811775598593 0 At least 31 killed in Moscow bombing (AFP). Pray Allah for their souls...
01/24/11 12:28:36 EST 29591410520887296 0 #OMG blast @ Moscow Airport at #16:41GMT
01/24/11 12:35:28 EST 29593139371053056 0 Bomb blast rips thru moscow,thankfully i stayed in nairobi 2day
01/24/11 12:39:08 EST 29594061170020352 0 I slept with the radio on and thought the Moscow Bombing were just a dream. They aren't! #dream #nodream #nozakuboy #terrorism #news #bombs
01/24/11 12:55:39 EST 29598216701870080 0 Moscow airport blast kills 29 http://bit.ly/NBRfreeBlast
01/24/11 13:03:49 EST 29600271831146496 0 Ouch -> RT @edsbs: Well, it's going to be extra fun replaying the Moscow airport massacre section of Call of Duty 2 now, isn't it? #noitwont
01/24/11 13:07:03 EST 29601084968275968 0 @thepottydiaries not at all, was nursing sad thoughts of you at Moscow airport In Trouble so am glad it was fab!
01/24/11 13:20:52 EST 29604562616713216 0 is very sad to hear a friend was caught up in the attack at Moscow's Domodedovo Airport, they are now in hospital fighting for their life!
01/24/11 13:39:42 EST 29609302670970880 0 this so upsets me @ItsMashaBitch Damn, suicide bombing at the Moscow airport. This world has so much evil it's overwhelming.
01/24/11 13:43:20 EST 29610215150194688 0 I think so too...R.I.P. RT @H_Kovalainen: Shocked about the Moscow bombing...
01/24/11 14:40:22 EST 29624567966212096 0 Suicide bomb attack on Moscow airport.....
01/24/11 14:49:54 EST 29626970346102784 0 Moscow airport blast an acid test for the news media. Hats off to Twitter (users), news agencies and radio #Domodedovo #Russia
01/24/11 14:52:34 EST 29627638700048384 0 Live Updates: Moscow Airport Explosion - http://tinyurl.com/6794ynu
01/24/11 14:54:51 EST 29628213319704576 0 Moscow bombing horror http://bit.ly/gK4bq6
01/24/11 14:56:51 EST 29628716652957696 0 Togetter - 「Blast at Moscow's Domodedovo airport: モスクワ、ドモジェドヴォ国際空港で爆発」 http://bit.ly/dSMuL5
01/24/11 15:15:14 EST 29633345809809409 0 Michele Kearney's Snuffysmith's Blog: Russian authorities: Terrorist bombing at Moscow a... http://t.co/gDxCcSa
01/24/11 15:34:11 EST 29638112900546561 0 Blast Strikes Main Moscow Airport - NYTimes.com http://goo.gl/4yhHF
01/24/11 15:35:05 EST 29638339707543553 0 Stay tuned for continuing live coverage of Moscow #Domodedovo airport #bombing w/ analysts, expert discussions, breaking news as it develops
01/24/11 15:38:06 EST 29639098226442242 0 @ImTommyHill What's all this about Andy Gray blowing up Moscow Airport?
01/24/11 16:28:10 EST 29651699689857024 0 Twitter: Togetter - 「Blast at Moscow's Domodedovo airport: モスクワ、ドモジェドヴォ国際空港で爆発」 : http://bit.ly/eMe0Dx
01/24/11 17:07:45 EST 29661660247490560 0 The bomb in Moscow is of course tragic, but please, no more fucking waste-of-time vox pops describing a 'big bang'
01/24/11 17:10:31 EST 29662355864428545 0 Shocking news re: bomb blast in Moscow Airport - so sad

Figure 3.12: Example identified "irrelevant" tweets according to TREC 2012 Microblog Filtering relevance judgment for the topic MB036 "Moscow airport bombing".

Chapter 4

Microblogging Novelty Detection

As explained in Chapter 1, in order to save user efforts for consuming on-topic microblog posts following the microblog filtering, a novelty detection is needed to reduce microblog posts in the relevant microblog stream that report redundant information given past posts. Thus, in this chapter, I focus on the microblogging novelty detection problem. The chapter starts with an introduction to the problem, then a list of detailed novelty measurements, and ends with an evaluation design and result analysis.

4.1 Introduction

Novelty detection is one of the fundamental problems in signal processing, and has been considered a challenging task in several areas because in practice, it is hard to distinguish between unknown normal objects and novel objects [119]. In general, the task refers to the identification of novel or abnormal patterns from normal data, where Dubravko defined novelty as a pattern in the data that does not conform to the expected behavior, also called an anomaly, outlier or exception [130]. In information retrieval and filtering, because of a common demand for further distinction between documents containing new and relevant information and documents containing information that is relevant but already known, the study of this problem

has come a long way from its early inception by Carbonell and Goldstein's Maximal Marginal Relevance (MMR) [36], and later TREC's Novelty track [78, 174, 173]. Despite its maturity, there is still no foolproof solution because of the difficulty of accurately defining what does is meant by "new" [7, 115]. Therefore, in this chapter, I first explore an effective feature set to represent a microblog post's novelty. Then, I investigate to make a joint novelty decision by utilizing this feature set.

Motivated by studies in automatic summarization, it was quickly surmised that working in a batch fashion is beneficial to the task at hand. Usually, this is done by first grouping similar information together, and then documents or segments containing unique (novel) information can be identified. As demonstrated in the TREC 2014 Microblog track Tweet Timeline Generation (TTG) task, clustering based methods achieved one of the best results in selecting tweets reporting unique perspectives for a given query [107]. The advantage of this approach is the use of all possible relevant documents to assign normal documents in dense clusters, which helps to define what is normal. However, the decision regarding each document's novelty can only made after seeing all relevant documents. Since the ultimate goal of this thesis is to produce a real-time microblogging temporal summary given a query in good quality, inspired by the design of the TREC Temporal Summarization track [13], a delay-discounted measurement is proposed and applied to compare online and offline novelty detection approaches.

Therefore, as detailed in the introduction, I will examine the following three research questions:

- What are the most potential features for representing a microblog post's novelty?
- Is the ensemble learning approach helpful for the novelty detection effectiveness?
- Can a batch mode approach (i.e. clustering-based) be an effective novelty detection method with the consideration of delayed prediction?

4.2 Novelty Measures

Given a set of normal microblog posts, this section lists methods that can be used to measure a new microblog post's novelty. For development purposes, I created a set of novelty annotations from the training data of the TREC 2014 Microblog TTG task. This training data contains 10 queries and manually-created tweet clusters according to their content similarity judged by the TREC assessors. These queries were randomly chosen from TREC 2011 and 2012 Microblog ad-hoc retrieval tasks, and the clustered tweets were known relevant tweets for each query also judged by TREC assessors for the previous years' retrieval evaluation. Some descriptive statistics of these ground truth clusters for each query are shown in Table 4.1.

In order to derive novelty labels from these clusters, I designed the following steps:

Step 1: create an empty sample pool;

Step 2: collect known relevant tweets for each query and their ground truth

Query ID	#Relevant Tweets	#Cluster	Avg.Tweets per Cluster	%Redundancy in Relevance	%Unary Cluster
3	38	20	1.90	47.40%	65.00%
21	155	46	3.37	70.30%	69.50%
22	148	45	3.29	69.60%	84.40%
26	144	102	1.41	29.20%	85.30%
42	34	11	3.09	67.60%	54.50%
51	61	52	1.17	14.80%	92.30%
57	104	66	1.58	36.50%	74.20%
66	190	133	1.43	30.00%	80.50%
68	165	86	1.92	47.90%	73.30%
88	269	87	3.09	67.70%	74.60%
Ave.	130.8	64.8	2.225	48.10%	75.36%

Table 4.1: Statistics for TREC 2014 Microblog TTG training queries and ground truth clustering.

clustering information;

Step 3: given a query from each cluster, label the first published tweet as novel (a positive example) with a contextual normal tweet set of all relevant tweets published before it;

Step 4: given a query, for the rest of the relevant tweets not labeled in step three, label them as normal (negative examples), with a contextual normal tweet

set of all relevant tweets published before it;

Step 5: put all the positive and negative examples as well as their contextual normal tweet set and query information into the sample pool.

Note that for each known relevant tweet, it can actually be used to create more than one positive and negative sample by changing its contextual normal tweet set (adding or removing normal tweets). However, at the moment, let us focus on this current step to more closely simulate the real novelty detection environment for each tweet. Following this procedure, I got a nearly balanced sample set, which is composed of 648 positive and 660 negative examples. From this set, I randomly selected 100 positive and 100 negative examples for validation, and used the rest as training. For this test, I used another set of samples created from the TREC 2014 Microblog TTG evaluation queries (55 in total), which will be discussed in detail in Section 4.5. With regard to the evaluation metric, in this section, I use accuracy - the percentage of correct prediction for both novel and normal tweets. In Section 4.4, I discuss another evaluation metric which takes into consideration the delayed time in making a decision.

4.2.1 Nearest Neighbor Based

One of the most commonly used novelty detection approaches is nearest neighbor based. The idea is based on the assumption that novel objects should be far from normal objects [80]. If we control the novelty decision by a threshold, then the k-Nearest Neighbors (k-NN) algorithm can use the k nearest neighbors to make a

Distance Measure	k**-NN**	**Threshold**	**Accuracy**
Cosine Similarity	1	0.51	0.764
Bi-gram Dice Coefficient	1	0.54	0.736
Bi-gram Jaccard Coefficient	1	0.37	0.736
Tri-gram Jaccard Coefficient	1	0.24	0.759
Jaro-Winkler Distance	1	0.80	0.684
Elapsed Minutes	3	292	0.603

Table 4.2: Accuracy of distance-based novelty detection approaches on 200 validation tweets.

final decision by majority vote. Miljkovic summarized two variants of this approach: distance-based approach and density-based approach [130].

For a distance-based approach, I applied six types of distance measures as listed in Table 4.2. For each measure, I tuned the decision threshold and reported their optimized novelty detection accuracy on validation.

For a density-based approach, the intuition is that the density around novel objects is significantly different than the density around its neighbors. Local Outlier Factor (LOF) uses the relative density of an object compared to its neighbors to indicate the degree of an object being an outlier (novel) [32]. Specifically, let us denote $dist_k(o)$ as the distance between an object and its k-th nearest neighbor, denote $N_k(o)$ as the set of objects that is within distance of $dist_k(o)$,[1] and define

[1] $N_k(o)$ could be larger than k if exists objects with identical distance to o

reachability distance from o to another object o' as Equation 4.1:

$$Rdist_k(o' \leftarrow o) = max\{dist_k(o'), dist(o, o')\} \tag{4.1}$$

Let us define local reachability density of o, $lrd_k(o)$ as Equation 4.2:

$$lrd_k(o) = \frac{|N_k(o)|}{\sum_{o' \in N_k(o)} Rdist_k(o' \leftarrow o)} \tag{4.2}$$

Then, LOF of object o can be calculated as Equation 4.3.

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{|N_k(o)|} = \frac{\sum_{o' \in N_k(o)} lrd_k(o')}{|N_k(o)|} / lrd_k(o) \tag{4.3}$$

To understand this $LOF_k(o)$ score, a value below or equal to 1 indicates a denser or similar dense region of object o comparing with its neighbors, which suggests a normal object, while a value greater than 1 suggests an outlier (novelty). On validation set, k is tuned to 7, which can achieve a novelty detection accuracy of 0.684. Note that, when the contextual normal tweet size less than k, I use all the neighbors for my calculation; and when there is only one contextual normal tweet, I use the cosine similarity with a threshold of 0.51 (as seen in Table 4.2) to make a decision.

4.2.2 Information Theory Based

In addition to measuring the distance and density based on differences between a tweet and a normal set of tweets, we can also measure the difference in the amount of information content brought by the tweet under examination. The idea behind

the information theory based approach is that if introducing a tweet alters the information content of the normal tweets dramatically, then the new tweet can be detected as novel [41, 130]. Grounded in this idea, I apply two commonly used information theoretic measures, cross entropy and relative entropy (KL divergence), to novelty detection.

In information theory, entropy is one of the key measures of information, as defined in Equation 4.4. The entropy H of a discrete random variable X quantifies the amount of uncertainty in order to predict the value of X. When using X to represent the vocabulary, p to represent the vocabulary usage probability distribution on a set of normal tweets, and $p(x_i)$ to represent the probability of seeing a word $x_i \in X$ used in this set, then, choosing logarithmic base $b = 2$ expresses an expected number of bits needed to encode one of these normal tweets.

$$H(X) = -\sum_i p(x_i) \log_b p(x_i) \tag{4.4}$$

For a tweet from a vocabulary usage probability distribution of q, cross entropy $H(q, p)$ as defined in Equation 4.5, can be used to measure the expected number of bits needed to encode q with encoding schema optimized on p. If q and p are different distributions, which suggests the new tweet is novel, then a higher cross entropy can be expected than if they have the same distribution, which suggests the new tweet is one of the normal tweets.

$$H(q, p) = -\sum_i q(x_i) \log_{b=2} p(x_i) \tag{4.5}$$

For the experiment, I used perplexity $2^{H(q,p)}$ to evaluate an examined tweet, with q estimated from the tweet words, and p from the contextual normal tweets of maximum likelihood. I also applied Jelinek-Mercer smoothing [43], as seen in Equation 4.6, to estimate p with a uniform model, where the vocabulary X is to all words used in the normal set tweets and the examined tweet. When tuning $\lambda = 0.997$, this method can achieve a novelty detection accuracy of 0.782 on validation set with a $threshold = 315$.

$$p(x_i) = \lambda \frac{c(x_i)}{\sum_i c(x_i)} + (1 - \lambda) \frac{1}{|X|} \tag{4.6}$$

In addition to cross entropy, which selects the threshold, regardless of the difference between tweets under examination, I also try the relative entropy (KL divergence or information gain), as defined in Equation 4.7 to measure the expected additional bits needed to encode a tweet of q when using a encoding schema optimized by p compared to optimized by q directly. The smaller this number is, the more similar q is to p. Note that Equation 4.7 is different than Equation 3.11 defined in Chapter 3, which is an optimized version applying to the information retrieval/filtering task. When tuning $\lambda = 0.990$ and $threshold = 5.20$, this method can achieve an accuracy of 0.764 on the validation set.

$$KL(q \parallel p) = H(q,p) - H(q) = \sum_i q(x_i) \log_{b=2} \frac{q(x_i)}{p(x_i)} \tag{4.7}$$

4.2.3 Statistical Approach

Finally, I also exploit a statistical novelty detection approach. The basic assumption behind this approach is that the normal data is sampled from an underlying probability distribution, parameters of which can be estimated from the observed normal data (density estimation). Then, by hypothesizing that a test tweet is sampled from the same distribution, a probability can be inferred. With a preset critical value ϵ if this test probability is smaller than ϵ, we can reject the hypothesis, because a low probability situation has occurred, which suggests that the test tweet is novel. Otherwise, we can accept the hypothesis, and conclude that tweet is normal. In this thesis, I first tried to assume this underlying probability distribution was a n-dimension diagonal covariance Gaussian distribution, where n is the size of vocabulary X.

Specifically, suppose a normal set contains m tweets $\{\vec{x}^1, \cdots, \vec{x}^m\}$, with each $\vec{x}^i \in \mathbb{R}^n$. By assuming a Gaussian distribution on each word $x_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, $p(\vec{x})$ can be modeled as Equation 4.8.

$$p(\vec{x}) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sigma_j \sqrt{2\pi}} \exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}) \tag{4.8}$$

By using the maximum likelihood, the parameters $\mu_1, \cdots, \mu_n, \sigma_1^2, \cdots, \sigma_n^2$ can be estimated using Equation 4.9 from the normal tweet set. Similar to the information theory approach, the vocabulary is composed of all words used in the normal set and the test tweet. When tuning $\epsilon = 0.62$, this method can achieve an accuracy of 0.782 on the validation set.

