ENGAGING MAINTREAM MEDIA FOR EFFICIENT CONTENT DISTRIBUTION AND CREATION

By

Aleksandre Lobzhanidze

A Dissertation

Presented to the Faculty of the Graduate School of

The University of Missouri

in Partial Fulfillments

for the Degree of

Doctoral of Philosophy

The University of Missouri

May 2014

The dissertation Committee for Aleksandre Lobzhanidze certifies that this is the approved version of the following dissertation:

ENGAGING MAINSTREAM MEDIA FOR EFFICIENT CONTENT DISTRIBUTION AND CREATION

Committee:

____
Wenjun Zeng, Supervisor

____
Yi Shang

____
Jianlin Cheng

____
Grant Scott

____

Dedication

To my family

ACKNOWLEDGEMENTS

To my family, who has given me the best possible knowledge and education to make it through difficult times in life.

I would especially like to acknowledge the assistance of my advisor Dr. Wenjun Zeng, who provided me with thoughtful insights, interesting challenges, and wonderful working environment. His guidance and patience has led me to the right direction of identifying and describing the issues that my dissertation deals with. His criticism and insights have helped me overcome lots of problems. Without his guidance, knowledge, commitment, and help my thesis would hardly be done and would have been deterred many times.

I would like to thank Dr. Grant Scott, Dr. Yi Shang, and Dr. Jianlin Cheng for taking time from their extremely busy schedule to serve on my PhD thesis committee and contribute valuable insights.

Lastly, I want to thank all my friends at the University of Missouri. Five years in a graduate school were tough and challenging, sometimes stressful and depressing. They have made my stay at Mizzou a great experience and thanks to them I have many memorable memories to take with; these memories will last my whole life.

Table of Contents

ACKNOWLEDGEMENTS ii

List of Figures vi

List of Tables ix

Abstract xi

Chapter 1 1

Introduction 1

1.1 Online Video Retrieval and Caching 7

1.2 News Item Suggestion & Automatic Content Creation 11

1.3 Organization of the Dissertation 16

Chapter 2 18

Background and Related Work 18

2.1 Online Video Caching 18

2.2 Propagation in Social Network 19

2.3 Video Popularity Prediction 20

2.4 Topic Modeling 21

2.5 Frequent patterns, Associations, and Correlations 25

2.6 Frequent Itemset 28

2.6.1 The Apriori Algorithm: Finding Frequent Itemset Using Candidate Generation 30

2.7 Semantic Web 33

Chapter 3 .... **35**

Effective Trend Detection in the Media **35**

3.1 Trends in the Media **36**

3.2 Social Media vs. Mainstream Media **38**

3.3 Methodology of Data Collection **39**

3.4 Trend Identification in the Media **42**

3.4.1 Detection and Analysis of Common Trends 42

3.4.2 Overlap of Trends 43

3.4.3 Trend Origins 45

3.4.4 Story Development Process in Media 48

3.5 Summary **51**

Chapter 4 .... **53**

Trend-aware Video Popularity Prediction and Proactive Caching **53**

4.1 Overview of the Proposed Proactive Caching System **54**

4.2 Popular Trend Detection **57**

4.3 Video Query Keyword Extraction **58**

4.4 Selecting Video for Proactive Caching **62**

4.4.1 Cache Candidate Selection 62

4.4.2 Cache replacement 65

4.5 Experimental Results **66**

4.5.1 Data Collection 67

4.5.2 Evaluating Keyword Selection 67

4.5.3 Evaluating Caching Candidate Selection 70

4.5.4 Efficiency of Proactive Caching ... 71

4.6 Summary ... 78

Chapter 5 ... 79

An Automated News Item Suggestion System ... 79

5.1. Related Work ... 81

5.2 System Overview ... 84

5.3 Exploring Frequent Patterns in News Articles ... 89

5.4 Semantic Entities and Relationships ... 91

5.5 Newsworthiness ... 94

5.5.1 Technical perspective of assessing news worthiness ... 96

5.6 Labeling Semantic Relationships ... 98

5.7 News Item Suggestion ... 102

5.8 Experimental Results & Evaluation ... 107

5.8.1 Methodology of Data Collection ... 107

5.8.2 News Stories Examined for Evaluation ... 109

5.9 Summary ... 112

Chapter 6 ... 115

Conclusion and Future Work ... 115

References ... 119

VITA ... 126

LIST OF FIGURES

FIG. 1.1 CORRELATION OF TRENDING TOPICS ACROSS MULTIPLE MEDIA PLATFORMS 8

FIG. 1.2. (A) GRADUAL INCREASE OF VIDEO POPULARITY AFTER VIDEO HAS BEEN EMBEDDED IN EXTERNAL WEBSITE. 10

FIG. 1.2. (B) SUDDEN VIEW-COUNT JUMP. PHENOMENA NOT EXPLAINED BY YOUTUBE INTERNAL KNOWLEDGE ALONE. 10

FIG. 2.1 THE PLATE REPRESENTATION OF LATENT DIRICHLET ALLOCATION........ 24

FIG. 2.2. SEMANTIC WEB REPRESENTED AS GRAPH........ 34

FIG. 3.1. PERCENTAGE OF TWEETS WITH URL VS. WITHOUT URL........ 46

FIG. 3.2. PERCENTAGE OF ASSOCIATION OF USERS WHO ORIGINATED TRENDING TOPICS FOR FIVE DIFFERENT MONTHS. FROM AUGUST 2012, THROUGH DECEMBER 2012. TRENDS WERE EXTRACTED ON THE 15TH OF EVERY MONTH... 46

FIG 3.3. PERCENTAGE OF STORY (METEOR CRUSH) IN MEDIA VS. TIME (15-HOUR STRETCH). DATA PRESENTED IN 3-HOUR INTERVAL, STARTING 9PM, FEB-14, 2013, PACIFIC TIME. 50

FIG. 3.4. PERCENTAGE OF STORY (REEVA STEENKAMP'S DEATH) IN MEDIA VS. TIME (ONE WEEK STRETCH, STARTING FEB 14TH, 2013)........ 50

FIG. 4.1. ARCHITECTURE OF THE PROPOSED CROSS-PLATFORM PROACTIVE CACHING SYSTEM 55

FIG. 4.2. THE KEYWORD SELECTION PROCESS FOR QUERYING ONLINE VIDEOS RELATED TO POPULAR TOPICS........ 60

FIG. 4.3 NETWORK TOPOLOGY USED IN THE EXPERIMENT. THE MAP SHOWS THE LOCATIONS OF BACKBONE NODES AND YOUTUBE SERVERS. 68

FIG. 4.4. VIDEO RELEVANCE TO THE TOPIC (%) VS. # OF FEEDS (IN THOUSANDS) 72

FIG. 4.5 ACCURACY OF PREDICTING VIDEO WITH MOST TRAFFIC (%) VS. # OF FEEDS (IN THOUSANDS) 72

FIG. 4.6. LOCAL CACHE HIT RATIO VS. CACHE CAPACITY (PERCENTAGE OF TOTAL VIDEO SIZE) 76

FIG. 4.7. NORMALIZED AVERAGE DELAY VS. CACHE CAPACITY (PERCENTAGE OF TOTAL VIDEO SIZE) 76

FIG. 4.8. NORMALIZED AVERAGE BANDWIDTH CONSUMED VS. CACHE CAPACITY (PERCENTAGE OF TOTAL VIDEO SIZE) 77

FIG. 4.9. SHARE OF TRAFFIC (%) BETWEEN VIDEOS LISTED AS TRENDING BY YOUTUBE AND ALL OTHER VIDEOS USED IN OUR SIMULATION DURING 30-DAY PERIOD 77

FIG. 5.1. WORKFLOW OF NEWS ITEM SUGGESTION FRAMEWORK FOR THE TRAINING AND RECOMMENDATION PROCESS 88

FIG. 5.2. POPULAR NEWS CATEGORIES IN THE U.S. 92

FIG. 5.3. SEMANTIC RELATION BETWEEN TWO ENTITIES 99

FIG. 5.4. WEIGHTING SEMANTIC RELATIONS 100

FIG. 5.5. EXTRACTING RULES OF ASSOCIATION FROM MAINSTREAM MEDIA 102

FIG. 5.6. NEWS SUGGESTION SYSTEM WORKFLOW 104

FIG. 5.7. OVERLAP (%) BETWEEN SUGGESTED NEWS ITEMS AND TOP RANKING NEWS ITEMS....................111

LIST OF TABLES

TABLE 5.1. CONDITIONS FOR NEWS WORTHINESS.. 95

TABLE 5.2. NEWS VALUES AND SUB-CRITERIA MAPPED TO NAMED ENTITIES, TOPICS AND ANALYSES USING SEMANTIC WEB TECHNOLOGIES 103

“Whether you think you can, or you think you cannot, you are right.”

- Henry Ford

Engaging Mainstream Media for Efficient Content Distribution and Creation

Aleksandre Lobzhanidze

Dr. Wenjun Zeng, Dissertation Supervisor

ABSTRACT

Artificial Intelligence (AI), when machines act intelligently like human, has emerged in many different fields, including journalism. The interaction between journalism, the Internet and social media has been intensely discussed, helping us understand how journalism can help increase our collective intelligence. In this thesis, we study how AI techniques may contribute to effective information distribution and creation, and network resources utilization. By leveraging mainstream media knowledge, crowd opinions (collective intelligence) and smart algorithms for contextual analysis, we explore a number of novel schemes for efficient content distribution and creation.

We first study trend detection and story development process in the media, and discuss why mainstream media is the tool of our choice. The types of information may vary from textual to visual, among which effective video distribution is one of the most challenging issues. Modern Internet faces new challenges with a growing demand on video; therefore our focus first falls on online video. We propose a mainstream media driven trend detection and proactive caching framework that transits the knowledge of detected trends in news to online video sharing portals, to detect emerging popular videos, and pre-cache them at strategically deployed caching nodes. We explore a combination of topic modeling and frequent pattern mining to design a cross-platform video popularity prediction scheme. We further propose a trend-aware and reputation-based video-ranking algorithm to select correct caching candidates among a large array of redundant content for proactive caching by the Internet Service Providers (ISP). Experimental results show that the proposed proactive caching framework can significantly outperform conventional caching methods that are based on the historical popularity.

Lastly, we discuss the design of a framework that empowers association rule mining by linking semantic entities in the mainstream media to facilitate the creation of an automated news item suggestion system for news generation that could operate as a mainstream media outlet, or serve as a guiding tool for human journalists.

CHAPTER 1

Introduction

From the early nineties of the twentieth century interdisciplinary research efforts have been made to develop efficient ways to automatically retrieve information and knowledge from multi-media journalistic content. The main objective is to let users find the relevant information quickly and effectively. Nowadays, major search engines yield millions of links to any request and cannot answer consumer request expressed by simple keywords. The community of researchers involved in the multimedia information retrieval and dissemination research cover topics such as Human Computer Interaction (HCI), Information Theory (IT), Data Mining (DM), Statistics, Pattern Recognition, Psychology, and recently, the Social Sciences. Substantial multi-disciplinary research into information retrieval is done from the perspective of consumer-initiated search for information and knowledge. Our objective is to study the prospects of new phenomenon where content is automatically retrieved and suggested to the specific group of consumers.

When it comes to information search, retrieval, and recommendation we cannot avoid the mentioning of Big Data. It is a popular phrase used to describe a massive volume of both structured and unstructured data that is so large that it is difficult to process with traditional database and software techniques. The major distinguishing characteristics of Big Data are so called “three Vs”: more volume, more variety, and higher rates of velocity. In the earlier ages of the Internet small volume of data was produced and made available through a limited number of channels. Today massive amount of data is regularly being generated and flowing from various sources, through different channels, every minute in today's Internet. Some have described the current digital age as "afloat in Data Ocean." In a broad range of application areas, data is being collected at unprecedented scale. The data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and from cell phone GPS signals. It is the speed and frequency with which data is emitted and transmitted on the one hand, and the rise in the number and variety of sources from which it emanates on the other hand, that jointly constitute the data surge. The amount of available digital data at the global level grew from 150 exabytes (1eb = 1 billion gb) in 2005, to 1200 exabytes in 2010. It is projected to increase by 40% annually in the next few years [79], which is about 40 times the much-debated growth of the world's population. This rate of growth means that stock of digital data is expected to increase 44 times between 2007 and 2020, doubling every 20 months.

Big Data is often discussed in popular media, business, and computer science. There is a broad recognition of the value of data, and products obtained through analyzing it. Popular news media now appreciates the value of Big Data. As an evidence, in 2008 *Wired* magazine opened special section on "The Petabyte Age" by stating "Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grow, so will the opportunity to find answers to fundamental questions." In 2010 *The Economist* started its special report "Data Everywhere" with the phrase "the industrial revolution of data", and then went to note that "the effect is being felt everywhere, from business to science, from government to the arts."

The discussions in the popular media usually do not define big data in quantitative terms. However in computer science the term has a more precise meaning: "Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time" [104].

While the potential benefits of Big Data are real and significant, it also imposes many technical challenges that must be addressed to fully realize this potential. The heterogeneity and incompleteness of the data creates some of the major challenges for data retrieval and dissemination. There are multiple steps to the data analysis pipeline; at each step there is work to be done.

The first step is data acquisition. Some data sources, such as sensor networks, can produce staggering amounts of raw data. Much of this data is of no interest, and it can be filtered and compressed. One of the challenges of big data is to define these filters in such a way that they do not discard useful information. Frequently, the information collected will not be in a format ready for analysis.

The second step is the information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. For example a news report will get reduced to a concrete structure, such as a set of tuples, or even a single class labels, to facilitate analysis.

Furthermore, we cannot assume Big Data is always telling us the truth. We have to deal with erroneous data: some news reports are inaccurate.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understand-able, and then robotically resolvable. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and Big Data computing environments.

In short, there is a multi-step pipeline required to extract value from data. Heterogeneity, incompleteness, scale, timeliness, privacy, and process complexity give rise to challenges at all phases of the pipeline. Furthermore, this pipeline is not a simple linear flow. There are frequent loops back as downstream steps suggest changes to upstream steps. While there are more than enough problems for the research community in retrospect of big data, we lay our focus on two particular problems. In the following paragraphs we present the overview of the problems explored in this thesis, and provide a high-level review of the proposed solutions.

The unprecedented increase of the data size on the Internet can partially be attributed to the blossom of social media platforms, and ubiquitous availability of powerful mobile devices that enable the Internet users to generate vast amounts of data on a daily basis. As we already mentioned, exponentially growing data on the Internet is offering both opportunities with unprecedented insights and information discovery challenges. Information overload is one of the main challenges of the modern Internet. Applying techniques from text mining, machine learning and statistical analysis can help reduce the overload of information.

In today's information-driven society, the ability to efficiently find, critically analyze, and intelligently use reliable information is a major factor. However, over exposure to that critically valuable resources, leads to *information overload* and its detrimental effects. Human beings of the 21st century need information continuously and increasingly. To satisfy the innate need of information, we have equipped ourselves not only with perceptual and cognitive functions, but also developed auxiliary facilities, from books to modern digital media. The success of these artifacts has made us overwhelmed by the very systems that should help us master the world. The irony lies in a fact that we tend to solve the problem of information overload, with more information. The information overload applies not only to the textual content, but also visual data. Various types of media platforms exist on the Internet, some publish content covering only specific topics while some others share in real time and provide crowd-sourcing opinions. Some media provide content in visual form, while others in textual form.

The existence of various media platforms has created opportunity to explore cross-platform design to leverage the knowledge from multiple media platforms. Our goal is to design effective information suggestion system. As we will show in this thesis, one of the characteristics of the modern Internet is the correlation among different platforms. The key to improve prediction-based applications in one domain is to understand why the domain behaves in certain ways. One crucial factor for predicting the developments in a single domain is to realize that different domains do not act independently. In fact the spread of information in one domain often propagates to other domains. Exchanging information across multiple platforms enhances the accuracy of prediction and gives us bigger picture to make more intelligent decisions.

Another observation made in the thesis applies to the correlation between similar events. If we classify news stories reported by mainstream media into certain categories, and pay close attention to the story development process, it will become evident that history often repeats itself. For example some popular events (presidential election, world cup competition, music award ceremony, etc.) often generate very similar emotions among the public. These emotions generate demand on certain type of information. Exploring the common patterns, and correlations between similar events that happened at different times, can help us understand and predict the future developments. The prediction of future popularity can be helpful in many applications. In this thesis we explore the popularity prediction and recommendation from the journalistic perspective. Mining information from mainstream media may help us uncover some hidden insights. In the era of information overload, any system that can accurately predict and suggest the relevant information in a given context would provide tremendous help not only to news content creators, but also news consumers.

In this thesis, we address two major problems: leveraging the cross-domain correlation to help proactive online video caching, and exploring the correlation between similar historical events for automated news item suggestion. We elaborate on these two problems in Section 1.1 and 1.2 respectively.

1.1 Online Video Retrieval and Caching

The rapidly growing popularity and data volume of modern web is attributed to the ease of use of sharing applications. The emergence of large-scale social web communities enabled even the inexperienced users to generate, upload and share the content through the social media platforms. Videos are no longer produced by a few centralized content providers, but by all individual users. In less than six years publishing of User Generated Content (UGC) has significantly reshaped both dissemination and accessibility of information. Social networking has become ubiquitous in the web culture throughout applications such as digital video sharing, micro-blogging, image sharing, and wikis. The boom of social networking and video sharing applications generated immense amounts of data. Accordingly, social media related research has evolved into a vast and variegated field with areas of research ranging from the study of social network recommender systems to the investigation of the socially aware replication for video contents. Several research papers published in the past few years confirm the correlation between textual and visual trends [116][117][91].

The topical trends that appear in social networking websites and/or mainstream media often correspond to the ones in video sharing portals (such as YouTube, Dailymotion, Vimeo, etc). Fig. 1.1 presents News Reference Volume and Google Search Volume Index of “Jeremy Lin” between February 5th and 13th, 2012. We can observe the correlating growth of interest in both media.

Recent studies have shown video streaming continues to dominate the broadband traffic, holding 42% share of all global bandwidth in 2011. Global Mobile Broadband Traffic Report [109] indicates that in the second half of 2011, YouTube, the largest video sharing portal, accounted 57% of all global video streaming traffic, meaning YouTube alone holds 24% share of global bandwidth. YouTube has hit a new record of 100 hours worth of videos uploaded by users per minute [117]. The huge number of user-generated content requires a large amount of storage and network resources.

Caching has been an effective method for reducing the bandwidth usage and the delay between video servers and end users. However challenges of modern web make traditional service paradigms ineffective [116][117]. Conventional caching methods may work well for videos with skewed and stable popularity profiles. [117] reveals a key observation: social videos, unlike regular videos, do not propagate randomly. It is difficult to predict the popularity of the videos and their distribution

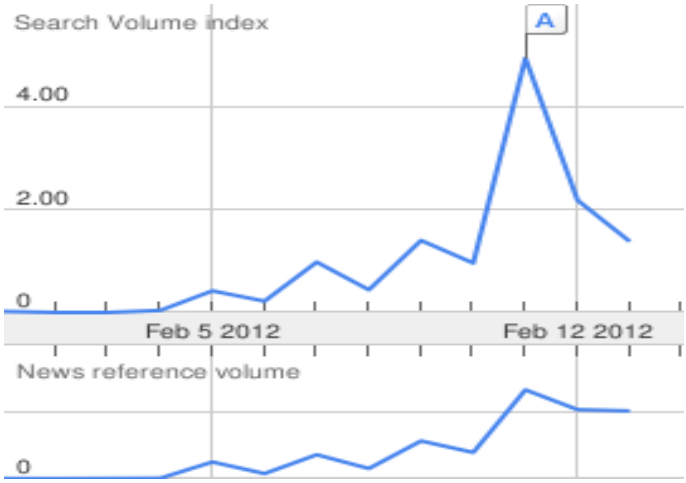


Fig. 1.1 Correlation of trending topics across multiple media platforms

pattern by exploring single platform based information only. The demand for a specific video varies in different geographic regions at different times. Recent research has discovered YouTube videos often exhibit sudden burst in popularity [117][91]. This effect cannot be properly captured by established video popularity prediction schemes alone. Sudden rise of popularity is often attributed to the trends in mainstream and/or social media. For example Fig. 1.2(a) and (b) show YouTube videos with certain associated events A-I. In Fig 1.2(a) event E represents video being embedded into Facebook, after which it gained gradual popularity increase. However YouTube cannot estimate the influence of the website by its internal information alone. Fig. 1.2(b) shows sudden jump in popularity, although there is no associated event with it. Such phenomenon can only be explained by looking at the broader picture of topical trends across multiple domains.

Popularity of the media content on the Internet is unevenly distributed. For example, the popularity of video is judged by its view-count, but not how prominent the topic to which it belongs is. Users browsing trending topics in mainstream media are often influenced to search for related information in different platforms, such as social media, and video sharing portals. In this thesis, we explore the idea of cross-platform information exchange to design a trend aware, proactive video caching system. Our approach is justified by the observation that trends correlate across multiple media platforms. For the purpose of trend detection, we use mainstream media. It is often common for people to browse videos about the interesting topics they read in the news. Mainstream media disseminates information via the largest distribution channels, which represents what the

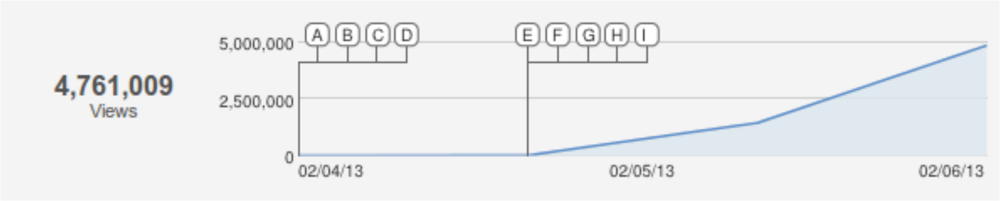


Fig. 1.2. (a) Gradual increase of video popularity after video has been embedded in external website.

http://youtu.be/JRWox-i6aAk

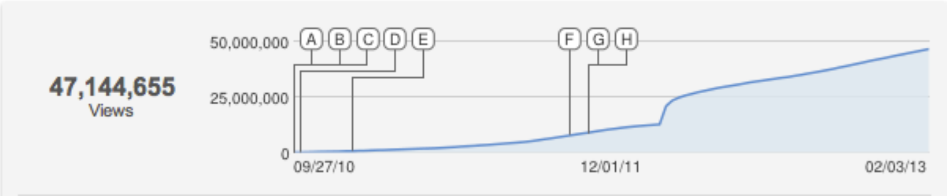


Fig. 1.2. (b) Sudden view-count jump. Phenomena not explained by YouTube internal knowledge alone.

http://youtu.be/FxYw0XPEoKE

Video popularity evolution in YouTube. Some videos exhibit unexplained phenomena of sudden jump of view-count. Existing traditional video popularity prediction systems cannot address the issue of drastic view-count increase, without considering trends in mainstream/social media.

majority of consumers are likely to encounter. News reporters are always interested in reporting most relevant, interesting, and attractive stories to the public. In contrast with social media, mainstream has broader platform of spreading the news, not only web, but also television, radio, newspaper, etc.

