Leveraging User-Generated Content for Enhancing and personalizing News Recommendation.

Youssef Meguebli

CentraleSupélec

N° d'ordre : 2015-07-TH

CentraleSupélec

ECOLE DOCTORALE STITS
« Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

THÈSE DE DOCTORAT

DOMAINE : STIC
Spécialité : Informatique

Soutenue le 27 Mars 2015

par :

Youssef Meguebli

Leveraging User-Generated Content for Enhancing and personalizing News Recommendation

Directeur de thèse :	Bich-Liên DOAN	Professeur Adjoint (CentraleSupelec)
Co-directeurs de thèse :	Fabrice POPINEAU	Professeur (CentraleSupelec)
	Mouna KACIMI	Maître de conférences (Free University of Bozen)

Composition du jury :

Rapporteurs :	Anne BOYER	Professeur (Université de Lorraine)
	Mohand BOUGHANEM	Professeur (Université de Toulouse 3)
Examinateurs :	Benjamin PIWOWARSKI	Chargé de recherche (CNRS)
	Josiane MOTHE	Professeur (Université de Toulouse 2)
	Nicolas SABOURET	Professeur (Université Paris-Sud)

Abstract

Online news websites are becoming one of the most popular and influential social media platforms allowing people to easily access information about daily life topics, share their opinions on different issues, and give feedback on published content. The tremendous increase of published news requires effective recommendation techniques that help users to find interesting news articles that match with their interests. Thus, users are continuously encouraged to participate to online news websites and keep sharing their opinions, which represent a valuable source of social information. In this thesis, we have investigated how to exploit user-generated-content for personalized news recommendation purpose. The intuition behind this line of research is that the opinions provided by users, on news websites, represent a strong indicator about their profiles. By mining such content, we can extract valuable information about the domains of interests of users, their inclination towards a certain version of news articles, their political orientation, their favorite sport teams, their preferences, and many other interesting features. Furthermore, such content can also be used to enrich the content of news articles, particularly for those describing controversial news articles that can reveal various aspects that are not well described or even not found in their content. Thus, user-generated-content is the core component of our work. This thesis is divided into three main parts, as described in the bellow, which represent the different steps of developing a news recommendation system based on user-generated-content.

In the first part, we have developed a fine-grained model that captures both user's and article profiles. The profile of each user is extracted from all the opinions and the reactions that are provided on the news websites, while the profile of an article is extracted from its content. A profile is mainly composed of the entities, the aspects, and the sentiments expressed in the corresponding content. While the extraction of entities is a well-established problem, aspect extraction often relies on supervised techniques, which are domain dependent. For a more general solution, we have proposed an unsupervised technique for aspect extraction from opinions and articles. We have investigated two types of models in three different applications. The first model, called a sentiment-dependent profile, exploits the sentiments related to each entity and aspect to define the orientations of users towards a specific trend. For this purpose, we have built a knowledge base of trends, more specifically of political orientations, that guides the extraction of profiles in an unsupervised manner. We have assessed the accuracy of the extracted profiles on two datasets crawled from CNN[1] and Al-Jazeera[2] and the results show that our approach gives high quality results. The second model, called a sentiment-independent profile, focuses only on entities and aspects and is used on the purpose of news recommendation. This model was used to define both users' interests and the content of news articles. We have test it on a large test collection based on real users' activities in four news websites, namely The Independent[3], The Telegraph[4], CNN and Al-Jazeera. The results show that our model outperforms baseline models achieving high accuracy. In the third application, we have used a combination of the two former models for news recommendation purpose: the sentiment-independent profile model to define users' interests is combined with the sentiment-dependent profile model to describe the content of news articles. The main goal of this application was to give a method that deal with the problem of redundancy on the list of recommended news articles. For this purpose, we have used a diversification model on news articles profiles to reduce the redundancy of the list of recommended news articles.

[1] www.cnn.com
[2] www.aljazeera.com
[3] www.independent.co.uk
[4] www.telegraph.co.uk

We have tested our approach on real users' activities on four news websites CNN, Al-Jazeera, The Telegraph, and The Independent. The results show that diversification improve the quality of recommended news articles.

In the second part, we have focused on how to enrich the article profiles with user-generated-content. The idea behind is to exploit the rich structure of opinions to tailor the articles to the specific needs and interests of users. The main challenge of this task is how to select the opinions used for profile enrichment. The large number and the noisy nature of opinions calls for an effective ranking strategy. To achieve this goal, we have proposed a novel-scoring model that ranks opinions based on their relevance and prominence, where the prominence of an opinion is based on its relationships with other opinions. To find prominent opinions, we have (1) suggested a directed graph model of opinions where each link represents the sentiment an opinion expresses about another opinion (2) built a new variation of the PageRank algorithm that increases the scores of opinions along links with positive sentiments and decreases them as well as links with negative sentiments. We have tested the effectiveness of our model through extensive experiments using three datasets crawled from CNN, The Independent, and The Telegraph news websites. The experiments showed that our scoring model selects meaningful and insightful opinions.

In the third part, we have focused on the development of a recommendation technique that exploits the results of the previous part and use them to enrich the content of news articles. We have tested various methods of leveraging opinions on the content of news articles. Concretely, we have worked on two main aspects. Firstly, we have only focused on sentiment-independent profiles, which consist on entities and aspects, and investigated of thoroughly the profile construction process. Secondly, we have enhanced the opinion ranking strategy described earlier by proposing an opinion diversification model based on authorities, semantic and sentiment diversification. The goal is to deal with redundant information and have a wide coverage of topic aspects. We have tested our approach by running large experiments on four datasets crawled from CNN, The Independent, The Telegraph, and Al-Jazeera. The results show that our model provide effective recommendation, particularly when enriching the content of news articles with a diversified set of opinions.

Acknowledgements

This thesis would not have been possible without the enormous support that I received during my PhD.

I am deeply thankful to Dr. Bich-liên Doan for guiding me while giving me the freedom to choose my research topic and explore various ideas. Her supervision has taught me how to combine ideas from different areas to inspire new creations, while her attention to detail has enabled this work to flourish.

I thank Professor Fabrice Popineau for his continuous support and gorgeous collaboration during the last three years. Further, he commented on various drafts of this thesis and gave me a lot of freedom to develop ideas.

Research under Dr. Mouna Kacimi supervision has been one of my most fruitful experiences. His passion in research is really admirable. I am really thankful to her for the time she devoted to me and the innumerable lessons she tought me. I would also to thank her for inviting me to visit the KRDB Research Centre for Knowledge and Data group at Free University of Bolzano from February to July 2014.

I would also like to thank the reviewers of my thesis, Professor Anne Boyer, University of Lorraine, and Professor Mohand Boughanem, University of Toulouse Paul Sabatier, for their thoughtful and detailed comments. Likewise, I am grateful to the examiners of my thesis, Professor Nicolas Sabouret, University of Paris-Sud, Professor Josiane Mothe, University of Toulouse Paul Sabatier, and Dr. Benjamin Piwowarski, researcher at CNRS.

I would like to thank my parents, whose love and support made it possible for me to complete this work. They have always encouraged me to follow my dreams. Many thanks to all members of my family, my brothers Yassine, Issam and Mohamed, my grandparents, my close Faycal, Ibrahim, Zied, Sofienne, and Ali.

I am greatly thankful for the friendship and support in surviving the PhD years to Haykel, Ezzedine, Atif, Arjumand, Noura, Hiba, Dorsaf, and all the CentraleSupelec people, which are too many to mention.

Preface

This thesis is the Ph.D. work done under the supervision of Dr. Bich-Liên Doan, Dr. Fabrice Popineau who work at Supelec in Gif-Sur-Yvette-France and Dr. Mouna Kacimi at the Free University of Bolzano in Italy and from January 2012 to December 2014. It focuses on how opinions can be exploited to enhance the accuracy of personalized news recommendation.
The work presented in this thesis resulted in a number of publications which are listed below:

1. Youssef Meguebli, Mouna Kacimi, Bich-liên Doan, and Fabrice Popineau. How hidden aspects can improve recommendation? In Social Informatics, pages 269–278. Springer, 2014.
2. Youssef Meguebli, Mouna Kacimi, Bich-Liên Doan, and Fabrice Popineau. Unsupervised approach for identifying users' political orientations. In Advances in Information Retrieval, pages 507–512. Springer, 2014.
3. Youssef Meguebli, Mouna Kacimi, Bich-Liên Doan, and Fabrice Popineau. Building rich user profiles for personalized news recommendation. In Proceedings of 2nd International Workshop on News Recommendation and Analytics (UMAP Workshops), 2014.
4. Youssef Meguebli, Mouna Kacimi, Bich-Liên Doan, Fabrice Popineau, "Exploiting Social Debates for Opinion Ranking". To appear in the proceedings of KDIR 2014, the International Conference on Knowledge Discovery and Information Retrieval, 2014.
5. Youssef Meguebli, Mouna Kacimi, Bich-Liên Doan, Fabrice Popineau, "Stories Around You: a Two-Stage Personalized News Recommendation". To appear in the proceedings of KDIR 2014, To appear in the International Conference on Knowledge Discovery and Information Retrieval, 2014.
6. Youssef Meguebli, "Classification non supervisée de profils d'utilisateurs en fonction de leurs orientations politiques". In the proceedings of Coria 2014, Conférence en Recherche d'Information et Applications 2014, 2014
7. Jao Barros, Zeno Tofano, Youssef Meguebli, and Bich-Liên Doan. Contextual query using Bell tests. In Quantum Interaction, pages 110-121. Springer, 2014.
8. Youssef Meguebli, Fabrice Popineau, and Bich-Liên Doan, Yolaine Bourda. A novel architecture for a smart information retrieval system based on opinions engineering. In Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, Cambridge, Massachusetts, 2012.

Contents

List of Figures

List of Tables

Part I

Overview and Background

Chapter 1

Introduction

In this chapter, we explain the motivations behind the research issues addressed in this thesis together with a description of the research questions. Further, we highlight the main contributions of this work and we conclude by presenting the structure of this thesis.