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$
$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2 \tag{4.9}$$

4.3 An Ensemble Learning Approach for Novelty Detection

Now that we have multiple ways to discern a tweet's novelty and each can be considered as a threshold-based binary classifier, in this section, I investigate combinations of these single classifiers. The following notations are used: let us denote novelty detection task as predicting $y \in \{0, 1\}$; given various feature sets $\vec{x}_l$ computed for each tweet against a set of normal tweets, we can derive a model set $\vec{\mathcal{L}}$, including methods proposed in Section 4.2, where each method can make the novelty detection decision when setting $y_l = \mathcal{L}_l(\vec{x}_l)$. In this section, the research interest is in a joint model that can predict $y = F(y_l) = F(\mathcal{L}_l(\vec{x}_l))$.

In machine learning, it is well known that the ensemble learning approach, which combines multiple classifiers, can provide more effective solutions to a task than single classifiers. Although there is a risk of over-fitting, it is worth to explore the using of all the powers of each classifier, and provide a more reliable and sophisticated novelty detection solution. Generally, there are four popular types of ensemble approach: voting, bagging, boosting and stacking.

Voting is a simple but strong baseline approach. Let

$$C_{j=\{0,1\}} = Count(l; \mathcal{L}_i(\vec{x}_l) == j) \tag{4.10}$$

Classifier	**Baseline**	**Bagging**	**AdaBoost**
L2-Regularized Logistic Regression	0.7989	0.8103	0.7989
SVM using Gaussian Kernel	0.7701	0.7816	0.7759
C4.5 Decision Tree	0.7874	0.8103	0.7989
Naive Bayes (NB)	0.7816	0.7874	0.7816

Table 4.3: Accuracy of ensemble learning novelty detection approaches on 200 validation tweets.

Then, a majority vote method predicts $y = \arg\max_j C_j$. On the validation set, this method can achieve an accuracy of 0.7989.

Bagging (Bootstrap Aggregating) tries to learn the combination of single classifiers from training data [31]. In order to promote model variance, so that to overcome the over-fitting issue, this method firstly draws randomly on a subset from the training P-percentage of the total training, then trains a model f with this subset. By repeating this process multiple times with I-iterations, a final predication model can be aggregated (e.g. majority vote) from each trained models f_i. In Table 4.3, I tried several types of base models as the method of combination, and listed their predication accuracy (baseline) compared to the results applying bagging method on the validation data. Note that for methods producing the same accuracy, their predictions are totally independent.

Boosting is a different approach than Bagging in that it learns f_i sequentially. The idea is that new model f_i is trained by focusing on examples that early models

$f_{\{1,\cdots,i-1\}}$ get wrong, and thus can enlarge the margin and enhance the prediction accuracy of a classifier. AdaBoost is the most popular boosting method with an algorithm described by algorithm 1 [67]. Its accuracy when applied to various combination models is shown in Table 4.3. However, as was noticed, this method is vulnerable to random classification noise, because the algorithm attempts to correctly classify these mis-labels poorly, and thus fails to produce a model with good prediction accuracy [111]. In the case of our novelty detection training data, unfortunately, the mis-classified examples exist because of the ambiguous nature of the manual clustering, which explains the under-performance of boosting over bagging.

Algorithm 1: AdaBoost [67].

$\vec{w} = \frac{1}{m}$, where $m = |\vec{w}| = |\vec{Y}|$

for $i = 1 : I\{$BOOSTING ITERATION$\}$ **do**

- $f_i = \text{TRAIN}(\mathbf{L(X)}, \vec{Y}; \vec{w})$
- $\vec{\tilde{Y}} = \text{PREDICT}(f_i, \mathbf{L(X)})$
- $\epsilon_i = \vec{w} \cdot (\vec{Y} != \vec{\tilde{Y}})'$
- $\alpha_i = \frac{1}{2} \log \frac{1-\epsilon_i}{\epsilon_i}$
- $\vec{w}.* = \alpha_i . * \vec{Y} . * \vec{\tilde{Y}}$
- $\vec{w} = \frac{\vec{w}}{\sum \vec{w}}$

return MAJORITY VOTE(f_i)

Stacking is another supervised approach for ensemble learning. The idea is to train a meta-classifier to learn how to combine different types of first-layer combination models [197]. As a typical setup, I apply logistic regression as the meta-classifier to combine the aforementioned 4 types of first-layer combination models:

L2-Regularized Logistic Regression, SVM, C4.5 Decision Tree, and Naive Bayes (NB). On validation data, this approach achieves a novelty detection accuracy of 0.7931.

4.4 Clustering-based Novelty Detection

As mentioned at the beginning of this thesis, one of the first difficulties in determining microblog novelty detection is the unclear definition of what should be considered as new and normal. In the sections above, I tried to avoid this issue by finding solutions from different methods, and hoping that these various perspectives would narrow down the answer. This is an approach called "the wisdom of crowds", which according to Surowiecki, can result in a better final decision than each aggregated individual decision maker would make alone [177]. In this section, I investigate a different approach, clustering-based novelty detection, which utilizes future information to confirm a microblog post's novelty.

This approach is widely adopted in the TREC 2014 TTG track, where a clustering method is applied in order to group tweets into topical clusters, so that tweets within each cluster represent an unique topical perspective (novel information) of a query. This is also the foundation of the procedure I designed to create the novelty detection label data from the manually-created topical clusters. The benefit of this approach is that future tweets can help to better define a normal set, and thus help to detect novel tweets. However, the drawback of the approach is that the novelty decision is delayed until seeing the future tweets. Therefore, this section is dedicated

to study this approach and the effect of the decision latency.

4.4.1 Globally Fixed Threshold Hierarchical Clustering

Hierarchical clustering with a globally fixed threshold (GFT) is the most commonly used approach in the TREC 2014 TTG track. Among the 13 participate teams, 7 adopted this approach. The idea is to apply a hierarchical clustering and to input "relevant" tweets with a globally fixed threshold tuned to optimize the clustering effectiveness of the training data. A classical hierarchical clustering algorithm is described in algorithm 2, which produced the top 1 TTG effectiveness in the TREC 2014 TTG track [114]. In my implementation, I follow this same algorithm, calculating the distance between clusters by the complete-linkage of $distance = 1-$ cosine similarity.

Algorithm 2: Hierarchical clustering.

$\mathcal{R} \leftarrow$ Relevant tweets

$\beta \leftarrow$ Threshold

$\mathcal{C} \leftarrow \{[R_1], [R_2], \cdots, [R_n]\}$

repeat

  $(C_i, C_j, \text{MinDistance}) \leftarrow \text{GetMinDistance}(\mathcal{C})$

  $\text{MergeCluster}(C_i, C_j)$

until $\text{MinDistance} < \beta$;

return $\mathcal{C}$

In order to compare with previous novelty detection methods' effectiveness, I first tuned the threshold to optimize the accuracy on the same 200 validation tweets, which gave me 0.46 for the threshold, and 0.708 for the accuracy. However,

clustering method works at query level and makes novelty decisions for all relevant tweets of each query together instead of making independent decision for each tweet. Therefore, a more appropriate training and validation split should be performed at the query level. I adopt a leave-one-query-out cross-validation approach. Suppose there are Q queries, leave-one-query-out cross-validation treats each one query as validation for Q iteration, and the rest $Q-1$ queries as training. The final effectiveness performance is measured by taking the average of the validation effectiveness over iterations. In my experimental case, I have $Q = 10$ queries. If we stick to using accuracy as the effectiveness measurement metric, a mean validation accuracy of 0.692 can be achieved with each optimized GFT threshold and validation accuracy listed in Table 4.4, where I also list the number of novel and normal tweets per query for reference.

4.4.2 Query Optimal Threshold Hierarchical Clustering

As pointed out, one noticeable issue of the globally fixed threshold hierarchical clustering method is that the threshold is chosen by ignoring specifics of a particular query and the relevant tweet set it triggers. This is problematic because the cohesion of relevant tweet sets would vary dramatically from query to query, and thus would require different clustering thresholds in order to distinguish them into sub-topical groups. As shown in Table 4.4, if we manually tune this threshold to optimize each query's accuracy, then we can see how it mostly different from the globally fixed threshold as tuned from the training queries. According to two-tailed

Query ID	#Novel	#Normal	GFT Threshold	GFT Accuracy	QOT Threshold	QOT Accuracy	Optimal Threshold	Optimal Accuracy
3	20	18	0.54	0.816	0.48	0.816	0.56	0.816
21	46	109	0.51	0.697	0.60	0.742	0.60	0.742
22	45	103	0.51	0.716	0.59	0.764	0.63	0.777
26	102	42	0.54	0.681	0.47	0.701	0.36	0.715
42	11	23	0.51	0.706	0.69	0.824	0.66	0.853
51	52	9	0.54	0.475	0.61	0.475	0.61	0.475
57	66	38	0.54	0.760	0.47	0.750	0.51	0.770
66	133	57	0.54	0.753	0.46	0.768	0.50	0.784
68	86	79	0.54	0.588	0.42	0.703	0.37	0.764
88	87	182	0.54	0.725	0.62	0.751	0.62	0.751

Table 4.4: Accuracy of clustering-based novelty detection approaches on validation queries.

paired t-test, the P value equals 0.022, which is a statistically significant gap of the accuracy between the validation results. Therefore, a query specific clustering threshold sounds reasonable if we consider the fact that novelty will be considered differently given query topics. For example, for topics about which people tend to tweet similar content, a more elaborate distinguishing effort (a lower threshold according to the proposed clustering method) is a better choice in order to identify tweets bringing new and subtle information.

Given a query and its set of relevant tweets, in order to automatically estimate its unique threshold, I propose using a linear regression model with the following features: (1) the globally fixed threshold from the training topics; (2) the check-

ing query's specific within query pairwise tweet distance; (3) the mean of averaged within query pairwise tweet distance calculated from the training queries; (4) the difference between (2) and (3); (5) the checking query's specific averaged tweet distance from the center of the query, where the center is represented by all relevant tweet of the query except the checking tweet; (6) the mean of the averaged tweet distance from the query center as calculated by (5) across the training queries; (7) the difference between (5) and (6); and (8) the ratio of (4) divided by (7). For all distances mentioned above, they are measured by using (1− cosine similarity) between two vector of uni-grams. The resulting validation query optimal threshold and accuracy for each query is listed in Table 4.4, which shows a statistically significant improvement over the GOT method on this small validation set, with a two tailed P-value of 0.028 according to paired t-test.

4.4.3 Delay-Discounted Accuracy

One key drawback of the clustering-based novelty detection is its delay in making decisions. Therefore, in order to fairly compare its effectiveness with the real-time novelty detection methods as discussed in Section 4.2 and Section 4.3, the latency must be penalized.

For this purpose, following work in the TREC Temporal Summarization, a latency discount is introduced into the effectiveness measurement [13, 53]. Given a tweet published at time t_p, and the system's novelty decision time t_d, the latency penalty function $L(t_p, t_d)$ can be described as Equation 4.11, which is a mono-

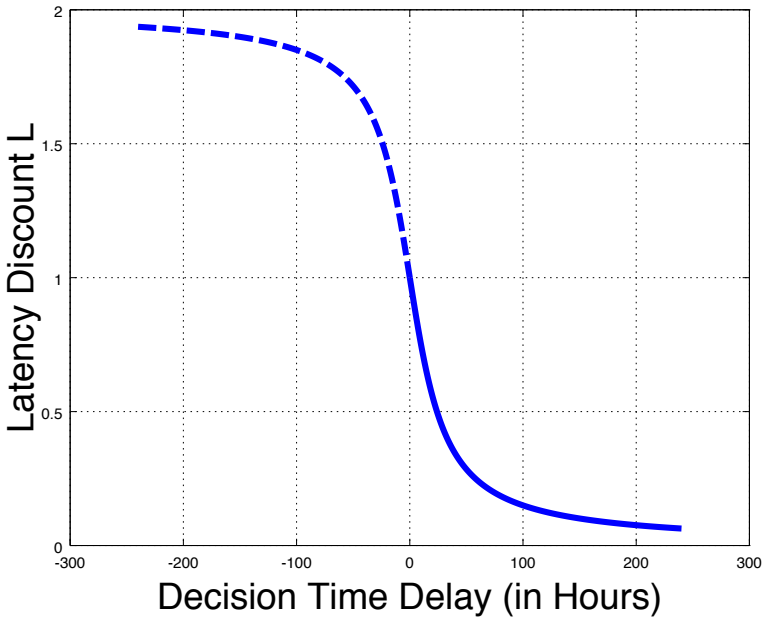


Figure 4.1: Latency discount function $x \in [0, +\infty)$.

tonically decreasing function of $t_d - t_p$. Because the system decision time cannot proceed the tweet's publishing time, the maximum value of the function is 1, which is achieved by making a real-time decision (no latency), and thus having no discount on the effectiveness. On the other side, the function is approaching 0 if the decision time is too long after the tweet is published, and when the value flattens is controlled by a parameter α, which if set as $3600 * 24$, means the latency-step is 24 hours (1 day), and the value is almost 0 after approximately 13 days. The function can be depicted as shown in Figure 4.1.

$$L(t_p, t_d) = 1 - \frac{2}{\pi} \arctan(\frac{t_d - t_p}{\alpha}) \qquad (4.11)$$
$$\alpha = 3600 * 24$$

By applying this delay discount, an Expected Latency Accuracy (ELA) of the microblogging novelty detection can be defined as Equation 4.12, where $t \in TP$

(true positive) means a true novel tweet is correctly predicted by the system as positive in the novelty detection, and $t \in TN$ (true negative) means a true normal tweet is correctly predicted by the system as negative. Therefore, for a system that can make a real-time decision, because the $L(t_p, t_d)$ will always be equal to 1, $ELA == Acurracy$; for a system that makes delayed decisions, each decision is penalized by the delayed time $t_d - t_p$ according to $L(t_p, t_d)$.

$$\text{Expected Latency Accuracy} = \frac{\sum_{t \in \text{TP or TN}} L(t_p, t_d) \times 1}{\#\text{of relevant tweets}} \tag{4.12}$$

The delay penalized ELA of the clustering-based novelty detection methods for the 10 validation queries, as well as their original accuracy are listed in Table 4.5 as an update from the scores in Table 4.4. According to the results, we can see a serious penalty was given to the clustering-based methods. This is because on average there are 72.4 hours of delay in decision time, with each validation query's decision delay time listed in Table 4.5. Of course, the choice of α, the latency-step, is arbitrarily set as 24 hours, which with a different value, could adjust the ELA to a certain level. However, the tuning of α is out of the scope of this thesis, and I assume 24 hours is a reasonable value considering the application scenario, where a news update can be captured in a reasonable time period after the occurrence.

4.5 Evaluation

In order to fully evaluate the effectiveness of the microblogging novelty detection methods proposed in this chapter, TREC 2014 Tweet Timeline Generation

Query ID	GFT Accuracy	GFT EL Accuracy	QOT Accuracy	QOT EL Accuracy	Optimal Accuracy	Optimal EL Accuracy	Avg Decision Delay (Hours)
3	0.816	0.155	0.816	0.142	0.816	0.179	125.8
21	0.697	0.201	0.742	0.207	0.742	0.207	50.6
22	0.716	0.410	0.764	0.438	0.777	0.443	15.6
26	0.681	0.233	0.701	0.244	0.715	0.243	63.9
42	0.706	0.095	0.824	0.132	0.853	0.160	152.0
51	0.475	0.182	0.475	0.182	0.475	0.182	59.2
57	0.760	0.466	0.750	0.462	0.770	0.477	12.4
66	0.753	0.340	0.768	0.343	0.784	0.355	30.2
68	0.588	0.156	0.703	0.203	0.764	0.227	38.6
88	0.725	0.087	0.751	0.087	0.751	0.087	174.9

Table 4.5: ELA of clustering-based novelty detection approaches on validation queries.

(TTG) track queries and manual clusterings are used. In this section, I describe the evaluation setup and analyze the results to answer the research questions raised at the beginning of this chapter.

4.5.1 Evaluation Setup

In total, there are 55 queries in the TREC 2014 TTG track, with each one labeled with on average of 190 relevant tweets by the TREC assessors. This is out of 243 million tweets of a Tweets2013 corpus crawled from the public Twitter sample stream between February 1st and March 31st, 2013 [107]. The query topics are also selected to cover news events overlapping within the same period. Some example

topics include "Ron Weasley birthday", "merging of U.S. Air and American", and "election of Hugo Chavez successor". For all the relevant tweets in a given query, a manual topical clustering is conducted by assessors from the University of Maryland and the University of Illinois. This procedure is strictly controlled and the effects of assessors' difference are reported in the TREC paper [107]. On average, 89 clusters are created per query.