We address two major problems, popularity prediction, and caching. Caching itself draws three major questions: 1) What to cache? 2) When to cache? and 3) Where to cache? We propose a framework to measure the influence of trending topics in a certain geographic region, and by leveraging this information identify videos related to the current trending topics to pre-cache at strategically deployed nodes. Our work can bring major benefit to the Internet Service Providers (ISP). Many institutions in charge of providing fast speed Internet have limited resources.

Content caching has been an effective way of reducing network load, and improving end-user experience. Our system provides opportunity for the ISP to evaluate the mainstream media, and more intelligently manage their caching servers. In order to learn trending topics, in our design we leverage several data mining techniques, namely latent Dirichlet allocation (LDA) [71] and the frequent pattern mining algorithm Apriori [65]. By mining news articles it automatically captures the thematic patterns and identifies emerging topics of text streams and their changes over time.

In this thesis we also conduct measurements on real-world data to study the content popularity, and view-count developments in online video sharing portals. We explore important characteristics, including popularity of the topic in mainstream media, to which the video belongs, and geographical and temporal locality. The framework proposed by our research can be universally used by any ISP. It can be applied to different media sources and video sharing portals. Our experimental results show that cross-domain proactive video caching performs significantly better than classical caching algorithms that are based on the historical popularity.

1.2 News Item Suggestion & Automatic Content Creation

In the second part of the thesis we intend to generalize the information retrieval and suggestion process, and consider any type of information that may be interesting to the user. Our aim is to design a framework that can automatically explore trending topics, and suggest to users what may become popular in the near future.

It has been long recognized that human satisfaction with search for knowledge and information in multimedia content involves several dimensions: a mixture of rational as well as emotional dimensions. The consumer search takes place in a certain context and emotional state. The same person based on his/her emotional state may view content of the news articles differently at different times. Background, education and values of the society also affect satisfaction with the news articles.

In the digital age we have seen tremendous transformation of the field of journalism. Until now journalism and the media were synonyms. Journalism was symbolized by the infrastructure for mass communications and vice versa. We consider the definition of journalism by separating it from the media, connecting journalistic principles based on the relation between journalism and its audience, rather than on its relation to the communications medium.

"*Journalism is the production of news and feature stories, bringing public attention to issues that interest the public. Journalism gets its mandate from the audience*." [111]

Journalism must act on behalf of its audience. Grabbing attention from the users is the major driving motivation for the owners of the mainstream media outlets. Journalists' agenda is to research the public opinion, identify the emerging topics, and provide the most interesting information to the public. We could easily think of this process as a problem of designing effective recommendation system, where we take trends as input, and provide a list of topics as output that matters to the public. Presenting such information would grab the attention of news consumers.

The Internet has no dearth of content. The information is provided in various formats. We intend to design a system that is focused not just on videos, but also on general news items. One of the particular problems of the modern Internet is searching information in the haystack of overwhelming amount of news articles and blogs. The big data on the Internet is providing challenges for both: news creators, as well as news consumers. On the one hand it takes enormous amounts of efforts to explore and decide what information should be included in the news articles. On the other hand users are facing the challenge of finding the right information, something that will answer current information needs, something user would like to read, listen, or watch. Search engines help query and retrieve some information, however in many cases user may not even know what to look for. In the case of news, the readers don't even have query keywords; instead they end up browsing news portals in hope to find something interesting. In such cases we would like to design an effective news item suggestion framework that makes recommendations to a user based on the existing trends, and news browsing patterns observed in past.

Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. The key challenge of news websites is to help users find the articles that are interesting to read. Considering

2.4 billion people are now connected to the Internet, the demand on information is growing with rapid pace. With large number of the Internet users, the diversity of the information is also accumulating. Recommending relevant news content to the target groups is a challenging problem. Automatic news item suggestion system could be a very promising direction not only for the researchers in the field of computer science, but also journalistic studies. Journalists spend long hours exploring the trends, and making decisions on what type of information should be included in the news articles. Automatic systems that suggest prominent news stories leave no question on usability. Our challenge is to design such a system with accurate prediction scheme, to reduce the workload for journalists.

Many news recommendation systems have been proposed over the course of last three decades. Many of them focus on personalized news suggestion by exploring the user browsing history. However none, to the best of our knowledge, considered exploring the existing patterns and hidden insights within the news articles. Frequent pattern-mining searches for recurring relationships in a given data set. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis. However applying such methods for news suggestion systems is novel, and the design is not as straightforward as in marketing, or customer shopping scenarios.

In Chapter 5 we present the motivation, news insights, and the framework for automated news item suggestion system. Our work is mainly inspired by the observation of similar events in the mainstream media. Major news production companies often report the news categories that are followed by large number of audience. Typically each news story can be classified as a member of a certain category. Based on our observation, similar news stories (belonging to the same category) tend to generate similar reactions and emotions among the public. Inspired by the observation, we explore the semantic relationships between the semantic entities mentioned in the news articles. We mine the semantic relationships to identify common patterns, and co-occurring items. The new insights, uncovered from the data mining process allow us to design novel news items suggestion system. The details of the proposed system and experimental results are presented in Chapter 5.

1.3 Organization of the Dissertation

We organize the thesis into the following chapters:

Chapter 1 introduces reader to the scope of the research and motivation for our study. We present the problems that users face with the current developments in the Internet. Our focus lies on two particular problems: a) efficient video retrieval and distribution, b) design of news item recommendation and suggestion system.

Chapter 2 reviews background concepts of the data mining techniques and state-of-the-art algorithms used in the field of Information Retrieval. We review works related to video popularity prediction, topic modeling, news propagation in social network, frequent pattern mining, and semantic web. In this chapter we present the tools used in our research, and their limitations. In the following chapters we clearly point out our contributions, and extensions to improve current state of the art.

Chapter 3 studies trend detection and story development process in the media, and discusses why mainstream media is the tool of our choice. We present comparison between social media and video sharing portals from the angle of overlapping trends, and compare the performance against mainstream media and video sharing portals.

In Chapter 4 we present a cross-platform framework that considers input from mainstream media and leverages the knowledge for efficient video caching. Further we unveil our novel keyword selection algorithm and compare its performance against existing methods. We lay out details of reputation system based video ranking, and demonstrate the advantage of proactive caching over reactive methods.

Chapter 5 introduces the concept of a news item suggestion framework. It presents the idea of empowering association rule mining in mainstream media for information retrieval. The concept of incorporating semantic web knowledge into the design of artificial mainstream media outlet is discussed and the details are presented, along with the experimental results.

Finally, in Chapter 6 we conclude the dissertation by summarizing how our work tackles the information retrieval and distribution problems. We suggest some applications of our work and future research in this field.

CHAPTER 2

Background and Related Work

In this chapter we focus on the technical background required prior to diving deeper in the details of the dissertation. We first discuss a number of related works that study online video caching, video popularity prediction, topic modeling, video propagation in online social network, and socially aware video popularity prediction. Later, we present the frequent pattern mining algorithms, and review the concept of semantic web.

2.1 Online Video Caching

For Internet-scale social video services, replicating the videos at different geographic locations is a promising approach to provide good quality of service [102]. Traditional approaches of content caching mainly take historical popularity into account. A popularity distribution among multiple multimedia objects models the static popularity relationship between the content items offered by a multimedia service. It can be used to derive the probability that the content item with a specific popularity index will be requested. Many models have been proposed for modeling the popularity distribution of a multimedia service, including Zipf [14], Zipf–Mandelbrot [101], stretched exponential [33], Zipf with exponential cut-off tail [16], power-law with exponential cut-off tail [17], and log-logistic [7]. Traditional strategies assume that what was most popular in the past will also be most popular in the future. However, the popularity of multimedia content is known to be highly dynamic [26]. Consequently, caching efficiency can be further increased by taking these dynamics into account and actually trying to predict future popularity instead of directly applying historical information, as will be proposed in the dissertation.

2.2 Propagation in Social Network

The year of 2005 was marked by the launch of two (nowadays very popular) video sharing portals, YouTube and Dailymotion. Since then huge amount of user generated content has been uploaded to the Internet. The creation of social video sharing platforms created new opportunities for researchers, and changed the assumptions of traditional content caching algorithms. Exploring the users' social connections can greatly help with effective video fetching [113]. After online social network is widely used to access online content, user relations and influence has become useful information that can reflect how content might distribute. In more recent works, the content distribution is shifted from "centralized" to "edge" based approach. [103] proposes to allocate cloud servers at the network edges to distribute the multimedia contents to the users. [116] is an attempt to design a social video replication scheme based on traces from Weibo re-shares. This approach proposes to explore the geographic location of the Weibo users, and replicate the videos in the nearby local cache servers. [117] is another approach based on the Youku and Weibo traces to predict video popularity. [66] investigates the difficulties of scaling online social networks, and proposes to design a social partition and replication middle-ware where user friends' data can be co-located in the same server. [115] observes that a social network can be used to help predict the content access pattern in a standalone on-demand system.

2.3 Video Popularity Prediction

There have been several efforts towards analyzing the popularity of online content. For example, [16] studied the popularity cycle of YouTube videos, finding that they have, on average, long life times, and the most popular video tends to be the one that has been recently uploaded. [28] characterized the popularity evolution patterns of YouTube videos and studied the impact of different types of referrers on such patterns, whereas [100] analyzed the characteristics of groups of duplicates of the same YouTube video, finding that these duplicates often have different popularities. [85] analyzed the popularity evolution patterns of YouTube videos, identifying four main classes, which they explained in terms of endogenous and exogenous effects. According to [85], the majority of the videos experience no marked peak in popularity: they either attract little attention or experience some popularity fluctuation that can be explained through a simple stochastic process. They referred to these videos as Memory-less. In contrast, the other videos experience bursts in popularity, being further categorized into: (1) Viral videos, which experience precursory word-of-mouth growth resulting from epidemic-like propagation through online social networks; (2) Quality videos, which experience a very sudden burst of popularity (due to some external event, such as being featured on the first page of YouTube); and, (3) Junk videos, which experience a burst of popularity for some reason (e.g., spam, chance, etc), but do not spread through the social network. The last three categories are the major subjects for our study. Behavior of such videos can be predicted by external events, and we design cross-platform information exchange system, to facilitate reducing the network traffic by pre-caching such videos at strategically deployed nodes.

In general, the key for socially aware content caching algorithms is the content access pattern. With respect to understanding how users access contents in the online social network, the related work presented above [116][117][103][66][115] considers user relations and content re-shares independent of its content. However video content can greatly determine the amount of traffic it will generate in the future. It would be interesting to consider the content of the video when predicting its future popularity. In order to address the problem of providing effective content distribution to users, we consider topical trends in our work. We learn trends in the mainstream media and select related videos for relevant geographic regions to better aid the video caching system.

2.4 Topic Modeling

Recently, topic modeling has gained a lot of popularity in analyzing semantic context in textual data [106]. Topic modeling originates from the impression that the construction of any sentence entails a mixture of topics [27]. Each word that the writer chooses to be part of the sentence is drawn from a mixture of topics in his head. Consequently each sentence, composed of these words, will also develop a membership towards some topics and not so much towards others. Thus, we can consider the mixture of topics to be the cause behind the generation of the entire document. Each document is a distribution over topics [12]. Topic modeling aims to uncover this inherent distribution of topics that guide the creation of the document.

A topic is an abstract concept. It is a collection of words, which when grouped together make some semantic sense. Another word for 'showing semantic sense' is to exhibit 'semantic coherence'. There are several popular methods to uncover the underlying topic distribution given a set of documents, such as Latent Dirichlet Allocation [13], Probabilistic Latent Semantic Analysis [12], Hierarchical topic models [98], Latent Semantic Analysis [27] etc. As mentioned earlier, the goal of topic modeling is to generate several clusters of words. Each cluster represents a particular topic. The result of topic modeling is to generate two distributions, namely the topic word distribution $P(w|z)$ and the document-topic distribution $p(z|d)$. Before we dive into LDA, we will first briefly discuss its predecessors, PLSA and LSA and the general vector space model.

The vector space model is a technique of representing documents as vectors of terms, usually the words in the documents. Each dimension corresponds to a separate term and when a term occurs in a document, its value in the vector is non-zero. Then, the cosine of the angle between two vectors indicates the similarity between documents represented by the corresponding vectors.

LSA falls into the category of vectorial semantics, where the features in a natural language sentence is represented by its words. Again, each document can be represented by a vector of words. The goal of LSA is to detect words that are semantically close. Given a collection of documents, each document can be represented as a column and the each word can represent the row. This generates a term-document matrix containing where each cell contains the number of times the word occurred in the document. Following this, a mathematical technique called Singular Value Decomposition is used to reduce the number of columns while preserving the similarity structure among rows [32]. Then, the cosine similarity between two rows represents how similar the two words are. Values close to 1 indicate very similar words while values close to 0 indicate dissimilar words.

The disadvantage of LSA is its inability to detect polysemy. It also assumes that words and document hold a joint Gaussian probability model, however research has shown this distribution is often Poisson [88]. The alternative to this is using a multinomial model, which is the basis of PLSA.

Probabilistic Latent Semantic Analysis is also statistical technique to understand co-occurrence of words in textual data. Unlike LSA which reduces a term-document matrix using linear algebra, PLSA uses a mixture decomposition from latent class models. As with LSA, let us assume there is a co-occurrence (w, d) of words and documents. PLSA models the probability that each co-occurrence was a mixture of conditionally independent multinomial distributions, i.e.

$$P(w,d) = \sum_{c} P(c)P(d|c)P(w|c) = P(d)\sum_{c} P(c|d)P(w|c)$$

where the latent class is c. More popularly, PLSA and LDA is often represented by the plate notation, shown in Fig. 2.1, where M is a set of documents, d is the document index, c is the word's topic drawn from the document's topic

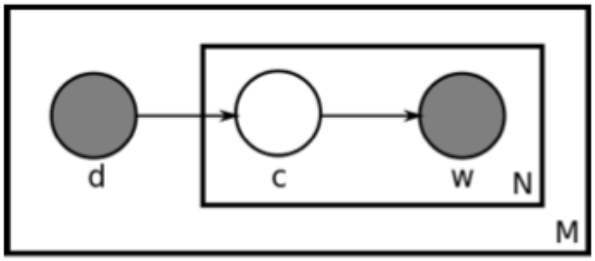


Fig. 2.1 The plate representation of Latent Dirichlet Allocation

distribution $P(c|d)$ and w is the word drawn from the word -topic distribution $P(c|z)$. The shaded circles (w and d) are observable whereas the un-shaded topic (c) is the latent variable. Of course, the number of parameters to learn equals cd + wc. These parameters can be learned using the Expectation Maximization algorithm [78].

In essence, LDA is very close to PLSA in terms of how terms and documents are treated. The major difference is that LDA is completely generative model overlapping a Hierarchical Bayesian model. In other words, PLSA does not maintain a prior probability on the parameters to be learned. But LDA assumes these parameters are itself variables and thus can be treated as hyper parameters with prior probabilities. *This prior is drawn from a Dirichlet distribution*, owing LDA its name.

LDA introduces two prior probabilities alpha and beta, which affect how the per-document topic distribution and the per-document word distribution respectively behaves. The limitations of current topic modeling algorithms include the scalability with streaming or bursty set of documents. There have been several attempts to uncover insights from Twitter by topic modeling [114]. Mining social streams for patterns and trends is a burgeoning area of research [23][73]. Topic modeling over social streams is explored in [9][10]. LDA is a widely used scalable topic-modeling tool for the statistical analysis of document collections [71][69]. Online Stream Latent Dirichlet Allocation (OSLDA), proposed in [91], is a preferred tool for our work. Unlike [71], OSLDA can handle a stream of documents. Although [69] proposes incremental topic modeling, their document stream is neither noisy nor dynamic as tweets, and therefore their technique of updating the Dirichlet prior with time leads to inferior topic modeling results in the social space [91]. The prior word-topic distribution changes too dynamically in social media. Mainstream media, unlike social media, does not introduce too much noise. However our framework has to process stream of incoming news feeds, therefore we employ OSLDA as a topic-modeling tool in both social and mainstream media.

2.5 Frequent patterns, Associations, and Correlations

Frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

Frequent pattern mining searches for recurring relationships in a given data set. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis. However applying such methods for information retrieval systems are somewhat novel, and design is not as straightforward as in marketing, or customer shopping scenarios.

In this section we present a typical example of frequent itemset mining in market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase item A also tend to buy item B at the same time is represented in Association Rule:

A => B [support=2%, confidence=60%] (2.1)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule means that 2% of all the transactions under analysis show that items A and B are purchased together. A confidence of 60% means that 60% of the customers who purchased a item A also bought item B. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Users or domain experts can set such thresholds. Additional analysis can be performed to uncover interesting statistical correlations between associated items.

2.6 Frequent Itemset

Below we describe the process of association rule generation formally.

Let $I = \{I1, I2, \dots, Im\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \phi$.

The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the union of sets A and B, or say, both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A). That is,

support(A=>B) = P(A => B) (2.2)

confidence(A=>B) = P(B|A) (2.3)

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min_conf) are called strong.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set {*computer*, *antivirus software*} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset. Note that the itemset support defined in Equation (2.2) is sometimes referred to as relative support, whereas the occurrence frequency is called the absolute support. If the relative support of an itemset I satisfies a pre-specified minimum support threshold (i.e., the absolute support of I satisfies the corresponding minimum support count threshold), then I is a frequent itemset. The set of frequent k-itemsets is commonly denoted by L_k.

From Equation (2.3) we have

$$\textit{confidence}(A{=>}B) = P(B|A) = \textit{support}(A \cup B) \,/\, \text{support}(A)$$

$$= \textit{support\_count}(A \cup B) \,/\, \textit{support\_count}(A) \qquad (2.4)$$

Equation (2.4) shows that the confidence of rule $A \Rightarrow B$ can be easily derived from the support counts of A and $A \cup B$. That is, once the support counts of A, B, and $A \cup B$ are found, it is straightforward to derive the corresponding association rules $A \Rightarrow B$ and $B \Rightarrow A$ and check whether they are strong. Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

1) Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min sup.

2) Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

Additional interestingness measures can be applied for the discovery of correlation relationships between associated items, as will be discussed in Chapter

5. Because the second step is much less costly than the first, the overall performance of mining association rules is determined by the first step.

2.6.1 The Apriori Algorithm: Finding Frequent Itemset Using Candidate Generation

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 [83] for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see following. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k + 1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database.

To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space. We will first describe this property, and then show an example illustrating its use.

Apriori property: *All nonempty subsets of a frequent itemset must also be frequent*.

The Apriori property is based on the following observation. By definition, if an itemset I does not satisfy the minimum support threshold, *min_sup*, then I is not frequent; that is, $P(I) < min\_sup$. If an item A is added to the itemset I, then the resulting itemset (i.e., $I \cup A$) cannot occur more frequently than I. Therefore, $I \cup A$ is not frequent either; that is, $P(I \cup A) < min\_sup$.

This property belongs to a special category of properties called antimonotone in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well. It is called antimonotone because the property is monotonic in the context of failing a test.

To understand how Apriori property is used in the algorithm, let us look at how L_{k-1} is used to find L_k for $k \geq 2$. A two-step process is followed, consisting of join and prune actions.

1) The join step: To find L_k, a set of candidate k-itemsets is generated by joining L_k-1 with itself. This set of candidates is denoted C_k. Let l_1 and l_2 be itemsets in L_{k-1}. The notation $l_i[j]$ refers to the jth item in l_i (e.g., $l_1[k-2]$ refers to the second to the last item in l_1). By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the $(k-1)$-itemset, li, this means that the items are sorted such that $l_i[1] < l_i[2] < \ldots < l_i[k-1]$. The join, $L_k-1 \mid\times\mid L_k-1$, is performed, where members of L_{k-1} are joinable if their first $(k-2)$ items are in common. That is, members l_1 and l_2 of L_{k-1} are joined if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$. The condition $l_1[\mathrm{k}-1] < l_2\ [k-1]$ simply ensures that no duplicates are generated. The resulting itemset formed by joining l_1 and l_2 is $l_1[1], l_1[2], \ldots, l_1[k-2], l_1[k-1], l_2[k-1]$.

2) The prune step: C_k is a superset of L_k , that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). C_k, however, can be huge, and so this could involve heavy computation. To reduce the size of C_k, the Apriori property is used as follows. Any (k − 1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k − 1)-subset of a candidate k-itemset is not in L_k−1, then the candidate cannot be frequent either and so can be removed from C_k. This subset testing can be done quickly by maintaining a hash tree of all frequent itemsets.

Generating association rules from frequent itemsets:

$$\textbf{Confidence(A}\Rightarrow\textbf{B) = P(B|A) = } \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

For every frequent itemset X, generate all non-empty subsets of X. Since all itemsets considered are frequent, all rules will satisfy *min_sup*.

The discovery of frequent patterns, association, and correlation relationships among huge amounts of data is useful in selective marketing, decision analysis, and business management. A popular area of application is market basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence). Association rule mining consists of first finding frequent itemsets (set of items, such as A and B, satisfying a minimum support threshold, or percentage of the task-relevant tuples), from which strong association rules in the form of $A \Rightarrow B$ are generated. These rules also satisfy a minimum confidence threshold (a pre-specified probability of satisfying B under the condition that A is satisfied). Associations can be further analyzed to uncover correlation rules, which convey statistical correlations between itemsets A and B. In Chapter 4, Section 4.3 we present a new keyword selection algorithm that leverages the Apriori algorithm.

2.7 Semantic Web

The Semantic Web is a collaborative movement led by international standards body the World Wide Web Consortium (W3C) [112]. The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF) [107].

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". The term was coined by Tim Berners-Lee for a web of data that can be processed by machines.