1.1 Motivation

Emergence of news media

The need for news and announcements on daily life events has been satisfied in various ways in different eras and cultures by diverse technical means. In ancient Rome, announcements were carved on stone or metal and were posted in public places. With Gutenberg's invention of the movable printing press, the printed word became a dominant medium for mass communication. Thus, newspapers, being a product of the printing press and the only medium for mass communication, enjoyed the privilege of monopolizing the mass media market for centuries until the advent of radio and television. The first attested newspaper was published in 1605 by Johann Carolus in *Strasbourg* and the first English daily, the Daily Courant, was published from 1702 to 1735 [ALM]. With the advance of the World Wide Web (WWW), reading news has changed from a traditional model of news consumption using physical newspaper to a digital model of consumption using news websites. Consequently, the newspaper industry has joined the Internet to increase its number of readers and advertisers. Online newspapers occurred in the middle of the 1990*s* when McAdams created an online version of The Washington Post (McAdams, 1995) [McA95]. At present, users spend more and more time on news websites as the web provides access to news articles from thousands of sources around the world. Even more, news websites have become one predominant source of information and opinion about public affairs and daily life events.

In the U.S., the digital audience with U.S. news websites increased 17% from 142 million in October 2013 to 166 million in October 2014 (More details in figure 1.1). It is an additional 24 million users. In the ten months from January to October 2014, the number of unique visitors engaged with newspaper digital content increased by 20 million. In addition, eight in ten (80%) of U.S. adults who were online in October 2014 engaged with news websites content[1]. Furthermore,

[1]The data is recorded based on an analysis of data gathered from more 300 U.S. news websites done by the media measurement firm comScore through its Media Metrix Multi-Platform and Mobile Metrix services.

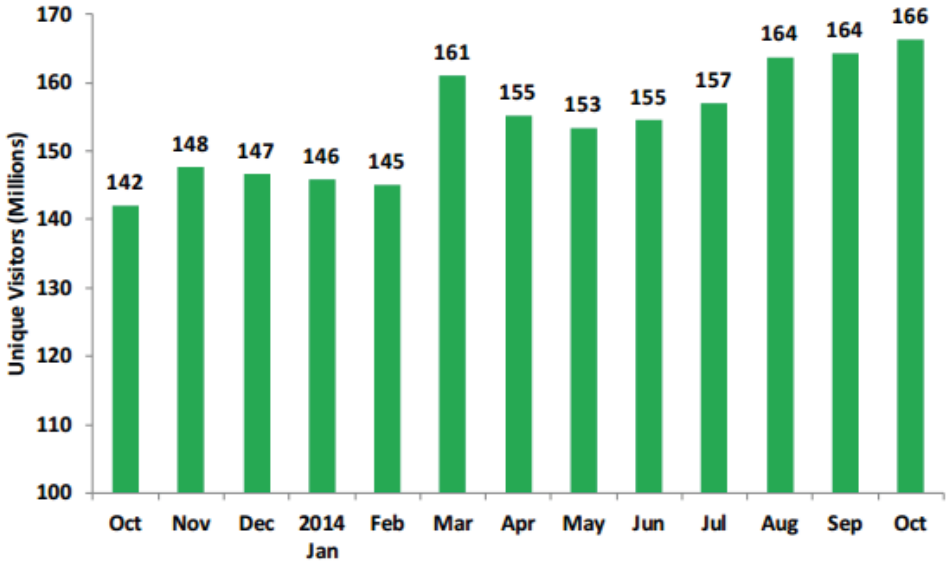

Figure 1.1: News website audience in the U.S. from October 2013 to October 2014

CNN gathered more than 840 unique millions visitors around the world and more than 1.9 billion viewed pages on its online news website in 2013. In the same year, the French news website Le point.fr received more than 200 unique millions visitors, implying around 17 million in average per month.

In last decade, there was a growing amount of new online websites resulting an exponential amount of daily published news articles. Moreover, those news websites cover diverse topics such as politics, sports, culture and entertainment. With so many online news articles, it becomes crucial to help users finding interesting articles that match with their interests and thus deal with the problem of information overload in general.

Two key challenges stand out in particular for online news websites :(i) helping users find news articles that match with their interests and preferences as much as possible, and more significantly (ii) keeping regular users and their participation, by making use of both news content and users' information.

Personalized news recommendation

With information overload problem, recommendation of news articles has become a promising service for news websites to improve user's satisfaction, as the Internet provides fast access to real-time information from multiple sources around the world. In last years, news recommendation became an important service available on most news websites [LWL$^+$11]. Such service generates recommendation to help users to discover relevant information and also to simplify personalized online information access using various types of knowledge and data on users, news articles, and previous transactions stored in customized databases. In other words, news websites recommend news articles that might be of interest or value to the user based on his/her interests and his/her previous activities. Thus, accurate profiles of users' current interest and news articles are crucial for the success of recommendation systems.

Despite a few recent advances, personalized news recommendation remains challenging for at

least three reasons. Firstly, the relevance and recency of news articles change dramatically over time, which differentiates news articles from other web objects, such as products and movies, making traditional collaborative filtering methods inapplicable [KR12, CTFLH12]. Secondly, accuracy is the missing element in profile models used to describe users' interests and the content of news articles. Thus, it becomes difficult to predict the list of recommended news articles accurately. Thirdly, in many cases, the content of news articles is not enough to have a clear idea on the subject of the news article, in particular if that user is not familiar with the news article topic. Thus, it is not obvious for a user to discover all related aspects to a given news article using only its content. Up to now, popularity (e.g., most shared and most commented), recency,

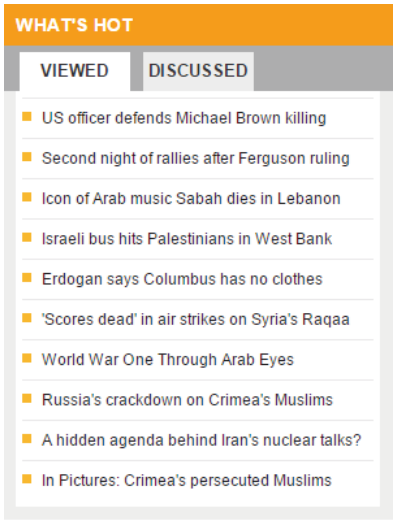

Figure 1.2: What's Hot on Al Jazeera English

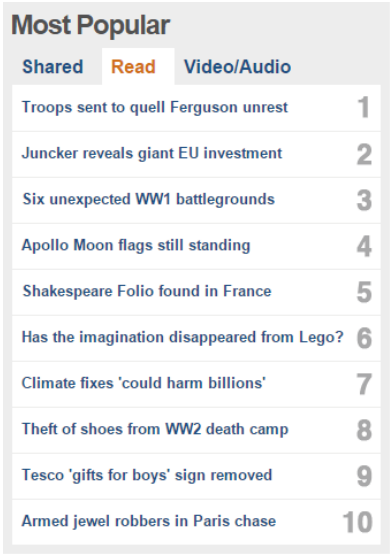

Figure 1.3: News Popularity on the BBC

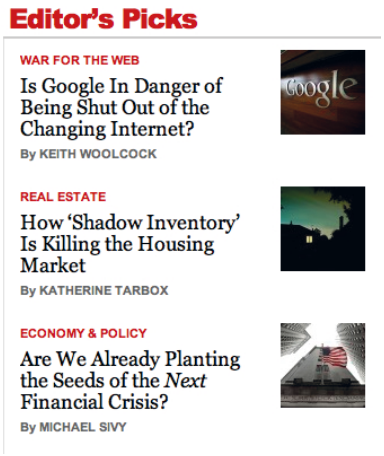

Figure 1.4: Editor's Picks on CNN

and (manual) editors' picks (based on daily hot topics) are still the most used techniques by popular news websites to recommend news articles.

Figures 1.2, 1.3, and 1.4 show screen-shots of related articles on popular news websites Al-Jazeera, CNN, and BBC, respectively. As it can be seen in the examples, a variety of tables that reflect different semantics such as "Most Viewed" and "Most watched" enable the exploration of articles related to the one currently being browsed.

User generated content

Previous years have seen more and more users freely expressing their opinions/sentiments and discussing around various topics in news websites. In fact, besides reading news articles, many news websites provide commenting areas for their users. An illustration of such spaces are shown in figures 1.5 and 1.6. Unlike the content provided by news websites publishers, such content contributed by users is collectively called the user-generated-content because it is created outside of professional practices. Nowadays, it is well known that the user-generated-content contains valuable information that can be exploited for many applications including trend detection [MK10], public mood and emotion mining [BMP11, DCCG12], interests and expertise mining [GAC+13, VJN13], identifying political orientation of users [MKDP14b], and personalized news recommendation [MKDP14a]. Furthermore, sometimes, the article content itself is not

enough to form a complete view over a topic. Thus, opinions are a valuable resource that complements the article and represents the "wisdom of crowds". In the news recommendation scenario, several states of the art approaches use the profiles of news articles to compare it with either the profile of users or other news articles profiles in order to select the list of news articles that should be recommended to a given user. Thus, the accuracy of the news article profile is crucial for an accurate news recommendation. Opinions might play a significant role on enhancing the quality of news articles profiles by revealing novel aspects which are not clearly present on the content of news articles. We called hidden aspects such aspects, which are extracted from the content of opinions and not clearly presented on the content of news articles. However,