In order to utilize this data set to evaluate the effectiveness of the proposed microblogging novelty detection methods, 2 experiments were set up:

For experiment I, I generated tweet novelty labels from the clusters as directed by the procedure described at the beginning of Section 4.2, and then compared each methods' novelty detection accuracy. This experiment used known relevant tweets as input. The results are listed in Section 4.5.2 with analysis.

However, because novelty detection is designed as a consecutive step after microblog relevance filtering (as shown in Figure 1.3), a stress test is necessary in order to understand the effect of noisy non-relevant tweets introduced from this former step. Taking this into consideration, a second experiment was set up using a simulated relevance filtering input stream instead of ground truth relevant tweets for the test. Section 4.5.3 describes details of this simulation and shows the results as well as analysis.

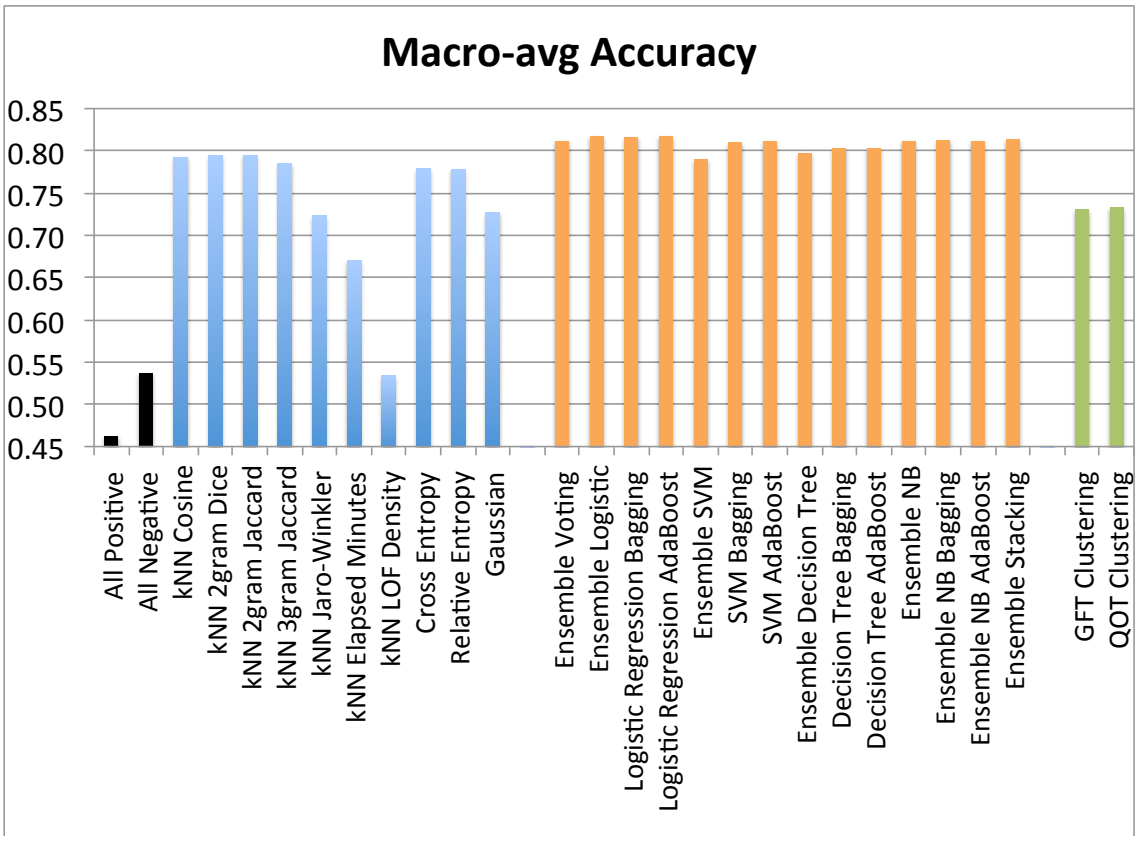


Figure 4.2: Accuracy of novelty detection approaches on TREC 2014 Microblog TTG evaluation with known relevant tweets as input.

4.5.2 Experiment I Result Analysis

For experiment I, Figure 4.2 summarizes the macro-averaged accuracy of each proposed microblogging novelty detection method with all-positive and all-negative prediction as baseline (black bar). If applicable, each method uses the hyper-parameter tuned by the validation data as described above.

Note in Figure 4.3, the y-axis starts from accuracy of 0.45, and the results can be divided into 3 groups by empty bar: (1) single novelty predictors; (2) ensemble learning based predictors; and (3) clustering-based predictors. From the results, we

could have the following observations addressing the 3 research questions:

(1) Nearest Neighbor based methods using both Dice Coefficient and bi-gram Jaccard Coefficient as the distance measurement with considering the $k = 1$ neighbors can achieve the highest macro-averaged novelty detection accuracy across queries. Although the information theory based method using relative entropy to measure information gain achieves slightly better micro-averaged novelty detection accuracy, according to two-tailed paired t-test, the difference is not statistically significant. Thus, we could conclude that the two Nearest Neighbor-based methods are the most effective feature for capturing a tweet's novelty. Information theory based methods can also be effective in terms of measuring a tweet's novelty. All the proposed novelty measures in Section 4.2 can statistically significantly outperform simple all-positive and all-negative predictors with no latency in prediction.

(2) In general, ensemble learning is more effective than single predictor. According to the results from the 4 ensemble approaches with 4 types of bass classifiers (14 combination in total), there are at most 2.3% in absolute improvement of both the macro-averaged and micro-averaged accuracy from using the logistic regression as the base classifier and AdaBoost as the ensemble approach. By conducting a paird t-test to compare this best ensembled predictor and the best single predictor (bi-gram Jaccard Coefficient kNN), a two-tailed P-value equals to 0.0035, which is very statistically significant improvement. Among the 14 ensemble learning combinations, only when using SVM to simply combine single predictors generates a slightly lower accuracy than the best single predictor, all the rest produce higher accuracy. When comparing ensemble learning approaches, voting and stacking are

more robust and are easier to implement than bagging and boosting. As discussed in Section 4.3, although Adaboost achieves the highest accuracy in the evaluation, however, it is more vulnerable with mis-labelled noisy examples in the training data.

(3) However, clustering based methods failed to be more effective in microblogging novelty detection task. According to the results, even without the latency discount, neither the GFT nor the QOT based hierarchy clustering can produce higher accuracy even than single real-time novelty measures. Although, the accuracy is better than some single novelty measures, when considering decision time latency, the GFT clustering method got an $ELA = 0.169$ and the QOT got an $ELA = 0.170$. Therefore, the methods cannot show their values in making in-time novelty decision. An additional observation is that QOT outperforms GFT by all evaluation metrics and shows a moderate improvement. When reviewing the methods, there are still limitations in my design and implementation. For example, I only use $(1 - \textit{cosine similarity})$ to measure distance; there is very limited number of training queries to learn the QOT decision model; and there are better clustering algorithms available. However, I will leave it for future work to further examine improvements to these methods.

4.5.3 Experiment II Result Analysis

For experiment II, in order to simulate a microblogging relevance filtering input stream, a retrieval step is first conducted for corpus Tweets2013 through a corpus API (the "evaluation as a service" model) [105]. Because Twitter's terms

of service,[2] which prohibit redistribution of tweets, this is the only way to access the corpus. Then, from the top 1000 returned tweets, I ordered them according to publication time and input them one by one into a relevance filtering process as developed in Chapter 3. A binary relevance prediction was made for each tweet, and only the predicted relevant tweets can enter the novelty detection process. On average, each topic gets approximately 174 input tweets for novelty detection with precision = 0.520, recall = 0.401, $F_{\beta=1} = 0.371$, and $T11SU = 0.393$.

Figure 4.3 shows each novelty detection's effectiveness measured by $F_{\beta=1}$ and $T11SU_4$. For each method, 2 bars are shown in the figure with the first bar for $F_{\beta=1}$, and the second bar for $T11SU_4$. The reason to report $F_{\beta=1}$ is because the input stream now contains unbalanced positive/negative data with higher a proportion of irrelevant tweets, accuracy is no longer an appropriate evaluation metric under this circumstance. For utility test T11SU, as defined in Equation 3.15, it is because this metric can help to tell the usefulness of the novelty detection output by comparing with a zero effort value of 1/3 that can be achieved by returning nothing to the user to read. In the figure, I use lighter grey color to denote methods that cannot exceed zero effort value, and darker grey color to mark methods that can exceed this value, and thus produce useful novelty detection results for user. Note that, one modification made from the original $T11SU$ score here is the doubled credits awarded for a successfully detected novel tweet. In relevance filtering, it is only awarded 2 credits for a identified relevant tweet, and penalizes 1 credit for a identified irrelevant tweet. Since novelty detection requires additional effort upon relevance

[2]https://twitter.com/tos?lang=en

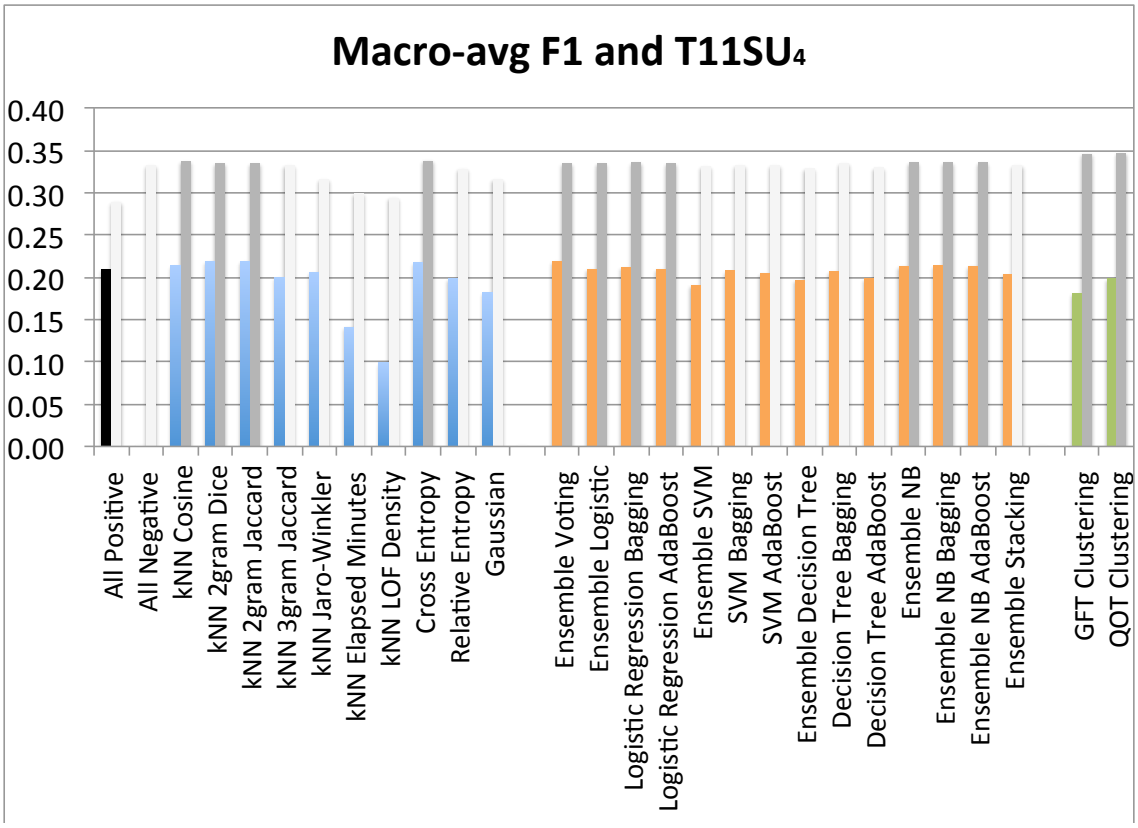


Figure 4.3: Effectiveness of novelty detection approaches on TREC 2014 Microblog TTG evaluation with predicted relevant tweets as input.

filtering, it is reasonable to credit more if one novel and relevant tweet is successfully identified, which I reward it with 4 credits.

As can be observed from the results, half of the proposed methods can produce useful predictions (> 0.333). 4 methods can outperform simple all-positive prediction statistically significantly by around 1% in absolute $F_{\beta=1}$. Although ensemble techniques do not show a decent improvement over single predictors as in experiment I, most of the top effectiveness are still achieved by ensemble techniques. Clustering-based approaches still does not competitive. However, QOT method again shows promising improvement over GFT, and achieves the highest utility score.

4.6 Conclusions

In this chapter, I investigated various novelty measurements, ensemble learning techniques, and clustering based microblogging novelty detection approaches. In addition to globally optimizing hierarchical clustering thresholds from the training data, I also proposed a query-optimized-threshold model using global features as well as local query features. I set up two experiments by using TREC 2014 TTG evaluation with relevance labeled tweets and noisy relevance filtering results as inputs of novelty detection, and compared the effectiveness of each proposed method under these two input conditions. Throughout this chapter, I demonstrated an effective novelty detection process.

Chapter 5

Salience Detection

Ignoring trivial updates will further improve the entire system's function. This Chapter focuses on the microblogging salience detection problem. After an introduction to the problem, this chapter begins with a discussion of the meaning of topical salience, and describes an application in journalism and the procedure of data labeling. Then, following investigation of methods for identifying a microblog post's topical salience, the chapter reports evaluation effectiveness of the data created and analyzes the results. As a comparison, a tweet "Dutch government is cutting off subsidies for renewables" is more likely to be detected as salient, as well as "British government provided financial support for two Fatah security forces linked to torture", which are actually filtering results from the current system.

5.1 Introduction

After novelty detection, in order to further filter microblog posts to provide readers with a more succinct summary that covers the major topic aspects, a topical salience detection is designed in the temporal summarization system pipeline as depicted in Figure 1.3. This salience detection attempts to extract microblog posts that can provide important information about the topic in addition to determining whether the post is relevant and novel. Extracting important content from text

document(s) is also one of the major approaches adopted by most automatic summarization works. However, similar to novelty detection, one serious question must be addressed is to determine what is salient in the source being summarized. According to Hahn and Mani, the produced summaries, and the perspective of salience, differ depending on the function (indicative, informative, or critical) and target reader (generic or domain-focused) [74].

Existing works in the microblogsphere tried to address this issue in various ways. For example, O'Connor, et al. clustered tweets into sub-topic groups, and considered salient sub-topic to be those with good coverage of frequent terms/phrases, diversity from other sub-topics, and containing large size tweets [144]. Therefore, a representative tweet (the central one) in a sub-topic group should be extracted in a final summary. TREC Microblog Tweet Timeline Generation task followed this idea and devised a weighted version of the evaluation that favors topical clusters with more tweets, especially those with more relevant tweets given a query topic.[1] Chakrabarti and Punera introduced a more sophisticated underlying sequential event/topic structure to address long-running structure-rich events (e.g. they use sports game tweets as study objects). They trained a Hidden Markov Model (HMM) to capture this structure according to tweet bursitiness and change of term distribution over time, and then selected the key tweet from each "sub-event" [141]. Similar work was conducted by Nichols, et al, with an extra focus on the summary sentence generation method that applied Sharifi's phrase-graph algorithm [172] and evaluation of the output summary against human-generated summary [141].

[1] https://github.com/lintool/twitter-tools/wiki/TREC-2014-Track-Guidelines

Due to the streaming setup of the temporal summarization, it introduces new challenge of identifying the focal without seeing the complete information. In addition, a distinct perspective is to consider a tweet's social influence as criteria for salience. For example, the Web Information System Engineering (WISE) 2012 Challenge [2] organized a microblog propagation prediction task, which tries to identify tweets given a query topic, that will be highly re-tweeted or viewed. Some corresponding methods tend to focus more on measuring a microblog post's quality, e.g. length, whether or not it contains Web links, an existing re-tweet rate, etc. [28, 182, 113].