It is important to note that the term Semantic Web is often confused with 'Semantic Networks'. The former is a standards movement, which involves designing web pages in formats that can be easily machine-readable. It uses Resource Description Formats (RDF) as units to describe data. Semantic networks on the other hand, are generic graph that describe concept relations. The latter can be engineered by using data from the semantic web.

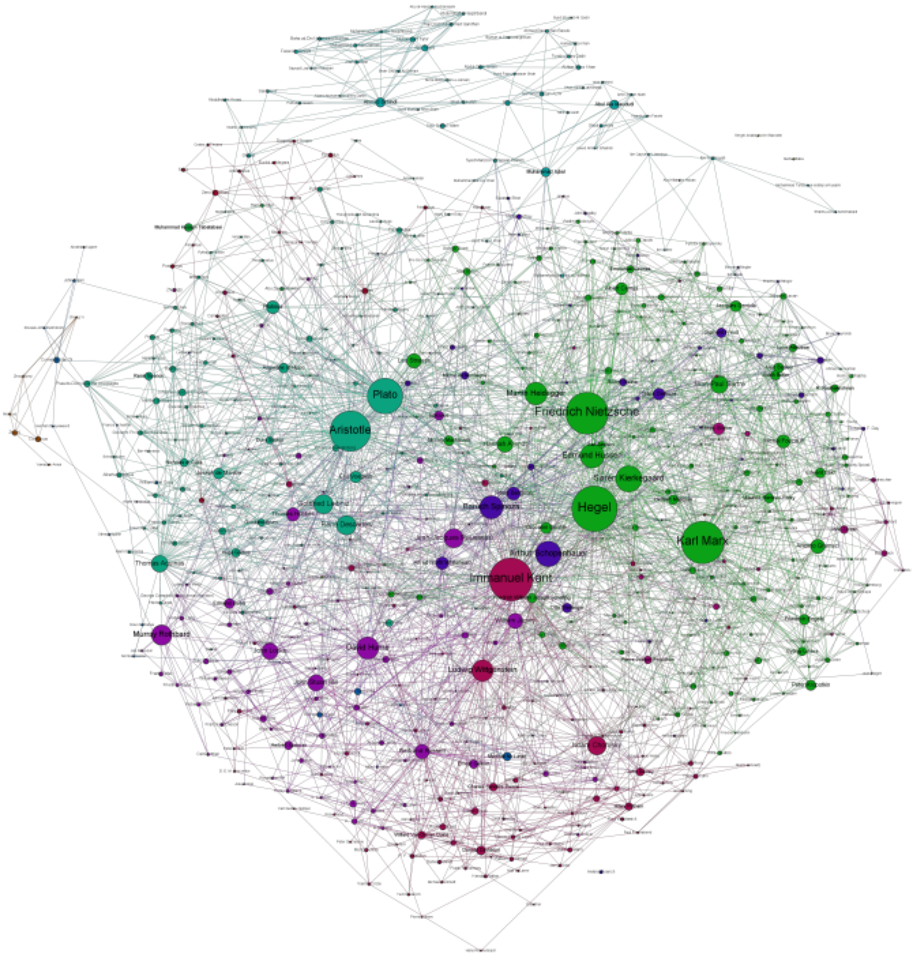


Fig. 2.2. Semantic Web represented as graph

CHAPTER 3

Effective Trend Detection in the Media

In this chapter we present the study that explores the influence of mainstream and social media over Internet users and try to understand which platform provides better input for video traffic prediction systems. In the digital age, human interaction with computers and browsing various news articles and short videos on the Internet is a daily business. Internet reach-out to the public is becoming more and more significant. It is almost undisputed mainstream media greatly influences its audience and shapes their interests, world vision, political, and philosophical beliefs. Recent years have witnessed the blossom of social networks; people are more connected these days than ever. Social network possesses some unique features, for example we can extract public opinion about different controversial issues, occurrence of unexpected events, the degree of interest regarding different political parties, reality shows, movies, and many others. Due to the unprecedented growth of social networking popularity, people often overlook conventional media. In this chapter we would like to explore the opportunities provided by social media, and compare it with the mainstream media. Both types of media have their strengths and weaknesses. We weigh each and conclude which one can be utilized more effectively for the purpose of video popularity prediction. There are a number of reasons to extract trending topics from the media. Social stream mining to make video recommendations based on the trending topics has been one of the active directions in the research community [116][117][91]. Another interesting application is predicting network traffic based on the trending topics. It is common for people to browse videos about interesting topics they read in the news. Such a trend creates a wave of data traffic in the network. Being prepared for such waves could greatly help with providing quality of service.

The rest of this chapter is organized as follows. In Section 3.1 we present the usefulness of trending topics, and how we can benefit from such information. In Section 3.2 we compare social and mainstream media by exploring their strengths, and weaknesses. Methodology of data collection and evaluation is presented in Section 3.3. Section 3.4 presents how the trends are detected in both media platforms and evaluates their usefulness for our target applications. Finally, Section 3.5 summarizes the chapter.

3.1 Trends in the Media

It is crucial to have first-hand access to trending topics, if we would like to understand which videos would become popular in the near future. Several prior works [116][91] have claimed social media hits the web faster than mainstream media. The papers point to examples such as shootings in Aurora, or death of Osama Bin Laden. Several tweets have appeared in social micro-blogging websites prior to mainstream media reports regarding the two events. Despite its ubiquity and ease of access, social media does not influence as big audience as mainstream media. We have looked into another work [90] where researchers found that the mainstream media, including organizations like New York Times, CNN, and BBC act as “feeders” for news topic, helping to amplify and in turn make something into a trend on the social network. Findings from [90] state that mainstream media drives a disproportionate number of Twitter trending topics. From a sample of over 16 million tweets, they identified 22 users who were the source of the most re-tweets while a topic is trending. Of those, 72% were Twitter accounts from mainstream media. Another interesting report that caught our attention was statement made in [50] that claimed featured videos listed by YouTube had very little correlation with trends reported by Twitter, meaning that users do not always tweet what they watch on YouTube. [3] explores the speed of trend detection in mainstream and social media. It also explores the origins of the trends. The experimental results presented in the paper show many trends that last longer than average in Twitter are initiated from the accounts that have association with some mainstream media outlets. Given all the reports, we decided to conduct an experimental study and evaluate which media platform (mainstream or social) provides better input for trends prediction in video sharing portals.

The major question presented before us is how much correlation exists between visual and textual trends. In order to answer the question we explore video sharing portal, social, and mainstream media. YouTube, the world’s largest video sharing portal is chosen as a representative of visual media. Twitter is the most popular social blogging platform; therefore we explore trends in Twitter. Mainstream media in our experiment is represented by several sources;

3.2 Social Media vs. Mainstream Media

The content of this chapter is mainly based on a number of experiments we conducted that allowed us to draw several interesting conclusions. As mentioned in Section 3.1 it is important to have first hand access to the trending topics, if we would like to understand which videos would become popular in the near future. Some research papers (such as [91]) claim social media is the fastest tool to detect the trends, while another study [90] presents report that shows majority of the trends in social media are attributed to the mainstream media. In order to address the problem we conduct our own experiments. We explore the virtues of each and evaluate which one would provide better input for the purpose of online video popularity prediction and caching.

Social media has many useful features, but it also imposes a number of difficulties that need to be dealt with. It is challenging to extract relevant information from social networking websites (Twitter, Facebook, etc.) and correlate across different domains. The data from the social stream tends to be very noisy (unstructured, grammatically incorrect, misspelled) and significant efforts need to be made for data polishing. Another characteristic of social media is the short length of the messages, whether it's a tweet, or status update, they often tend to be short. In fact, Twitter restricts users to 140 characters only. Given documents are short and noisy, large amounts of data are required to hatch out something meaningful from the social stream. Data arriving in high volume creates scalability issues, which additionally need to be addressed.

Mainstream media- Major news companies post articles related to the popular topics/events daily. In the age of social network, the news can spread very fast through the web and reach the farthest corners of the globe. Major credible news companies (CNN, BBC, AP, etc.) have always carried the slogan “give people what they want”. Mainstream media is interested in reporting stories that attract readers; therefore, it is reasonable to assume most news articles will reflect what’s popular in the society. Unlike social data, news articles are structured documents, written with high grammatical accuracy. We also have to keep in mind some newspapers might only follow a limited number of topics (showing emphasis only in sports, showbiz, technology, etc.), and present the story from a bias angle..

As we have drawn the differences and identified some strengths and weaknesses for both types of media, we have to decide which one provides better input for video popularity prediction schemes. Social media, regardless of its noisy nature, can still convey interesting information. We would like to know how better, or worse it performs against mainstream media. Intuitively, social media has the ability of detecting trends. However in the mainstream media there are a number of trained professional journalists who specialize in identifying and reporting top news. We can certainly take advantage of this characteristic. *The purpose of our experiment is to understand how many topical trends identified in Mainstream and Social media also become trends in YouTube*.

3.3 Methodology of Data Collection

Comparison of YouTube, Twitter and Mainstream media trends is not straightforward; therefore, we have collected data in a similar fashion and used some normalization methods to keep things fairly balanced.

Data from mainstream media - The news websites are collected based on their popularity. [108] provides global web traffic analysis, where we have identified some of the most popular news websites, and subscribed to the corresponding RSS feeds [59]. Unlike Twitter data, there is no need for data polishing. We pile collected feeds into a text file and feed it to a topic modeler [91]. According to our collected data, websites such as CNN, BBC on average post 171 RSS feeds per day. We have collected all available RSS feeds from the selected news sources during the month of September 2013.

Data from social media (Twitter) - User posts in Twitter are publically available, thus it was a natural choice to use this system as a data source. Twitter also provides API services, however it comes with many limitations. Since we are also exploring trends in popular mainstream media outlets, we identify some popular news related twitter accounts [47] and extract a list of all their followers. Once we acquired the list of the followers (and removed redundancies) we extracted their tweets. For the purpose of tweet extraction we have not used API services, instead implemented a crawler to process the HTML source of each user's Twitter page, and parse the source code to extract only tweets. The collected tweets were posted between 9/1/2013 and 9/30/2013. During this one-month period from each user up to 20 tweets could be collected, some have no tweet at all. Note that it is reasonable to believe that our process of sampling tweets gives social media some bias in the comparison (between mainstream media and social media) of trend overlapping with YouTube, as it will be demonstrated in Section 4.5 that a majority of YouTube popular videos are news related.

Data from Video Sharing Portal (YouTube) –Trends in YouTube are represented as featured videos. YouTube provides a list of popular videos; therefore we are not mining any data to identify emerging trends. Instead we observe the YouTube popular videos page [56], and record ID and metadata of each video. Popular videos can be retrieved based on the category. Currently YouTube has 9 categories: Music, Sports, Gaming, Education, Movies, TV Shows, News, Live, and Spotlight. Our focus lays on 'Music', 'Sports', and 'News'. Category 'Live' streams different channels with live video, and is beyond the focus of our work. 'TV Shows' and 'Movies' are payment-based services, and may not correlate well with dynamic trends in other media. 'Gaming' and 'Education' are lacking dynamics and may not benefit considerably from proactive caching. Categories excluded in our experiment account only for 10% of the traffic according to [55]. 'Spotlight' is a collection of trending videos from different categories, and we should be able to capture most of those under the 'Music', 'Sports', and 'News' categories.

We have implemented a crawler that updates itself every 15 minutes, and extracts HTML source code of the three target categories from YouTube website. Video IDs are found in the source code and recorded along with Title, Description, Tags, and timestamp. We remove redundant content (e.g. when featured videos are unchanged over a long period of time) and use the textual metadata of popular YouTube videos to find overlap with social and mainstream media trends. YouTube trends data is acquired in the same time range (9/1/2013 – 9/30/2013) as is done for mainstream media and Twitter.

3.4 Trend Identification in the Media

To identify popular topics in the social and mainstream media, we use OSLDA. The goal of topic modeling is to automatically discover the topics from a collection of documents. We feed OSLDA with RSS and Tweets to detect trending topics in the mainstream and social media respectively. News sources are broadly covering different topics/events throughout the world. Similarly the selected Twitter users are spread all over the world, discussing a large array of topics. We apply OSLDA to Twitter stream and RSS feeds in order to identify popular trends. Our Tweeter dataset consists of 10 million tweets, and we have acquired approximately 75,000 RSS feeds from the selected sources [59]. The number of topics in OSLDA is a fixed parameter. We did not know what number would be most optimal, therefore experimented with different numbers to derive the most appropriate one for the trends detection.

3.4.1 Detection and Analysis of Common Trends

The common trends are defined as follows. Popular topics in YouTube, in the context of our system, are defined as the textual information provided by the user when posting his/her content. We represent each topic as a single string and make no distinction between the various fields in a given post. For example, YouTube posts contain three text fields (title, tags, and description), while Twitter posts only contain one text field. At any time, for each category, up to 15 videos can be extracted from the popular YouTube videos website [56]. During a 24-hour time period at most 1000 videos may be identified as trending in YouTube. The time for emerged trends may vary in Twitter, YouTube, and Mainstream media. The textual trends typically are more diverse and much more dynamic, and precede visual trends in time. These trends are examined to find the overlap with YouTube in the next 24-hour time interval. For example if "Military Invasion" emerges as a trending term in Twitter, or Mainstream Media, and the same words appear in the title, description, or tags of YouTube trending videos within a 24-hour time interval, we consider this as a common trend. *Our aim is to identify the percentage of YouTube trends that can also be detected by Twitter and/or Mainstream Media*.

Twitter in its nature is very noisy and terms sometimes may be represented in the form of acronyms, abbreviations, or concatenated words with hash-tag sign. If we apply direct string comparison some common trends may be missed (e.g. #EFC and Everton Football Club). We used Natural Language Processing (NLP) [45] tool to map meaningless Twitter words and YouTube tags into useful terms (when possible) that can be found as entities in DBpedia [37]. This method allows us to better link trending terms between YouTube and Twitter

3.4.2 Overlap of Trends

In this subsection we present experimental results. By detecting popular topics in the collection of Tweets and RSS feeds, we identify how many YouTube trends have also been identified in Mainstream and Social media. The number of topics parameter in our experiment was initially set to 50. Then we increased it to 100 and mined the same data through OSLDA. We kept increasing the number to up to 500. The results suggest 200 is the closest to the optimal for Twitter. Having a larger number of topics will not influence the outcome. For mainstream media 100 is still optimal.

Once we established the optimal selection for the number of topics, we observed common trends between YouTube and Twitter, and YouTube and Mainstream media. Only 10% of YouTube trending videos in the news section could be detected by Twitter, while sports and music have even worse results, 6 and 4 percent respectively. According to Pear Analytics, the majority of text in Twitter (77 percent) accounts for pointless babble and conversational tweets [44]. On the other hand, Mainstream media excludes the noise, and refers mostly to some real world events. From our observation it identified up to 59% of the news related trends that were listed among popular YouTube videos. The sports section had slightly worse results, catching 52% of the common trends with YouTube, while the music section (the least dynamic from our observations) identified 62% common trends. From the experiments conducted, we found out that the news section is the most dynamic; popular videos are replaced with new popular videos under this category with higher pace than in the other two sections (sports and music). Thus proactive caching may be exploited most effectively with news related videos.

We have also checked the trend overlap between Twitter and Mainstream media. From the collection of Tweets and RSS feeds that we acquired, OSLDA has not detected a trend in Twitter that also became a trend in YouTube and was not among the trends in the Mainstream media. This suggests that collaborative trend detection with both mainstream and social media may not help too much in predicting YouTube trends. Based on the experiments conducted, we conclude that mainstream media might be better input to the video popularity prediction system.

We have also checked the trend overlap between Twitter and Mainstream media. From the collection of Tweets and RSS feeds that we acquired, OSLDA has not detected a trend in Twitter that also became a trend in YouTube and was not among the trends in the Mainstream media. This suggests that collaborative trend detection with both mainstream and social media may not help too much in predicting YouTube trends. Based on the experiments conducted, we conclude that mainstream media might be better input to the video popularity prediction system

3.4.3 Trend Origins

We also explore how big the influence of mainstream media is over the social media. We have defined three types of events: *scheduled event* (e.g. sport competition, political debate, concert, etc.), *breaking news* (unexpected events), and news *first announced through social media* (e.g. Rafael Nadal tweeting he is not fit for the US open). The third type of events mostly comes from the well-established celebrities. We have looked into the accounts of the most followed users and manually extracted events that became trending topics later.

The experiment presented in Fig. 3.1 attempts to study the influence of mainstream media over social media. We select the above mentioned 3 types of events, with 20 different stories for each type. The tweets are extracted within 7 hours after the event occurs. We check how many tweets contain reference URL links to some news websites and plot the results. Due to the large amount of Twitter data, we have only extracted 1500 random tweets for each story, and identified how many of them contained URL. The findings suggest the majority of the tweets are referring to the news sources.

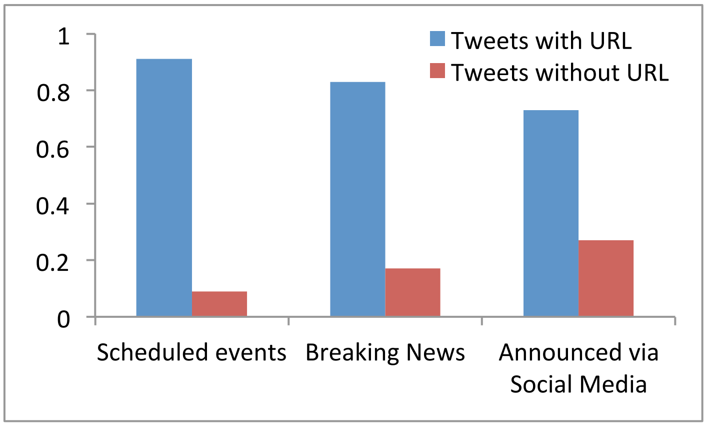


Fig. 3.1. Percentage of tweets with URL vs. without URL

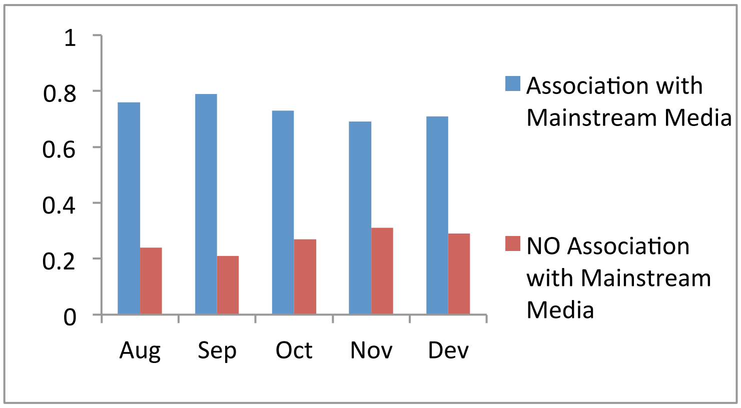


Fig. 3.2. Percentage of association of users who originated trending topics for five different months. From August 2012, through December 2012. Trends were extracted on the 15th of every month.

Our second experiment attempts to study the origins of trends. Some twitter accounts are associated with mainstream media. Our interest lies in identifying how many Twitter users belonging to mainstream media generate trending topics. Analysis on Twitter [90] has revealed that the majority of trends in Twitter will not outlive 20-40 minute popularity, and 34% of the topics do not maintain continuous occurrence in the trending list and reappear multiple times. We would like to consider topics that outlive at least 60-minute popularity. We experimented with different time intervals and the justification of our time interval selection is attributed to the fact that the trends that spend less than an hour in trending list are less likely to have impact on video sharing portals.

We collected list of trending topics from August 2012 through December 2012. Once we have a popular topic, through the chain of re-tweets we track down the originator of the tweet. We store these users in the database and later manually classify the ones that belong to mainstream media brands.

Fig. 3.2 presents the plotted results of association of Twitter users who originated trending topics. The percentages of Twitter users belonging to mainstream media outlets are shown in blue columns. Users that we could not associate with any mainstream media brand are in red columns. As we can observe mainstream media plays the role in generating the majority of trends, serving as a feeder to social media. The percentage of Twitter users associated with mainstream media varies between 69% and 76%. As our experimental results show, for information dissemination, the number of followers is not as critical.

3.4.4 Story Development Process in Media

In this section, we present the story development process in the mainstream and social media. We collected data over 3 consecutive days, i.e., February 13th, 14th, and 15th, 2013. By mining the data through the topic modeler LDA [71] we identify emerging popular topics. During the 3-day period we observed two events that attracted the attention of large audience: meteor crush in Chelyabinsk, Russia and the shooting scandal involving Paralympics athlete Oscar Pistorius [48].

We have been collecting sample tweets and news articles from both media outlets. The meteor crash was first observed around 18:30 GMT. Within 2 hours the topic had gained wide popularity, in about 15 hours videos depicting meteorite crash in Russian city Chelyabinsk had over millions of views. We believe reports in media had greatly boosted the view-count increase; however our task is to identify the media platform that can detect trend quickly, and efficiently. Fig. 3.3 presents the percentage of stories about meteor crush in mainstream and social media. From the pool of 30-million twitter users, we extract their tweets with 30-minute interval. Tweets collected over the 15-hour period had been passed to the detector, which identifies tweets that have the presence of words, or hash-tags related to the meteor crush. Similarly with mainstream media, we parse the text to detect words in the news articles related to the event. According to Fig. 3.3 it seems that at the beginning of the 15-hour period, at least half of the news sources had reported the event, while in social media nearly 10% of users have mentioned the meteor crash.

As time progresses the mainstream media outlets are quick enough to report the story, while social media attracts more users with slower pace.

Another trending news in the media appeared on Feb. 14th, 2013. Scandal surrounding Paralympics athlete Oscar Pistorius broke out as he was charged with the death of Reeva Steenkamp. The story dominated media for several days. Videos related to Pistorius and Steenkamp showed sudden jump of view-count. Here we present the story share in the media with one-week stretch.

We evaluate the percentage of story in the media with 12-hour interval, starting Feb. 14th 2013. According to Fig. 3.4 mainstream media was the first to report the story. If we were to identify this event as quickly as possible, mainstream media would contribute better than social media.

In order to provide statistical knowledge of the story development process in social and mainstream media, we take multiple trending topics and compare the trend detection time, to provide an average performance of mainstream and social media. Again Twitter is used to represent the social media, and various news sources are selected for mainstream media, to cover global and local scope stories around the globe.