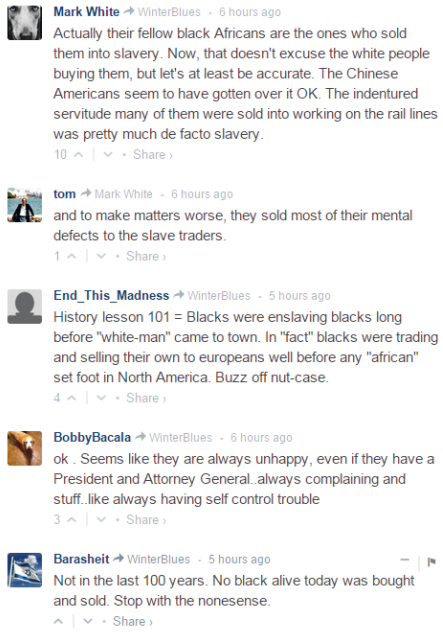

Figure 1.5: CNN opinions space

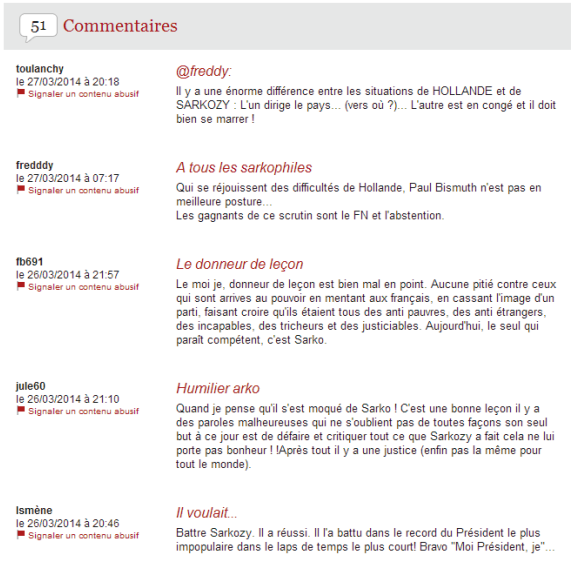

Figure 1.6: Le point.fr opinions space

such opinions might be a subject to a lot of noise and redundancy. Thus, a pre-processing step to organize and eventually extract only relevant opinions is very important to be able to use them on such task. Organizing and extracting only relevant opinions for a given news article can also be valuable for other scenarios. For instance, an archivist that needs to archive web information/resources (news articles) with complementary information (users' opinions). This process would help, e.g. future journalists who are trying to review past events, to gather as much information on a topic as possible, in order to present an objective view of the topic. In this thesis, we propose to investigate how opinions, which are the user-generated content we are interested in during this thesis, are organized and then exploited to enhance the effectiveness of personalized news recommendation.

1.2 Research questions

Based on the motivations stated in the previous section, our main goal during this thesis was to investigate how opinions can be exploited to enhance and personalize the accuracy of news recommendations.

To answer this major goal, we have identified three related research questions which are as follows:

- **Research question 1.** How should users' interests and news articles be described ?
  To answer to this question we need to investigate three main points:

  1. The first point is to examine the needed information to describe either users' interests or news articles. For example, which information should we exploit from the available data to be able to describe users' interests accurately.
  2. In literature, several techniques have been used to represent users' interests and news articles content. Three of them are the most used. Firstly, approaches that use a set of weighted keywords. Secondly, approaches that use ontologies to extract the most used concepts and thus describe users' interests and news articles content using a list of weighted concepts. Thirdly, approaches that use a set of weighted entities which represent well-defined concepts such as persons and locations. We should investigate more deeply which strategy we should exploit to represent the users' interests and the content of news articles accurately.
  3. The last point deals with investigating whether the sentiments that a user or the content of a given news article expresses towards an entity or an aspect should be taken into account or not, in order to enhance the quality of their profiles.

- **Research question 2.** How to deal with the problem of noise on opinions ?
  To answer this research question, we detailed the following related points:

  1. We need to define a set of features that will be used to measure the score of a given opinion. Those features can be deduced from the content of opinions, reaction to those opinions, or even be defined and based on the authors of these opinions.
  2. We need to investigate whether the expertise of the opinion author can play a significant role on defining its quality or not. For this reason, we can examine the importance of taking into account such feature on defining the score of opinions.
  3. In the last point, we investigate the results per topic and examine whether there is a difference on the quality between categories or not. In other words, the feature might be more relevant for some topics compared to others. Thus, it is important to select topics or categories that give best results and, subsequently, analyze the reasons behind it.

- **Research question 3.** How to improve the effectiveness of personalized news recommendation with opinions ?
  This general research question leads to the following detailed points:

  1. Firstly, we should examine the impact of enriching the profiles of news articles by its related opinions. We need to study the impact of such task on the accuracy of personalized news recommendation.

2. Secondly, we need to analyze the impact of using only a subset of ranked opinions to define the profiles of news articles. Subsequently , we should test the impact of removing noise from opinions on the accuracy of the profiles of news articles. We have to deal also with the problem of opinions redundancy. In fact, when only an opinion ranking strategy is applied we cannot deal with the problem of opinions redundancy.

1.3 Contributions

This thesis focuses on exploiting opinions in the context of personalized news recommendations. The main contributions are as follows:

- **Profile model.** We propose an effective profile model that formally describes and represents users' interests and news articles profiles. The model is based on three components:(1) entities which reflect well defined concepts such as persons, locations, organizations, objects, etc., their related aspects representing entity attributes or any abstract object, and the sentiments expressed for each entity and/or aspect. We evaluate our model through three different applications. Firstly, by using a sentiment-dependent model for users to identify their political leaning using an unsupervised approach. For this reason, we have also created a knowledge base taken from Wikipedia to define different political leanings and thus be able later to automatically classify users of news websites according to their political orientations. We have tested our approach on two groups of users from US and Egypt crawled from CNN and Al-Jazeera. The experiments showed that our approach provides high quality results to classify US users into Republican/Democrat leanings and Egypt users into secular/Islamist leanings. Secondly, we have tested a version of sentiment-independent model to define users' interests and news articles in the context of classic news recommendation. We have defined each profile by a set of weighted tuples <entity, aspect>. The user profile is defined through the set of opinions he/she provides in the news websites, while the article profile is extracted from its content. These profiles are then matched to recommend to each user the list of articles that correspond to the user's interests and the current article he/she is reading. Thirdly, we have used a combination of sentiment-dependent and sentiment-independent profiles on the context of news recommendation. We have used like in the second application, a sentiment-independent profile to define users' interests. However, we have used a sentiment-dependent profile to define the profiles of news articles mainly because it will be applied on diversification strategy whose main goal reducing the number of redundant recommended news articles. In fact, we define the dissimilarity between news articles when they discuss about different entities and/or aspects or have different sentiments about those entities and/or aspects. We have tested our approach on real users from CNN and Al-Jazeera and the results show that diversification improves the quality of recommendation. This contribution addressees the first research question which will be more discussed in chapter 3.

- **Opinion ranking.** Opinions on news websites can play an important role in improving the accuracy of news articles profiles. However, these opinions are not structured and might

diverge from the main topic of the news article or even be subject to a lot of noise. Thus, there is a need for selecting only relevant opinions to each given news article. To this end, we have proposed a novel scoring model that ranks opinions based on two components namely their relevance to a given aspect and their prominence. We define the prominence of an opinion using its relationships with other opinions. To this end, we (1) create a directed graph of opinions where each link represents the sentiment an opinion expresses on another opinion, then we (2) propose a new variation of the PageRank algorithm that boosts the scores of opinions along links with positive sentiments and decreases them along links with negative sentiments. We have tested the effectiveness of our model through extensive experiments using three datasets crawled from CNN, The Independent, and The Telegraph Web sites. Our experiments showed that these debates, enhanced by explicit feedbacks, are definitely valuable and should be taken into account for ranking opinions. We have shown that our model can effectively exploit the large amount of debates and reactions to select the best opinions that are relevant to the user's query. Our proposed approach achieves its best performance when the query topic is highly controversial or popular. It is also clear that our model does not perform well with categories and news articles related to unpopular topics. This contribution is our attempt to solve the second research question which will be more detailed in chapter 4.

- **Opinion diversification.** Diversification of opinions can play a significant role to increase the coverage of new topic aspects and overcome the problem of opinions redundancy. Thus, in case of using opinions to enrich the profile of news articles, diversifying opinions have a direct impact on enhancing the quality of news article profiles and subsequently on the accuracy of news recommendation. To this end, we have used a variation of an existing diversification approach [KG11]. In this approach, we consider two opinions as dissimilar if (1) they discuss different aspects, and/or (2) they exhibit different sentiments about a given aspect, including positive, negative, and neutral sentiments. This technique has proven to be effective in reducing the problem of redundancy and thus extracts various set of aspects from a set of opinions related to a given news article. This contribution is a part of our solution to solve the third research question which will be more detailed in chapter 5.

- **Leveraging opinions.** We deeply investigate on how opinions can be exploited on the context of personalized news recommendation. To this end, we have tested different approaches to leverage opinions on defining news articles. To describe either users' interests or news articles, we have used the sentiment-independent profile model described above. We have leveraged the content of news articles using different approaches including taking into account all opinions, a list of topk opinions using only an opinion-ranking strategy, or a set of prominent and diverse opinions. We obtain an improvement on the accuracy of personalized news recommendation on only two cases: (i) by employing only topk opinions using opinions ranking strategy, and (ii) using a set of diverse opinions. We have also observed that taking all opinions into account is not a good idea since opinions are subject to noise and redundancy and some of them might even deviate from the topic of interest, and thus this approach showed the worst performance. Our study was conducted through an extensive set of experiments using four real datasets extracted from CNN, The

Telegraph, Al-Jazeera, The Independent. It showed that diverse opinions record the best results compared to baseline approaches. This contribution is a part of our solution to solve the third research question which will be more discussed in chapter 5.

1.4 Thesis Outline

The rest of this thesis is structured as follows.

- **Background and state of the art.** In this chapter, we survey prior work, which is linked to the areas of information retrieval, recommender systems, and we conclude with a survey on opinion mining. Section 2.1 describes with concrete examples some basics of information retrieval. Section 2.2 highlight the related work in the area of recommender systems and more particularly on news recommendation and the key challenges that were not properly addressed in previous works. Section 2.3 concludes by surveying prior works on opinion mining, opinion ranking and opinion summarization.