According to the application scenario designed in this thesis, a domain specific consideration for salience is focused on the professional journalist's perspective. That is, given a query's topic, which microblog post is so important that a journalist would include it in tomorrow's news article. However, this does not attempt to give a definition for "newsworthiness", which remains an open question in journalism [143, 166, 65, 19, 77, 66, 148]. Instead, this work attempts to provide journalists with information resources from which they could derive new insights from microblog posts. A data-driven approach is therefore adopted to establish such a prediction model from objective real-world data. In particularly, the following research questions are addressed in this chapter:

- How effective is using a microblog post's quality measurements in salience pre-diction?

[2] http://www.wise2012.cs.ucy.ac.cy/challenge.html

- Can features extracted from past relevant news reports be helpful in deciding a microblog post's salience?

5.2 Data Collection

Because a data driven approach is applied and this thesis selects journalists as the target readers of the resulting temporal summary, it is necessary to first ensure the existence of labeling data from the real world about which microblog posts journalists considered to be important given a particular topic. Unfortunately, such data does not exists nor is it easy to collect. Therefore, this section describes how the data is created step by step.

The basic idea is that given a topic, future news articles can be utilized to inform which content journalists would consider newsworthy. Thus, if we could identify which microblog post possesses the content presented in a future news article, then we can assume it indicates that the post is important. It appears to be a straightforward solution, however there are three prerequisites that must be fulfilled in order to validate this assumption. The first two requirements are that the post under examination must be relevant and novel. The relevance requirement is because human labor is expensive and should be economized by filtering out irrelevant posts, which surely will not be salient. The novelty requirement is because a post's content is important only the first time it appears. The second time is redundant. A third requirement is that a novelty judgment must be created according to previous relevant posts and a past news article from the same news source as the future news

article (e.g. written by the same journalist or published by the same news agency). There are three options for making a novelty judgment:

- A judgment is made according to only the previous relevant microblog posts with no past news articles;
- A judgment is made according to the previous relevant microblog posts and any past news articles;
- A judgment is made according to the previous relevant microblog posts and the past news article from the same news source as the future article used for the salience judgment.

Firstly, the obvious relevant microblog posts are necessary because they are the basic definition of microblogging novelty detection. Secondly, almost every news article provides background information, which is duplicate information for journalists but is useful for an article's audience. Thus, a past article can help to eliminate valueless content considering journalists are the target readers of summary. As shown in Figure 5.1, the information valuable to journalists is shown in a future article. And if a post can help bring such information before the future article is published (here we ignore the news agency's publication procedure and schedule), we can assume it is valuable. Lastly, because different news agencies publish articles in their own story line, choosing past and future news articles from the same news source can guarantee the consistency in the reports and ensure the difference captures additional valuable content in between.

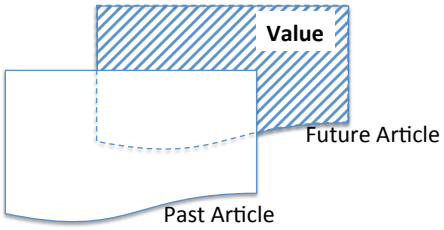


Figure 5.1: News value in future news article.

Based on these three prerequisites, the following steps are designed to create salience label data, which is also detailed in Figure 5.2.

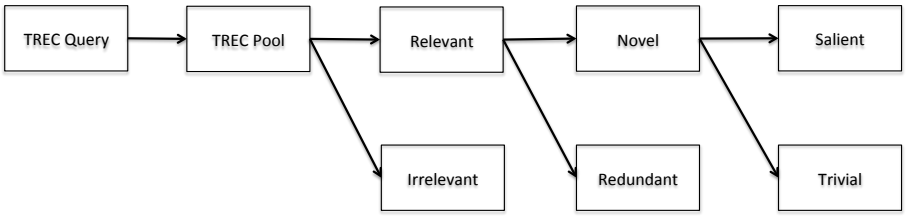


Figure 5.2: Local assessment from TREC Microblog track relevance assessment.

- TREC 2011 and 2012 Microblog track queries and relevance judgment are utilized as the input source. Because these queries correspond to a particular news topic, this step fulfills the relevance requirement.

- The relevant tweets given a query are sorted according to their publication time, and embedded external Web links are extracted from each tweet if available.

- If we could find two Web links from the same news source (the same base URL) and two different Web pages, then we can assume these two Web pages are two news reports (past and future) from the the same news source.[3] Figure 5.3

[3]If multiple Web pages are identified from the same source, the first and last one are used.

lists the 43 queries that meet this requirement and the time-stamp for both the past (initial report) and future news article (follow-up report), as well as the news source. I also show the heat map over the days of the number of tweets published in between.

- The novelty of each selected tweet is then judged by human assessors according to whether the content appears in the initial news article and any previously relevant tweets. Section 5.2.1 shows details about this step.

- From the novel tweet judged by the human assessors, salience is then judged by human assessors according to whether the content appears in the follow-up news article. Section 5.2.2 details this step.

5.2.1 Novelty Assessment

Given a topic, an initial news article, and a set of known relevant tweets sorted according to their publication time, this assessment aims to label each tweet with a binary label to indicate whether the tweet brings new information against the news article as well as all relevant tweets published before it. In total, 9 assessors were recruited. 7 of them are graduated students either from information science or computer science programs, and the remaining 2 are undergraduate students. One of the assessors is from Syracuse University, and the remaining 8 are from the University of Maryland, College Park.

For each topic from the 43 selected topics, 2 assessors are randomly assigned, which gives each assessor approximately 8-10 topics to assess with around 110 tweets

Query ID	1/23/11	1/24/11	1/25/11	1/26/11	1/27/11	1/28/11	1/29/11	1/30/11	1/31/11	2/1/11	2/2/11	2/3/11	2/4/11	2/5/11	2/6/11	2/7/11	2/8/11	News Organization	Initial Report	Follow-up Report	Elpased Hours	Relevant Tweets In-between
1		4	11	31	5													bbc.co.uk	1/24/11 13:13	1/27/11 5:42	64	48
3	3	2	1	0	0	1	0	1	3	10	2	2	0	2				haitilibre.com	1/23/11 20:45	2/5/11 16:00	307	24
7					8	15												bbc.co.uk	1/27/11 13:36	1/28/11 8:48	19	14
8	18	8	5	10	23	4	1	2	4	2	7							bbc.co.uk	1/23/11 11:38	2/2/11 18:43	247	82
9				64	8	2	1	0	3	0	1	1	1	0	0	3	51	bbc.co.uk	1/26/11 6:17	2/8/11 20:58	326	125
19										18	8	3	3	1	2	10		cnn.com	2/1/11 17:08	2/7/11 14:29	141	32
21		47	48	13	11	29	1											nytimes.com	1/24/11 19:10	1/29/11 3:03	103	124
22									78									reuters.com	1/31/11 6:05	1/31/11 23:31	17	73
23					14	2	2	1	0	0	10	7	1	1	2	9		cnn.com	1/27/11 10:18	2/7/11 22:46	276	46
24															22	42	17	sports.yahoo.com	2/6/11 22:59	2/8/11 15:29	40	64
26					20	6	2	2	4	2	11	28	25					cnn.com	1/27/11 11:09	2/4/11 14:06	194	94
29			8	3	8	6	3	4	3	6	8	10	7	6	12	16		nytimes.com	1/25/11 1:58	2/7/11 18:56	328	94
30																	80	nytimes.com	2/8/11 2:09	2/8/11 15:43	13	34
32	2	0	1	31	14	5												huffingtonpost.com	1/23/11 0:53	1/28/11 20:47	139	53
33		3	4															cnn.com	1/24/11 16:13	1/25/11 14:06	21	5
34			2	2	5													huffingtonpost.com	1/25/11 17:23	1/27/11 15:09	45	7
36		148																bbc.co.uk	1/24/11 14:29	1/24/11 21:10	6	128
37	21	10	1	31	11	1												cnn.com	1/23/11 17:49	1/28/11 1:51	104	60
42							18	3	1	1	2	0	1	0	0	3		bbc.co.uk	1/29/11 10:30	2/7/11 20:47	226	26
45			8	9	18	12	7	7										nytimes.com	1/25/11 4:20	1/30/11 5:56	121	56
48							2	40										usatoday.com	1/29/11 17:45	1/30/11 18:18	24	27
51		4	0	2	2	1	1	1	7	2	5	4	2	1	13	7	8	guardian.co.uk	1/24/11 1:20	2/8/11 20:48	379	59
54					27	8	1	0	1	4	151	67	22					cnn.com	1/27/11 19:15	2/4/11 23:51	196	261
56			8	1	5	1	0	3	3	1	4							bbc.co.uk	1/25/11 10:45	2/2/11 23:47	205	25
57									7	33	62							reuters.com	1/31/11 16:20	2/2/11 19:31	51	93
59				2	9	8	1	2	0	2	8	2	5	4	1	1	5	huffingtonpost.com	1/26/11 0:40	2/8/11 4:47	316	47
66				10	4	12	4	2	10	0	19	78	49					guardian.co.uk	1/26/11 15:02	2/4/11 19:58	220	182
67				7	1	5	3	8	1	1	0	0	0	1	4			sports.espn.go.com	1/26/11 1:26	2/6/11 23:32	286	30
68					16	55	40	9	23	21								tmz.com	1/27/11 20:41	2/1/11 17:47	117	158
71						39	14	90										bbc.co.uk	1/28/11 8:16	1/30/11 11:52	51	138
78		35	32	32	37	21	21	32	30	22	35	46	35	26	28	32	55	forbes.com	1/24/11 17:41	2/8/11 19:00	361	491
79	15	7	4	0	40	15	9	3	3	0	38	45						cnn.com	1/23/11 2:56	2/3/11 19:18	280	171
84	1	2	5	0	4	2	2	7	1	4	3	5	7					nytimes.com	1/23/11 6:29	2/4/11 15:57	297	37
86		15	3															bbc.co.uk	1/24/11 4:56	1/25/11 22:10	41	15
88	24	2	89	22	3	5	5	22	67									bbc.co.uk	1/23/11 13:41	1/31/11 9:10	187	211
91			3	10	4	2	5											huffingtonpost.com	1/25/11 18:36	1/29/11 23:21	100	23
92		3	3	5	3	2	4	0	5	5	4	7	2	1	1	5		cnn.com	1/24/11 15:07	2/7/11 15:07	336	47
95				25	16	16	3	2	4	3	2	6	13					cnn.com	1/26/11 19:30	2/4/11 15:29	211	72
98	13	23	7	5	15	15	8	4	7	8	11	9	9	2	4	2		abc.net.au	1/23/11 7:25	2/7/11 22:23	374	140
99													22	9	61	203		huffingtonpost.com	2/4/11 5:00	2/7/11 16:15	83	250
103		10	4	1	7	2	1	1	0	0	0	0	1	0	0	2		huffingtonpost.com	1/24/11 2:17	2/7/11 23:58	357	29
107								9	7	5	5	3	2	3	9			bbc.co.uk	1/30/11 4:52	2/6/11 12:01	175	36
110	6	4	2	2	5	0	1	1	5	1	8							reuters.com	1/23/11 21:18	2/2/11 3:57	222	26

Figure 5.3: Heat map of number of relevant tweets between two same-source news articles. More **red** denotes higher number of relevant tweets published in the day.

per topic. For each topic, an assessor must first read the initial news article, and then assess the novelty for each tweet sequentially according to their publication time. For each novelty judgment, the assessor needs to remember what he/she read from the article and the previously viewed relevant tweets on the topic, and assign a novelty label to the tweet to indicate whether it brings new information that has not yet been seen.

As discussed in Chapter 4, the definition of "new" is ambiguous. Therefore,

a pilot assessment was conducted using the common sense definition of the world "new". After a discussion with the two assessors involved in the pilot assessment, the issue of what is considered new arises. For example, is more specific information considered new? This leads to the following operational definition of "new" specified for the assessment. Given these constraints, a macro-averaged inter assessor agreement measured by Cohens Kappa, $CK = 0.427$, which suggests a moderate agreement $[0.41, 0.60]$.

- It must be information that has not been read thus far;
- The new information must be relevant to the query topic;
- Personal comments or opinions are not considered to be new;
- New values, like the casualty number after an earthquake, are considered as new;
- More specificity is new, while more generality is not.

In order to finalize the judgment for each tweet, the simple majority vote strategy cannot be applied. Because each novelty judgment is made that is sensitive to the context of the assessor's previous judgments, either we can use them all given a single topic or we can use none of them. Therefore, a measure of the assessor's authority is conducted to indicate which assessor's judgments are more reliable compared to the other assessor's judgments assigned to the same topic. A graph $G := (V, E)$ is first constructed where each vertex v represents an assessor, and an edge $e_{(i,j)}$ is the macro-averaged $(1 - CK_{(i,j)})$ between the two connected

assessors v_i and v_j over co-assigned topics. Then, an eigenvector centrality [192] is calculated for each assessor from the adjacency matrix of G, with the ith component of the eigenvector, which corresponds to the greatest eigenvalue of the ith node in G, gives its centrality score. Table 5.1 lists the centrality score for each assessor. A higher score suggests higher influence of the node in the graph. The graph is constructed according to the Cohen's Kappa score, which means the center of the graph represents the common sense of novelty among the 9 assessors. Thus, the assessor with the higher centrality score is given higher authority in terms of novelty judgment. A similar assessor scoring strategy has also been applied by Liu, et al in their work of assigning reviewers to papers, where they used a Random Walk with Restart algorithm to calculate centrality [109]. Parisi et al. applied the same eigenvector centrality to rank multiple independent classifiers, so that to construct a meta-classifier [152]. Iterative likelihood maximization procedure, pioneered by Dawid and Skene [50], has also been popular recently, especially in crowdsourcing applications [88], where the notion of "authority" is handled using full confusion matrices for each assessor. Whitehill et al. furthered this idea, by employing a Bayesian model of the assessing process that simultaneously considering the assessor's accuracy as well as item difficulty [191]. Carpenter, on the other hand, introduced a hierarchical model of relevant documents in multiple topics, and learned assessor's sensitivity and specificity in the model through a semi-supervised approach [38].

On average, 30 out of 102 relevant tweets were judged as novel. Compared to the novelty detection data used in Chapter 4, the percentage of novel tweets among relevant tweets drops, 28.9% compared to 46.8%. This is because of the

Assessor	EigenVector Centrality
M	0.522
X	0.486
R	0.408
B	0.295
T	0.272
Y	0.261
N	0.206
J	0.187
A	0.134

Table 5.1: Novelty assessors' authority measured by eigenvector centrality.

consideration of an initial news article, where many tweets' content has already been mentioned in the article.

5.2.2 Salience Assessment

Salience assessment is a much simpler task than novelty assessment because of the independence of each judgment. The task for each assessor is that given a tweet (relevant and novel) and a future news article, a binary judgment is made regarding whether content mentioned in the tweet appears in the article. For this task, only 5 out of 9 assessors from the novelty assessment task were recruited because of their

availability.

The assessment procedure is designed as follows:

- 2 assessors are assigned to each topic at random;
- A macro-averaged Cohen's Kappa score of 0.6375 is achieved over topics, which suggests a substantial agreement;
- A third assessor is assigned to each topic;
- A majority vote strategy is applied to make the final decision for each tweet's salience label.

On average, this salience assessment labels 11 out of 30 novel tweets as salient. The data is then split into 2 groups for a 2-fold cross-validation purpose with one group containing 22 topics, and the other 21 topics. Before the split, the topics were shuffled.

5.3 Methodology

This section provides details of the investigated salience detection methods. I begin with some standard features to measure a microblog post's writing quality, which serves the real-time decision purpose well, but is also needed to examine their effectiveness. Then, I attempt to utilize prior knowledge gathered from Web pages on past relevant event collected from Google Search to improve the prediction.

5.3.1 Microblog Quality Measurements

To determine which microblog post is important, the quality of the written post must be considered. There has been a great deal of uneven quality in microblog posts, which downstream applications are required to tackle appropriately. For example, given a microblog post, e.g. tweet, there are several features, either from the post's metadata or content, that could help to infer its quality. In this study, I try out the following 6 quality (QA) features for the experimental Twitter data:

- Whether a tweet contains at least one Web link;
- Whether a tweet contains at least one hashtag;
- The length of the tweet, in words;
- The proportion of "informative" words in a tweet (e.g., excluding common English stopwords and tweet-specific stopwords such as "rt" or "http");
- The average length, in character, of the informative words;
- The proportion of out-of-vocabulary words by checking an English vocabulary with a size of 274,926 words, composed from Letterpress[4], the English Open Word List[5], and other word lists that are publicly available with minor local refinements.