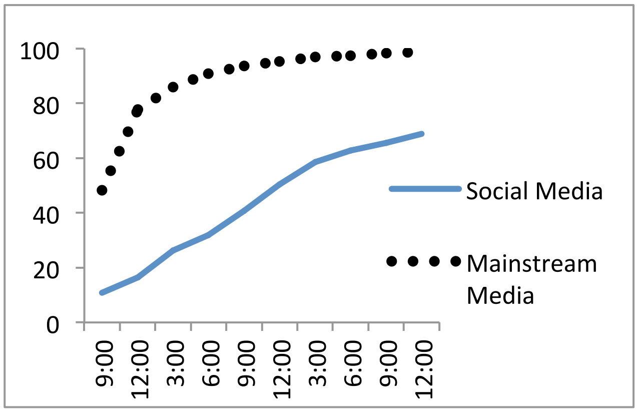


Fig 3.3. Percentage of story (meteor crush) in media vs. Time (15-hour stretch). Data presented in 3-hour interval, starting 9PM, Feb-14, 2013, Pacific Time.

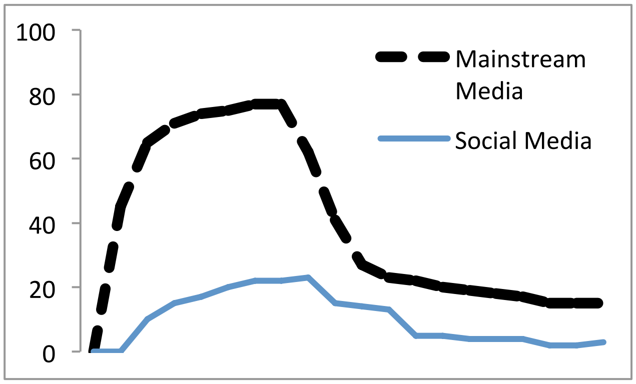


Fig. 3.4. Percentage of story (Reeva Steenkamp's death) in media vs. Time (One week stretch, Starting Feb 14th, 2013)

Trends such as #SpringBreak, #MentionSomebody, etc. are excluded from the comparison process. We identified the overlap between emerging topics (common trends in both media) and tested which media source detects story as trending quicker. We mine RSS feeds and tweets through LDA and define the top 20% as trending (as described in Chapter 4). On the course of 3 days, we identified 137 common topics, related to some news. We focus on the story development process in the media for approximately the next 24 hours after it has been posted. It was found that 79% of the time mainstream media outperformed the social media by detecting the trend quicker. In some cases social media detected a story as trending quicker. This can be explained by the lack of interest from mainstream media to certain stories, or that the news is re-shared/re-tweeted so many times by Twitter users that it gains coverage in the mainstream media.

3.5 Summary

In this chapter, we present experimental studies based on the real data that show the story development process in the mainstream, social media, and video sharing portals. To the best of our knowledge, there have been no thorough studies, or experimental results on how to make a choice between these two platforms (mainstream and social) when it comes to cross-domain application systems design that learn trending topics from textual media to make predictions for video sharing portals. Social networks have significantly reshaped the news consumption among web users by speeding up the news spread. Even though social network has gained unprecedented popularity, it is still reasonable to consider the mainstream media as the primary source to make predictions for video popularity. Data from social networks tends to be noisy, and comes in large volume. Mainstream media is mostly structured. On top of that, journalists are trained to learn, explore, and present information in a laconic way. By using mainstream media, we can remove data filtering, and use a smaller amount of data for trend detection.

Based on the experiments, we have learned mainstream media might be better input to the video popularity prediction scheme. We can exploit the structured, well formatted, and grammatically correct input from mainstream media, and leverage the information for building common topic space between mainstream media trends, and YouTube videos. The next chapter explores designing an application for proactive online video caching system that will learn trending topics from mainstream media and make predictions of video popularity for the social video portals, to replicate the popular content into relevant geographic regions.

CHAPTER 4

Trend-aware Video Popularity Prediction and Proactive Caching

In the recent years, popularity of social media and online video sharing services has grown at an unprecedentedly fast pace. Massive amount of information is being generated and uploaded to the Internet every day. Modern Internet faces new challenges with a growing demand on video. Caching content has been an effective method for improving quality of service for online video. As presented earlier in Chapter 3, correlation between mainstream media and video-sharing portals exists. In this chapter we propose a mainstream media driven trend detection and caching framework that transits the trend knowledge to online video sharing portals, to detect emerging popular videos, and pre-cache them at strategically deployed caching nodes. In particular, we employ a combination of topic modeling tools and data mining techniques, to design a cross-platform video popularity prediction scheme. We further propose a trend-aware, and reputation-based video-ranking algorithm to select correct caching candidates among a large array of redundant content for the Internet Service Providers (ISP). Experimental results in Section 4.5 show that the proposed proactive caching framework can significantly outperform conventional caching methods that are based on historical popularity.

The remainder of this chapter is organized as follows. In Section 4.1 we present high-level overview of the proposed proactive caching system. In Section 4.2 we discuss the trend detection process. Section 4.3 goes into the details of automatic query keyword selection. In Section 4.4 we present the caching video selection strategy. Section 4.5 highlights the effectiveness of the proposed system by demonstrating experimental results. Finally, in Section 4.6 we summarize the contributions presented in this chapter.

4.1 Overview of the Proposed Proactive Caching System

Our work tackles two important problems, video popularity prediction and caching. The main contribution of our work is the design of an automated system that predicts which videos will generate large amounts of traffic, and then pre-cache those videos for better user experience. We employ the idea of media cross-correlation to design a framework that learns trends in mainstream media and predicts popularity of related videos in video sharing portals. The key component of our work is to link popular topics to videos that carry topical relevance and are likely to have a large view-count, and then pre-cache those videos at strategically deployed nodes. Fig. 4.1 presents an overview of our framework.

Typically ISP possesses the information about local traffic. Based on these traces, we can identify popular news media sources among the Internet users. We mine the articles from the selected news sources to evaluate which topics are most popular. [116][117][91] confirm the correlation between the textual and the visual media. We try to take advantage of this correlation and design a framework for proactive caching. As shown in Fig. 4.1, ISP builds the list of news sources relevant to its users. To mine mainstream media, we use Really Simple Syndication (RSS) feeds. The majority of news websites support RSS. Thus, documents are formatted XML files that contain a short summery of the article, publication date, title and the link to the original article. We examine mainstream media through a topic modeler to build a list of popular topics. Each topic consists of a set of highly popular topical words. We then explore the combination of the OSLDA [91] and frequent patter mining algorithm [65] to derive accurate query keywords associated with the

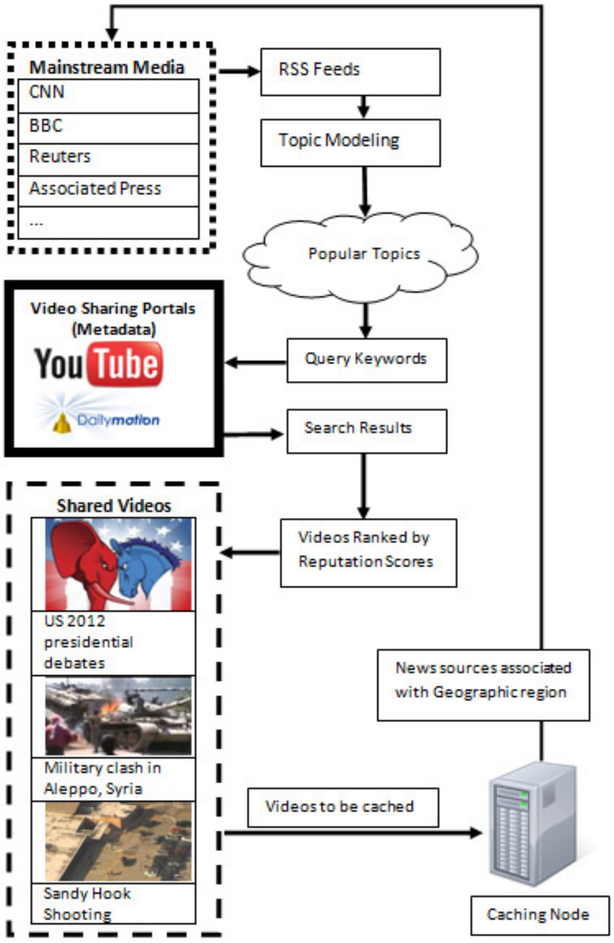


Fig. 4.1. Architecture of the proposed cross-platform proactive caching system

popular topics, which are then submitted to the video sharing portals to retrieve relevant online videos.

The emergence of large-scale social web communities has enabled users to share online vast amounts of multimedia content. The leading social video sharing platforms reveal a large amount of redundancy, in the form of videos with overlapping or duplicated content. The redundancy applies to metadata as well, such as video titles, descriptions, and tag words. The task for our proactive video caching system is to identify, from a set of trend-based candidate videos retrieved from the video portal as described above, the ones that will attract most traffic. To that end, we introduce a reputation score system, where each individual channel is ranked based on its historical performance. A channel on YouTube is the home page for an account. It shows the account name, the account type, the public videos uploaded to the channel, and user information. The decision for selecting videos for caching is based on the combined knowledge of the channel popularity (to which the candidate video is uploaded), the channel's reputation, and the request rate of the video.

The entire framework of proactive video caching system can be broken down into two parts. The first part tries to identify proper keywords that will be used to query video portals. Sections 4.2 and 4.3 provide the details. Due to large amount of redundant content in the social video sharing platforms, as discussed in Section 4.4, we need additional steps for caching candidate selection. In the second part, we unveil our reputation score system for video channels and lay out further details of the caching candidate selection process.

4.2 Popular Trend Detection

To identify popular topics in the news, we use OSLDA, which extends LDA to handle streaming data. LDA is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if the observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word creation is attributable to one of the document's topics. In generative probabilistic modeling, data is treated as arising from a generative process that includes hidden variables. The generative process defines a joint distribution over the observed and hidden random variables. The observed variables are words in the document, and the hidden variables are the topic structure. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, i.e., the conditional distribution of the hidden variables given the document.

In order to describe the original LDA more formally, let's define some notations. The topics are denoted as $\beta_{1:K}$ where each topic β_k is a distribution over the vocabulary. The topic proportion for the d_{th} document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignment for the d_{th} document are z_d where $z_{d,n}$ is the topic assignment for the n_{th} word in document d. D is the total number of documents and N is the number of words in a document.

Finally, the observed words for document *d* are w_d, where $w_{d,n}$ is the n_{th} word in document *d*, which is an element from the fixed vocabulary. Then we have.

$$\boldsymbol{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}$$
$$\boldsymbol{= \prod_{i=1}^{K} p(\beta i) \prod_{d=1}^{D} p(\theta d)\, (\prod\nolimits_{n=1}^{N} p(z_{d,n}|\theta_d) \mid p(w_{d,n} \mid \beta_{1:K}, z_{d,n}))}$$

The formula above represents the generative process for the LDA joint distribution of the hidden and observed variables. The distribution specifies a number of dependencies. The topic assignment $z_{d,n}$ depends on the per-document topic proportions θ_d. The observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and all of the topics $\beta_{1:K}$. Our goal here is to have LDA compute the popularity of each topic given the documents. Multiple topics can be present in a single document. We assume the document is a member of the topic that has the highest popularity in this particular document. Once we have the list of popular topics, we can take the next step, linking popular topics from mainstream media to videos in the online video sharing portals.

4.3 Video Query Keyword Extraction

The OSLDA implementation assumes topics are distributions over a fixed vocabulary. Limited words from the dictionary do not allow us to have a rich textual description of a topic. The vocabulary typically consists of well-defined words (no

First-Last names or specific words). When linking topics with videos, it is important to have very precise words for querying videos. For example in March 2012, a song named “Blue Jeans” performed by singer Lana Del Rey was among the most popular topics. OSLDA can identify words such as “Blue” and “Jeans” among the most popular words; however, “Lana Del Rey” is not listed in the dictionary. This could potentially create significant problems while linking topics with videos. The word “Blue Jeans” is a very broad term and could apply to the trouser cloths, rather than a song. [92] attempts to address the above problem by identifying emerging terms through the use of dictionary learning. Another obstacle of the statistical models (such as LDA [71], OSLDA [91]) is the presence of poor quality topics that mix unrelated or loosely related concepts. Lack of semantic coherence may group irrelevant words under the same topic that makes video query inefficient. While composing articles, journalists often employ auxiliary words to describe the story in details. Statistically, the total share of such words may overwhelm the important (event specific) keywords in the article that make query more effective and precise. Although broad terms still compose semantically coherent topic, they leave ambiguity. In order to address the above problems, we propose to combine Online Stream LDA with frequent pattern mining. Fig. 4.2 shows the entire process for selecting query keywords.

Earlier in this section we describe how topic assignment ($z_{d,n}$) is computed in OSLDA. We select the documents assigned to the most popular topic, and extract only the titles of these articles. Our move is justified by the observation that in the domain of journalism, the most interesting and crucial information is included in the title [111]. We mine article titles and identify the most frequent itemsets. Note that mining the entire text of RSS feeds would significantly reduce the chances of most critical terms to be identified as part of frequent itemset. The text presents the story in details employing auxiliary words that account for a large portion of the text. Thus we only use the title of the RSS feeds to derive the most important keywords to make video query more effective.

Among many frequent pattern-mining algorithms, we chose Apriori [65]. It is an iterative approach known as level-wise search, where k-itemset is used to explore (k+1)-itemset. The iteration of generating itemset is terminated once a minimum support threshold is reached. Apriori is a perfect fit for the purpose of finding frequent words in the title. For example, if we refer back to the "Lana Del Rey" example, many of the news feeds were posted with the title "Lana Del Rey leads charts with Blue Jeans". OSLDA would identify words "Blue" and "Jeans" as highly popular, however not "Lana Del Rey." Please note, the dictionary may contain word "Del Rey" (which translates as "The King" from Spanish) and the term may lead to videos related to "Copa Del Rey", which is a very famous soccer competition in Spain. Apriori can observe "Lana Del Rey," and "Blue Jeans" as frequently co-

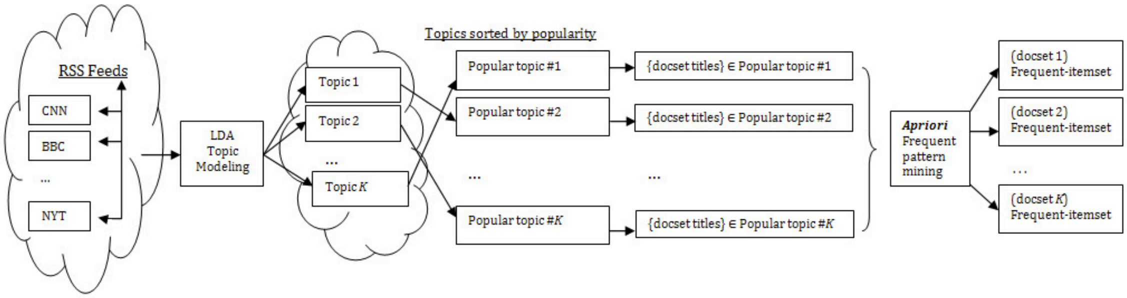

RSS Feeds
CNN
BBC
...
NYT
LDA Topic Modeling
Topic 1
Topic 2
...
Topic K
Topics sorted by popularity
Popular topic #1
Popular topic #2
...
Popular topic #K
{docset titles} ∈ Popular topic #1
{docset titles} ∈ Popular topic #2
...
{docset titles} ∈ Popular topic #K
Apriori Frequent pattern mining
(docset 1) Frequent-itemset
(docset 2) Frequent-itemset
...
(docset K) Frequent-itemset

Fig. 4.2. The keyword selection process for querying online videos related to popular topics.

occurring items. Querying videos in websites such as YouTube with words “Del Rey” or “Blue Jeans” can return results that are completely irrelevant to the song. However, “Lana Del Rey Blue Jeans” produces more relevant results. Thus, selection of frequent itemset from the titles of news articles that belong to a single topic identifies more precise keywords than OSLDA alone.

The combination of OSLDA and frequent pattern (FP) mining in the current form provides unique solution for extracting query keywords. One may wonder if FP alone would produce acceptable results. In the large collection of RSS feeds we cannot derive accurate keywords for single event only. The share of each story is diluted among all other stories mentioned in the RSS feeds. OSLDA helps us collect feeds only related to a single topic, thus FP can find frequent itemset of words belonging only to a single topic.

Our goal is to design real-time, scalable system. The complexity of the proposed proactive caching system depends on the efficiency of OSLDA and Apriori. OSLDA was designed for real-time topic learning from social streams. Unlike our system, social stream introduces much larger amounts of noisy data. Major factor that affects the performance of Apriori is the size of the transactional database. In our case it is limited to the documents related to the trending topics only. Moreover, RSS feed titles only are considered for mining to detect the correlations and associations of keywords. Related issues have been studied and are presented in [24][65].

4.4 Selecting Video for Proactive Caching

Proper keyword selection is an important part of our proactive caching scheme; however, it alone is not sufficient. In addition to abundant content, the social video sharing websites (such as YouTube) provide a rich set of videos associated to the same metadata (title, description, tags). In order to design an efficient proactive caching scheme, we need to identify the ones among a large number of such redundant videos that will generate the most traffic.

4.4.1 Cache Candidate Selection

We use YouTube as an example to explain our caching candidate selection method. Once keywords are selected for query, we retrieve search results from YouTube (20 video links per page). Each video on YouTube is uploaded to a certain channel (as defined in Section 4.3) that has subscribers. Large number of subscribers of a channel does not always guarantee video uploaded to this channel will attract the most view-count in competition with other similar videos, nor does it guarantee YouTube will select the video for trending list. We thus add a simple reputation system on top of the subscriber number and build a database of reputation scores for channel ranking. The videos from the first page of the YouTube search results are re-ranked according to the following formula

$$v = \arg\max(R_i * S_i). \tag{1}$$

where R_i is the reputation score of the channel i, and S_i is the number of subscribers to the same channel. The reputation scores are built on the fly. Initially, all scores are 0. The reputation scores are updated based on the observed views from different channels. Each YouTube channel that presents video related to the popular topic competes for large traffic. We periodically (with a time window of 1 hour) monitor the view-count of all candidates. The query is issued to the video-sharing portal and the results are sorted by view-count in descending order. In the case of news category, additionally we apply time constraint (i.e., results are limited to the time frame from when the topic was first detected in mainstream media as trending, to the present time). Every time our algorithm selects a video for caching that does not attract the most view-count, we increase the reputation score by one point for the channel that presented the video with the largest traffic, with respect to the same keywords.

Before placing an item in the cache, we identify it as a caching candidate. *The video is placed in the cache only if it passes the threshold of a minimum requests rate*. In order to select the threshold for caching, we take a similar approach as [74] does for the trending terms in social media. Typically trending videos are requested with higher frequency and deviate from normal behavior. We will assign a score (i.e., normalized request rate) to a YouTube video according to their deviation from an expected behavior. Assume that *U* is the set of all URLs collected by us, *R* is the URL of the video to which we wish to assign score, and *h*, *d*, and *w* present an hour of the day, a day of the week, and a week, respectively. We then define $U(R, h, d, w)$ as the set of all Rs such that R was requested during hour h, day d, and week w. With this information we can compare the amount of requests in a specific day/hour in a given week to the same day/hour in other weeks.

To define the score for R precisely, let $Mean(R, h, d) = (\Sigma_{i=1 \ldots n} |U(R, h, d, w_i)|) / n$ be the number of URLs requested each week on hour h and day d, averaged over the previous n weeks (w_1 through w_n). Correspondingly, $SD(R, h, d)$ is the standard deviation of R requested each week on day d, hour h, over the previous n weeks. Then the score of R over specific hour h, day d, and week w is defined as

$$Score(R,h,d,w) = (|U(R,h,d,w)| - Mean(R,h,d)) / SD(R,h,d). \quad \textbf{(2)}$$

To determine the threshold score for caching, we look at the past caching candidates, which became listed as part of YouTube‘s featured videos list in our training dataset. We compute the average score of all the above identified caching candidate videos for the hour during which they started to be on YouTube’s featured video list. This average score is used as the threshold (i.e., minimum normalized request rate) to make the final decision on which videos to store in the cache. We observe the scores of caching candidate videos in real time. Should any of them pass the threshold score, we place it in the cache.

4.4.2 Cache replacement

We apply the proactive approach not only to caching candidate selection, but also to cache replacement. Given the view-counts of videos observed from YouTube, and the popularity of the topic in the mainstream media to which this video belongs, we design a proactive cache replacement algorithm. The method consists of three major attributes: (a) Popularity of the topic in the mainstream media, to which the i_{th} video belongs P^t_i, (b) Latest popularity of a video R^T_i, i.e., view-count in the past T hours (in our experiment we set T similar to [16]), (c) Average popularity duration of the category to which the video belongs C_i. YouTube currently has 15 different categories. Each video has a category field provided by users. We grouped videos in our dataset by their categories and calculated the average popularity duration (how long they remain in the trending list) for all trending videos. The necessity to keep video i in the cache is calculated as follows.

$$N_i = (P_i^t) * (R_i^T) * (C_i) \tag{3}$$

Videos with smallest N_i are more likely to be dumped out of the cache. The replacement process is called only when cache is full, and new popular videos are arriving. Once the selected caching candidates pass the popularity (i.e. minimum request rate) threshold, the algorithm will start dumping out videos with smallest N_i until sufficient space is available to accommodate all emerging videos marked for caching.

4.5 Experimental Results

To evaluate our proactive caching scheme, we simulate the network performance in NS-3 [110]. In order to draw a realistic scenario, we replicate the Internet backbone topology into our simulation. Every ISP has traces from its own network at its disposal. Although we do not possess such data, there are a number of sources (such as [108]) that provide information about global web traffic. Please note that analysis of web traffic is done offline. ISP can analyze the history of browsing and identify which news sources are most popular among its end users. Selection of the most popular news portals maximizes the chance of detecting popular trends that will drive user interest in video sharing portals.

The largest provider of mobile telephone services in the US, AT&T, has published its IP backbone network topology [18]. Under our NS-3 simulation environment, we deploy nodes, assign link capacities, and specify propagation delay according to the Internet backbone topology [18][72]. Fig. 4.3 shows the topology of the Internet backbone used in the simulation, along with the placement of YouTube servers [46].