- **Building fine-grained user and article profiles.** In this chapter, we first highlight the importance of having a fine-grained description of either users' interests or news articles in the context of news recommendations. In the section 3.2, we present previous techniques used to define users' interests and the content of news articles. We describe in the same section the three components of our model. In the section 3.3, we use a sentiment-dependent version of our profile model to identify the political orientation of users. We describe the experiments used to test the effectiveness of this profile model. The results show that our approach outperforms state of the art techniques used to identify the political orientation of users. In third section 3.4, we present a sentiment-independent version of our profile model used on the context of classic news recommendation. The model is based on a set of entities and their related aspects. We explain the conducted experiments on four datasets. Results show that our model gives high quality performance compared to baseline profile models. In the section 3.5, we describe the profile models used to describe users' interests and the content of news articles. We explain the diversification model used to diversify the list of recommended news articles. The results show that diversification of news articles using our profile give better results compared to entity-profile model.

- **Extraction of prominent opinions from news sites.** In this chapter, we present our approach of opinions ranking. Section 4.2 introduces the debate-based scoring model which explains the two main components of measuring the score of each opinion namely opinion relevance and opinion prominence. Section 4.3 explains the adopted algorithm to compute the score of opinion prominence. It is a variation of PageRank algorithm which gives more importance to opinions receiving positive reactions than opinions receiving negative reactions. Section 4.4 defines how we measure the expertise of each user. The user expertise is defined not only based on the explicit ratings, but also on implicit ratings the user has for his actions. Section 4.5 explains conducted experiments on three different datasets crawled from The Independent, The Telegraph, and CNN.

- **Enriching news articles using hidden aspects.** In this chapter, we explain how opinions can be used to improve the effectiveness of news recommendation. To this end, we test different strategies of leveraging opinions on the content of news articles. Section 5.2 describes the model used to leverage opinions on the content of news article for the purpose of news recommendation. Section 5.3 presents the technique used to extract hidden aspects from either users' opinions or the content of news articles. Section 5.4 explains the diversification model used to diversify the list of opinions related to each news article. Section 5.5 describes conducted experiments and results on four datasets extracted from four well-known news websites namely CNN, Al-Jazeera, The Telegraph, and The Independent.

- **Chapter 6 - Conclusions:** In the last chapter, we give an answer to the research questions, and we finalize this dissertation by presenting planned future directions of research.

Chapter 2

Background and State-of-the-art

In this chapter, we briefly describe fundamental techniques in the research area of information retrieval, which are useful for understanding our contributions in the following chapters. Then, we give an overview about recommender systems: we present the main strategies used for recommendation and we conclude by giving a study about the main news recommendation techniques used over the previous years. Finally, we present a brief study about the opinion mining field by giving an overview of the state-of-the-art techniques on sentiment analysis, opinion summarization, and finally opinion ranking.

2.1 Information Retrieval

2.1.1 Indexing

In order to efficiently retrieve documents from a huge document-based corpus while answering to a given query, an indexing process should be done [SM83]. During indexing, data structures, also termed index, are created to summarize the content of all information items that will be searched for later. These data structures are generated for an efficient access to the list of documents containing the query term. In this thesis, we have used indexing to create index structures either for news articles and/or opinions. To deeply explain the indexing process, we select a sentence from a CNN news article and we apply indexing techniques on it. The chosen sentence was extracted from the first paragraph of a CNN news article entitled *Reporter's notebook: Scenes from the ground in Gaza*, and was published on *July, 27, 2014*. The content of this sentence is :

A temporary truce Saturday between Israel and Hamas provided a precious few hours for hundreds of people, who had fled the fighting, the opportunity to learn whether they had a home to return in Gaza

- **Tokenization:** Tokenization is the first step in the processing index. Accordingly, all tokens for a query text are identified based on some boundary conditions such as white-space characters. All punctuation, are also erased from the text during processing. Thereafter, the above text can be seen as:

  A temporary truce Saturday between Israel and Hamas provided a precious few hours for hundreds of people who had fled the fighting the opportunity to learn whether they had a home

to return in Gaza

- **Stopwords Removal:** Some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called Stopwords.
  [Luh57] explained how the resolving power of a world exhibits a normal distribution with respect to the rank of its frequency. These words, which are called stop words, include articles such as *a, an,* and *the*, prepositions like *at, by, in, to, from,* and *with* and conjunctions like *and, but, as,* and *because*.
  Meaning that, the more frequently a word is employed, the less informative it turns out. For example, the words "with" or "the" cannot be employed to discriminate between documents as nearly all documents contained. Such words are commonly referred to stopwords, and are discarded from the list of potential indexing terms [BYRN99b]. [LHO05] showed that a stopword list can be expanded by looking for the most frequent or least informative terms in the documents. Moreover, removing stopwords serves also for reducing the size of the final index structures. After such process, the formal sentence might look as follows:

  temporary truce Saturday Israel Hamas provided precious hours hundreds people fled fighting opportunity learn home return Gaza

- **Stemming:** Most words, generally, show several forms (e.g., take took taken). Thus, the matching between queries and documents corpus can be attenuated if they use different forms even in case of semantic similarity between words. To that reason, a commonly known solution is to refer to these words back into their root forms, which is known as conflation. Conventionally, conflation is performed through what is known stemming, where syntactical suffixes are removed based on a set of logical transformation rules. A typical example of a stem is the word collect, which is the stem of collects, collected, collecting, collection, collections.
  [Lov68] presented the first stemming algorithm and this influenced much of the later work, among which Porter's stemming algorithm for English [Por80] is probably the best known. Stemmers now exist in different languages. In our thesis, we have used the well-known package Tartarus which is the release of Porter's Snowball project that gathers stemmer for 14 common languages. By applying Porter's stemming algorithm to our sentence, it is as follows:

  temporari truce saturday israel hama provid precious hour hundr peopl fled fight oppportun learn home return gaza

  It is to say that not all words are changed such as the words truce, precious and return, some are taken to their root form like the term learn, while some others are transformed into forms that do not correspond to real English words such as the words temporari and opportun. However, this is not a problem as users' queries will be equally transformed such that a mapping can be obtained. In most IR systems, a bag-of-words strategy is adopted to store the final text. In this case, the ordering of terms is ignored, instead, only the

frequency occurrences of tokens within the text is counted to determine how frequently each term occurs in the bag. Usually, the set of terms in a document with their respective frequencies can be pointed to as document-posting list.

- **Index Data Structures:** For efficient documents retrieval, the index structures need to be as compact as possible to provide fast random access and store meta information about documents. Inverted index [FBY92] is the core of any IR system. For each term, the inverted index contains a term-posting list, which lists the containing documents. In our context, the term-posting list contains either (i) containing opinions when the task is related to opinions such as our contribution on opinion ranking (More details in chapter 4), or (ii) containing news articles when the task is related to news articles such as our contribution on the context of news articles recommendation. This is the transpose of the document-posting list, which lists the terms of each document. Typically, unique identifier (docids) are allocated for each document in a collection. For each term, the posting list can be represented as a series of ascending integers representing the document identifiers (docids) and a series of small integers representing the term frequencies of the term in each document (tf). The size of inverted index dependents heavily on the number of documents that are indexed. In order to have low disc usage and fast access, compression is widely applied to the inverted index posting lists [CMO14, ZM06]. That is fulfilled by storing only the delta-gaps between each pair of docids [ZM06] in combination with a variable length encoding. Particularly, docids are stored into ascending order and afterwards the distances between docids are stored, these distances are usually much smaller than the original docid. As to derive benefits from the knowledge that we are storing small integers, an efficient variable length encoding like Elias gamma encoding [Eli75] is employed to minimize the amount of space required. Variable length encoding uses a variable number of bits of storage depending on the size of the query integer. The term frequencies will always consist of small integers (bounded by the document size in tokens) and so an encoding that is further optimized for storing very small integers like Elias unary encoding [Eli75] is often applied. In addition to that, other structures are created as to store the extra information needed to perform matching. These include:

  1. *Lexicon:* A structure that consists of detailed information regarding each specific term. Meaning that, it may hold the total number of documents in which it appears, or the total number of occurrences of the term in the collection. The lexicon points into the main inverted index by attributing to each term a pointer to its posting list.
  2. *Direct/forward Index:* The "inverse of the inverted index". This structure holds term information (term id and frequency) for each document in the collection [HSDL12]. Similarly to the inverted index, compression can be an option to save storage space. A direct index is called when a listing of all terms contained within a document is required.
  3. *Meta Index:* This index structure carries additional information about each indexed document, sometimes referred to as document meta data. The meta index eases look up on this information given the docid of the document in hand. For instance,

one might store the time at which the document was created, or the author of that document in the meta index. Generally, the meta index is used to store document information to be used later by the user. These structures represent the core of a search engine index [BP98]. Once these structures have been created for the query document corpus, afterwards the system is ready to answer to user's queries. However, in a large-scale Web search configuration, both the structure of the documents to be indexed and the efficiency of the indexing process itself need to be taken into account.

2.1.2 Query processing

To express their information need on the web, users generally use queries containing a few keywords which represent usually only one piece of information about what they are looking for. For instance, a recent study shows that the average length of the queries in the MSN search log [1], a sample of about 15 million queries collected over one month, is 2.4 words [BC09]. For this reason, two basic components were exploited in query processing to improve the quality of queries namely Query preprocessing and Query refinement. In query preprocessing stage, a query should be processed the same way as we have mentioned for the content of documents to be able to match with the index terms. For instance, a query can be tokenized, stop-words can be removed from the query, and also query terms can be stemmed or lemmatized. In most cases, a query is not pre-processed extensively as it only contains few number of keywords. The main goal of Query refinement is the reduction of the scope of search results or alternatively expanding the search to other related terms with the hope of improving precision. The query refinement is optional and dependent on IR application, and has as main goal improving retrieval performance by suggesting similar terms to the original user's query. For instance, the query containing the term plane cannot match with the document containing the term aircraft, and similarly the query containing the term car cannot match with the document containing the term auto because documents do not exactly contain the term queries. This is one of two main known problems in natural languages: synonymy and polysemy. Synonymy refers to a case where two different words have the same meaning while polysemy is the case where a term has several meanings like the term java which can refer to programming language, island in Indonesia, or a coffee.
In literature, two main strategies have been adopted to tackle these problems [MRS08] :

- **Global methods:** These methods are based mainly on reformulation of the original query by expanding it with other semantically similar terms, usually independently to the initial retrieved results. Like examples of global methods, we can mention query expansion/reformulation using thesaurus, spelling correction, and query suggestion.