With these features, in addition to the relevance prediction score and novelty prediction score, a neural network is trained with backpropagation algorithm [81]

[4]http://www.atebits.com/letterpress/

[5]http://dreamsteep.com/projects/the-english-open-word-list.html

Prediction Model	Macro-averaged Accuracy
All-Positive	0.549
All-Negative	0.451
2-layer Neural Network with QA Features	0.554

Table 5.2: 2-fold cross validation accuracy of salience detection on local assessed queries with tweet quality features.

through WEKA machine learning toolkit [75]. Two hidden layers with 5 and 3 nodes respectively are added to the network. The parameters (number of hidden layer and number of nodes at each layer) are tuned by 10-fold cross validation on training data. This model can achieve a macro-average salience detection accuracy of 0.554 with a prediction ROC area of 0.627, which suggests a fair prediction. As shown in Table 5.2, although the model can perform better than all-positive and all-negative predictions, it is not statistically significantly according to two-tailed paired t-test (note that because of the relatively small sample size, this conclusion is biased). This is because the features are not designed specifically for salience. In the next section, more designated salience features are explored.

5.3.2 Learning from Past Relevant Web News

Existing studies of microblog summarization mainly focus on retrospective problems. Therefore, a common step is to identify sub-topics either by clustering or using topic modeling approaches [124, 188, 56, 186]. However, due to the focus

on real-time prediction, these methods would not be effective considering the delay in prediction, as discussed in Section 4.4.3. Therefore, in this section, I study how to utilize prior knowledge learned from past news reports to predict which concerns should be important given a query topic.

The idea is that given a topic on a news event, because there are similar past news events that occurred, which were well reported in the news and on the Web, these reports could teach us how to report this similar future topic, and discern microblog posts containing important information. For example, one of the training topics is the "Chicago blizzard," which happened on Feburary 2nd, 2011. Because this type of weather occurrence is not unusual in Chicago, there are plenty of past relevant Web pages or news, that could be used to learn how journalists would report on this topic.

In order to acquire historical Web pages and Web news, Web search engines, like Google, provide a convenient way to search and access the news. They even cache Web sites in case the original pages are no longer available. Howell discussed the value of using these internet archived resources in a legal context [86]. In this study, I investigate their effectiveness in detecting salient microblog posts.

Providing a query topic expressed by a few keywords and interested query time, I first conducted a Web search facilitated by the Google Custom Search Engine API, similar to that described in Section 3.3, but with an additional 14 days prior to the query start time in order to avoid getting the current Web pages for the topic. In addition, I added the word "news" at the end of the query keywords so that the result would be more likely to return Web news. Then from the returned search

results, I picked the top 10 pages, fetched and extracted their textual content,[6] which is the source for the modeling (learning) process. In order to check the quality of the collected content, I manually checked the resulting text for 5 randomly selected topics, which gave me an averaged of 64% in precision. I noticed that the quality varied depending on the topic. For topics with rich historical relevant events that attracted good attentions, e.g. "U.S. unemployment", "Chicago blizzard", and "Charlie Sheen's rehab", a 100% precision was achieved. However, for topics about unique events, the precision was relatively low. One of such example topics was about a U.S. diplomat arrested in Pakistan and charged for murder. The other topic got relatively low precision is the topic of "the daily". The topic is actually about the launch of an ipad newspaper called "the daily". However, without enough context, the query terms are too ambiguous to get any useful results from the Google search.

After collecting the text from the searched Web pages, I computed the verb-noun probability distribution from the text to estimate real verb-noun usage probability distribution when reporting a particular kind of topic. More specifically, given the Web page text, the following natural language processing (NLP) steps were conducted using Factorie NLP toolkit [123]: (1) sentence segmentation; (2) tokenization; (3) lemmatization (WordNet [62] and Porter [194]); (3) part-of-speech (POS) tagging; and (4) dependency parsing. Then, noun type of tokens with a dependency parse tree parent token as verb type were extracted for the probability

[6]If any exception happens during this process, the process skips to the next result, but will keep the total number of 10 Web pages returned

Noun Tag	Description	Verb Tag	Description
NN	Noun, singular or mass	VB	Verb, base form
NNS	Noun, plural	VBD	Verb, past tense
NNP	Proper noun, singular	VBG	Verb, gerund or present participle
NNPS	Proper noun, plural	VBN	Verb, past participle
		VBP	Verb, non-3rd person singular present
		VBZ	Verb, 3rd person singular present

Table 5.3: Noun and verb token tags by Penn Treebank POS tag.

distribution estimation. To identify noun and verb type of tokens, the decision was made according to the Penn Treebank POS tag [118], as shown in Table 5.3:

For each topic, by ranking the extracted verb-noun pairs according to their frequency in sentences, we can observe some good examples of news-report preferred verb-nouns and confusing examples. Table 5.4 lists the top 10 most frequent verb-nouns (lemmatized) for the topic about "U.S. unemployment", which includes Web pages about relevant historical events, and another topic about "Pakistan diplomat arrest murder", which is a more unique event that has appeared less frequently in past relevant Web pages.

For probability distribution estimation, I made a similar n-dimensional diagonal covariance Gaussian distribution assumption, where n is the size of the verb-noun lemma vocabulary extracted from each topic's past relevant Web pages. Then, fol-

U.S. Unemployment		Pakistan Diplomat Arrest Murder	
Verb-Noun Lemma	Sentence Frequency	Verb-Noun Lemma	Sentence Frequency
drive economi	10	kill peopl	11
lost job	9	said claussen	4
ha degre	9	becom minist	4
is rate	8	said lawyer	4
drive point	8	said akbar	4
remain rate	7	found bodi	3
ad job	6	conduct basra	3
fell rate	5	conduct oper	3
increas employ	4	wa offici	3
wa rate	4	wa assassin	3

Table 5.4: Top-10 verb-noun lemmas for 2 example topics.

lowing the equation as defined by 4.8 and 4.9, I estimated parameters $\vec{\mu}$ and $\vec{\theta^2}$ for each topic, and computed the probability that an investigated tweet is draw from the same distribution. Each tweet is also processed through the same NLP pipeline as mentioned above.

Because of the vocabulary gap between words, I also applied the BOEW representation introduced in Section 3.4 to calculate the cosine similarity of the Bag-of-Embedded-Verb-Noun (BOEVN) vector between a topic's historical Web pages

Prediction Model	Macro-averaged Accuracy
2-layer Neural Network with QA Features	0.554
2-layer Neural Network with QA+Gaussian	0.570
2-layer Neural Network with QA+Gaussian+BOEVN	0.607

Table 5.5: 2-fold cross validation accuracy of salience detection on local assessed queries with verb-noun features.

and a tweet. Since word embedding is designed to map a single word to a low-dimensional real-value vector, in order to represent a verb-noun pair, I aggregate the relevant two word vectors by averaging the values.

By adding these two new features, Table 5.5 shows the macro-averaged salience detection accuracy improvement achieved by each of the feature. The neural network classifier uses 2 hidden layers with 6 and 4 nodes respectively. According to the two-tailed paired t-test, the combined accuracy of 0.607 is statistically significant improved from only using the QA features.

5.4 End-to-End Temporal Summarization Evaluation

So far, I have proposed and evaluated methods for microblogging relevance filtering, novelty detection, and salience detection respectively. In this section, I am interested in determining the end-to-end system's effectiveness in producing a temporal summary. It also serves as a stress test for salience detection because the input is noisy relevance filtering and novelty detection results, which contains

```
<top>
<num> Number: MB001 </num>
<title> BBC World Service staff cuts </title>
<querytime> Mon Jan 24 12:02:01 +0000 2011 </querytime>
<querytweettime> 29509222337085440 </querytweettime>
<querynewesttweet> 34952194402811904 </querynewesttweet>
</top>
```

```
<top>
<num> Number: MB001 </num>
<title> BBC World Service staff cuts </title>
<querytime> Mon Jan 24 13:13:09 +0000 2011 </querytime>
<querytweettime> 29527121936261121 </querytweettime>
<querynewesttweet> 30500781002063872 </querynewesttweet>
</top>
```

Figure 5.4: Local topic modification from TREC Microblog Filtering track.

irrelevant or duplicated tweets.

5.4.1 Evaluation Setup

Because only 43 topics were annotated for salient tweets, these topics are used in this evaluation. For each topic, a beginning and ending news report from the same source was selected to serve as the boundary for novelty and salience annotation. The query start and ending time is therefore set accordingly to the corresponding tweets. Because the news articles are extracted from external Web link mentioned in a tweet, the tweet's publication time can be used to approximate the time. This modifies the original TREC Microblog filtering queries as illustrated in Figure 5.4, the highlighted sections are the modifications. According to the introduction in Section 3.6.1, TREC uses a tweet ID, a 16-digit identifier[7] to express the query start and ending time, which is not the practical input for real application, but is used only for system evaluation purpose. Note that, the updated timestamp and start and ending tweet ID are modified according to the boundary news reports selected.

Because TREC Microblog track query relevance judgments are used, I also ad-

[7]https://dev.twitter.com/overview/api/twitter-ids-json-and-snowflake

justed this ground truth by only considering the judgments in-between the beginning and ending tweets.

With regard to evaluation metrics, because given one query, we still have extremely unbalanced positive and negative examples. following work in Section 4.5.3, $F_{\beta=1}$ and $T11SU$ evaluation metrics are used, with precision and recall reported for reference. Note that, methods are evaluated in a harsh way, which means a latter detected salient tweet is considered as a false prediction, even though it is partially correct because of an earlier salient tweet is missed, and thus the latter one should be novel and bring important information. Another setup is that, I reward it with more points in the calculation of $T11SU$ by giving 6 points for a successful salience detection, which is also a true positive prediction in the temporal summary, because it triples efforts for salience detection compared to relevance filtering, which awards 2 points per successfully found relevant tweet. For the lower boundary of the $T11SU$ utility calculation, I continuously use the same MinU as defined in Equation 3.15, which gives a zero effort $T11SU = 0.333$, and can be achieved by returning users no tweets to read.

5.4.2 Results and Analysis

For relevance filtering with modified queries, by using the relevance filtering system developed in Chapter 3, the following macro-averaged score can be achieved across the 43 topics: $precision = 0.5700$, $recall = 0.3699$, $F_{\beta=1} = 0.3996$, and $T11SU = 0.4530$. For novelty detection, I applied the simple ensemble vote to

	Macro-Avg Precision	Macro-Avg Recall	Macro-Avg $F_{\beta=1}$	Macro-Avg $T11SU_6$
All Positive	0.147	0.284	0.156	0.238
QA	0.202	0.207	0.159	0.316
QA+VN	0.416	0.192	0.205	0.407

Table 5.6: 2-fold cross validation effectiveness of salience detection approaches on local assessed queries with predicted relevant and novel tweets as input.

combine individual novelty measurements, which results in $precision = 0.2294$, $recall = 0.2712$, $F_{\beta=1} = 0.2131$, and $T11SU_4 = 0.3631$. Note that, I use $T11SU_4$ to denote 4 credits for a successfully novelty detection. We can observe from the scores that the salience detection process got a reasonable input tweet stream with both relevance filtering and novelty detection working properly compared to test effectiveness that was observed in the previous chapters.

For salience detection, 3 runs were performed: all-positive prediction, neural network using 8 quality measuring features (QA), and neural network using 2 additional features computed based on verb-noun pairs' usage propensity learned from past relevant news reports (VN). Table 5.6 lists their effectiveness.

In this table, $T11SU_6$ denotes the 6 award points used for the salience detection. According to this utility test, only predictor using additional verb-noun based features can produce useful results above zero effort. According to the two-tailed paired t-test, although the 8 tweet quality features can outperform all-positive pre-

diction, this is not statistically significant. The improvement from the 2 verb-noun based features got a two-tailed p-value of 0.1097. Therefore, it is also not statistically significant, despite the nearly 5% absolute $F_{\beta=1}$ improvement. As shown in Figure 5.5, this can be explained by the relatively high standard error of the VN predictor. Therefore, I conducted topic-level error analysis as described in Section 5.4.3.

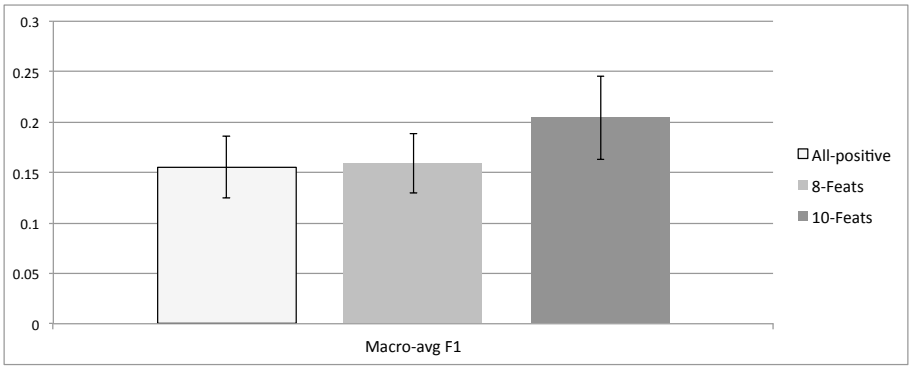


Figure 5.5: Error analysis for salience detection.

5.4.3 Focused Error Analysis

I picked the topic with median $F_{\beta=1} = 0.202$, and its extracted tweets through the three steps, as shown in Figure 5.6. The topic is on "Emanuel residency court rulings", discussing the legitimacy of Mr. Rahm Emanuel to run for mayor of Chicago because of his residency in 2011. As shown in the system produced twitter temporal summary, the highlighted 3 tweets are labeled as salient, and according to the system design, these should be the tweets that may interest a journalist. Because totally, 12 tweets are selected by the system, which results in $precision = 3/12 = 0.250$; 17 tweets are manually labeled as salient, which results in $recall =$

$3/17 = 0.176$.

Query: "Emanuel residency court rulings"
Query Start Time: Mon Jan 24 19:10:09 +0000 2011
Query End Time: Sat Jan 29 03:03:30 +0000 2011

01/24/11 19:26:28 EST 29621072722661376 0 Court tosses Emanuel off Chicago mayoral ballot (Reuters) http://ow.ly/1b0yyG
01/24/11 19:33:24 EST 29622817754447872 0 Courts rule that Rahm Emanuel isn't a Chicago resident. #notthemayor
01/24/11 20:20:37 EST 29634699332685824 0 Court Says Emanuel Not Eligible to Run for Chicago Mayor (NY Times) http://nyti.ms/fuU3QK pas habité sur place assez longtemps
01/24/11 20:50:02 EST 29642099909468160 0 Court rules Rahm Emanuel ineligible for Chicago mayoral race http://bit.ly/gEzifU #law #p2 #il #chicago #rahm @mayoremanuel @RahmEmanuel
01/24/11 21:18:17 EST 29649212366917632 1 Emanuel Seeks Reversal of Ruling Upending Chicago Race: Rahm Emanuel, President Barack Obama'... http://bit.ly/hah8hD #FinancialNews #fb
01/24/11 22:45:08 EST 29671069405151232 1 BREAKING: Rahm Emanuel files motion to STAY Illinois Appellate Court ruling with Illinois Supreme Court. Will file appeal with court TUES.
01/25/11 09:25:02 EST 29832102534971392 0 Rahm Emanuel Residency Twitter Reactions (PHOTOS) http://bit.ly/fa0HJz
01/25/11 18:12:21 EST 29964806458970112 1 Darn! IL supreme court rules to stop printing ballots without Emanuel's name.
01/25/11 21:21:22 EST 30012375142764544 0 Illinois high court will hear Rahm Emanuel appeal - http://yhoo.it/hQE1AJ Panty waist congressional shower stalker gets an appeal.
01/26/11 07:54:09 EST 30171620752494592 0 http://myprops.org/ujzf8 - Illinois Supreme Court hands Emanuel initial victory [Facebook Most Shared Politics]
01/26/11 13:14:59 EST 30252359724572672 0 NEWSMAX #POLITICS: Ill. High Court Mulls Rahm Emanuel's Mayoral Run: http://bit.ly/fASX0B #uknews #news #rt
01/27/11 23:30:33 EST 30769661930905601 0 #Rahm Emanuel ok to run for #Chicago Mayor in #Illinois Supreme Court ruling #xmnr #chicagomayor #rahmemanuel see #http://exm.nr/hwilyz

Figure 5.6: System produced Twitter temporal summary for an example topic MB021.