For the end user locations, we use 7 large cities of US (New York City, NY; Los Angeles, CA; Chicago, IL; Houston, TX; Seattle, WA; Jacksonville, FL; Kansas City, MO;). We also use London, England and Sydney, Australia. The three major English-speaking countries were chosen because we mine text (news articles) written in English language only. These countries are located in different parts of the world, and the cities are located in different corners of continental US, thus we evaluate our caching algorithm based on the transcontinental data transmission traces. The intermediate routing nodes in the simulation are randomly deployed. A caching server is deployed for each city, which is attached to the gateway routing node. The number of end users in the simulation is approximately 50,000.

4.5.1 Data Collection

For all cities we have selected RSS feeds [59]. The selection was based on the web traffic traces from Alexa [108]. We learn which city is using which news source most, and subscribe to the corresponding news feeds. The feeds are collected with 30-minute interval. For online video content we use YouTube. We have implemented crawlers to collect traces of YouTube daily trends for 30-day period (Sept. 1 – 30, 2013), and view-count information of each video listed in the YouTube's trends list is stored in a log file with 15-minute time interval. Note that in our proactive caching framework, we have only used publically available data that could be acquired by anyone. List of YouTube's featured videos, and RSS feeds used in our experiment can be found online [47].

4.5.2 Evaluating Keyword Selection

We evaluate the performance of our keyword selection scheme by mining news articles from different news sources with different number of RSS feeds. The first test was conducted by mining 10,000 RSS feeds. We identified popular topics through OSLDA (The total number of topics is 100), and selected keywords for the popular topics by applying Apriori. We doubled the number of feeds for the next experiment and kept increasing the number of feeds until the total number of documents reached 100,000. The experiment compares the performance of our scheme, combination of OSLDA and Frequent Pattern mining (OSLDA+FP), against OSLDA alone or FP alone. We send query keywords derived above to YouTube. We select the first 5 videos from the search results page. [36] presents the click-through rate of top 10 search results in Google. Apparently, on the average, results below top 5 generate less than 5% click rate. We thus focus on the top 5 videos in the YouTube search results page. The video content that carries relevance to the popular event from mainstream media, defined by keywords is marked with 1, otherwise 0. A maximum of 5, minimum of 0 point can be assigned per search results when evaluating the keyword selection algorithms. The relevance in our experiment is defined similar to [62]. The marking above is done by a human. Three undergraduate students have contributed to the manual labeling. Fig. 4.4 shows that the combined effort of OSLDA and Apriori can achieve much better results than OSLDA alone or FP alone, especially when the number of feeds is small. The quality of the keywords generated by our algorithm can be verified by the experiments presented earlier in Section 4.5. Indeed, not every topical trend detected by OSLDA has related popular YouTube video, however this does not hinder our goal of

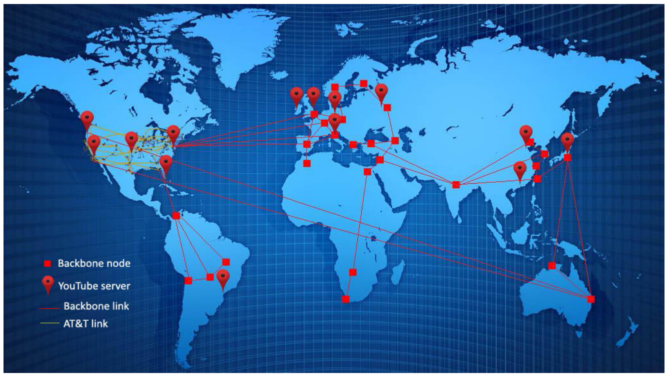


Fig. 4.3 Network topology used in the experiment. The map shows the locations of backbone nodes and YouTube servers.

predicting the future popular videos. During the 30-day period, on average approximately 35% of trends detected by OSLDA could be linked to popular videos in YouTube, however this accounts for 60% of videos listed by YouTube as trending. As shown in Section 4.5, the majority of the YouTube featured videos are news related. We also look at the quality of topics generated by OSLDA. [22] presents a range of topic scoring models, which evaluate the LDA topic coherence. Among several approaches discussed in the paper, we used WordNet [60] for OSLDA topic coherence evaluation. WordNet is a lexical ontology that represents word sense via "synsets", which are structured in a hypernym/hyponym hierarchy (nouns) or hypernym/troponym hierarchy (verbs). WordNet additionally links both synsets and words via lexical relations including antonymy, morphological derivation and holonymy/meronym. A number of computational methods for calculating the semantic relatedness/similarity between synset pairs have been developed. These methods apply to synset rather than word pairs, so to generate a single score for a given word pair. Each topic in OSLDA is represented by 54 distinct words [71]. We look up each word in WordNet and exhaustively generate scores for each sense pairing defined by them, and calculate their arithmetic mean. For our experiments over WordNet, we use the WordNet::Similarity package. We categorize each topic as one of the three types: "good", "intermediate", or "bad", similar to [21]. After evaluating topics generated by OSLDA, less than 10% were classified as "bad". We should also note that our keyword generation process does not depend solely on OSLDA output. It is only an intermediate tool to help us narrow down a collection of documents belonging to the same trending topic. Frequent pattern mining identifies most co-occurring terms among the RSS feed titles, which are used as the final product for video query. If certain documents are misclassified as member of an irrelevant topic, the words in the title should have week correlation with all other titles, therefore their share will be small and will not appear in frequent itemsets.

4.5.3 Evaluating Caching Candidate Selection

Our next experiment shows the performance of our reputation system. As mentioned in Section 4.4.2, proper keywords alone do not guarantee that the selected video will attract most traffic. Once keywords are identified, we query YouTube, re-arrange the first page of search results based on Eq. (1) and select the top five videos as caching candidates. We observe the view-count development of these videos (with 15-minute time interval), and if any of them passes the minimum request rate threshold we mark it as video for caching. After five days, we query YouTube with the same keywords, and sort results by view-count in descending order. If the video with the most view-count is the same as the video selected by our caching algorithm, we mark the result as 1, otherwise 0. The averages are plotted in Fig. 4.5. As we observe the performance of OSLDA+FP without reputation system is not very effective. The improvement in the performance with a larger number of feeds is explained by better input to Apriori. With a huge number of news articles, Apriori can derive frequent-4, or frequent-5 itemset that makes video search more effective. With the reputation system incorporated, the video selection algorithm performs with 90% accuracy or above. Our proactive caching algorithm is intended for the Internet Service Providers (ISP). Some ISPs are serving small neighborhoods, where its subscribers might be affected by small number of media sources. Therefore, it is important to have a system that works reliably with a few thousand feeds.

4.5.4 Efficiency of Proactive Caching

We evaluate our proactive caching algorithm (proactive cache candidate selection and proactive cache replacement) in comparison with two widely used caching algorithms in video service systems. 1) Popularity based caching in which video is being cached based on its historical popularity (the number of views in the recent period) [76]. We evaluate this method with two different cache replacement algorithms: Least Frequently Used (LFU) and Least Recently Used (LRU) [1]. 2) Random approach, where videos are randomly cached and randomly replaced. This algorithm does not require keeping any information about the access history [93]. We use Local Cache Hit Ratio, Average Delay, and Bandwidth Consumed, as a function of cache capacity, to evaluate the performance. Additionally, we define "Oracle" caching that considers a posteriori best-case caching strategy and compare it with our proactive caching method. Based on the collected traces from YouTube, we can "foresee" the future development of video view-counts, thus the videos are placed in cache right before the demand is about to grow. Such caching system does not exist, however it allows us to see how far our proactive method is from the ideal caching.

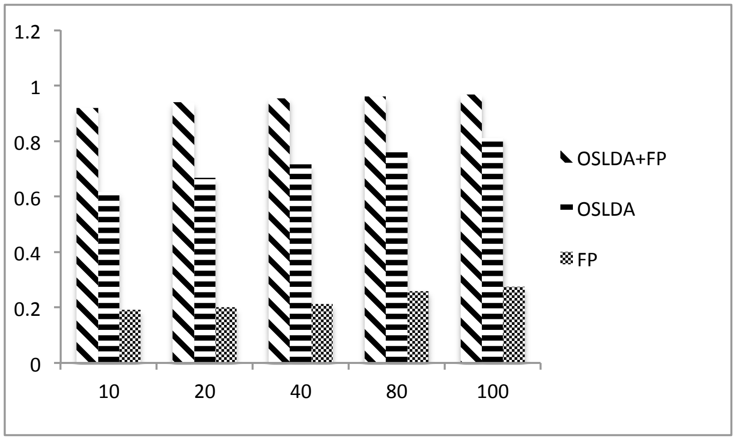


Fig. 4.4. Video relevance to the topic (%) vs. # of feeds (in thousands)

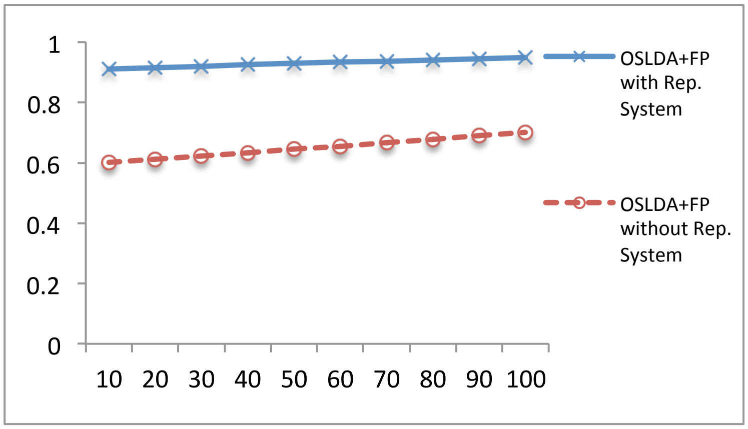


Fig. 4.5 Accuracy of predicting video with most traffic (%) vs. # of feeds (in thousands)

The total number and frequency of video requests in the simulation are generated according to the data we have crawled from YouTube. We do not possess data from ISP with detailed web requests, such as originator IPs, requested URLs, and packet routes, however we wanted to replicate somewhat realistic traffic into our simulation. We have observed the view-count development of all videos identified by our caching candidate selection algorithm. We have also monitored the YouTube trending videos list [56] and recorded the view-counts of all trending videos with 15-minute intervals. YouTube does not provide the statistical information regarding how much traffic the trending videos are responsible for. However, for the videos used in our dataset, we have crawled the detailed view-counts information and present the total share of traffic generated by YouTube's trending videos. The ratio between the number of videos listed by YouTube as trending, and all other videos is 30%. The traces have been recorded over the course of 30 days (September 2013). The requests are coming from random sources in our simulation, as the origins (real geographic location) of requests are unknown to us.

Local cache hit ratio is defined as the fraction of requested videos that can be directly downloaded from the local caching nodes. Fig. 4.6 illustrates the local cache hit ratio versus the capacity of the caching nodes. We observe that our proactive caching algorithm can achieve up to 32% improvement over the reactive method that uses historical popularity. The advantage of proactive caching is mostly perceptible with cache-size ranges 0.4% (16GB) and 3.15% (128GB) of the total video size The observed growing gap when size of caching nodes increases (but less than 3.15% of total video size), is attributed to the fact that larger cache size enables proactive method to accommodate more videos and restrain from replacement, which could potentially remove the videos, with small Ni, that still attract some attention. Our system always gives priority to newly uploaded videos with emerging popularity. Our decision is influenced by the report in [29], which states videos tend to receive most of its views earlier in their lifetimes. In a static environment, where videos obtain gradual increase of view-count slowly, proactive method may find it hard to demonstrate the advantage. However our approach can capitalize strongly in a dynamic environment, such as YouTube, where videos worth of 100 hours are uploaded every minute, and trends change quickly over time [117]. When caching capacity exceeds 6.26% (256GB) the performance of proactive and reactive methods starts converging. Such behavior is explained by the fact that majority of the future popular videos may have been cached and increasing caching capacity does not give advantage to proactive caching. 50 GB is roughly 35% of the total size of all trending videos that may appear during a 24-hour time interval in YouTube's featured video list. 200GB is approximately 5% of all videos used in our experiment. 4096GB is considered "infinite caching capacity" since it can fit entire YouTube video dataset used in our experiment. The gap between proactive and "Oracle" caching on average is 15%. The random approach has the worst performance as expected.

Normalized average delay (ratio of average and the maximum delay observed in the three simulation scenarios) reflects the end-to-end average delay of the network. Fig. 4.7 shows that the performance of LRU and LFU seems to be very similar. Random cache replacement shows the worst performance. With the larger caching capacity, the performance gain of the proactive method becomes more and more perceptible. The advantage of proactive caching starts to decrease when the caching capacity exceeds 6% of the total video size used in the experiment. Our method can improve over the traditional reactive caching methods (LRU/LFU) by up to 27 percent. On average it trails the "Oracle" caching by 18%.

Normalized average bandwidth consumed (the average consumed bandwidth normalized by the maximum consumed bandwidth observed in the three simulation scenarios) refers to the amount of bandwidth consumed by the system. Major Internet backbone service providers (IBP) that allow access to the Internet backbone for small-scale ISPs charge for service based on the traffic consumed. It is important to minimize the bandwidth consumption, especially when it comes to trans-continental transactions. Fig. 4.8 presents the performance details. The proposed proactive scheme shows up to 25 percent reduction in bandwidth consumption over the reactive approach. The random approach shows a large gap from our approach. Proactive caching on average is 16% behind ideal caching.

In Fig. 4.9, we present the percentage of the total traffic generated by the YouTube trending videos in our simulation, which accounts to about 70%. Targeting these videos and placing them into caching nodes in time can greatly reduce the delay and bandwidth consumption.

As we have observed from Fig. 4.6, 4.7, and 4.8, our proactive approach demonstrates significant advantage over the reactive methods. The smallest caching capacity used in the experiment 0.02% (1GB) of the total video size has shown 7% improvement of local cache hit ratio for the proactive approach, when compared against the reactive methods. The largest performance gap (32% improvement) observed in local cache hit ratio is at the caching capacity 0.78% (32GB) of the total video size. The advantage starts shrinking when the caching capacity exceeds 6.25% (256GB) of the total video size. Normalized average delay and normalized average bandwidth consumed show similar performances. The advantage of proactive

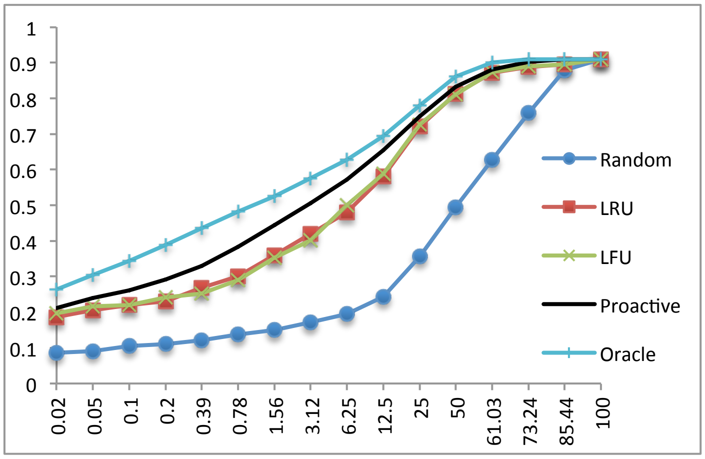


Fig. 4.6. Local Cache Hit Ratio vs. Cache Capacity (Percentage of total video size)

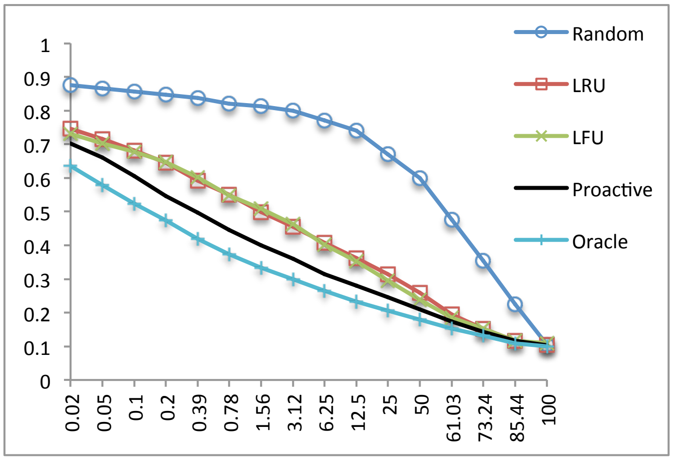


Fig. 4.7. Normalized Average Delay vs. Cache Capacity (Percentage of total video size)

caching over the reactive approaches is mostly perceptible when the caching capacity ranges between 0.4% (≈16GB) and 6% (≈256GB), achieving 24 and 21 percent improvement on average, for normalized average delay and normalized average bandwidth consumed respectively. This work addresses the challenges exposed by the modern online video service systems. We design a proactive caching scheme to aid quality of service in consuming online video. Our novel method explores the opportunities provided by the trends observed from mainstream media. Due to the media cross correlation, we can identify popular topics in the mainstream media, and pre-cache related videos on caching servers in a geographic region, where the videos are expected to attract large amount of traffic.

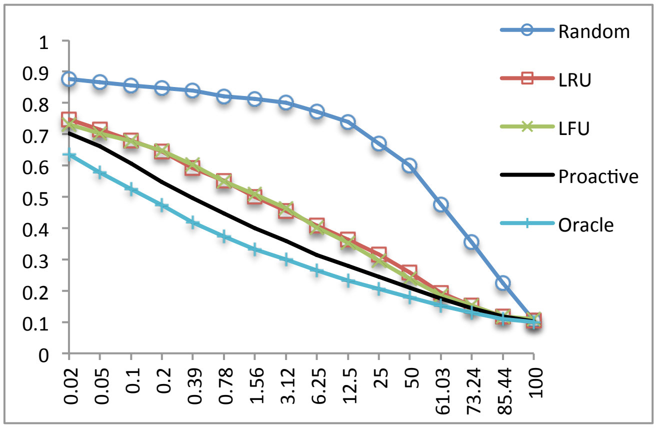


Fig. 4.8. Normalized average bandwidth consumed vs. cache capacity (Percentage of total video size)

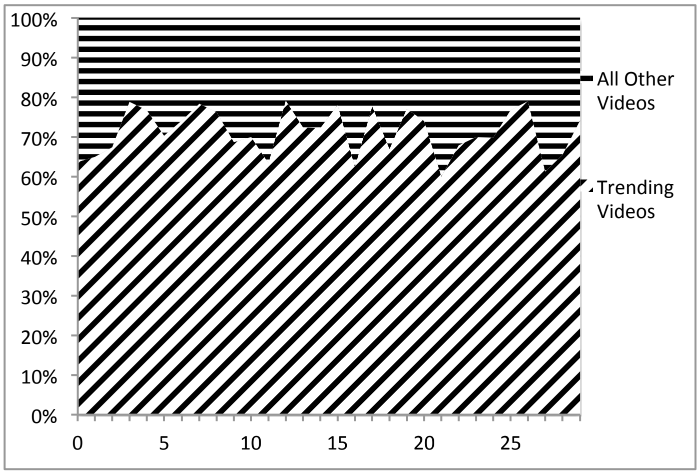


Fig. 4.9. Share of traffic (%) between videos listed as trending by YouTube and all other videos used in our simulation during 30-day period

4.6 Summary

In this chapter we present our work that addresses the challenges exposed by the modern online video service systems. We design a proactive caching scheme to aid quality of service in consuming online video. Our novel method explores the opportunities provided by the trends observed from mainstream media. Due to the media cross correlation, we can identify popular topics in the mainstream media, and pre-cache related videos on caching servers in a geographic region, where the videos are expected to attract large amount of traffic. Proactive caching is mainly designed for and is limited to accommodating the quality of service for videos that can be linked to the trending topics in mainstream media. We show the effectiveness and significant performance improvement of our proactive caching algorithm over the traditional methods. The proactive online video caching system provides opportunities for the ISPs to evaluate mainstream media, and more intelligently utilize the caching capacities. The proposed framework considers statistical co-occurrence of the terms that are composed by human journalists with semantic coherence and belong to the trending topics. Such approach may have some value for keyword suggestion systems. Our intention is to further test the system in real network environments in the future. The simulation shows promising results, however real network environment may exhibit slightly different characteristics, and fine-tuning of certain components may be required. The performance results obtained from the real network will help us better evaluate the performance and measure the gains of our system.

CHAPTER 5

An Automated News Item Suggestion System

In this chapter we propose an automated news item suggestion system. Our objective is to facilitate the creation of a mainstream media outlet that will automatically explore the existing trends in the media, make decisions on what news topics should be covered, build the list of items that should be included in the news, and generate article for the viewers. The system could also serve as a guiding tool for the journalists.

In order to achieve our goal we explore the virtues of Artificial Intelligence (AI). Interaction between journalism, the Internet, and social communities has been intensely discussed. In this chapter we discuss how AI will add to the picture. Before we dive into the details of the framework, we would like to present our motivation, analysis of the decisions made during the process, and justify the time and efforts spent along the direction of incorporating AI into journalism.

Every day, journalists have to choose from numerous items to decide which events qualify for inclusion in the media, as they cannot all fit the available space. The media act as 'gatekeepers', who control and moderate the access to news and information. In todays' society, connected by the Internet, the amount and the flow of information and communication have increased dramatically. In addition, users become involved as creators and distributors of content. The abundance of the information makes it difficult for journalists to manage and assess the events in terms of their alleged newsworthiness.

In the modern age of journalism, many decisions are guided based on the ratings. Ratings are important in the current society. Any recognized rating method influences societal development. Even if people find ratings annoying and counterproductive, it will still influence the system, given that people think others recognize the ratings. Many mainstream media outlets focus their attention on stories that interest the public. In order for journalism to remain meaningful, it should engage the audience. Advertisers demand full validation of consumer ratings. Competition has always been fierce among news production companies for selling the consumer attention to advertisers, whose ROI (Return On Investment) will determine the fate of the channels for advertising, including journalism paid by ads, across all media formats. Traditional media spend hugely to measure readership, estimating their size and attention probabilities and creating statistics and probabilities for advertisers. On the Internet, the new media offer content producers and advertisers not probabilities but hard data: which user looked at what content, when and for how long. Journalistic content is undergoing major changes via interactive platforms that make media content available continually, everywhere. Until recently, the mass media for distributing content were controlled by the same companies that produced the content. However the rise of the social media has changed the traditional perception of the journalism.