- **Local methods:** These methods reformulate the original query by examining the initial results returned. Like examples of local methods, we can cite relevance feedback and pseudo relevance feedback.

Relevance feedback is a strategy which involves the user in improving the final results of the IR system. First, the user gets a list of initial results in response to a given user query. Then, the

[1] http://www.msn.com

user can provide feedbacks by labeling each document in the initial result set as relevant or not relevant. Finally, the user feedbacks are used to reformulate the original query and return the final results based on the modified query.
Pseudo relevance feedback does not require the involvement of the user. The topk retrieved documents are considered to be relevant to the query and without asking additional input from the user. Both pseudo relevance feedback and relevance feedback have proved they improve the retrieval effectiveness. However, they can conduct to query drift for some queries with too few relevant documents in the topk retrieved results.
In a response to a given query, the user receive a list of document that should contain information that match with the query terms. The relevance of each document depends on user's assessment, thus a document is considered as relevant if the user perceives it as containing information of value that match with his/her personal information need [MRS08]. Further, the degree of relevance between a document d and a query q is computed using an IR system and depends on the retrieval model that the system employs. More details about retrieval models are presented in the next section.

2.1.3 Document or Entity Retrieval

The goal of Document retrieval, which is the core process of IR, is to retrieve a ranked list of documents in response to a given user's query. Usually, ranking is done in decreasing order of predicted relevance. Document weighting model is used to perform the ranking of documents by taking into account the user's query and a document to produce a score for each document. The score is a prediction of the relevance for that document to the user's query [BYRN99b]. Note that relevance is subjective, varying from user to user [VH$^+$05] and no document weighting model is perfect for all cases. In fact, certain document weighting models may be appropriated for some kind of queries such as short queries or for specific tasks like homepage finding. Retrieval models differ from each other in many aspects including query interpretation, document representation, and document scoring and ranking algorithms employed. We describe in the following subsections three of the most known retrieval models.

Boolean Retrieval Model:

The boolean retrieval is the simplest IR model where the query is a combination of terms and boolean operators such as AND, OR and NOT. The document is represented by a bag of words and each term in the document is represented using binary weighting 1 or 0: 1 if the term exists and 0 if the term does not exist in the document. The degree of relevance is ignored as the boolean retrieval model assumes that there exist only two degrees of relevance: relevant and non-relevant.
In other terms, let f(d,q) be the function giving the relevance score. The score of this function will equal 1 if the document d is relevant to the query q and equal 0 if the document is not relevant to the query q.
For instance, in response to the boolean query *(Restaurant AND Paris) NOT Chinese*, the results should be these documents which contain both terms Restaurant and Paris but not the term chinese. Practically, the model retrieves all exactly matched documents with the query

term without ordering the documents. To have high quality results, the user should formulate a complex query which is too difficult for non expert users.

Vector Space Model (Term Frequency- Inverse Document Frequency (TF-IDF))

In vector space model, documents are ranked and retrieved according to their degree of relevance. The degree of relevance is measured, on the basis of the similarity between the query terms and the content of documents. The most common and used technique in practice is *tf-idf* [Sal71]. For a term m and a document d, tf is the term frequency of the term w, which is normalized by the total term frequency in d. Thus, tf is computed as follows:

$$tf(w,d) = \frac{freq(w,d)}{\sum_{j=1}^{n_d} freq(w_j,d)} \tag{2.1}$$

where $freq(w,d)$ is the term frequency of the term w in d and n_d is the number of distinct terms in d, and tf captures the importance of a term w in a document by assuming that the higher tf score of w, the more importance of w with respect to d. Intuitively, terms that convey the topics of a document should have high values of tf. idf is the inverse document frequency weight of a term w. It measures the importance of w with respect to a document collection. idf can be seen as a discriminating property, where a term that appears in many documents is less discriminative than a term which appears in a few documents. idf can be computed as:

$$idf(w) = log\frac{N}{n_w} \tag{2.2}$$

where n_w is the number of documents in which a term w appears, and N is the total number of documents in the collection. The *tf-idf* weigh of a term w in a document d can be computed as follows:

$$tf - idf(w,d) = tf(w,d) \cdot idf(w) \tag{2.3}$$

Finally, a query q and a document d can be describes as vectors of all terms in the vocabulary, given such as:

$$\vec{q} = \langle \psi_{1,q}, \ldots, \psi_{n,q} \rangle \tag{2.4}$$

$$\vec{d} = \langle \psi_{1,d}, \ldots, \psi_{n,d} \rangle \tag{2.5}$$

where $\psi_{i,q}$ is *tf-idf* weight of a term w_i in q and $\psi_{i,d}$ is *tf-idf* weight of a term w_i in d. The similarity of the term-weight vectors of q and d can be computed using the cosine similarity as follows:

$$\begin{aligned} sim(\vec{q},\vec{d}) &= \frac{\vec{q}\cdot\vec{d}}{\mid\vec{q}\mid \times \mid\vec{d}\mid} \\ &= \frac{\sum_{i=1}^{n} \psi_{i,q} \times \psi_{i,d}}{\sqrt{\sum_{i=1}^{n} \psi_{i,q}^2 \times \sum_{i=1}^{n} \psi_{i,d}^2}} \end{aligned} \tag{2.6}$$

Compared to the Boolean retrieval model, the main strengths of the vector space model are: (1) firstly, the degree of similarity allows partially matching documents to be retrieved, (2) secondly, it employs term weighting which increases the retrieval effectiveness, and (3) thirdly, allows for a fact and easy implementation. However, there are some weakness of the vector space

model over the Boolean retrieval model. Firstly, it makes no assumption about term dependency, which might lead to poor results [BYRN99b]. Moreover, the vector space model makes no explicit definition of relevance.
In other terms, there is no assumption whether relevance is binary or multivalued, which can impact the effectiveness of ranking models. We have used TF-IDF as a feature about opinion-query and news articles-query pairs in Chapter 3 , 4 and 5.

Probabilistic Model

In this section we introduce the probabilistic models where the first model was proposed by Robertson and Jones [RSJ88] in 1988. This model is based on the probabilistic theory to capture the uncertainty in the IR process. Thus, ranking of documents is done according to the probability of relevance between documents and a given query.
The two key assumptions are considered as the basis of this model and are: (i) the relevance of a document is a binary property, so, a document is either relevant or non-relevant. (ii) The relevance of a document does not depend on other documents.

Let's consider R and $\bar{R}$ respectively the set of relevant documents and the set of non-relevant documents for a given query q.
The similarity of q and a document d can be computed using the odd ratio of relevance as:

$$sim(d,q) = \frac{P(R|d)}{P(\bar{R}|d)} \tag{2.7}$$

To better explain the calculation, Baye's theorem is applied and gives the following formula:

$$\begin{aligned} sim(d,q) &= \frac{P(R|d)}{P(\bar{R}|d)} \\ &= \frac{P(R)\cdot P(d|R)}{P(\bar{R})\cdot P(d|\bar{R})} \\ &\approx \frac{P(d|R)}{P(d|\bar{R}} \end{aligned} \tag{2.8}$$

where $P(R)$ is the a prior probability of a relevant document, and $P(\bar{R})$ is the a prior probability of a non-relevant document. For a given query q, it is assumed that both prior probabilities are the same for all documents, so they can be ignored from the calculation. $P(d|R)$ and $P(d|\bar{R})$ are probabilities of randomly selecting a document d from the set of relevant documents R and the set of non-relevant documents $\bar{R}$ respectively. In the probabilistic model, a document d is represented as a vector of terms with binary weighting, which represents term occurrence or non-occurrence.

$$\overrightarrow{d} = \langle \psi_{1,d} \dots \psi_{n,d} \rangle \tag{2.9}$$

where $\psi_{i,d}$ is the weight of a term w_i in a document d, and $\psi_{i,d} \in \{0,1\}$. To compute $P(d|R)$ and $P(d|\bar{R})$, it assumes the Naive Bayes conditional independence [MRS08], that is, the presence or absence of a term in a document is independent of the presence or absence of other terms in

the given query. Thus, the similarity can be simplified as:

$$sim(d,q) \approx \frac{P(d|R)}{P(d|\bar{R})} \approx \prod_{i=1}^{n} \frac{P(w_i|R)}{P(w_i|\bar{R})} \tag{2.10}$$

where $P(w_i|R)$ is the probability that a term w_i occurs in relevant documents, and $P(w_i|\bar{R})$ is the probability that a term w_i occurs in non-relevant documents.

Using probability to model the relevance of documents makes the probabilistic model theoretically efficient compared to the the vector space model and Boolean retrieval model. Nevertheless, a drawback is an independence assumption of terms, which is contrary to the fact that any pair of terms can be semantically related. What is more important in our thesis is that the probabilistic model is very difficult to implement because the complete sets of relevant news articles and non-relevant news articles are not easy to obtain. In other terms, to compute $P(w_i|R)$ and $P(wi|\bar{R})$, it is necessary to guess prior probabilities of a term w_i by retrieving top-n relevant documents and then perform iterative retrieval in order to recalculate probabilities. This makes it tough to implement the model. In addition, the probabilistic model does not take into account the frequency of terms in a document which was very important on our different contributions. For this reason we haven't used this model in our different contributions.