According to this example topic, an issue of low recall rate is noticed. Figure 5.7 illustrates the macro-averaged recall drop through the 3 temporal summarization sub-processes, where we can observe that after initially input tweets from the corpus API, relevance filtering step loses the most "true" tweets, corresponding to the biggest recall drop, then novelty. Note that (1) in the figure, the bar chart on the right denotes the averaged number of true positive and false negative tweets, which together is the number "true" tweets; and (2) the meaning of "true" varied from step to step, which corresponding to relevant, novel, and salient respectively at each step.

Among the 43 test topics, 4 topics got zero salience recall, thus zero in precision and $F_{\beta=1}$, 2 of these topics lost all the salient tweets from the relevance input stream, and 1 topic even from the initial corpus API. This topic is on "British government cuts", it also has the lowest relevance filtering effectiveness with 5 relevant tweets out of 683 found tweets, and totally there are 59 tweets labeled as relevant, which results in $precison = 0.007, recall = 0.085$, and $F_{\beta=1} = 0.014$.

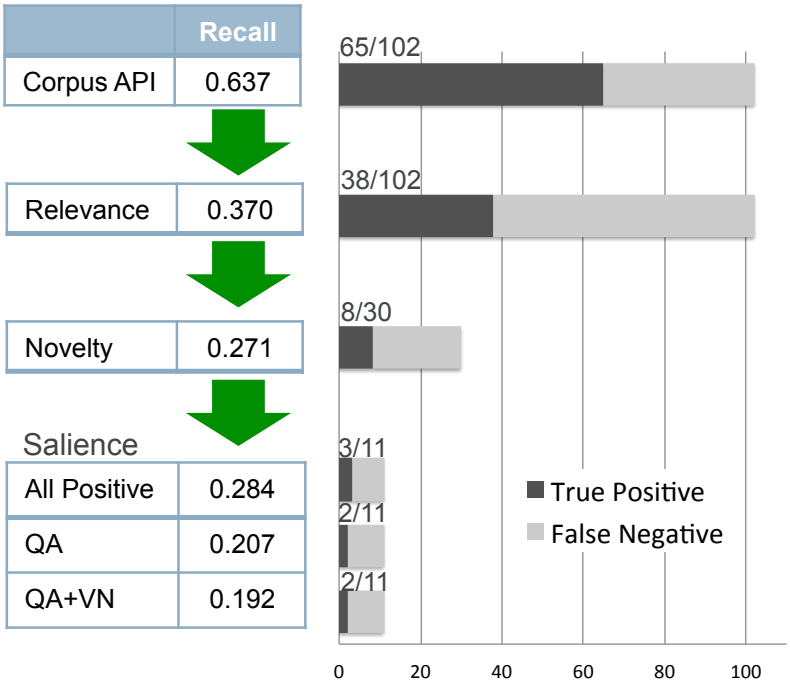


Figure 5.7: Recall drop through temporal summarization sub-processes.

A deeper investigation reveals that this low relevance filtering effectiveness is not because this is a topic with small number of relevant tweets (in fact 59 relevant tweets ranks the topic as the 21st largest topic among the 43 topics), nor because the topic has an ambiguous query, but because the diversity in the sub-topics. When looking at the labeled relevant tweets, it is noticeable that most of the relevant tweets are talking about concrete public services that were affected because of this government cuts. For example, one tweet said "Harrogate Theatre at risk over council cuts", and another tweet said "Women's groups struggle amid funding cuts". Although, both tweets mentioned "cuts", however, that single term, even with consideration of the expansion technologies introduced in Chapter 3, is not enough to distinguish relevant tweets from a large number of irrelevant ones (think of the tweets could be found by "British" and "government"), especially considering

the current employed methods cannot bridge "Harrogate Theatre" and "Women's groups" with"British government".

Meanwhile, the noisy relevance filtering results of the topic also affects its novelty detection effectiveness, which gives zeros in precision, recall and $F_{\beta=1}$. The effect is in two directions: (1) lower precision means more irrelevant tweets are input to novelty detection, which are more easily to be identified as novel; and (2) lower recall means more earlier novel tweets cannot be identified because they are not in the input, and thus causes latter redundant tweets be considered as novel because the system did not see earlier tweets. Therefore, next section investigates this pipeline effect caused by the preceding sub-process(es), and also identifies to current pipeline's bottleneck.

5.4.4 Pipeline Analysis

To show the pipeline effect, a Pearson Correlation Coefficient analysis is conducted [193]. The correlation coefficient between results of the three sub-processes: relevance filtering, novelty detection, and salience detection (with QA+VN features), as well as the initial corpus API are listed in Table 5.7.

According to definition, Pearson correlation coefficient ranges from $[-1, 1]$, a positive value suggests a positive correlation, and a negative value suggests a negative correlation. By convention, absolute value ranges from [0,0.3] is a weak correlation, (0.3,0.6] is a moderate correlation, and a value larger than 0.6 suggests a strong correlation. In the table, we can observe that: (1) salience detection strongly

	Relevance	Novelty	Salience
Search API	0.235	-0.333	-0.369
Relevance		0.668	0.476
Novelty			0.782

Table 5.7: Pearson correlation coefficient between temporal summarization sub-processes.

depends on novelty detection effectiveness; (2) novelty detection strongly depends on relevance filtering effectiveness; (3) salience detection moderately depends on relevance filtering effectiveness; and (4) all the three processes relatively weakly depend on the corpus API's effectiveness and sometimes negatively. These observations confirm the previous error analysis, and suggest that the cascade framework design, although is efficient, however, costs robustness because a latter sub-process depends on its proceeding sub-process's effectiveness heavily.

Under this circumstance, in order to identify the pipeline's bottleneck, an upper bound analysis is conducted. The idea is to use each step's ground truth as input for the next step, and see which step can cause the most improvement of the whole pipeline's effectiveness. For relevance filtering, a perfect result can generate salience detection's effectiveness of $precision = 0.304, recall = 0.333, F_{\beta=1} = 0.308, T11SU_6 = 0.474$. With a perfect novelty input, the salience detection effectiveness is $precision = 0.574$, $recall = 0.637$, $F_{\beta=1} = 0.562$, and $T11SU_4 = 0.684$. By plotting the $F_{\beta=1}$ improvements in a pie chart, we can see that the bottleneck of

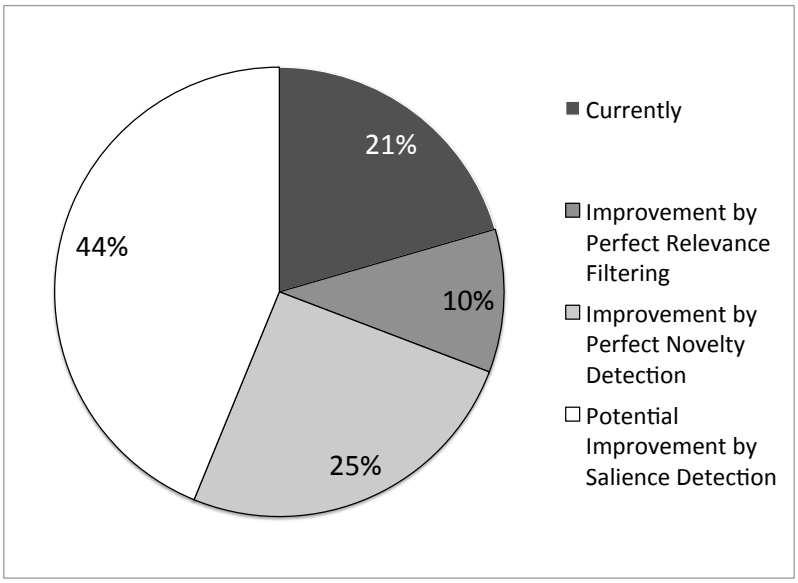


Figure 5.8: Upper-bound analysis for microblogging temporal summarization pipeline.

the current system is salience detection itself, then novelty detection, then relevance filtering.

As one example to illustrate the difficulty in salience detection, a topic is on "Pakistan diplomat arrest murder", which, as mentioned in Section 5.3.2, has few relevant past news reports. In total, the topic is labeled with 4 novel tweets, as listed in Figure 5.9. And all these 4 tweets are also labeled as salient. However, through salience detection with writing quality and verb-noun features, only the 3rd novel tweet is predicted as salient. Thus, the system, even with perfect relevance and novelty input, can only achieve $recall = 0.250$ and $F_{\beta=1} = 0.400$.

Query: "Haiti aristide return"
Query Start Time: Thu Jan 27 13:36:55 +0000 2011
Query End Time: Fri Jan 28 08:48:02 +0000 2011

01/27/2011 15:07:12 EST 30642989122461696 Truth is not what it seems: http://bit.ly/eu8lAN via @addthis My stand on Pakistanis killed by US diplomat Rayan David's > pl read & comment
01/27/2011 15:22:41 EST 30646884209000448 Holy smokes...US consular employee shot down 2 pakistani muggers, then consular van hits & kills a motorcyclist - lahore http://is.gd/Syulf9
01/27/2011 16:04:51 EST 30657495219306497 Pakistan News Armed US diplomat kills two in Pakistan: A U.S. consulate employee shot and killed two people in t... http://bit.ly/hUKJ5y
01/27/2011 20:26:44 EST 30723400687165441 Uh oh, American diplomat charged with double murder in Pakistan: http://yhoo.it/fG04tY He should have just called in a drone.

Figure 5.9: Ground truth Twitter temporal summary for an example topic MB003.

5.5 Conclusion

In this Chapter, I created microblogging salience detection annotation data and proposed two salience detection methods, one with eight features to measure a post's writing quality, and the other relying on verb-noun usage propensity learned from past relevant news report collected through a Web search engine. However, only the second method shows statistically significant effectiveness improvement with perfect relevance and novelty inputs. I also evaluated the proposed temporal summarization system's effectiveness using the local annotated data, and analyzed the effects of each sub-component: relevance filtering, novelty detection and salience detection in the system. Finally, I conclude that salience detection is currently the system's bottleneck and is worth more attention in the future.

Chapter 6

Conclusions

Microblogs have played an important role in delivering live news reports to interested audiences. They cover all ranges of topics happening in our world, and inspire people to read and share opinions and information. Especially for breaking news, microblogs are a quick way to answer the question "what just happened?". Motivated by the story that Twitter was the first to break the news of Bin Laden's death, I used this thesis to propose a microblogging temporal summarization system to help people utilize the power of the human sensors on the Web. This allows a valuable microblog post to be recommended to an interested user in real time, and saves the user time by keeping him or her from digesting the noisy amount of microblogging input. In this chapter, I summarize major approaches investigated by highlighting findings and contributions. Then I address the limitations of this study and point out future work.

6.1 Summary of Findings

Based on work in TREC 2013 Temporal Summerization [198] and TREC 2014 Microblog track [199], I presented a microblogging temporal summarization framework to systematically study the three involved sub-problems: microblog filtering, novelty detection and salience detection. In this section, I summarize major ap-

proaches for each sub-problem.

6.1.1 Microblog Filtering

For the microblog filtering, I focused on the word expansion techniques and word embedding to tackle the data sparsity issue of the task at hand. This led to investigation of various word expansion techniques, including initial query expansion utilizing Web search, incrementally expanding a query from previous filtering decisions, and tweet expansion from its linked Web content. This also led to exploration of the Bag-of-Embedded-Words model, which represents and measures similarity between a query and microblog post in semantic space. In addition, a new machine learning-based combination of the two complementary approaches is studied.

Given a query and a microblog post, and the proposed multiple ways to discern their relevance, I analyzed microblog filtering in a train/validation/test framework leading to a quantitative expression of the difference between various models. This allowed me to determine the best trade-off between filtering precision and recall, as demonstrated by a state-of-the-art result of the TREC 2012 Microblog Real-time Filtering evaluation.

6.1.2 Novelty Detection

Differential novelty detection techniques were investigated to identify novel microblog posts from relevant ones. Because what is considered "new" depends on the context, accurately representing novelty is difficult. In contrast, features based

on texture cues make the detection task easier because it is very likely that novel microblog posts use different words. This reduces the ambiguities and allows us to use various neighborhood-based, density-based, information theory-based, or statistical approaches for the detection. Furthermore, I also investigated four ensemble learning techniques: voting, bagging, boosting, and stacking. As a comparison, I used hierarchical agglomerative complete-link clustering based approach as benchmark with an improved version to dynamically decide the clustering threshold with consideration of a query's specificity.

The TREC 2014 Microblog track Tweet Timeline Generation task's evaluation queries and manual created tweet clustering are used to compare effectiveness of the proposed methods. Because novelty detection is designed as a consecutive step after relevance filtering, in addition to examining these methods with a perfect relevant tweet stream as input, a stress test is also conducted, which uses the proposed microblog filtering system to input noisy relevant tweets from the raw Twitter corpus. Under the two test environments, the most effective and robust novelty detection method is identified, which uses a Logistic Regression model as the base classifier to ensemble the power of single novelty predictors. When comparing with the clustering-based baseline, especially considering the delay in decision time, this novelty detection method demonstrated its effectiveness.

6.1.3 Salience Detection

Two research efforts are taken in the salience detection study: (1) creating a salience detection evaluation data; and (2) devising effective salience detection technique.

In order to create evaluation data, TREC 2011 and 2012 Microblog track queries and relevance judgment are utilized. Local novelty assessment and salience assessment is then conducted with 9 assessors by checking the presence of a tweet's content in a beginning and an ending same-source news articles. An inter-assessor agreement measured by Cohens Kappa reports moderate agreement (0.427) for the novelty assessment, and strong agreement (0.638) for the salience assessment.

For the salience detection methods, two types of feature sets are explored with consideration of real-time feature calculation: a tweet's writing quality is measured in 6 ways; and prior verb-noun pairs usage propensity is learned from past news reports on related events of the query topic. A two-layer Neural Network binary classifier is then learned from training topics to predict whether a tweet is reporting important update regarding a topic. Focused error analysis and pipeline analysis is performed on 2-fold cross validation results of all the local assessed data (43 topics). According to the analysis, salience detection is identified as the bottleneck of the current microblogging temporal summarization pipeline, which heavily depends on effectiveness of novelty detection, and moderately on relevance filtering.

6.2 Limitations and Future Work

Even though this thesis attempts to provide users with a succinct and real-time report for breaking news, the current design is only capable of determining what can be seen and known instantaneously from millions of live microblogs, instead of what is right. In other words, the system favors immediacy over accuracy, and is even less sensitive when it comes to editorializing. Another important aspect of the temporal summarization that I did not address is the limited post processing necessary to synthesize the output into an appropriate human-readable format. It is often the largest source error in algorithms, and a well-formed, cohesive, and coherent summary can further reduce the reading and understanding time and cost that a human reader must possess. The original microblog posts generally offer a good trade-off between linguistic quality and informativeness. Recent literature has described natural language generation trying to produce human-like text, which is an interesting topic for further study [26, 159].

Nowadays, microblogging has become a more common publication tool among journalists and news agencies. Soon, it will be considered a legitimate source of news information. Because the immediately accessible raw microblog data, for example, through Twitter streaming API, the implementation of a microblogging temporal summarization system will be possible with little effort. The main challenges that remain for the system's effectiveness are the noisy quality and shortness of the microblog posts. Additionally, the real-time prediction requirement makes the selection of methods more limited. Thus, it is important to identify the three critical sub-

tasks and explore possible solutions that address each one. The stronger each part functions, the better final temporal summary the whole pipeline can produce. This legitimizes this thesis's work on microblog filtering, novelty detection and salience detection. I believe the techniques and framework have great potential and can be extended to many additional tasks, especially with the advent of creating similar short text, including Facebook statuses, Youtube captions, and Pinterest descriptions. The desires for real time recommendation offers new opportunities to apply this study's results in ways that are unimaginable today.

Bibliography

[1] Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. UMass at TREC 2004 : Novelty and HARD. In *Proceedings of the 13th Text REtrieval Conference*, pages 1–13, 2004.