Within the research community of the journalism studies, the selection of news by the media is said to be influenced by a range of factors. Several criteria have been determined over time to assess the news value. Most of the analysis in the process of determining what news should be included and covered in the news is purely dependent on the human judgment. At the time they were created, large-scale empirical evaluation of these criteria would have been extremely labor intensive. Nowadays it would be infeasible, given the amount of information generated on the Internet on a daily basis. Recent developments in the field of AI enable us to revisit the process of newsworthiness evaluation. Our goal is to design an automated approach to help journalists in assessing the news value. The framework of news item suggestion system considers enabling journalists to search trending and interesting topics in the pool of large information. Note that so-called 'reader targeting' is an important component in the field of journalism.

The reminder of this chapter is organized as follows. In Section 5.1 we review some related works of news recommendation and suggestion. Next we present an overview of the proposed system in Section 5.2. The details of exploring correlations and associations in mainstream media are presented in Section 5.3. Section 5.4 dives into details of extracting semantic relationships between news items. The newsworthiness is discussed in Section 5.5. The process of semantic relation labeling is presented in Section 5.6. The details of final news items suggestion are covered in Section 5.7. We evaluate the proposed system in Section 5.8 and present experimental results. Finally, Section 5.9 summarizes the chapter.

5.1. Related Work

Two different technologies are commonly used in recommender systems: content-based recommendation and collaborative filtering. The content-based approach recommends information based on profiles; these profiles are built by analyzing the content of items that the user accessed and favored in the past. In contrast, the collaborative filtering approach does not consider the content of items, but uses the opinions of peer users to generate recommendations. Although our work does not depend on any of the above methods, we would like to review some of the news recommendation systems, analyze similarities and differences, and evaluate the advantages of our system.

The content-based approach has been applied to provide personalized selection of news information in various forms such as personal news agents [19], news readers for wireless devices [20][84], and web based news aggregators [5]. These systems build user profiles from information explicitly provided by the user or implicitly observed in user activities. The profiles are then compared with the content of news articles to generate personalized recommendations.

[5] presents a personalized news system, named Personalized Information Network (PIN). It retrieves and ranks news articles according to the user's profile, which is initially defined by the user as a list of keywords and then learned from user feedback using neural network technology. When interacting with PIN, users provide explicit feedback by rating the articles. A similar news recommendation system [19] is based on the series of feedback options such as "interesting", "not interesting", "already know this", etc. WebClipping2, a PDA news browser [84] uses a Bayesian Classifier in order to calculate the probability that a specific article would be interesting to the user. Rather than requiring users to provide explicit feedback, WebClipping2 observes the total reading time, number of lines read and some other

characteristics of user behavior to infer the user's interests. Another personal news agent [80] uses a proxy to collect user's page clicks and the browsing time, in order to construct a "personal view" that reflects user interest. [80] is applied and evaluated to provide personalized news.

Unlike the personalized news systems, our system does not record any user activity. There are no personal ratings or negative votes to gauge what the user dislikes. It is important to design a news recommendation system that does not implicate issues related to the user's privacy protection. For example Google News does not record detailed information about the clicks, such as the amount of time spent on the page. Thus, the system needs to make reasonable prediction with the limited and noisy information of user activity on the website.

As we mentioned the second technology often employed in recommendation systems is collaborative filtering. It has been applied to personalized news reading applications as well, such as [63]. The firs version of Google News [6] recommender was based on the collaborative filtering.

The content-based approach and collaborative filtering have their advantages and limitations [84]. Some research tried to combine both methods and achieved encouraging results [84][70]. The hybrid method benefits from both methods, providing early predictions that cover all items and users, and improving the recommendations as the number of users and rating increases.

The type of recommendation system we are trying to design is different in nature from the methods discussed above. Majority of the recommender systems simply try to aggregate the news articles and present the results to the users. Our approach despite having the same target of providing relevant information to the target groups, analyzes the problem from a different angle. Typically, journalists are presented with two problems before composing the article: what topic should they write about, and which news components should they include in the article. We attempt to design a system that will be a guiding tool for the human journalists, so that they can compose and present to the audience the most relevant and interesting information.

Journalism has existed for a very long time, and large numbers of news articles are available in the digital form. Even articles from the 19th century can be found in news archives in digital form. We try to understand how semantic entities in the news articles are connected, and what is the news value of the item, that makes it interesting for the public. We use data mining techniques to uncover hidden insights in the collection of news articles, and based on the observations make predictions on what type of information will be demanding in the near future. The next section presents a high-level overview of the system, which is followed by the detailed discussion of each component of the system in the subsequent sections.

5.2 System Overview

Our work is an attempt to design an automated news item suggestion framework. In the long run, we are hoping to create a prototype of the artificial mainstream media outlet that generates and posts articles online automatically. In this thesis we only present a partial solution. The artificial mainstream media outlet can be broken down into two parts, from the research point of view. The first part consists of news item selection for the content, and second is automated content creation. The latter is beyond the scope of our research. In this thesis we only focus on the news item selection process.

Our work is based on the mining of frequent patterns, association, and correlations. Frequent pattern mining searches for recurring relationships in a given data set. It has been long established that the discovery of interesting associations and correlations between itemsets in a transactional and relational dataset can lead to uncovering hidden insights. These insights are used for making intelligent business decisions. Although from the first sight the usefulness of association rule mining in news items is not apparent, we still tried to explore and experiment with the idea, to see if this method could be helpful for the purpose of recommendation and suggestion.

Market Basket Analysis [64] is a very famous and widely used example to demonstrate the usefulness of frequent pattern mining. Our work is inspired by the observations from this example. We try to mine news articles, as part of training process, to understand which association and correlations are strong, and based on the results, make predictions of which news items should be put together in the news article. Much like planning the shelf space as explained in [64].

One major obstacle for mining plaintext of news articles is the diversity of the news items. For example, in the case of "*Market Basket Analysis*" the product names are used in the mining process. Names such as: *milk*, *bred*, etc. always remain same for every transaction in the super market, however in case of news articles the names change from story to story. Our task is to find the similarity measure between entities and events that can be grouped under similar category. As we will see in Section 5.4 news are classified into different categories. Each news category generates different reactions and emotions among the news consumers. Therefore we would like to build rules of associations and correlations for each category. For example, the presidential elections in the U.S. take place once every four years. In each election period news agencies provide extensive coverage of the whole cycle. Most of the news articles provide very similar information (party affiliation of the candidate, educational/professional background, etc.) to the public for different presidential candidates. We try to generalize the description of the items from news articles so that each similar entity reported in the news can be described with similar label.

In order to achieve our goal we explore the Semantic Web. The Semantic Web is a collaborative movement led by the international standards body: the World Wide Web Consortium (W3C). The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF) [107]. Another tool used in our system is DBPedia [37]. DBpedia is a project aiming to extract structured content from the information created as part of the Wikipedia project. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related datasets. In our system the relationship between two entities is the key metric for generalizing the labels. For example, Woodrow Wilson, Franklin Delano Roosevelt, and Barak Obama all share the same relationship label with Democratic Party. We extract the semantic relations from DBPedia to make general/universal description of the relationship between items. In Section 5.4 we lay out the details on how the relation labels are extracted.

Mining the dataset of entity relationships gains the insight on which connections are the most interesting and often used by the journalists. The relationships are later used to extract concrete semantic entities from DBPedia. Although strong associations of semantic relationships give us good insight on what to recommend, often news items must be treated case by case. For example movie actors/actresses get wide coverage by the media. It is common for news articles to review the movies/TV Show they played in, however mostly, only the popular ones are mentioned in the news articles. Our task in this step is to understand how prominent is the news item, and if it qualifies to be included in news article. We turn to the journalistic studies to find some clues on how the process is performed. [43] presents the common standard, agreed upon by the experts in the field of journalism that is employed for generating cumulative news value, which is used for determining if news item is sufficiently popular to make the cut in the article. We automate the process of news value assessment by exploring the Named Entity Recognition (NER) services. The details of this process are presented in Section 5.5.

Our process can be broken down into two parts. The first is the training process, which explores data mining techniques, semantic web, and DBPedia to build the relationship dataset, and identify strong correlations and associations. This process is done off-line and does not need to deal with real-time scalability issues. The second step is applying these rules to existing trending topics, to extract news items, with expected popularity. This process is done on-line and needs to constantly monitor the mainstream media to identify trending topics, and make suggestion of which items should be included in news articles. The second step is very "light" computationally and does not require much processing power, while the training process is very "heavy", since it is applies to a large data set. The workflow of the entire process is presented in Fig. 5.1.

As we can see from Fig. 5.1 in the training process each news story is classified as a member of certain news category. Further we can see the notation

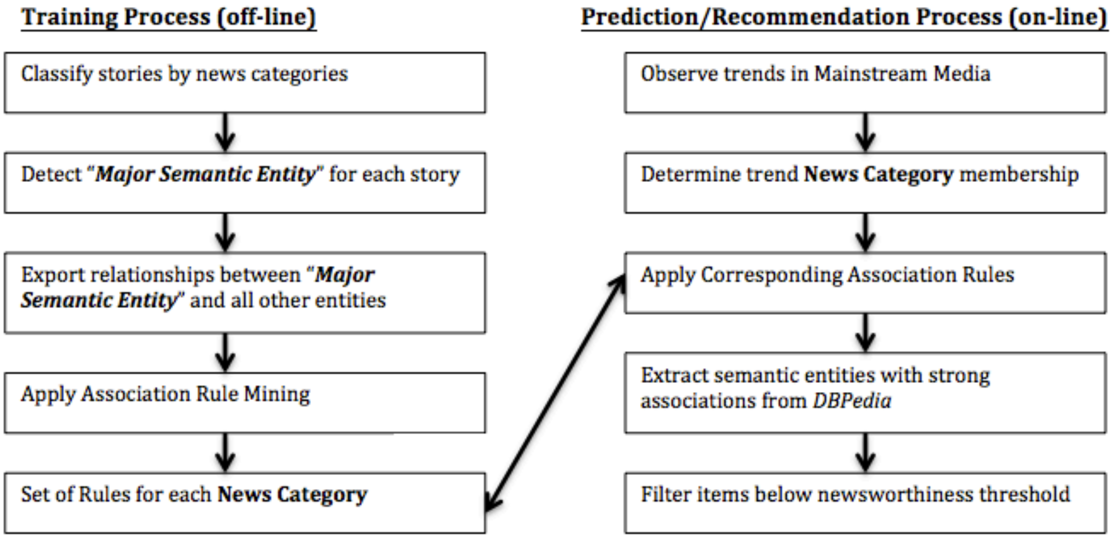

Training Process (off-line)
Classify stories by news categories
Detect "Major Semantic Entity" for each story
Export relationships between "Major Semantic Entity" and all other entities
Apply Association Rule Mining
Set of Rules for each News Category
Prediction/Recommendation Process (on-line)
Observe trends in Mainstream Media
Determine trend News Category membership
Apply Corresponding Association Rules
Extract semantic entities with strong associations from DBPedia
Filter items below newsworthiness threshold

Fig. 5.1. Workflow of News Item Suggestion Framework for the training and recommendation process

"***Major Semantic Entity***", which represents the center of the attention about the particular event. In Section 5.4 we present the list of news categories used in our system, as well as the explanation on how "***Major Semantic Entity***" is selected. This entity is used for building the semantic relationship with all other entities mentioned in the news articles. Since we employ NER services and DBPedia, our training process is limited to the items that have representative DBPedia articles, and can be recognized by NER service.

5.3 Exploring Frequent Patterns in News Articles

Given the fact that ratings determine greatly what content should be covered, and certain criteria exist for choosing newsworthy items, we explore the patterns in the news articles. In this section we describe examples where frequent patter mining may be helpful, and draw parallel, how such techniques could be helpful for achieving out goal. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational set. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patters from their databases. The discovery of interesting correlations relationships among huge amounts of business transaction records can help in business decision-making processes, cross-marketing, and customer shopping behavior analysis.

A typical example of frequent itemset mining is the market basket analysis. This process analyses customer-buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying certain item, how likely are they also going to buy another item on the same trip to the market? Such information can lead to increased sale by helping retailers do the selective marketing and plan their shelf space.

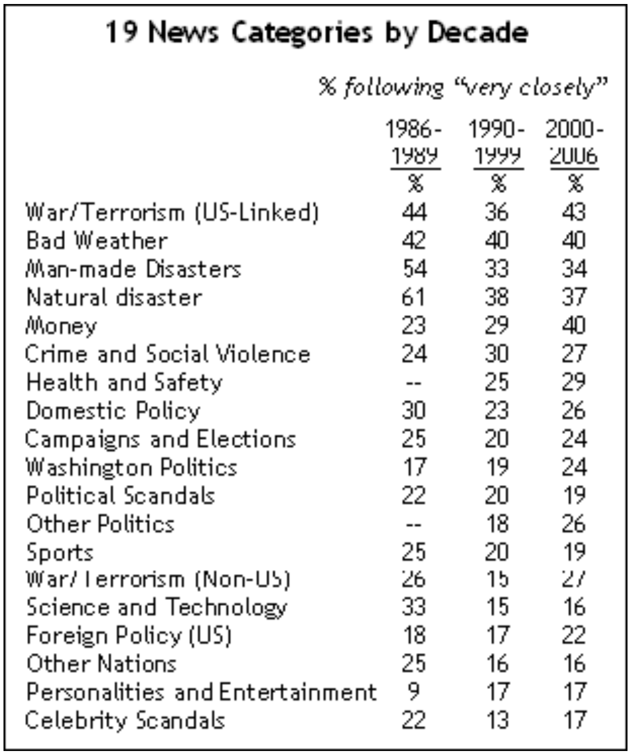
The example above gives an idea, how frequent itemset mining can help make decision and plan the order of items. Learning customer habits and adjusting business strategies based on the frequent pattern mining can help boost the revenues. We try to think of the process of composing articles, as process of planning the shelf (in the grocery or clothing store). The link may not be very obvious. Grocery stores, and clothing shops provide different type of service for the customers, than mainstream media outlets. Unlike supermarkets, mainstream media does not have the customer transactional database. The only dataset we can mine are the news articles composed in the past by the human journalists. This information we believe is interesting enough to observe the associations and correlations among the entities mentioned in the article. Typically, news articles are composed following a certain journalistic standards [75]. Each article is a collection of names that describe person, country, organizations, event, etc. Observing frequent patterns in the news article may provide us some insight about the frequently co-occurring names that are related to the same story. Obviously exploring frequent patterns in the collection of random news articles will not provide any helpful information. Even if we group related stories (e.g. all articles related to War), we will not be able to observe meaningful patterns. There are many reports about wars that happened in different times, at different places. The names involved in the story are different, thus making it difficult to obtain any meaningful co-occurrence information.

5.4 Semantic Entities and Relationships

In this section we introduce the idea of exploring the semantic relationships between two named entities. In the last three decades major players in the field of news reporting have been presenting news in the digital form. In the past few years availability of social media and blogging platforms have boosted so called "citizen journalism". Website such as Tumblr claims to have over 27,000 blogs generated every minute [39]. There is huge amount of information available on the Internet, and it is growing day by day. All the news articles generated through the years are out there, just waiting to be leveraged. If we think of the news articles as a database of transactions, we could employ association rule mining to extract hidden knowledge from the mainstream media. Obviously, unlike market analysis scenario, the transactions are not available for us in the same format as it is in the stores. We think of news articles as collection of semantic entities. Every article contains name of objects, people, places, etc., which are defined as objects in the semantic web.

Any news in the media belongs to a certain category. Each category is a collection of similar types of news. These types of news consist of similar types of events that involve different people, places, etc. Our aim here is to design a system, where we can convert each news article into a collection of semantic relations. These relations must be described into a general form that is universal for any event. Once we have a list of such relations, we can treat it as a large dataset. Applying frequent pattern mining will lead us to the discovery of associations and correlations among items in large relational data sets.

When news appears in the media, we can classify it as member of a certain category. A list of categories is shown in Fig. 5.2. For each category we have rules of association that can suggest, when similar event appears, what type of information should be reported to the public. The system should use topic modeling algorithm to identify emerging topics, detect "major semantic entity" via our keyword selection method (discussed in Chapter 4) and through association rules, we can obtain a list of topics that will matter to the public and gain much attention in the media.



19 News Categories by Decade

% following "very closely"

	1986-1989	1990-1999	2000-2006
	%	%	%
War/Terrorism (US-Linked)	44	36	43
Bad Weather	42	40	40
Man-made Disasters	54	33	34
Natural disaster	61	38	37
Money	23	29	40
Crime and Social Violence	24	30	27
Health and Safety	--	25	29
Domestic Policy	30	23	26
Campaigns and Elections	25	20	24
Washington Politics	17	19	24
Political Scandals	22	20	19
Other Politics	--	18	26
Sports	25	20	19
War/Terrorism (Non-US)	26	15	27
Science and Technology	33	15	16
Foreign Policy (US)	18	17	22
Other Nations	25	16	16
Personalities and Entertainment	9	17	17
Celebrity Scandals	22	13	17

Fig. 5.2. Popular news categories in the U.S.

A technical challenge here is how to convert semantic entities into some general terms. For example, let us look at the sentence “Alex Ferguson retires from Manchester United”. Here we have “*Manager*”, “*Football Club*”, and “*Retirement*”. By putting the semantic relation labels together, we should be able to mine the news articles (only related to this story) and discover the associations. Given the event, what information was reported in the media? (e.g. Clubs Alex Ferguson managed before Man United, trophies he won, manager who will replace him, new manager’s career highs and lows, etc.). When other managers retire we collect related news articles and apply the process of generalization. The identified semantic entities are converted into general entity definition labels (e.g. David Moyes -> Club Manager, Everton -> Football Club, Glasgow -> Birthplace etc.). Apart from this we need to learn the semantic relations between the entities that appear in the news articles. A single document is treated as a single transaction table, with multiple semantic entities appearing together (e.g. Glasgow -> Birthplace of A. Ferguson.). The semantic relations are the key components for creating association rules. Given a large dataset of news articles, we put together all the “transactions” (entities appearing together in news article) and apply frequent pattern mining algorithm to discover confidence and support threshold. These two values will be used to determine what’s worth recommending.

5.5 Newsworthiness

Human journalists, as discussed earlier in this chapter, can manually evaluate the news worthiness based on the guidelines, agreed upon journalism domain experts. First we would like to discuss the sociological perspective of newsworthiness assessment, and then we present technical perspective.

Sociological research on the process of news selection in newsrooms has resulted in various overviews of "news values" or "news selection criteria". The concept of newsworthiness is built on the assumption that certain events get selected by media above others based on the attributes or "news values" they possess. The more of these news values are satisfied, the more likely an event will be selected. If an event lacks one news value, it can be compensated by another attribute. Hence, journalists criteria for selecting the news are cumulative, making stories significant based on their overall level of newsworthiness. The earliest attempt for a systematic approach of determining newsworthiness by news value, is a taxonomy [30] that triggered both scholars and practitioners in examining aspects of events that make them more likely to receive coverage [34][95]. For analytical purposes, the concept of news values is valuable to understand that news selection is more than just the outcome of journalists' 'gut feeling'. To decide what is news and what is not, journalists consciously and unconsciously use a set of selection criteria that help them assess the newsworthiness of a story or an event [82][94]. In Table 5.1 we present an overview of the different news values that are available in the literature of journalistic studies [43].

Table 5.1. Conditions for News Worthiness

News values
Frequency
Negativity
Unexpectedness
Unambiguity
Personalization
Reference to elite nations
Reference to elite persons
Conflict
Consonance (media readiness to report)
Continuity
Composition
Competition (story by a rival)
Co-optation (marginally newsworthy)
Prefabrication
Time constraints
Logistics

We would like to give brief review of the most important studies in this domain. In the 1960's, Galtung & Ruge [30] published a theory of news selection, which provided a taxonomy of 12 news values that define how events become news. More specifically, they discerned (1) *frequency*: the time-span of the event to unfold itself, (2) *threshold*: the impact or intensity of an event, (3) *unambiguity*: the clarity of the event (4) *meaningfulness*: the relevance of the event, often in terms of geographical proximity and cultural similarity, (5) *consonance*: the way the event fits with the expectations about the state of the world, (6) *unexpectedness*: the unusualness of the event, (7) *continuity*: further development of a previous newsworthy story, (8) *composition*: a mixture of different kind of news, (9) *reference to elite nations*, (10) *reference* to elite persons, (11) *reference to persons*: events that can be made personal, (12) *reference to something negative*. Bell [11] used Galtung and Ruge's list as a starting point, but redefined some of them and added the value of facticity: a good story needs facts (e.g. names, locations, numbers and figures). Harcup and O'Neill [34] also made a revision of Galtung and Ruge's criteria for the contemporary news offer, with special attention for the entertainment offer available in newspapers. More specifically, the following values were added: (1) Events with picture opportunities, (2) *Reference to sex*, (3) *Reference to animals*, (4) *humor*, (5) *showbiz/TV* related events and (6) *good news* (e.g. acts of heroism). In addition, concerning elite persons they made a distinction between *power elite* (e.g. Prime minister) and celebrities (e.g. Pop Stars). McGregor [77] also proposed news values reflecting the modern way of news selection, with a focus on TV news. He referred to events that are visually accessible and recordable. In addition, when events involve tragedy, victims or children, it is likely to appeal to the emotions of the audience. Concerning negative events, conflict, scandal and crime are highlighted. Bekius [105] added competition as a value, which explains the extensive coverage of sports. Finally, Shoemaker and Cohen [95] extended the notion of significance, by demarcating four sub-dimensions: *political* significance, *economical* significance, *cultural* significance and *public* significance.

5.5.1 Technical perspective of assessing news worthiness

While the aspects of news described in the previous section provide valuable guideline for the theoretical analysis of news, it still assumes human involvement, which could be labor intensive and prone to human error. With the processing capabilities of the modern technology, social network, and the semantic web, we see it possible to make news worthiness assessment automatic. In this section we present the details on how scores for each attribute are calculated and how news score is computed. Our aim is to provide users (such as human journalists) information about the newsworthiness of semantic entity. It is infeasible to generate single newsworthiness score for any item. We take a similar approach to [97], and assign a normalized score to a semantic entity that ranges between 0 and 1.

Any entity, considered for grading (with newsworthiness score) must have representative article in Wikipedia. Typically, wiki articles provide background information about the entity, highlighting the most interesting facts/aspects. We take the Wiki article as input and analyze it in different ways. The results of the analysis are tracked back to the news values. After the analysis we normalize the score and return as final result.