2.1.4 Evaluation

IR systems are evaluated through two main aspects: Effectiveness and Efficiency. Effectiveness measures the the quality of the system relevance ranking while efficiency computes a system response time and space usage. In this thesis, we only focused on the retrieval effectiveness aspect and more specifically on how many relevant news articles or opinions are retrieved or recommended and at what ranks. In the IR research community, it is very popular to evaluate an IR system using a test collection like TREC datasets [GCC10, ZYM07, SHMO09] where the relevance judgments of some queries is already known. In other terms, these test collection contain usually a various document collections, a set of queries, and more importantly the relevance judgments for queries. Thus, using these collections they compute the effectiveness of the system on a query for which some relevant documents are known. In our thesis, we have suggested models that need special data such as nested opinions, opinions holders which are not available on public datasets. For this reason, we have crawled our proper data from a set of popular news websites such as *CNN*, *The Telegraph*, *The Independent*, and *Al-Jazeera*.

Evaluation measures:

Several evaluation measures have been suggested to evaluate IR systems. In section, we describe the main evaluation measures used in the context of Information retrieval and Recommender systems:

- **Precision and Recall.** To assess the effectiveness of an information retrieval or recommender system, two main related measures have been widely used: Precision and Recall.

Precision is the fraction of retrieved documents that are relevant and *recall* is the fraction of relevant documents that are retrieved. In other terms, precision measures how good our returned documents are and Recall measures how many correct documents are returned, in comparison to how many there were. Note that precision and recall are also used in a classification context as well as for ranking. Let's R be the set of documents and A be the set of retrieved documents (answer set) for a given query q. Precision and Recall can be computed as follow:

$$precision = \frac{\left|R \bigcap A\right|}{A} \tag{2.11}$$

$$recall = \frac{\left|R \bigcap A\right|}{R} \tag{2.12}$$

Also, metrics combining both precision and recall have also been proposed. For example, the metric F-measure is the weighted harmonic means of precision and recall and it is computed as follows:

$$recall = \frac{2 \cdot P \cdot R}{(P + R)} \tag{2.13}$$

In this thesis, we have used other metrics for measuring retrieval effectiveness. Precision at top-k documents, so called $P@k$, focuses on only top documents and it is easy to compute. For instance, precision at $top - 5$, 10 and 20 are denoted as $P@5$, $P@10$ and $P@20$ respectively. More thorough descriptions in retrieval evaluation can be found in [BYRN+99a, CMS10, MRS08].

- **Mean Average Precision.** Mean average Precision (MAP) has been used for several years to measure the effectiveness [VH06] of IR systems. It is an extension of Precision metric which take into account the position of relevant document in a list of retrieved documents, i.e. the position k in a list n of recommended items. More specifically, it measures the mean of the average precision (AP) values for all queries. The AP for a query is the average of all precision values computed after each document is retrieved [VH+05]. The mean average precision is computed as follows:

$$MAP = \sum_{q=1}^{Q} \frac{\sum_{k=1}^{n} Precision(R(q), k) \cdot Recall'(R(q), k)}{\left|Q\right|} \tag{2.14}$$

  Where $\left|Q\right|$ is the number of queries, n is the number of retrieved documents, k is the rank within the retrieved documents $R(q)$, $Precision(R(q), k)$ is the precision at cut-off k and $Recall'(R(q), k)$ is the change in Recall between ranks $k - 1$ and k.

- **(Normalised) Discounted Cumulative Gain.** Mean average Precision is built upon the precision metric, which is binary by nature, i.e. each document is considered relevant or not, it is not useful when documents are evaluated with respect to multiple relevance grades such highly relevant, relevant and not relevant. For this reason, Discounted Cumulative Gain (DCG) metrics were proposed [JK02]. These approaches assume that documents

with higher relevance grades are more relevant than those with lower grades and that highly relevant documents are most useful when returned in the top ranks. Hence, DCG measures are compatible with multi-graded assessments and are top heavy in nature. DCG is computed as follows:

$$DCG = rel \sum_{i=2}^{p} \frac{rel_i}{log_2 i} \tag{2.15}$$

Where rel, is the relevance of the document at rank i. However, DCG is not sufficient for a typical IR evaluation, because not all result sets for a query are of the same length, i.e. fewer documents than the rank cutoff may be retrieved [JK02].
To take into account for this, Normalized Discounted Cumulative Gain (nDCG) was proposed, that normalizes the cumulative gain across queries. This is achieved by dividing DCG by the ideal DCG, i.e. that which would have been achieved by the perfect ranking according to the relevance assessments.

2.1.5 Diversification

Another related area of research to our work is search result diversification which has been extensively investigated following two different approaches. The first one is *taxonomy-independent* where no knowledge base is used to diversify search results [ZL06, RD06, SMO10, GS09a, WZ09]. Some of the works falling into this category include the work by Gollapudi et al., [GS09a] that uses a diversification model combining both novelty and relevance of search results. Radilinski et al. [RD06] use query expansion to enrich search results generating more relevant documents for various interpretations. The second approach to result diversification is *taxonomy-based*. [AGHI09, CKC+08, CC09]. Representative works include the work by Agrawal et al., [AGHI09] which makes use of a taxonomy for classifying queries and documents and create a diverse set of results according to this taxonomy. Clarke et al., [CKC+08] focus on developing a framework of evaluation that takes into account both novelty and diversity. Carterette et al., [CC09] propose a probabilistic approach to maximize the coverage of the retrieved documents with respect to the aspects of a query. In our thesis, we have adopted the technique proposed by Kacimi et al., [KG11], a *taxonomy-independent* approach either to diversify opinions or news articles. It uses three components namely authorities, novelty, and sentiment diversification to rank opinions and shows that it provides high quality results compared to existing techniques.

2.2 Recommender Systems

The study of recommender systems is a relatively new one compared to research on information retrieval [GNOT92, MR09, MM09, AM03] and has increased dramatically driven by the growing interests of highly rated Internet sites such as *Amazon*, *YouTube*, *CNN*, *Netflix*, *Yahoo*, and *Tripadvisor*. Recommender systems are basically software tools and techniques aiming to suggest items for users [Bur07, MR09, RV97]. Item is a general term generally used to refer to what the system should recommend: it can be a product to buy, a music to listen to, or such as in our case an online news article to be read. Several ways to classify recommender systems have been

proposed [CGT12, Bur07, JBE+13, CCCT09, PKCK12].
[Bur07] presents a very popular taxonomy which formed a classical way for distinguishing different recommender systems. According to him, recommendation approaches are divided into five basic categories: content-based, collaborative, demographic, utility-based and knowledge-based [Bur02, Bur07]. Jeckmans et al. [JBE+13] also follows [Bur07] and considers collaborative, content-based, demographic, and knowledge-based filtering approaches as the basic recommender types. [CGT12] considers personalization as key taxonomic criterion to distinguish between recommender systems and thus classifies approaches into two categories: personalized and non-personalized recommendations. Non-personalized recommender systems suggest items without considering the user's profile, while personalized recommender systems are performed by taking into account users' interests and preferences using a user's profile. Several recent papers [CCCT09, PKCK12] tend to simply the classification of recommender systems into two categories, namely collaborative and content-based filtering. Others such as [AT05] add to these two categories the hybrid category which contains approaches that combine collaborative and content-based filtering techniques.
Given the various categorization schemes and the fact that algorithmic approaches are not the key objective of this dissertation, we have adopted the taxonomy in which recommender systems are categorized into three categories namely content-based, collaborative filtering, and hybrid approaches.

2.2.1 Collaborative Filtering

Collaborative filtering-based recommendation attempts to identify similarities between entities (users and/or items) in order to suggest items to read, to purchase or to examine. Approaches using the similarity between users are known as user-to-user (or also user-based) strategies, and approaches using the similarity between items are known as item-to-item (or also item-based) strategies.
In other words, user-to-user approaches are based on similar users liking similar items, while item-to-item approaches are based on users liking items similar to the items they have shown preference for [JBE+13]. Thus, in both approaches, the idea is to generate recommendations based on users' interests with similar tastes to the current user in the past, i.e. to ignore the content of the item and exploit collective preferences of the crowd [CGT12].
Various metrics that typically have statistical origins are used to compute the similarity between users to find the k nearest neighbors [BOHG13, JBE+13]. The most commonly used traditional metrics include Pearson correlation, cosine, adjusted cosine, constrained correlation, mean squared difference, and Euclidean [BOHG13].
The same similarity measures can also be used to calculate item-to-item similarities [CGN+11, BOHG13]. Given their central position and importance, developing metrics to calculate similarities between users and between items is a recurring theme in collaborative filtering research [BOHG13].
While the user-to-user approaches require at least some preference data from the user, item-to-item recommendations do not need a user profile for generating recommendations; in fact, they can be generated as soon as the user has expressed interest in just one item [SKR99]. This strategy is widely used by well-known news publishers such *CNN* and *Al-Jazeera* to recommend their

news articles, significantly in cases where their users are unknown. While the classic user-to-user collaborative filtering approach provides high-quality recommendations, the challenge for it has turned out to be its slow speed when the numbers of users grow into hundreds of thousands or millions [KR12, LSY03].

Item-to-item filtering, developed as an alternative to the user-to-user approach, has a significantly faster online response time, especially if the item relationships are pre-computed, in addition to offering slightly better recommendations [KR12, LSY03, SFHS07, CTLM08, CCCT09]. Item-to-item approach builds correlations between pairs of items and then generates recommendations by finding items that are similar to the set of items that the current user has (implicitly or explicitly) shown liking [KR12, SFHS07, SKR99]. The strengths of the user-to-user and item-to-item versions of the K-nearest-neighbor (kNN) algorithm can also be combined [BOHG13].