[2] M-Dyaa Albakour, Craig Macdonald, and Iadh Ounis. On Sparsity and Drift for Effective Real-time Filtering in Microblogs. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 419–428. ACM, 2013.

[3] Arifah Che Alhadi, Thomas Gottron, Jérôme Kunegis, and Nasir Naveed. LiveTweet: Microblog Retrieval Based on Interestingess and an Adaptation of the Vector Space Model. In *Proceedings of the 20th Text REtrieval Conference*, 2011.

[4] James Allan. Incremental Relevance Feedback for Information Filtering. In *Research and Development in Information Retrieval*, pages 270–278. ACM, 1996.

[5] James Allan. Introduction to Topic Detection and Tracking. *Handbook of Information Science*, 12:1–16, 2002.

[6] James Allan, Rahul Gupta, and Vikas Khandelwal. Topic Models for Summarizing Novelty. In *ARDA Workshop on Language Modeling and Information Retrieval*, pages 1–6, Pittsburgh, PA, USA, 2001.

[7] James Allan, Hubert Jin, Martin Rajman, Charles Wayne NSA, Daniel Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. Topic-based Novelty Detection 1999 Summer Workshop at CLSP Final Report. pages 1–59, 1999.

[8] Sihem Amer-Yahia, Samreen Anjum, Amira Ghenai, Aysha Siddique, Sofiane Abbar, Sam Madden, Adam Marcus, and Mohammed El-Haddad. MAQSA: a System for Social Analytics on Mews. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 653–656. ACM, 2012.

[9] Enrique Amigo and Jorge Carrillo de Albornoz. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 307–322. Springer, 2014.

[10] Enrique Amigo, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martin, Edgar Meij, Maarten Rijke, and Damiano Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems.

In Pamela Forner, Henning Miller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 333–352. Springer Berlin Heidelberg, 2013.

[11] Massih-Reza Amini and Patrick Gallinari. Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–112. ACM, 2002.

[12] Karl Appel, Lauren Mathews, Darren Lim, and Sharon Small. Siena's Twitter Information Retrieval System : The 2012 Microblog Track. Technical report, Gaithersburg, MD, USA, 2012.

[13] Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Virgi Pavlu, and Tetsuya Sakai. TREC 2013 Temporal Summarization. In *Proceedings of the 22nd Text REtrieval Conference*, Gaithersburg, MD, USA, 2013. NIST.

[14] Roja Bandari, Sitaram Asur, and Bernardo a Huberman. The Pulse of News in Social Media: Forecasting Popularity. *Proceedings of the 6th International Conference on Weblogs and Social Media*, 2012.

[15] Ayan Bandyopadhyay, Mandar Mitra, and Prasenjit Majumder. Query Expansion for Microblog Retrieval. *Proceedings of the 20th Text REtrieval Conference*, 2011.

[16] Regina Barzilay and Lillian Lee. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 113–120, 2004.

[17] Marcia J. Bates. Information Search Tactics. *Journal of the American Society for Information Science*, 30(4):205–214, 2007.

[18] Nicholas J. Belkin and W. Bruce Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 35(12):29–38, 1992.

[19] Allan Bell. *The Language of News Media*. Blackwell Oxford, 1991.

[20] Michael Bendersky, W Bruce Croft, and Yanlei Diao. Quality-biased Ranking of Web Documents. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, page 95. ACM, 2011.

[21] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[22] Sabine Bergler. Generating Update Summaries for DUC 2007. In *Proceedings of the Document Understanding Conference*, 2007.

[23] Ms Bernstein, Osama Badar, David R Karger, Samuel Madden, Robert C Miller, and Adam Marcus. Tweets as Data: Demonstration of TweeQL and Twitinfo. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pages 1259–1261, 2011.

[24] David M. Blei, Michael I Jordan, and Andrew Y. Ng. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 1(45):993–1022, 2003.

[25] Stephen Blott, Oisin Boydell, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Gareth J F Jones, Noel Murphy, Noel O Connor, Alan F Smeaton, Barry Smyth, and Peter Wilkins. Experiments in Terabyte Searching , Genomic Retrieval and Novelty Detection for TREC-2004. *Proceedings of the 13th Text REtrieval Conference*, 2004.

[26] Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. Natural Language Generation in the Context of the Semantic Web. *Semantic Web*, 5(6):493–513, 2014.

[27] Florian Boudin, Marc El-b, and Avignon Cedex. A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization. *Proceedings of the 21st International Conference on Computational Linguistics: Posters*, (August):23–26, 2008.

[28] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, pages 1–10, 2010.

[29] W Brandenberg. *Selective Dissemination of Information: SDI 2 System*. International Business Machines Corporation, Advanced Systems Development Division, 1961.

[30] J Breese, D Heckermanx, and C Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[31] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[32] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: Identifying Density-based Local Outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 29(2):93–104, 2000.

[33] Andrei Broder. A Taxonomy of Web Search. *ACM SIGIR Forum*, 36(2):3, 2002.

[34] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 264–270, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics.

[35] Chris Buckley and Ellen M Voorhees. Retrieval Evaluation with Incomplete Information. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[36] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336. ACM, 1998.

[37] Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D Brown, Tom Pierce, and Xin Liu. CMU Report on TDT-2: Segmentation, Detection and Tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120, 1999.

[38] Bob Carpenter. A Hierarchical Bayesian Model of Crowdsourced Relevance Coding. In *Proceedings of the 20th Text REtrieval Conference*, 2011.

[39] Simon Carter. Twitter Hashtags : Joint Translation and Clustering. *Human Factors*, pages 1–3, 2011.

[40] Asli Celikyilmaz and Dilek Hakkani-tur. A Hybrid Hierarchical Model for Multi-Document Summarization. In *Proceedings of the 48th Annual Meeting on Association for Computational Linguistics*, number July, pages 815–824. Association for Computational Linguistics, 2010.

[41] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

[42] Ming-Wei Chang, Wen-tau Yih, and Christopher Meek. Partitioned Logistic Regression for Spam Filtering. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 97–105. ACM, 2008.

[43] Stanley F Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[44] Jaeho Choi, W. Bruce Croft, and Jin Young Kim. Quality Models for Microblog Retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 1834. ACM, 2012.

[45] S Clinchant and Florent Perronnin. Aggregating Continuous Word Embeddings for Information Retrieval. *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 100–109, 2013.

[46] Ronan Collobert and Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167. ACM, 2008.

[47] John M Conroy and Dianne P O'leary. Text Summarization via Hidden Markov Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 406–407. ACM, 2001.

[48] John M Conroy, Judith D Schlesinger, and Dianne P O'Leary. Topic-Focused Multi-Document Summarization Using an Approximate Oracle Score. In *Proceedings of the COLING/ACL Poster*, COLING-ACL '06, pages 152–159, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[49] Hal Daumé III and Daniel Marcu. Bayesian Query-focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 305–312, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[50] A P Dawid and A M Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[51] S Deerwester, S Dumais, G Furnas, T Landauer, and R Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[52] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. In *Proceedings of the 2010 IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2010.

[53] Fernando Diaz and Matthew Ekstrand-abueg. TREC 2014 Temporal Summarization Track Overview. In *Proceedings of the 24th Text REtrieval Conference*, pages 1–15, Gaithersburg, MD, USA, 2015. NIST.

[54] M Dork, D Gruen, C Williamson, and S Carpendale. A Visual Backchannel for Large-Scale Events. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1129–1138, 2010.

[55] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-yeung Shum. An Empirical Study on Learning to Rank of Tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, number August

in COLING '10, pages 295–303, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[56] Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P Xing. A Non-parametric Mixture Model for Topic Modeling over Time. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 530–538. SIAM, 2013.

[57] M Efron, A Kehoe, and P Organisciak. The University of Illinois' Graduate School of Library and Information Science at TREC 2011. In *Proceedings of the 20th Text REtrieval Conference*, pages 1–10, 2011.

[58] Ahmed Saad El Din and Walid Magdy. Web-based Pseudo Relevance Feedback for Microblog Retrieval. In *Proceedings of the 21st Text REtrieval Conference*, 2012.

[59] T El-Ganainy, Z Wei, W Magdy, and W Gao. QCRI at TREC 2013 Microblog Track. In *Proceedings of the 22nd Text REtrieval Conference*, 2013.

[60] Gunes Erkan and Dragomir R. Radev. LexPageRank : Prestige in Multi-Document Text Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[61] Güne Erkan and Dragomir R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, 2004.

[62] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[63] Paul Ferguson, Neil O Hare, James Lanagan, Owen Phelan, and Kevin Mccarthy. An Investigation of Term Weighting Approaches for Microblog Retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 3–6, Berlin, Heidelberg, 2012. Springer-Verlag.

[64] Christian Fluhr. Information Filtering. In LING LIU and M.TAMER ÖZSU, editors, *Encyclopedia of Database Systems*, pages 1481–1484. Springer US, 2009.

[65] Roger Fowler. *Language in the News: Discourse and Ideology in the Press*, volume 20. Routledge London, 1991.

[66] Bob Franklin, Martin Hamer, Mark Hanna, Marie Kinsey, and John E. Richardson. *Key Concepts in Journalism Studies*. Sage Publications Limited, 2005.

[67] Yoav Freund and Robert E Schapire. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[68] Fumiyo Fukumoto, Akina Sakai, and Yoshimi Suzuki. Eliminating redundancy by spectral relaxation for multi-document summarization. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-5, pages 98–102, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[69] Fumiyo Fukumoto and Yusuke Yamaji. Topic Tracking Based on Linguistic Features. In Robert Dale, Kam-Fai Wong, Jian Su, and OiYee Kwong, editors, *Natural Language Processing IJCNLP 2005*, volume 3651 of *Lecture Notes in Computer Science*, pages 10–21. Springer Berlin Heidelberg, 2005.

[70] Michael Gamon. Graph-based Text Representation for Novelty Detection. In *Proceedings of the 1st Workshop on Graph Based Methods for Natural Language Processing*, number June in TextGraphs-1, pages 17–24, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[71] Thomas Gottron and Arifah Che Alhadi. Bad News Travel Fast : A Content-based Analysis of Interestingness on Twitter. *Proceedings of the 3rd International Web Science Conference*, pages 1–7, 2011.

[72] Mena B. Habib and Maurice Van Keulen. Unsupervised Improvement of Named Entity Extraction in Short Informal Context Using Disambiguation Clues. *Workshop on Semantic Web and Information Extraction*, 925:1–10, 2012.

[73] Aria Haghighi and Lucy Vanderwende. Exploring Content Models for Multi-Document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, number June in NAACL '09, pages 362–370, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[74] Udo Hahn and Inderjeet Mani. The Challenges of Automatic Summarization. *Computer*, 33(11):29–36, 2000.

[75] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[76] Z Han, X Li, M Yang, H Qi, S Li, and T Zhao. HIT at TREC 2012 Microblog Track. In *Proceedings of the 21st Text REtrieval Conference*, pages 267–276, Gaithersburg, MD, USA, 2012. NIST.

[77] Tony Harcup and Deirdre O'Neill. What Is News? Galtung and Ruge Revisited. *Journalism Studies*, 2(2):261–280, 2001.

[78] Donna Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[79] Vasileios Hatzivassiloglou, Judith L Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. SIMFINDER : A Flexible Clustering Tool for Summarization. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

[80] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier Detection Using k-Nearest Neighbour Graph. In *Proceedings of the 17th International Conference on Pattern Recognition - Volume 03*, ICPR '04, pages 430–433, Washington, DC, USA, 2004. IEEE Computer Society.

[81] Robert Hecht-Nielsen. Theory of the Backpropagation Neural Network. In Harry Wechsler, editor, *Neural Networks for Perception*, chapter Theory of, pages 65–93. Harcourt Brace & Co., Orlando, FL, USA, 1992.

[82] A Hickl, K Roberts, and F Lacatusu. LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization. In *Proceedings of the Document Understanding Conference*, volume 7, 2007.

[83] Graeme Hirst. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press, 1992.

[84] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57. ACM, 1999.

[85] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 57–58. ACM, 2011.

[86] Ba Howell. Proving Web History: How to Use the Internet Archive. *Journal of Internet Law*, 9(8):3–9, 2006.

[87] Elena Ikonomovska, João Gama, and Sašo Džeroski. Learning Model Trees from Evolving Data Streams. *Data Mining and Knowledge Discovery*, 23(1):128–168, 2011.

[88] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality Management on Amazon Mechanical Turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, 2010.

[89] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65. ACM, 2007.

[90] Hubert Jin, Richard Schwartz, Sreenivasa Sista, and Frederick Wall. Topic Tracking for Radio, TV Broadcast, and Newswire. In *Proceedings of the DARPA Broadcast News Workshop*, pages 199–204. San Francisco: Morgan Kaufmann, 1999.

[91] Hideto Kazawa, Tsutomu Hirao, Hideki Isozaki, and Eisaku Maeda. A Machine Learning Approach for QA and Novelty Tracks: NTT System Description. In *Proceedings of the 11th Text REtrieval Conference*, pages 1–4, 2002.

[92] Edward F Kelly and Philip J Stone. *Computer Recognition of English Word Senses*, volume 13. North-Holland Amsterdam, 1975.

[93] Sang-Bum Kim, Hee-Cheol Seo, and Hae-Chang Rim. Information Retrieval Using Word Senses: Root Sense Tagging Approach. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 258–265. ACM, 2004.

[94] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):42–49, 2009.

[95] Wessel Kraaij and Martijn Spitters. Language Models for Topic Tracking. In W.Bruce Croft and John Lafferty, editors, *Language Modeling for Information Retrieval*, volume 13 of *The Springer International Series on Information Retrieval*, pages 95–123. Springer Netherlands, 2003.

[96] Robert Krovetz and W Bruce Croft. Word Sense Disambiguation Using Machine-readable Dictionaries. In *ACM SIGIR Forum*, volume 23, pages 127–136. ACM, 1989.

[97] Julian Kupiec, Jan Pedersen, and Francine Chen. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 68–73. ACM, 1995.

[98] Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger. From Word Embeddings To Document Distances. *Proceedings of The 32nd International Conference on Machine Learning*, 37:957–966, 2015.

[99] William Labov. *Principles of Linguistic Change, Cognitive and Cultural Factors*. John Wiley & Sons, 2011.

[100] Leah S Larkey, Fangfang Feng, Margaret Connell, and Victor Lavrenko. Language-specific Models in Multilingual Topic Tracking. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 402–409. ACM, 2004.

[101] Changki Lee, Gary Geunbae Lee, and Myunggil Jang. Dependency Structure Language Model for Topic Detection and Tracking. *Information Processing and Management*, 43(5):1249–1259, 2007.

[102] Michael Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. ACM, 1986.

[103] Y Li, Z Zhang, W Lv, Q Xie, Y Lin, R Xu, W Xu, G Chen, and J Guo. PRIS at TREC2011 Microblog Track. *Proceedings of the 20th Text REtrieval Conference*, 2011.

[104] Nut Limsopatham, Richard McCreadie, M-Dyaa Albakour, Craig Macdonald, Rodrygo L Santos, Iadh Ounis, and Others. University of Glasgow at TREC 2012: Experiments with Terrier in Medical Records, Microblog, and Web Tracks. Technical report, Gaithersburg, MD, USA, 2012.

[105] Jimmy Lin and Miles Efron. Overview of the TREC2013 Microblog Track. In *Proceedings of the 22nd Text REtrieval Conference*, 2013.

[106] Jimmy Lin, Rion Snow, and William Morgan. Smoothing Techniques for Adaptive Online Language Models : Topic Tracking in Tweet Streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 422–429. ACM, 2011.

[107] Jimmy Lin, Yulu Wang, Miles Efron, and Garrick Sherman. Overview of the TREC-2014 Microblog Track. In *Proceedings of the 23rd Text REtrieval Conference*, Gaithersburg, MD, USA, 2014. NIST.

[108] Ziheng Lin, Tat-seng Chua, and Min-yen Kan. NUS at DUC 2007 : Using Evolutionary Models of Text. In *Proceedings of the Document Understanding Conference*, 2007.

[109] Xiang Liu, Torsten Suel, and Nasir Memon. A Robust Model for Paper Reviewer Assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 25–32. ACM, 2014.