In Table 5.1 we present the list of the news value attributes that can be calculated automatically. A crucial step in our analysis is gathering as much descriptive metadata about the news article as possible. However, manual addition of medatada by the authors is often neglected. Therefore, we need to automatically generate these metadata ourselves. To do this, we use publicly available tools known as Named Entity Recognition (NER) services. These services accept plain text as input and output a list of linked Named Entities (NEs), detected in the text. OpenCalais [52] is a well-established and thoroughly tested tool that is publically available and provides NER services. For a thorough overview of NER services and their performance, we refer to the NERD framework [87]. At the moment OpenCalais is able to detect 21 types of entities from plain text, of which a selection is shown in Table 5.2. Documentation regarding lined data entities can be found at

[53]. Conforming to the mapping in Table 5.2 these entity types contribute to the *Frequency, Negativity*, *Unexpectedness*, *Personalization*, *Reference to elite nations*, *Reference to elite persons*, *Conflict*, *Continuity*, and *Competition*. The amount at which a detected entity contributes to its respective news value(s) is determined by the relevance score assigned by OpenCalais. This relevance score represents the importance of an entity in the text it occurs in. For each entity detected in the text, the relevance score is accumulated in the corresponding news value(s).

Types, such as *Consonance*, *Composition*, *Co-optation*, *Prefabrication*, *Time constraints*, and *Logistics* can be preset by the administrator of the automated news worthiness assessment system. For example, *consonance* refers to the stories that fit with the media's expectations. They receive more coverage than those that defy them (and for which they are thus unprepared). Another example is *logistics*, which refers to the availability of reporters in the geographic region where the story takes place. Scores for some attributes may be hard to assign automatically. Our system is designed to aid human journalists to narrow down the vast amount of information to specific topics that may be appealing to the public.

5.6 Labeling Semantic Relationships

In this section we describe the process of creating labels of relationships between semantic entities. Each news article consists of a collection of named entities. Articles belong to a certain category, and to a certain topic. For each topic we have a ***major semantic entity***. In order to understand how major semantic entity is selected we refer to an example. As discussed in Chapter 4, Section 4.3 our algorithm provides the query keywords to retrieve related video. We look at the query to detect a word that has dedicated DBPedia article. For example, the query "*Hillary Clinton running president*" contains three entities that have DBPedia article (Hilary Clinton, running, president). We look at the news value of each entity based on the technique discussed in the previous section. The entity with the highest score is considered to be the *major semantic entity*, in this particular case Hilary Clinton.

Every article related to the story is examined for determining the relationship between the *major semantic entity* and all other entities detected in the news article. For example if the word *Chicago* is mentioned in the article, we look at the relationship between *Hilary Clinton* and *Chicago*.

According to the DBPedia page of *Hilary Clinton*, Chicago is listed as the *birthplace*. (http://dbpedia.org/page/Hillary_Rodham_Clinton)

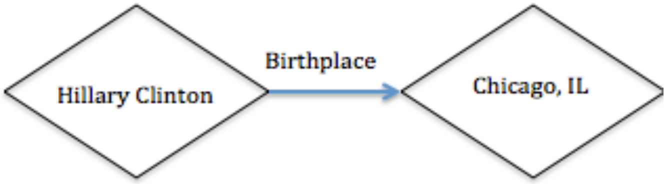


Fig. 5.3. Semantic relation between two entities

Thus the relationship label (birthplace in this case) will be stored in the database of transactions, which will be used for the frequent pattern mining later.

The example above considers the case when two entities have a direct link in the semantic graph (we refer to DBPedia as a graph of concepts, as discussed in Chapter 2). Now let's consider an example when two entities do not have a direct link. For example let's consider *Arjen Robben* as a major semantic entity. We scan articles related to the story (to detect all other entities that have representative DBPedia article) and detect *Rafael van der Vaart*. These two entities do not have direct link. DBPedia is unweighted bi-directional graph of concepts. If two entities do not have direct link in the graph we explore the shortest path between these two entities based on the *A** algorithm [35]. Since DBPedia is un-weighted graph we assign weight to the link between the two nodes based on the Normalized Google Distance (NGD) [86] score:

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where *M* is the total number of web pages searched by Google. *f*(*x*), *f*(*y*) are the number of hits for search terms *x* and *y* respectively. *f*(*x*,*y*) is the number of web pages on which both *x* and *y* occur.

As shown in Fig. 5.4, we compute the NGD score between R. van der Vaart

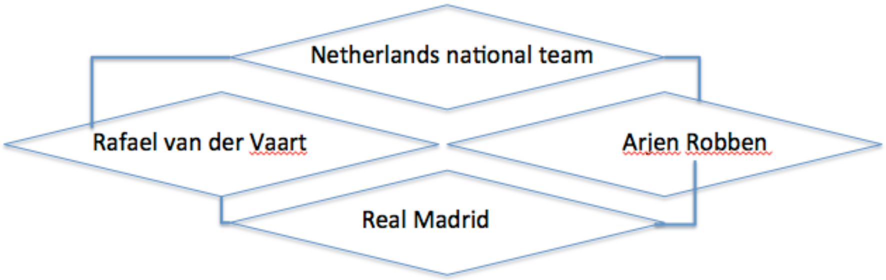


Fig. 5.4. Weighting Semantic Relations

and A. Robben through multiple links (R. van der Vaart->Netherlands national team-> A. Robben vs. R. van der Vaart->Real Madrid-> A. Robben). The path with the shortest score will be chosen and the relation between two entities will be defined as the list of semantic relations along the shortest path. In the example shown in Fig. 5.4 *Rafael van der Vaart* -> *Netherlands national team* -> *Arjen Robben* has shortest path, thus the relations will be labeled as: [football player, team, football player]

The reason we explore NGD for the purpose of finding relations between the entities is the nature of semantic web. It was not constructed for aiding journalism. The shortest path between two Wikipedia articles is defined as the minimum number of clicks required to move from one article to another. This can lead to discovery of connections between topics that are completely unrelated. For instance, if you are on the Wikipedia page of *Microsoft*, it would require two clicks to reach page *Saddam Hussein* (Microsoft -> 1990 -> Saddam Hussein). DBPedia is a structured content of Wikipedia, where each article is a graph node, and the hyper link between two articles is the undirected graph edge. If we look for shortest path between two entities without considering NGD, we may obtain relations that are irrelevant for the purpose of journalism. Google distance is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. Keywords with the same or similar meanings in a natural language sense tend to be "close" in units of Google distance, while words with dissimilar meanings tend to be farther apart. Thus we obtain the relationships that make some sense semantically. Later we explore strong correlations and associations among the relations listed in the transactional database. Fig. 5.5

presents the workflow process of the proposed framework to generate the rules for each news category (Fig. 5.2).

5.7 News Item Suggestion

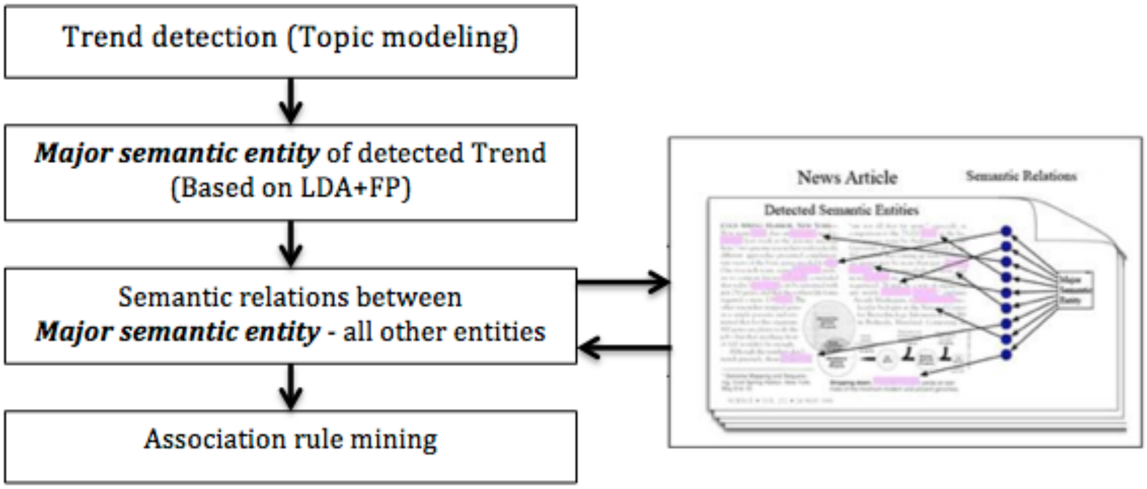

Trend detection (Topic modeling)
Major semantic entity of detected Trend (Based on LDA+FP)
Semantic relations between Major semantic entity - all other entities
Association rule mining
News Article
Semantic Relations
Detected Semantic Entities
Major Semantic Entity

Fig. 5.5. Extracting rules of association from mainstream media

The process of generating association rules for each news category is a prerequisite for the creation of automatic news item suggestion system. We should note that everything described up to this point is an off-line process, and does not need to cope with the real-time incoming stream of information. The larger the training dataset, the better will be the output. Assignment of weights to the DBPedia graph links is a very time consuming process, however the time complexity is not a reason for concern. Once rules of association and correlation are generated for each category, we can apply those rules to the real time trends and make suggestions on what kind of information should be included in the news articles.

Table 5.2. News values and sub-criteria mapped to Named Entities, Topics and analyses using Semantic Web technologies

News values	Sub-criteria
Frequency	
Unexpectedness	Novelty Unpredictability Uniqueness Normative deviance Social change deviance Statistical deviance
Recency	
Continuity	
Power Elite	Reference to elite nations Reference to elite persons Reference to elite institution
Showbiz/TV	
Celebrities	TV/Movie start Sports stars Pop Stars Royalty
Negativity	Conflict Crime Material or personal damage Tragedy (natural, or personal disaster) Scandal (allegations that damage reputation)
Good News	
Reference to sex	
Significance	Political significance Economical significance Public significance Cultural significance
Relevance	Geographical proximity Cultural proximity
Facticity	Numbers Graphs
Picture opportunities	
Competition	Numbers & Names
Logistics	

The process of news item suggestions consists of three steps. First we observe the mainstream media in real time. The tools discussed in Chapter 4, Section 4.2, such as OSLDA, can be used to monitor the trends in real time. When concrete trending topics are identified, the second step consists of classifying the trend into news category. This is required for making decision on which association rules should be used. As discussed earlier in this chapter, each news category generates different emotions among news consumers, and journalists are writing different types of articles for different types of events.

News categorization is a widely explored field. It can be treated as text categorization, where text is content of the news article. Various approaches have been proposed for the purpose of news categorization. Support vector machine (SVM) is a famous tool for categorization and classification purposes. The goal of text categorization is the classification of document in a fixed number or predefined categories. In Fig. 5.2 we show the list of predefined categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples, which perform the category assignment. This is a supervised learning problem. Since categories may overlap, each category is treated as a separate binary classification problem.

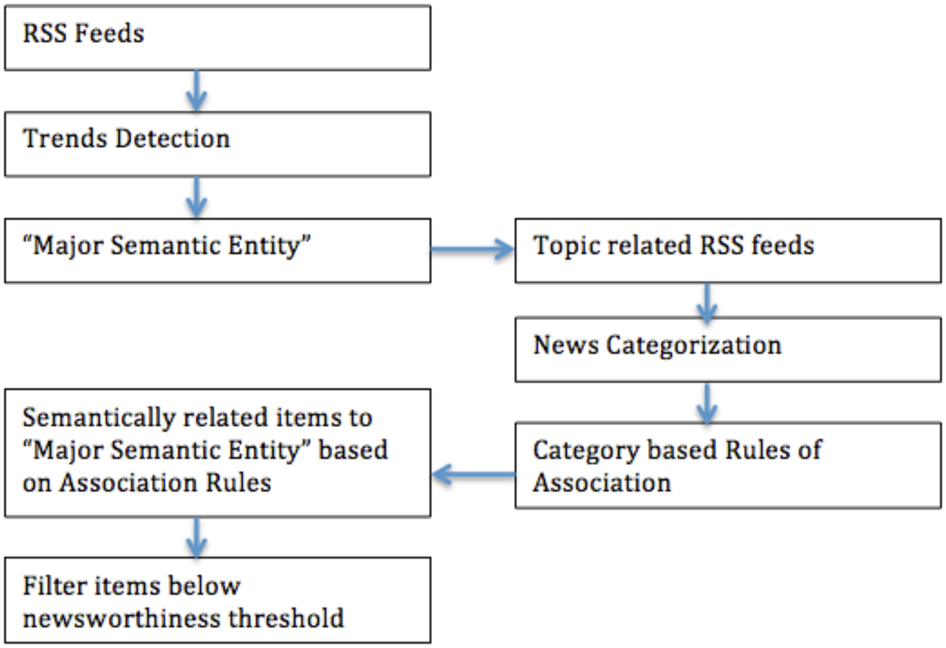

RSS Feeds
Trends Detection
"Major Semantic Entity"
Topic related RSS feeds
News Categorization
Category based Rules of Association
Semantically related items to "Major Semantic Entity" based on Association Rules
Filter items below newsworthiness threshold

Fig. 5.6. News Suggestion System Workflow

The first step in text categorization is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and classification task. Information retrieval research suggests that word stems work well as representation units and that their ordering in a document is of minor importance for many tasks. This leads to an attribute-value representation of text. Each distinct word w_i corresponds to a feature, with the number of times word w_i occurs in the document as its value. To avoid unnecessarily large feature vectors, words are considered as features only if they occur in the training data at least 3 times and if they are not *stop-words* (such as "*and*", "*or*", etc.).

[99] is a study that compares different categorization approaches and evaluates the performance. In our system we use approach described in [89] to decide the category of the news. As described in Chapter 4, OSLDA explores trends in real time, and produces the list of popular topics. We have a subset of the news documents, associated to each topic. These documents are used as the input to the text categorization algorithm, and based on the returned result we choose which association rules should be used to extract the news items. Fig. 5.6 presents the workflow of the process.

We will use the example to explain the process of item suggestion. Attentive reader may have noticed "newsworthiness threshold" as a new component of the system. Here we present the details regarding the threshold selection.

Based on the list in Fig. 5.2 we created the association rules for the categories. For example category "Celebrity" identified strong association between "Celebrity" and "Filmography/Discography". When we evaluated the mainstream media, one of the trending topics was “death of Whitney Houston”. In the DBPedia graph if Whitney Houston is selected as the *major semantic entity* there are several nodes that can be related to her through the Filmography/Discography label. Eight different movies were extracted through “filmography label“ (from Whitney Houston as starting node). However not all those movies were mentioned in the follow up articles that covered the death of a singer. We try to understand which items should be mentioned in the news article by incorporating the newsworthiness threshold.

There are no research studies regarding the newsworthiness value threshold selection with some analytical discussion or experimental data. Thus we employ the experiment-based data driven approach to select the threshold. In order to understand what kind of item would satisfy the requirements of mainstream media, we explore the articles that have attracted substantial number of viewers. *New York Times* and Associated Press (AP) provide API services [38][58] that help us find articles that have been popular among the viewers. We retrieve a list of articles that have been popular according to NY Times and AP API and detect every entity in the article that has representative article in the DBPedia. The text of DBPedia article is passed to OpenCalais tool to assign the news value score. We compute the average score of all the scores obtained from the OpenCalais and use the number as the threshold, to make a decision on what items are newsworthy and what are not.

5.8 Experimental Results & Evaluation

Until now, analysis of news and assessment of its newsworthiness was performed by domain experts, such as journalists and media researchers. To our knowledge there is no automated system that operates as mainstream media outlet. Consequently, there are no automated systems to compare to our approach. Therefore, we design our own evaluation method and present the findings.

As the first evaluation of the automated news item suggestion framework, we have collected the data via publically available sources. We have implemented the framework and created association rules for several news categories.

5.8.1 Methodology of Data Collection

Before we present the implementation details, we would like to provide the list of sources from which we collected the necessary data to perform the evaluation. In order to build rules of association, and derive the semantic relationships we need a large dataset of news articles. Reuters [54] and Associated Press (AP) [49] provide link archive from which we downloaded all articles. NY Times has a search tool [51] that allows searching and downloading articles related to a particular story. Additionally we have used the Internet search engines to search and download stories related to certain news categories. Further in this chapter we provide a full list of stories that were used for the purpose of framework evaluation.

In order to understand how our system performs, we have constructed a test set. Typically our system explores the trends in mainstream media, detects the "*major semantic entity*" and outputs the list of suggested items that should be covered in the news. We take this list and compare it to the items mentioned in articles that were posted as the follow up of the events, related to the same "*major semantic entity*". To achieve a fair and justifiable comparison we explore the frequency of the news items mentioned in the articles. The list of most frequent items featured in the actual articles is compared to the list of items suggested by our system.

To be more specific we have built rules of association for several news categories (as shown in Fig. 5.2). As the most followed category War/Terrorism was our first choice. We have implemented the framework to collect the articles related to War, and then we extracted the list of semantic relations and obtained the rules of associations and correlations.

War/Terrorism - as a category is a very broad term, and may be broken down into several sub-categories. For example "invasion" is part of the war. They can be related to "Terrorism", however generate different emotions in the public. Later in this section we shall see the importance of creating news sub-categories to achieve better results.

In our experiment we identified articles related to Wars. [42] provides the list of wars that took place between 1945 and 1989. We searched and retrieved articles from the news portals [47] related to all the war activities as shown in [42]. Additionally we searched the articles related to World War I and World War II. We also used all Wikipedia articles of [42]. The text collected was used as the training set to learn the semantic relations and obtain the rules of association.

We selected two stories [41][40] and made the predictions of the news items, based on the rules of association obtained from the training set. Later we extracted articles from [54][51][49] related to [41] and [40] and explored the TF-IDF [68] score of the entities that have representative article in DBPedia, and can be detected by NER [97] service. The last two requirements are necessary in order to understand the semantic relationships of the entities and be able to assign the newsworthiness score. Term-Frequency – Inverse Document Frequency (TF-IDF) [68] is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [4]. It is often used as a weighting factor in information retrieval and text mining.

In order to evaluate the performance of our system, we take the full list of suggested news items and compare it to the top list of items identified in news articles based on the TF-IDF score. For example our system may generate list of 10 items that it predicts should be covered in the news articles. Accordingly we take 10 items that ranked highest in the actual articles (with respect to the same news topic) and compared the overlap of the items.

5.8.2 News Stories Examined for Evaluation

In this section we present the list of the news categories for which we build the rules of associations, and also present the full list of news topics for which we made predictions and evaluations.

We should note that digital representation of the news articles started in the past decade, and very limited number of articles are available in an electronic plaintext form that were published before the eighties of the 20th century. We also observed that Wikipedia articles are a good summary of the story for which we tend to detect the semantic relationships. For every story used in the process of data mining we have used corresponding Wikipedia article.

As discussed in the previous section, we create the rules of association for the news category *War*. Initially we searched all articles in [54][51][49] with query word "War". The returned set of articles was used to explore the semantic relations between named entities, and we obtained a list of strong associations and correlations. We experimented with the two stories: "*Russo-Georgian war in 2008*" [41] and "*2014 Russian military intervention in Ukraine*" [40]. The results were very poor; up to 15% overlap could be achieved. We decided to sub-divide the *War* category into *Invasions* sub-category. We queried articles related to all the stories listed on [42], and created new dataset of articles. Again we explored the rules of association and correlation between the semantic entities mentioned in these articles. This time we achieved 75% overlap on the average of the two stories [42][41].

In order to understand how specific the sub-categories need to be, we experimented with the news stories that could be clustered as sub-groups of the news. A large number of articles are available about the celebrities. Thus we build association rules for the "Celebrity Scandal" category. We posted the full list of stories that were used in our experiments on an external link [48].

For the "*Celebrity Scandal*" category, we have obtained the list of articles from the sources ([54][51][49]) related to the celebrity scandals, and created the list of associations. We made predictions on three stories "*death of Philip Seymour Hoffman*", "*death of Paul Walker*", and "*death of James Gandolfini*"; All three stories are related to the death of a celebrity that have died with different causes (overdose, car crash, heart attack - respectively).

The initial set of articles consisted of various types of stories related to celebrity scandal (domestic violence, public misbehavior, involvement in fraud and criminal activities, etc.). The overlap between the suggested news items and the actually popular items observed in the articles related to the specific story was under 30%. Later we created the rules of association only based on the articles that are related to the death of celebrity.

Websites such as IMDB (imdb.com) provide tools to query artists that have died on certain year (imdb.com/search/name?death_date=2014). We query articles

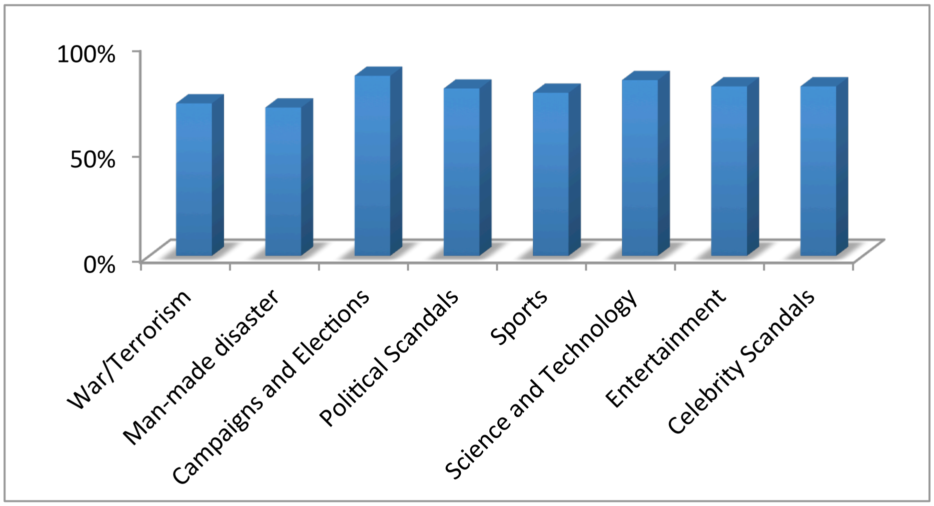


Fig. 5.7. Overlap (%) between suggested news items and top ranking news items

related to the specific death related story and build a new subset of the articles. Again we observe rules of correlation and association and generate a new list of suggested items. The performance of the prediction has significantly improved, producing 82% overlap between the suggested items and the top ranking items in the news articles.

We have further subdivided the "*death of celebrity*" sub-category into even more specific category "*death of celebrity caused by overdose*". The list of sources can be found at [48]. We made predictions again for the "*death of Philip Seymour Hoffman*" story. No improvement was observed in this case, suggesting there is a limit of sub-divisions, which makes no improvement by going further into smaller and more concrete categories.