One major challenge for collaborative filtering is data sparsity, i.e. a sparse user-item matrix resulting from few users rating the same item [SFHS07, Bur02, SK09, KMM$^+$97, PKCK12]. Collaborative filtering depends on overlap in ratings (or other preference data) across users, and if few users have rated the same items, correlations either cannot be calculated or can be skewed resulting in very little overlap, and hence, neighborhood formation is challenging and correlation based on too few common ratings results in spurious correlations [SFHS07, Bur02, KMM$^+$97]. Consequently, as far as data distribution is concerned, collaborative filtering works the best when there are many items and many ratings per each item, there are more users rating items than items to be recommended, and users rate multiple items [SFHS07, Bur02]. Data sparsity also leads to cold-start issues, i.e. the system is unable to make meaningful recommendation because of the lack of preference data for new users, new items, and new communities [BOHG13, SFHS07, Bur02, SK09, KMM$^+$97]. However, cold-start problems can be made less severe by carefully selecting which items are given for a user to rate, as this can have considerable effects on how quickly good recommendations can be generated to the user [KR12]. Scalability problems [SFHS07, SKR01, SK09, PKCK12] are another major challenge within a collaborative filtering recommendation framework. In naive implementations, the time and memory requirements of user-to-user algorithms scale linearly with the number of users and ratings, making it impossible to use such approaches e.g. on Amazon.com [SFHS07, SK09]. Such challenges have been tackled e.g. with subsampling, using a subset of users selected prior to prediction computing, and clustering, comparing a user to a group of users instead of to individual users and then selecting the nearest neighbors from the clusters most similar to the user [SFHS07]. However, the problem has not been entirely solved, and with various implementation approaches, tradeoffs between scalability and prediction performance tend to persist [SK09].

What is more important in our thesis that collaborative filtering needs items to persist and tastes to persist [KR12]. If items change quickly, it is challenging to have many users rating them to create an overlap in ratings [KR12]. An example where the collaborative filtering approach has difficulties is news recommendation, as the approach requires overlapping ratings but news articles are most interesting when they are new and fresh [KR12, CTFLH12]. With regards to tastes, if tastes do not persist over time but change quickly, then older ratings are not useful for collaborative filtering, and, again, generating neighborhoods is challenging [KR12].

2.2.2 Content-based Filtering

While collaborative filtering is based on the assumption that people with similar tastes rate things similarly, content-based filtering is based on the assumption that items with similar objective features, or attributes, are rated similarly [SFHS07, BOHG13]. In effect, content-based recommenders suggest to a user items the content of which is similar to the content of the items for which the user has rated positively, or has otherwise shown preference for, in the past [CGN+11, CGT12, BOHG13]. In other words, a content-based approach learns a profile of the user's interests based on the attributes present in the items that the user has rated positively [Bur02]. For example, if the user has liked a web page with words like car, engine, and gasoline, the system will suggest more pages related to automobiles [BOHG13] or, in case of videos, the content can be the title, the actors, the director, and the genre of the liked videos, as videos can be characterized by them and they are readily available [CGT12]. In other words, the item similarity is calculated based on explicit content attributes associated with the items being compared [CGT12, JBE+13]. Compared to collaborative filtering, content-based recommendation has many benefits: It is easy to implement when used collections are already described by metadata. Further, it is more appropriate for contents that change quickly such as the case of news recommendation. One of the challenges is to identify and extract the attributes, or meta-data, of items that are most predictive [SFHS07, BOHG13]. Text domains work well for extracting keywords and such meta-data but some content, e.g. sound file content (rather than associated metadata, e.g. artist), can be hard to analyze [CLA+03, SFHS07]. If some attribute cannot be automatically extracted from the item, it has to be added by hand or it cannot be used in content-based filtering [SFHS07]. In addition to extracting attributes, a user profile must be built to know preferred items that can then be compared to other items for attribute similarity [SFHS07, PKCK12]. While the basic approach to content-based filtering bases its similarity analysis on term-by-item occurrences, neglecting the semantic structure of the content, more advanced approaches, such as latent semantic analysis, attempt to also exploit semantic features [CGN+11]). In addition to keywords and other meta-data that can be extracted, content-based systems are increasingly incorporating social information on items that users in Web 2.0 provide, such as tags, posts, and reviews [BOHG13].
Content-based filtering approach in its pure form has numerous shortcomings [Bur07, BOHG13]. In certain domains, generating attributes for items is challenging but content-based systems are, by their very nature, limited to the attributes that are explicitly associated with the items they are to recommend [BOHG13, Bur02]. For example, a content-based movie recommender is limited to written materials about a movie [Bur02]. Consequently, a content filtering model can only be as complex as the content to which it has access [SFHS07]. In addition, because these systems suggest items the content of which is similar to the content of the items that the user has shown preference for in the past, content-based approaches also suffer from overspecialization problem [SFHS07, CGT12, PKCK12, BOHG13]. The user is often already aware of the items that are suggested or he could have figured them out easily, by searching, for example, movies with the name of their favorite actor or books by their favorite author [SFHS07, CGT12, BOHG13].

2.2.3 Hybrid Filtering

The final category of recommender system algorithms is, in fact, hybrid recommender systems, which combine collaborative and content information. Multiple techniques are achieved to avoid the weaknesses that each collaborative and content-based technique has [Bur07] . Firstly, we are going to consider those approaches that require building separate recommender systems using techniques that are specialized to each kind of information used, and then combine the outputs of these systems. For instance, the resulting scores can be combined using a weighted approach [CGM+99] or voting mechanism [Paz99], switching between different recommenders [Kob94, LC08a], and filtering or reranking the results of one recommender with another [Bur02]. A different approach consists of combining both content and collaborative features. By means of this combination a single unified technique might be used regardless of the types of information used [BHK98, GM09]. In such systems, a careful selection of the features is needed. There have been some works on using boosting algorithms for hybrid recommendations [MMN02, PPM+06]. These works attempt to generate new synthetic ratings in order to alleviate the cold-start problem. These new ratings can be obtained using various heuristics, based on content information (for instance according to who acted in a movie) or demographic information. After injecting these new ratings into the user-item matrix along with actual user ratings, a collaborative algorithm is used. The use of aspect models [Hof04] has been also extended to many types of meta-data (e.g. actors, genres, and directors for movies) [SPUP01]. A similar approach has been also used for music recommendations [YGK+08] and online document browsing [PUPL01]. Also related, the hybrid Poisson-aspect model [HCH04] approach combines a user-item aspect model with a content-based user cluster. [JBE+13] calls recommender types that build on basic types as improved recommender types. However, the term hybrid appears more popular and is therefore adopted here. Jeckmans et al.'s [JBE+13] terminology is mentioned here simply to demonstrate and underline that even today there is no systematic consistent way of naming the filtering algorithms used in recommender systems.

To sum up, content-based and collaborative filtering are complimentary [SFHS07]. In domains where content is scarce or difficult to obtain, collaborative filtering can work efficiently because users take care of the evaluation and little needs to be known about the item beyond this [SFHS07]. On the other hand, content-based filtering can work without ratings, e.g. for new items, for high-turnover items (e.g. news articles), and huge item spaces (e.g. web pages), as items can be evaluated based on attributes, while collaborative filtering without preference data on an item is unable to consider it [SFHS07]. Approaches can also be combined in multiple ways, and no consensus exists on how to do it the best [SFHS07].

2.2.4 Recommending News articles

With the large volume of news events happening every day, recommending news articles has become a promising research direction and one of the most important applications for major content publishers such as CNN, Al-Jazeera, and The Telegraph [CFMH12, RF13, MKDP14a, GDH04, LP07, LHH+10, KBV09, ZYZ11]. In this section, we have also classified proposed news recommender systems into three different categories namely collaborative filtering, content-based, and hybrid approaches.

Collaborative filtering. Collaborative news recommender systems assume that users with similar rating behaviors in the past usually have similar preferences to new news articles. For example, if two users are interested in the same topic, they would read similar news articles relevant to this topic. Such systems use historical user-item-rating combinations to provide recommendation services and most of them do not use the context or content of news articles. In practice, most collaborative filtering systems are constructed based on users' past rating behaviors, either using a group of users similar to the given user to predict news ratings [RIS+94, SKKR01], or modeling users' behaviors in a probabilistic way [Hof04, LXL+12, PHLG00]. Collaborative filtering systems can efficiently capture users' behaviors in case where overlap in historical consumption across users is relatively high and the content universe is almost static [SKR99]; however, in many web-based scenarios, the content universe undergoes frequent changes, with content popularity changing over time as well [LHH+10]. Typically, under two circumstances, collaborative filtering systems can efficiently predict the score of unrated items based on similar users' behaviors: (1) when there is relatively good amount of overlap in historic ratings on the item set, and (2) when the content universe is almost static [SKR99] which is not the case on the domain of news recommendation. Indeed, there is a growing amount of new published news articles every day with content popularity changing over time as well [LHH+10]. Consequently, traditional collaborative filtering methods are inefficient in such domains where the content is highly dynamic [KR12, CTFLH12]. Moreover, new relevant items with no historical ratings from users cannot receive high predicted scores, which is known as a cold-start problem [SPUP02]. Users prefer to peruse news articles that happen recently, instead of old articles on her interest topic. Hence the user similarity which grounds model based collaborative filtering is inaccurate for users who have different active periods.

Content-based. Content-based news recommenders models use basically the similarity between users' profiles and news articles profiles to recommend to users new interesting news articles [SFHS07, BOHG13]. Users' profiles define mainly preferences and interests of users while news articles profiles describe its content. Thus, the accuracy of describing users' interests and the content of news articles are very important to ensure an effective recommendation. In other words, content-based news recommenders suggest to a user news articles which is similar to the news articles that the user has rated positively, or has otherwise shown interested for, in the past [CGN+11, CGT12, BOHG13]. In certain domains, content-based filtering approach has some weakness [Bur07, BOHG13]. For example, a content-based movie recommender is limited to written materials about a movie [Bur02]. Consequently, a content filtering model can only be as complex as the content to which it has access [SFHS07]. However, it works well on text domains such as the case of news websites where the content is publicly exposed and it is simple to analyze and extract the content of such data. While the basic approach of content-based filtering uses mainly the similarity analysis on term-by-item occurrences, neglecting the semantic structure of the content, more advanced approaches, such as latent semantic analysis, attempt to also exploit semantic features [CGN+11]). In addition to keywords and other meta-data that can be extracted, content-based systems are increasingly incorporating social information on items that users in $Web2.0$ provide [BOHG13]. Opinions on news websites present one of the main features that can be used to define either users' interests or to enrich the content of news articles.