[110] Yuen-yee Lo and Jean-luc Gauvain. The LIMSI Topic Tracking System for TDT2001. *Proceedings of the DARPA Broadcast News Workshop*, pages 1–5, 2001.

[111] Philip M. Long and Rocco a. Servedio. Random Classification Noise Defeats all Convex Potential Boosters. *Machine Learning*, 78(3):287–304, 2010.

[112] S.A. Lowe. The Beta-binomial Mixture Model and its Application to TDT Tracking and Detection. In *Proceedings of the DARPA Broadcast News Workshop*, page 127, 1999.

[113] Zhilin Luo. Predicting Retweeting Behavior Based on Autoregressive Moving Average Model. In X.Sean Wang, Isabel Cruz, Alex Delis, and Guangyan

Huang, editors, *Proceedings of the 13th International Conference on Web Information Systems Engineering*, volume 7651 of *Lecture Notes in Computer Science*, pages 777–782. Springer Berlin Heidelberg, 2012.

[114] Feifan Qiang Runwei Fei Yue Lv Chao Fan and Jianwu Yang. PKUICST at TREC 2014 Microblog Track: Feature Extraction for Effective Microblog Search and Adaptive Clustering Algorithms for TTG. In *Proceedings of the 23rd Text REtrieval Conference*, Gaithersburg, MD, USA, 2014. NIST.

[115] Junshui Ma and Simon Perkins. Online Novelty Detection on Temporal Sequences. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 613. ACM, 2003.

[116] Andrew L Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics*, pages 142–150. Association for Computational Linguistics, 2011.

[117] Inderjeet Mani. *Automatic Summarization*, volume 3. John Benjamins Publishing, 2001.

[118] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[119] Markos Markou and Sameer Singh. Novelty detection: a reviewpart 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[120] Sergio Martin, Gabriel Diaz, Elio Sancristobal, Rosario Gil, Manuel Castro, and Juan Peire. New Technology Trends in Education: Seven Years of Forecasts and Convergence. *Computers and Education*, 57(3):1893–1906, 2011.

[121] L. Massey. Real-world Text Clustering with Adaptive Resonance Theory Neural Networks. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*, volume 5, pages 2748–2753. IEEE, 2005.

[122] Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 362–367, Berlin, Heidelberg, 2011. Springer-Verlag.

[123] Andrew McCallum, Karl Schultz, and Sameer Singh. Factorie: Probabilistic Programming via Imperatively Defined Factor Graphs. In *Advances in Neural Information Processing Systems*, volume 22, pages 1249–1257, 2009.

[124] Qiaozhu Mei and ChengXiang Zhai. Discovering Evolutionary Theme Patterns from Text: an Exploration of Temporal Text Mining. In *Proceedings of the*

11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 198–207. ACM, 2005.

[125] D Metzler and C Cai. USC/ISI at TREC 2011: Microblog Track. In *Proceedings of the 20th Text REtrieval Conference*, 2011.

[126] Donald Metzler and Tapas Kanungo. Machine Learned Sentence Selection Strategies for Query-biased Summarization. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Learning to Rank Workshop*, 2008.

[127] Rada Mihalcea and Paul Tarau. A Language Independent Algorithm for Single and Multiple Document Summarization. In *Proceedings of the International Joint Conference on Natural Language Processing*, volume 5, 2005.

[128] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111—-3119, 2013.

[129] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernock\‘y, and Sanjeev Khudanpur. Recurrent Neural Network Based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048, Makuhari, Chiba, Japan,, 2010.

[130] Dubravko Miljkovic. Review of Novelty Detection Methods. In *Proceedings of the 33rd International Convention MIPRO*, pages 593–598, 2010.

[131] Gilad Mishne and M De Rijke. A Study of Blog Search. In *Advances in Information Retrieval*, volume 3936 of *LNCS*, page 12. Springer, 2006.

[132] Andriy Mnih and Geoffrey Hinton. Three New Graphical Models for Statistical Language Modelling. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 641–648. ACM, 2007.

[133] Alexandros Moukas. Amalthaea Information Discovery and Filtering Using a Multiagent Evolving Ecosystem. *Applied Artificial Intelligence*, 11(5):437–457, 1997.

[134] M Naaman, J Boase, and CH Lai. Is it Really About Me? Message Content in Social Awareness Streams. In M Ac, editor, *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192. ACM, 2010.

[135] Naama Nagar. The Loud Public: Users' Comments and the Online News Media. *Online Journalism Symposium*, 2009.

[136] Rinkesh Nagmoti, Ankur Teredesai, and Martine De Cock. Ranking Approaches for Microblog Search. In X J Huang, I King, V Taghavan, and S Rueger, editors, *Proceedings of the 2010 IEEE/WIC/ACM International*

Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, volume 1, pages 153–157. IEEE Computer Society, 2010.

[137] Ramesh Nallapati. Semantic Language Models for Topic Detection and Tracking. In *Proceedings of Human Language Technologies: The 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics - Student Research Workshop*, number June in NAACLstudent '03, pages 1–6, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[138] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Searching Microblogs: Coping with Sparsity and Document Quality. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 183–188. ACM, 2011.

[139] R Nedunchelian. Centroid Based Summarization of Multiple Documents Implemented Using Timestamps. In *Proceedings of the 2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 480–485, 2008.

[140] Ani Nenkova and Lucy Vanderwende. The Impact of Frequency on Summarization. *Microsoft Research Redmond Washington Technical Report MSRTR2005101*, 2005.

[141] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing Sporting Events Using Twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, page 189. ACM, 2012.

[142] Douglas W Oard. Topic Tracking with the PRISE Information Retrieval System. In *Proceedings of the DARPA Broadcast News Workshop*, pages 209–211, 1999.

[143] Detecting Outlying Observations, Samples Author, Frank E Grubbs Source, American Society, Quality Stable Url, and FE Grubbs. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21, 1969.

[144] B O'Connor, M Krieger, and D Ahn. TweetMotif : Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, number May, pages 2–3, 2010.

[145] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 11:122–129, 2010.

[146] Vicki L O'Day and Robin Jeffries. Orienteering in an Information Landscape: How Information Seekers get from Here to There. In *Proceedings of the*

INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, CHI '93, pages 438–445. ACM, 1993.

[147] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. Overview of the TREC-2011 Microblog Track. In *Proceedings of the 20th Text REtrieval Conference*, 2011.

[148] Deirdre ONeill and Tony Harcup. News Values and Selectivity. *The Handbook of Journalism Studies*, pages 161–174, 2009.

[149] Aditya Pal and Scott Counts. Identifying Topical Authorities in Microblogs. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 45–54. ACM, 2011.

[150] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[151] Ron Papka, James Allan, and Victor Lavrenko. UMASS Approaches to Detection and Tracking at TDT2. In *Proceedings of the DARPA Broadcast News Workshop*, pages 2–7, 1999.

[152] Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and Combining Multiple Predictors without Labeled Data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):1253–8, 2014.

[153] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 12:1532–1543, 2014.

[154] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming First Story Detection with Application to Twitter. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[155] Saša Petrović, Miles Osborne, Victor Lavrenko, and Sasa Petrovic. Using Paraphrases for Improving First Story Detection in News and Twitter. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL HLT '12, pages 338–346, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[156] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, page 10. Association for Computational Linguistics, 2000.

[157] Dragomir R. Radev, Hongyan Jing, Magorzata Styś, and Daniel Tam. Centroid-based Summarization of Multiple Documents. *Information Processing and Management*, 40(6):919–938, 2004.

[158] K. Rajaraman, K Rajaraman, and Ah-hwee Tan. Topic Detection, Tracking and Trend Analysis Using Self-organizing Neural Networks. In David Cheung, GrahamJ. Williams, and Qing Li, editors, *Advances in Knowledge Discovery and Data Mining*, volume 2035 of *Lecture Notes in Computer Science*, pages 102 – 107. Springer Berlin Heidelberg, 2001.

[159] E Reiter and R Dale. *Building Natural Language Generation Systems*, volume 33. MIT Press, 2000.

[160] Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[161] Stephen E Robertson and Ian Soboroff. The TREC 2002 Filtering Track Report. In *Proceedings of the 11th Text REtrieval Conference*, volume 2002, page 5, 2002.

[162] Joseph John Rocchio. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323, 1971.

[163] He Ruifang, Qin Bing, Liu Ting, Liu Yang, and Li Sheng. Iterative Feedback Based Manifold-Ranking for Update Summary. *Polibits*, 37:5–14, 2008.

[164] Robert E Schapire, Yoram Singer, and Amit Singhal. Boosting and Rocchio Applied to Text Filtering. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 215–223. ACM, 1998.

[165] Barry Schiffman and Kathleen R Mckeown. Columbia University in the Novelty Track at TREC 2004. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

[166] Philip Schlesinger. *Putting 'Reality' Together: BBC News*. Methuen London, 1987.

[167] Hinrich Schütze and Jan O Pedersen. Information Retrieval Based on Word Senses. *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

[168] Da Shamma. Conversational Shadows: Describing Live Media Events Using Short Messages. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 331–334, 2010.

[169] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Tweet the Debates: Understanding Community Annotation of Uncollected Sources. In *Proceedings of the 1st ACM SIGMM International Workshop on Social Media*, WSM '09, pages 3–10. ACM, 2009.

[170] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Peaks and Persistence: Modeling the Shape of Microblog Conversations. In *Proceedings of the 2011 ACM Conference on Computer Supported Cooperative Work*, CSCW '11, pages 355–358. ACM, 2011.

[171] Upendra Shardanand and Pattie Maes. Social Information Filtering: Algorithms for Automating 'Word of Mouth'. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, volume 1, pages 210–217. ACM Press/Addison-Wesley Publishing Co., 1995.

[172] Beaux Sharifi, Mark Anthony Hutton, and Jugal Kalita. Summarizing Microblogs Automatically. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[173] Ian Soboroff. Overview of the TREC 2004 Novelty Track. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

[174] Ian Soboroff and Donna Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the 12th Text REtrieval Conference*, pages 38–53, 2003.

[175] Ian Soboroff, Donna Harman, and M. D. Gaithersburg. Novelty Detection: the TREC Experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, volume 2004, pages 105–112. Association for Computational Linguistics, 2005.

[176] Ian Soboroff, Dean Mccullough, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Richard Mccreadie. Evaluating Real-Time Search Over Tweets. *Artificial Intelligence*, pages 579–582, 2012.

[177] James Surowiecki. The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business. *Economies, Societies and Nations*, 2004.

[178] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. #TwitterSearch: a Comparison of Microblog Search and Web Search. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, page 35, Hong Kong, 2011. ACM.

[179] Tomoe Tomiyama, Kosuke Karoji, Takeshi Kondo, Yuichi Kakuta, Tomohiro Takagi, Akiko Aizawa, and Teruhito Kanazawa. Meiji University Web, Novelty and Genomic Track Experiments. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

[180] Flora S Tsai, Wenyin Tang, and Kap Luk Chan. Evaluation of Novelty Metrics for Sentence-level Novelty Mining. *Information Sciences*, 180(12):2359–2374, 2010.

[181] Flora S. Tsai and Yi Zhang. D2S: Document-to-sentence Framework for Novelty Detection. *Knowledge and Information Systems*, 29(2):419–433, 2011.

[182] Sayan Unankard, Ling Chen, Peng Li, Sen Wang, Zi Huang, Mohamed a. Sharaf, and Xue Li. On the Prediction of Re-tweeting Activities in Social Networks - A Report on WISE 2012 Challenge. In X.Sean Wang, Isabel Cruz, Alex Delis, and Guangyan Huang, editors, *Proceedings of the 13th International Conference on Web Information Systems Engineering*, WISE'12, pages 744–754. Springer Berlin Heidelberg, 2012.

[183] I Uysal and W B Croft. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2261–2264. ACM, 2011.

[184] Ellen M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 171–180. ACM, 1993.

[185] Xiaojun Wan and Jianwu Yang. Multi-document Summarization Using Cluster-based Link Analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 299–306. ACM, 2008.

[186] Chong Wang, David Blei, and David Heckerman. Continuous Time Dynamic Topic Models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, 2008.

[187] Dingding Wang and Tao Li. Document Update Summarization Using Incremental Hierarchical Clustering. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 279–288. ACM, 2010.

[188] Xuerui Wang and Andrew McCallum. Topics over Time: a Non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433. ACM, 2006.

[189] S Watanabe, T Iwata, and T Hori. Application of Topic Tracking Model to Language Model Adaptation and Meeting Analysis. In *Spoken Language Technology Workshop*, pages 378–383, 2010.

[190] Uriel Weinreich, William Labov, and Marvin I. Herzog. Empirical Foundations for a Theory of Language Change. 1968.

[191] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, 22(1):1–9, 2009.

[192] Wikipedia. Centrality, 2015.

[193] Wikipedia. Pearson Product-moment Correlation Coefficient, 2015.

[194] Peter Willett. The Porter Stemming Algorithm: Then and Now. *Program: Electronic Library and Information Systems*, 40(3):219–223, 2006.

[195] T.D. Wilson. Information Behaviour: an Interdisciplinary Perspective. *Information Processing and Management*, 33(4):551–572, 1997.

[196] T.D. Wilson. Models in Information Behaviour Research. *Journal of Documentation*, 55(3):249–270, 1999.

[197] David H. Wolpert. Stacked Generalization. *Neural Networks*, 5(2):241–259, 1992.

[198] Tan Xu, Paul Mcnamee, and Douglas W Oard. HLTCOE at TREC 2013: Temporal Summarization. In *Proceedings of the 22nd Text REtrieval Conference*, Gaithersburg, MD, USA, 2013. NIST.

[199] Tan Xu, Paul Mcnamee, and Douglas W Oard. HLTCOE at TREC 2014: Microblog and Clinical Decision Support. In *Proceedings of the 23rd Text REtrieval Conference*, Gaithersburg, MD, USA, 2014. NIST.

[200] Tan Xu and Douglas W. Oard. Wikipedia-based Topic Clustering for Microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[201] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. Van Mulbregt. Topic Tracking in a News Stream. In *Proceedings of the DARPA Broadcast News Workshop*, pages 133–136, 1999.

[202] Yiming Yang, Tom Ault, Thomas Pierce, and Charles W. Lattimer. Improving Text Categorization Methods for Event Tracking. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 65–72. ACM, 2000.

[203] Yiming Yang, Jaime Carbonell, Brown Ralf, Lafferty John, Thomas Pierce, and Thomas Ault. Multi-strategy Learning for Topic Detection and Tracking. In James Allan, editor, *Proceedings of the DARPA Broadcast News Workshop*, volume 12 of *The Information Retrieval Series*, pages 85–114. Springer US, 2002.

[204] Chengxiang Zhai and John Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342. ACM, 2001.

[205] Chengxiang Zhai and John Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *Proceedings of the 10th ACM International Conference on Information and Knowledge Management*, CIKM '01, pages 403–410. ACM, 2001.

[206] Jiayue Zhang, Sijia Chen, Yue Liu, Jie Yin, Qianqian Wang, Weiran Xu, and Jun Guo. PRIS at 2012 Microblog Track. Technical Report 1, Gaithersburg, MD, USA, 2012.

[207] Jin Zhang, Pan Du, Hongbo Xu, and Xueqi Cheng. ICTGrasper at TAC2009 : Temporal Preferred Update Summarization. In *Proceedings of Text Analysis Conference*, 2009.

[208] Qi Zhang, Jihua Kang, Jin Qian, and Xuanjing Huang. Continuous Word Embeddings for Detecting Local Text Reuses at the Semantic Level. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 797–806. ACM, 2014.

[209] Xianfei Zhang, Zhigang Guo, and Bicheng Li. An Effective Algorithm of News Topic Tracking. In *Proceedings of the 2009 WRI Global Congress on Intelligent Systems - Volume 03*, GCIS '09, pages 510–513, Washington, DC, USA, 2009. IEEE Computer Society.

[210] Yi Zhang. Using Bayesian Priors to Combine Classifiers for Adaptive Filtering. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 345–352. ACM, 2004.

[211] Yi Zhang, Jamie Callan, and Thomas Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 81–88. ACM, 2002.

[212] Le Zhao, Min Zhang, and Shaoping Ma. The Nature of Novelty Detection. *Information Retrieval*, 9(5):521–541, 2006.