In Fig. 5.7 we present the average percentage of overlap between the suggested news items and the items that have ranked highest in the set of articles.

On average our system produced 78% prediction accuracy. For each category we have created several sub-categories, and present the average score for each category. Totally 42 different stories from different categories were used for the evaluation. The full list of sub-categories, and list of news stories can be found at the external link [48].

5.9 Summary

To sum up the proposed news item suggestion system, we review some of the advantages and disadvantages of the system in this section. We also recall some of the key observations obtained during the process of experiments, and present the future work required for evaluating and enhancing the proposed system.

The initial impression of using association rule mining of semantic entity relationships seems positive. The major novelty of the proposed system is to design a system that treats news articles as transactional database, extracts relationships and applies frequent itemset mining to uncover the hidden insights. An additional step is to automate the news worthiness assessment to make final decision on which items should be suggested. The system presented in this chapter has only been experimented with limited number of stories, and needs further evaluation. Nowadays not many NER services are freely available, thus we only experimented with a single tool (OpenCalais). It would be highly desirable if professional journalists could provide manual labeling of the suggested items, to have better insight on the usability of the system. Another limitation we encountered in the training process was the collection of news articles related to a particular story. News articles from early twentieth century and earlier are very difficult to obtain in digital form, thus most of the stories in the training process are less than 50 years old.

The news categories used in the experiment are provided by the major news reporting agencies. During the process of experiments we discovered the list is very general and needs further sub-categorization. For example "*Celebrity scandal*" is a very broad term. It may include celebrity death, involvement of celebrity in criminal activity, adultery, and other topics. Mining all articles classified as "*Celebrity scandal*" may miss the useful associations due to diversity of the information. For example if we only mine articles related to "*celebrity death*" we are making news story more specific. Association and correlations obtained from such dataset will produce better results, versus the more general news category of "*Celebrity scandal*", where the probability of identifying key relationship is diluted. We also discovered infinite sub-division of news categories does not necessarily improve the accuracy of the prediction.

The news item suggestion system is a preliminary work to facilitate the creation of an artificial mainstream media outlet. Most of the research in the field of news recommendation assumes aggregation of the existing content, or observation of the user activity. Our work does not depend on monitoring user behavior to make recommendation decisions, nor does it aggregate the content from other news articles. Based on the trends observed in mainstream media, it extracts news items from the semantic web, and generates final list of suggested items based on the newsworthiness of each item. The current state of the system would only allow to guide journalists what topics they should write, and which news items they should include in the article. Our desire is to extend the work further by exploring automated article composition algorithms to create human independent mainstream media outlet.

CHAPTER 6

Conclusion and Future Work

In this dissertation we first address the challenges exposed by the modern online video service systems. We succeeded in building the application of proactive caching. The novelty of the system is to design a cross-platform framework that explores the information in mainstream media to aid the caching servers to optimize the network performance. We further discuss the possibility of creating an automated information retrieval and suggestion system that could facilitate the creation of the news item suggestion framework. Our approach explores the correlations between semantic entities in the news articles, and by finding strong associations we build system to automatically retrieve the items that are expected to generate interest in the public. Further research is required to accomplish the task of creating artificial mainstream media outlet.

In the first part of our research we design a proactive caching scheme to aid quality of service in consuming online video. Our novel method explores the opportunities provided by the trends observed from mainstream media. Due to the media cross correlation, we can identify popular topics in the mainstream media, and pre-cache related videos on caching servers in a geographic region, where the videos are expected to attract large amount of traffic. Proactive caching is mainly designed for and is limited to accommodating the quality of service for videos that can be linked to the trending topics in mainstream media. Through our experimental results we confirm the superiority of using mainstream media over social media for detecting common trends with YouTube. We show the effectiveness and significant performance improvement of our proactive caching algorithm over the traditional methods. The framework presented in the dissertation provides opportunities for the ISPs to evaluate mainstream media, and more intelligently utilize the caching capacities. The proposed framework considers statistical co-occurrence of the terms that are composed by human journalists with semantic coherence and belong to the trending topics. Such approach may have some value for keyword suggestion systems. We also would like to deploy and test the proactive video caching system in a real network environment, to better evaluate the performance and measure the gains of our system.

As for news item suggestion framework, while the initial impression of our approach is positive, a more extensive evaluation is necessary. Our first evaluation is limited to certain number of news categories. As we have seen in the experimental results section, further division of news categories, into sub-categories is necessary to improve the accuracy of prediction. The process of news worthiness evaluation also requires more careful and thorough evaluation, and exploration. It would be desirable if panel of domain experts could manually compose the newsworthiness profiles, and generate the list of items that they think should have wide coverage in the mainstream media. This way our approach could be compared against the list of items generated by the human journalists. However, before we initiate this extensive, time-consuming evaluation, further optimization of our proof-of concept software is required, to avoid excess repetition of the process. Once the technological optimization is complete, we would be able to asses the usability of the system with media practitioners in order to discern improvements in terms of ease of use (e.g. visualization of the suggested items, historical data and statistical analysis of the reports).

Our approach provides an opportunity to gain a new perspective in the field of information retrieval and automated content creation. The automated approach allows for a large-scale analysis of published works. This way, it becomes possible to verify whether the theoretically determined news item selection criteria still hold for today's media landscape. It also provides the possibility to explore new criteria.

As experienced during the development process, comparing automated and human news item selection mechanisms might reveal new components not yet identified in the literature of journalistic studies. Human journalists have to deal with tremendous amount of information that needs processing. The automated tools will certainly help reduce the workload on journalists. One specific use case that has recently emerged is the citizen journalism. As users become more informed, they often participate actively in news content generation. Examples of this sort of journalism are often found in social media platforms. We must also mention that traditional journalism has been defined by the set of rules, agreed upon the professional journalist. Nowadays, the increasing number of citizen journalists may follow different guidelines for news article composition. While we explore the associations and correlations in the articles posted during the past three decades, the number of digital articles is very limited. With the availability of the Internet many citizen journalists are contributing and generating large amount of content on a daily basis. It is expected that these citizen journalists will use a distinctly different set of criteria to determine whether something is newsworthy to them. Our approach allows to investigate this objectively.

REFERENCES

1. A. Balmash and M. Krunz, “An Overview of Web Caching Replacement Algorithms,” *IEEE Communications Surveys*, Vol. 6, No. 2, 2004.

2. A. Lobzhanidze and W. Zeng, “Proactive Caching of Online Video by Mining Mainstream Media,” *IEEE International Conference on Multimedia and Expo*, 2013.

3. A. Lobzhanidze, W. Zeng, P. Gentry, A. Taylor, “Mainstream Media vs. Social Media for Trending Topic Prediction – An Experimental Study,” *IEEE Consumer Communications & Networking Conf*. 2013

4. A. Rajaraman, J.D. Ullman, "Data Mining", *Mining of Massive Datasets*. pp. 1–17, 2011

5. A. Tan, and C. Tee, “Learning User Profiles for Personalized Information Dissemination,” *Proceedings of 1998 IEEE International Joint conference on Neural Networks*, pp. 183- 188, May 1998

6. A.S. Das, M. Datar, A. Garg, S. Rajaram, "Google news personalization: scalable online collaborative filtering," *Proceedings of the 16th international conference on World Wide Web*, 2007

7. Abhari A, Soraya M. "Workload generation for YouTube," Multimedia Tools Applications 2010.

8. B. A. Huberman, D. M. Romero and F. Wu. “Crowdsourcing, Attention and Productivity,” *Journal of Information Science*, 2009

9. B. OConnor, M. Krieger, and D. Ahn, “Tweetmotif: Exploratory search and topic summarization for twitter,” in *Inter. Conf. on Weblogs and Social Media*, 2010.

10. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *SIGIR*, 2010.

11. Bell, A.: The language of news media. Blackwell Oxford (1991)

12. Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.

13. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation," *the Journal of machine Learning research 3* (2003): 993-1022.

14. Breslau M, Baum L, Molter G, Rothkugel S, SP. "The implications of Zipf’s law for web caching," *Third international conference on web caching*, 1998.

15. Carley, K. M. "Dynamic network analysis," In *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 133-145). Committee on Human Factors, National Research Council. (2003.)

16. Cha M, Kwak H, Rodriguez P, Ahn Y-Y, Moon S. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," *ACM SIGCOMM conference on Internet measurement* (IMC), ACM, 2007.

17. Cheng X, Dale C, LJ. "Understanding the characteristics of Internet short video sharing: YouTube as a case study," *ACM SIGCOMM conference on Internet measurements*, 2007, pp. 28–36.

18. D. B. Masi and M. Fischer, "Modeling Internet Service Provider Backbone Networks," *The Tele-communications Review*, 2007.

19. D. Billsus, D. and M. J. Pazzani, "A hybrid user model for news story classification," In *Proceedings of the Seventh International Conference on User Modeling*, 1999

20. D. Billsus, D. and M. J. Pazzani, "User Modeling for Adaptive News Access, User Modeling and User- Adapted Interaction," *User Modeling and User-Adapted Interaction Journal*, v.10 n.2-3, p.147-180, 2000

21. D. Mimno, H.M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," *EMNLP*, 2011.

22. D. Newman, J. H. Lau, K. Grieser , T. Baldwin, "Automatic evaluation of topic coherence," Human Language Technologies: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p.100-108, June 02-04, 2010, Los Angeles, California

23. D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *4th Inter. AAAI Conf. on Weblogs and Social Media*, 2010.

24. D. Sontag and D. Roy. "Complexity of inference in Latent Dirichlet Allocation," *NIPS*, pp. 1008–1016, 2011

25. De Nies, T., Coppens, S., Van Deursen, D., Mannens, E., Van de Walle, R.: Automatic discovery of high-level provenance using semantic similarity. *In: Proceedings of the 4th International Provenance and Annotation Workshop IPAW* 2012, LNCS 7525, Springer, Heidelberg. (2012) 97–110

26. Devleeschauwer D, Laevens K. "Performance of caching algorithms for IPTV on-demand services," *IEEE Transactions on Broadcasting*, 2009.

27. Dumais, Susan T. "Latent semantic analysis." *Annual review of information science and technology* 38.1 (2004): 188-230.

28. F. Figueiredo, F. Benevenuto, and J. Almeida. "The tube over time: Characterizing popularity growth of youtube videos," *Conference of Web Search and Data Mining*, 2011.

29. F. Figueiredo, F. Benevenuto, J. M. Almedia, "The Tube over Time: Characterizing Popularity Growth of YouTube Videos," *WSDM*, 2011

30. Galtung, J., Ruge, M.: The structure of foreign news. *Journal of peace research* 2 (1965) 64–90

31. Girvan, Michelle, and Mark EJ Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences* 99.12 (2002): 7821-7826.

32. Golub, Gene H., and Christian Reinsch. "Singular value decomposition and least squares solutions," *Numerische Mathematik* 14.5 (1970): 403-420.

33. Guo L, Tan E, Chen S, Xiao Z, Zhang X. "The stretched exponential distribution of Internet media access patterns," *27th ACM symposium on principles of distributed computing*, 2008.

34. Harcup, T., O'neill, D. "What is news?" Galtung and Ruge revisited. Journalism studies 2 (2001) 261–280

35. Hart, P. E.; Nilsson, N. J.; Raphael, B. "Correction to "A Formal Basis for the Heuristic Determination of Minimum Cost Paths"". *SIGART Newsletter* 37: 28–29, 1972

36. http://bit.ly/15FPcux

37. http://dbpedia.org

38. http://developer.nytimes.com/docs/read/most_popular_api

39. http://digital-diva.tumblr.com/post/25083966633/how-much-data-is-generated-every-minute

40. http://en.wikipedia.org/wiki/2014_Russian_military_intervention_in_Ukraine

41. http://en.wikipedia.org/wiki/Georgian-Russian_war

42. http://en.wikipedia.org/wiki/List_of_wars_1945–89

43. http://en.wikipedia.org/wiki/News_values

44. http://en.wikipedia.org/wiki/Twitter

45. http://nlp.stanford.edu/software/

46. http://royal.pingdom.com/2008/04/11/map-of-all-google-data-center-locations/

47. http://wsn.rnet.missouri.edu/archives

48. http://wsn.rnet.missouri.edu/dataset

49. http://www.ap.org/products-services/news-archive

50. http://www.fastcompany.com/1489676/twitters-and-youtubes-trends-2009-couldnt-differ-more

51. http://www.nytimes.com/ref/membercenter/nytarchive.html

52. http://www.opencalais.com

53. http://www.opencalais.com/documentation/linked-data-entities

54. http://www.reuters.com/resources/archive/us/

55. http://www.sysomos.com/reports/youtube/

56. http://www.youtube.com/channel/HCPvDBPPFfuaM

57. http://www.youtube.com/channel/HCPvDBPPFfuaM/about

58. https://developer.ap.org/ap-content-api

59. https://sites.google.com/site/rsssources/

60. https://wordnet.princeton.edu/wordnet/download/

61. Iacobelli, F., Nichols, N., Birnbaum, L., Hammond, K. “Finding new information via robust entity detection. In: Proactive Assistant Agents”, (PAA2010) *AAAI 2010 Fall Symposium*. (2010)

62. Ivana Bosnic, Katrien Verbert, and Erik Duval, "Automatic Keywords Extraction – a Basis for Content Recommendation," In *Proceedings of the Fourth International Workshop on Search and Exchange*, pp. 51–60, 2010.

63. J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Group-Lens: Applying collaborative filtering to usenet news," *Commun. ACM 40*, 77-87. 1997

64. J. Han, and K. Harcourt, “Data Mining Concepts and Techniques,” *Second edition*, pp. 257-260, 2004

65. J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, “FreeSpan: frequent pattern-projected sequential pattern mining,” *KDD*, 2000.

66. J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The Little Engine(s) That Could: Scaling Online Social Networks," *SIGCOMM Comput. Commun. Rev.*, 2010.

67. Jeon, J. Laverenko, V. and Mammatha, R., “Automatic Image Annotation and Retrieval Using Cross Media Relevance Models”, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

68. K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, 1972

69. L. Alsumait, D. Barbara, and C. Domeniconi, “On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking,” in *IEEE ICDM*, 2008.

70. M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," In *Proceedings of ACM SIGIR Workshop on Recommender Systems*, 1999

71. M. Hoffman, D. Blei, and F. Bach, "Online learning for latent dirichlet allocation," Nature, vol. 23, pp. 1–9, 2010.

72. M. Liljenstam, J. Liu, and D. Nicol, "Development of an Internet Backbone Topology For Large-Scale Network Simulations," *WSC*, 2003.

73. M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *SIGMOD*, 2010, pp. 1155– 1158.

74. M. Naaman, H. Becker, and L. Gravano. "Hip and trendy: Characterizing emerging trends on twitter," *J. Am. Soc. Inf. Sci*., 62:902–918, 2011.

75. M. Schudson, "The objectivity norm in American journalism", *Journalism*, 2001

76. M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: YouTube network traffic at a campus network: measurements and implications," *Multimedia Computing and Networking*, 2008.

77. McGregor, J.: Terrorism, war, lions and sex symbols: Restating news values. What's news (2002) 111–125

78. Moon, Todd K. "The expectation-maximization algorithm." *Signal processing magazine*, IEEE 13.6 (1996): 47-60.

79. N. Ammu, M. Irfanuddin, "Big Data Challenges," *International Journal of Advanced Trends in Computer Science and Engineering*, Vol.2 , No.1, pp. 613 - 615, 2013

80. N. Good, J.B. Schafer, J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, J. Riedl, "Combining collaborative filtering with personal agents for better recommendations," *Proceedings of the 16th national conference on Artificial intelligence* and the *11th Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, 1999

81. Newman, Mark EJ., "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences 103.23*, 2006

82. O'sullivan, T., Hartley, J., Saunders, D., Montgomery, M., Fiske, J.: Key concepts in communication and cultural studies. Routledge London (1994)

83. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proceedings of the 20th International Conference on Very Large Data Bases*, p.487-499, September 12-15, 1994

84. R. Carreira, J. M. Crato, D. Gonçalves, J. A. Jorge, "Evaluating adaptive user profiles for news classification," *Proceedings of the 9th international conference on Intelligent user interfaces*, 2004

85. R. Crane and D. Sornette. “Robust dynamic classes revealed by measuring the response function of a social system,” *National Academy of Sciences*, 105(41), 2008.

86. R.L. Cilibrasi, P.M.B. Vitanyi, "The Google Similarity Distance", *Knowledge and Data Engineering*, *IEEE Transactions on* (Vol. 19 , Iss. 3, 2007

87. Rizzo, G., Troncy, R.: NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In: (ISWC’11) *Workshop on Web Scale Knowledge Extraction* (WEKEX’11). (2011)

88. Rosario, Barbara. "Latent semantic indexing: An overview." Techn. rep. *INFOSYS* 240 (2000).

89. S Tong, D Koller, "Support vector machine active learning with applications to text classification", *The Journal of Machine Learning Research*, 2002

90. S. Asur, B. A. Huberman, G. Szabo, and C. Wang. “Trends in social media - persistence and decay,” In *5th International AAAI Conference on Weblogs and Social Media*, 2011.

91. S. D. Roy, T. Mei, W. Zeng, and S. Li, “Empowering Cross-domain Internet Media with Real-time Topic Learning from Social Streams,” *Proc. of IEEE Inter. Conf. on Multimedia and Expo*, 2012.

92. S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani., "Emerging topic detection using dictionary learning," *CIKM*, 2011.

93. S. Podlipnig and L. Boszormenyi, "A survey of web cache replacement strategies," *ACM Comput. Survey*, 2003.

94. Schultz, I.: The journalistic gut feeling. *Journalism practice* 1 (2007) 190–207

95. Shoemaker, P., Cohen, A. “News around the world. Content, practitioners, and the public,” Recherche 67 (2006) 02

96. Snijders TA. "Models for longitudinal network data," Models and methods in social network analysis. Cambridge University Press, New York, pp 148–161. (1997)

97. T. D. Nies, E. D’heer, S. Coppens, D. Van Deursen, E. Mannens, S. Paulussen, and R. Van de Walle, "Bringing Newsworthiness into the 21st Century", *CEUR In Proceedings for Scientific Workshops*, 2012

98. T. Griffiths, "Hierarchical topic models and the nested Chinese restaurant process," *Advances in neural information processing systems* 16 (2004): 106-114.

99. T. Joachims, "Learning to classify text using support vector machines: Methods, theory and algorithms", *Springer, 2002 edition*, April 30, 2002

100. T. Rodrigues, F. Benevenuto, V. Almeida, J. Almeida, and M. Gonc ̧alves. “Equal but different: A contextual analysis of duplicated videos on youtube,” *Springer Journal of the Brazilian Computer Society*, 16(3), 2010.

101. Tang W, Fu Y, Cherkasova L, Vahdat A. "Modeling and generating realistic streaming media server workloads," *Computer Networks* 2007.

102. V. Adhikari, S. Jain, and Z. Zhang. "Where Do You Tube? Uncovering YouTube Server Selection Strategy," *In Proc. of IEEE ICCCN*, 2011.

103. V. Adhikari, S. Jain, Y. Chen, Z. Zhang, "Reverse Engineering the YouTube Video Delivery Cloud," *In Proc. of IEEE Hot Topics in Media Delivery Workshop*, 2011.

104. Viktor Mayer-Schönberger, "Big Data: A Revolution That Will Transform How We Live, Work, and Think," *Eamon Dolan/Houghton Mifflin Harcourt, 1st edition*, March 5, 2013

105. W. Bekius, "Werkboek journalistieke genres," *Coutinho 1st edition*, August 12, 2003

106. Wallach, Hanna M. "Topic modeling: beyond bag-of-words." *Proceedings of the 23rd international conference on Machine learning. ACM*, (2006).

107. World Wide Web Consortium (W3C), "W3C Semantic Web Activity," November 26, 2011.

108. www.alexa.com

109. www.marketingcharts.com

110. www.nsnam.org

111. www.nychsj.com/5.html

112. www.w3.org

113. X. Cheng and J. Liu, "NetTube: Exploring Social Networks for Peer-To-Peer Short Video Sharing," *In Proc. of IEEE INFOCOM*, 2009.

114. X. Jin, A. C. Gallagher, L. Cao, J. Luo, and J. Han, "The wisdom of social multimedia: using flickr for prediction and forecast," in *ACM Multimedia*, 2010.

115. Z. Wang, L. Sun, C. Wu, and S. Yang, "Guiding Internet-Scale Video Service Deployment Using Microblog-Based Prediction," *In Proc. of IEEE INFOCOM Mini-Conference*, 2012.

116. Z. Wang, L. Sun, C. Wu, and S. Yang, "Guiding Internet-Scale Video Service Deployment Using Microb log-Based Prediction," *IEEE INFOCOM*, 2012.

117. Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang, "Propagation-Based Social-Aware Replication for Social Video Contents," *ACM Multimedia*, 2012.

VITA

Alex Lobzhanidze is a recent graduate from the University of Missouri (MU) where he received a PhD degree in Computer Science. Prior to pursuing the PhD degree he obtained his Master's degree from the University of Missouri, and Bachelor's degree from Kutaisi State Technical University. After graduating from MU he will join Amazon to work as a software development engineer.

During studies at the University of Missouri, Alex worked as an intern at Technicolor research lab in Princeton, NJ. He assisted with writing a proposal to extend the International Standard Organization Base Media File Format and MPEG-2 Transport Stream to support multi-component media content HTTP streaming. He participated in developing the demo application. He also worked as a Research & Development Engineer at the Center for Geospatial Intelligence. His tasks included analysis, processing, and indexing of high dimensional data. Under the mentorship of Dr. Grant Scott he developed secure UI components for the access and monitoring of an image catalog database. Additionally, Alex also worked as a Teaching Assistant at MU, and held Research Assistant position at the Mobile Networking and Multimedia Communications lab, led by Dr. Wenjun Zeng.

As MU student, Alex was also involved in professional and social activities. For two years he served as the president of the Computer Science Graduate Student Council. He held the position of the Committee Chair at MU International Club. In recognition of his service he received the Excellence in Leadership award from the Missouri International Student Council.

After graduation, Alex will become a member of Amazon's Search and Discovery team. His main tasks will include exploring patterns of customer's repeated purchases to build new recommendation systems, and optimize the existing ones. As part of the team he will also be involved in creating innovative solutions to measure and improve the quality of information within Amazon's product catalog and influence the way millions of customers discover and buy products at Amazon.