Another related area of research to news recommendation is how to define users' interests which has been investigated extensively following different approaches [TLZ12, SCTB+12, LDP10, AAYIM13, SKKL12, AGHT11, GCP03, BWC+12, WLJH10, SAyMY08, CNN+10, HD10, MM10, LHH+10]. A first class of approaches builds user profiles based on search history [SCTB+12, TLZ12, GCP03, BWC+12, LDP10, DLB09, DTB10]. Sontag et. al., [SCTB+12] propose a generative model of relevance with user-specific parameters learned from user's long term search history. Similarly, Tan et. al., [TLZ12] show that useful patterns for recommendation can be extracted from long lasting search sessions and explorative behaviors. Gauch et. al., [GCP03] build a structured user profile, from browsing actions, as a weighted concept hierarchy to better reflect user's interests. Bennett et.al., [BWC+12] show that long-term (historic) behavior provides substantial benefits at the start of a search session, while short-term (session) helps in an extended search session. Daoud et al., [DLB09, DTB10] suggest composing graph-based profiles for all queries that are related to one search session. The initial query is mapped to the concepts of an ontology (ODP in this case) and used as a seed profile, that subsequently gets extended by the vocabulary attributed to each new arriving query. Liu et. al., [LDP10] build profiles of users' news interests based on their past click behavior.

A second class of approaches address the problem of extracting topics of interest in micro-blogging environments [CNN+10, AGHT11, HD10, WLJH10, MM10]. Chen et. al., [CNN+10] exploits user Tweets to build a bag-of-words profile for each Twitter user. Abel et al., [AGHT11] build hashtag-based, entity-based, and topic-based user profiles from Tweets, and show that semantic enrichments improves the variety and the quality of profiles. Hong et.al., [HD10] train a topic model on aggregated messages to improve the quality of topic detection in Tweets. Similarly, Weng et. al., [WLJH10] apply Latent Dirichlet Allocation (LDA) model to identify latent topic information from Tweets. Michelson et. al., [MM10] use a knowledge base to disambiguate and categorize the entities in user Tweets and then develop users profiles based on frequent entity categories. During this thesis, we do not fall in the two previous classes since (1) we do not have access to search history and (2) we exploit richer and longer opinions than Tweets which makes their related approaches unsuitable for us. Thus, we relate our work to the third class of approaches [AAYIM13, SKKL12] which exploit opinions on news websites to build user profiles. Shmueli et. al., [SKKL12] restrict user profile to a set of tags extracted from related opinions using a bag-of-words model. Abbar et. el., [AAYIM13] build the profile of each user by extracting the set of entities he has commented on and their related sentiments. While the proposed approaches are interesting, they do not exploit the different aspects of opinions. During this thesis, we have exploited opinions to describe accurately users' interests and the content of news articles. To this end, we have proposed a profile model based on three components namely *entities*, *aspects*, and *sentiments*. We have also investigated the impact of leveraging opinions on enriching the content of news articles.

Hybrid approaches. Traditional hybrid recommenders [Bur05, PTLMHV12] aim at combining both content filtering and collaborative filtering to provide more meaningful recommendation. In domains where content is difficult to obtain, collaborative filtering can work much better than content-based approaches [SFHS07]. On the other hand, content-based filtering can work without ratings as news articles can be evaluated based on attributes, while collaborative filtering

without preference data on a news article is unable to consider it [SFHS07]. In such cases, both approaches can be combined in multiple ways, and no consensus exists on how to do it the best [SFHS07]. Some works have proposed hybrid approaches on news recommendation domain [ZKL+10, LDP10, DGM08, AC10]. The combination can be implemented in various ways. In the first manner, collaborative filtering and content-based approaches are implemented separately, and predictions are merged. For example, the prediction in [ZKL+10] use weighted linear aggregation of a heat-spreading (HeatS) algorithm and probabilistic spreading algorithm. A topic category is learned for each article to predicted user interest content-wise, such a score is then multiplied with the collaborative score to generate final news recommendations [LDP10]. In the second manner, content-based and social characteristics are added to a collaborative filtering model. For example, fLDA [AC10] regularizes both user and item factors simultaneously through user features and the bag of words associated with each item. In [LY09, MKL09, MLK09] social and trust relations regularize rating matrix factorization. SCENE [LWL+11] performs a two-stage clustering on both contents and user access patterns. The inability of collaborative filtering to recommend news items is alleviated by combining it with content-based profiling [DGM08]. In the third manner, the recommendation framework is content-based, with features derived from content-based and collaborative filtering methods. For example, in [DGM08], attributes used for profiling users and items are assigned weights estimated from a set of linear regression equations obtained from a social network graph. While the proposed approaches are interesting, they are too complex to implement and more importantly too time-consuming on generating the list of news articles that should be recommended. In our thesis, we have adopted the strategy where a content-based framework is used with a feature derived from collaborative-based approaches. As recency is very important in news recommendation, we have first selected the list of candidate news articles that are published in a period of time surrounding the time of publishing the seed news article. It is to note that the seed news article is the one that it was read by a given user u in a given time t. From the list of these candidates we recommend a list of news articles that match with users' interests. In other words, news articles similarity feature based on their time of publishing is used as collaborative feature during the process of personalized news recommendation during this thesis. The collaborative-based feature is very quick and further very important on news recommendation domain.

2.3 Opinion mining

In this section we present a review of the existing and related works on sentiment analysis, opinion summarization, and opinion ranking proposed in the literature.

2.3.1 Sentiment analysis

Sentiment analysis, also called *opinion mining* domain is directly related to our thesis. This problem has been studied in the past few years [DLP03, PL05a, Tur02a, GACOR05, PE05, HL04a, KH06, WWH04, DLY08] by exploiting two main directions: (1) finding product features that have been commented on by reviewers and (2) deciding whether the opinions are positive or negative. In this section, we are more interested in the second direction which was exploited in

different tasks on our thesis. Two main categories for classifying sentiments have been explored: classification at the *document level* and a more fine-grained classification at the *sentence level*. As stated in [DLY08], most *sentence level* and *document level* classification methods follow one of two approaches: (1) *corpus-based* approaches, and (2) *lexical-based* approaches. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases, e.g., the works in [HW00, Tur02a, HW00]. Lexical-based approaches use synonyms and antonyms in WordNet to determine word sentiments based on a set of seed opinion words. Such approaches are studied in [AB06, DLY08, HL04a, KH06]. Representative works on classification at *document level* include [DLP03, Hea92, PLV02a, PL05a, RW03, Tur02a]. For example, Dave et. al., [DLP03] build a classifier based on information retrieval techniques for feature extraction and scoring. Hearst et. al., [Hea92] propose a sentence interpretation model where isolated portions of a text are interpreted then integrated with an information retrieval system to incrementally improve the text classification task. Riloff et.al., [RW03] learn linguistically rich extraction patterns for subjective (opinionated) expressions to identify more subjective sentences. Turney et.al., [Tur02a] present a simple unsupervised technique for classifying reviews based on the average semantic orientation of the phrases in the review that contain adjectives or adverbs. Other approaches address sentiment classification at *sentence level* [GACOR05, PE05, HL04a, HW00, KH06, WWH04, DLY08]. Gamon et.al., [GACOR05] combine a clustering technique with a machine-learned sentiment classifier, allowing for a visualization of topic and associated customer sentiment. Hu et.al., [HL04a] classify opinions and summarize costumer reviews based only on the features of the product on which the customers have expressed their opinions. Kim est.al., [KH06] address the problem of finding the sentiments expressed about a topic in each text, and identify the people who hold each sentiment. Wilson et.al., [WWH04] develop new syntactic clues for opinion recognition, as well as a variety of subjectivity clues from the literature. They demonstrated that these features can be adapted to the task of strength recognition, and that the best classification results are achieved when all types of features are used. Ding et.al., [DLY08] propose a holistic lexicon-based approach that exploits external evidences and linguistic conventions of natural language expressions. Most of the opinion mining techniques above described focus on product reviews while in our work we take a more general approach for classifying opinions about general topics which is far more challenging.

2.3.2 Opinion summarization

Summarization of opinions is crucial in helping users digest the different opinions expressed on the web. Previous studies have primarily focused on the task of generating highly structured summaries. This could be a simple sentiment summary such as 'positive' or 'negative' on a topic of interest [PLV02b, PL05b, Tur02b, WWH05, Tur02b] or a multi-aspect summary [LZS09, SB07, TM08, LHC05, HL04b]. For example, a concrete multi-aspect summarization for a mp3-player can be as follow: battery life: 1 star, scree:3.5 stars, etc. While structured summaries can be useful in conveying the general sentiments about a person, a product, or a service, such summaries lack the level of details that an unstructured textual summary could offer, often forcing users to go back to the original text to get more information. Moreover, most of multi-aspects summarization approaches assume that aspects related to opinions are already known which is not the case of

opinions related to daily life events such as opinions on news websites. Textual summaries are thus critical in conveying key opinions and reasons for those opinions at different granularities (i.e. entity level or topic level). Unfortunately, generating textual opinion summaries is a hard task. First, the summaries have to be representative of the key opinions (to avoid bias) and then, it has to be readable so that it can be easily understood by the user. Further, with the increased use of hand-held devices for various online activities such as shopping and finding places to eat, the conciseness or compactness of such summaries is also crucial. Indeed, there are many scenarios where concise summaries would be very beneficial. Consider shopping sites where there could be hundreds of reviews per product or news websites where we can have thousands of opinions chiefly when the content of news article concerns a controversial topics. Tata et al., [TDE10] produces an opinion summary of song reviews where for each aspect and each sentiment (positive or negative) they first select a representative sentence for the group. The sentence should mention the fewest aspects (thus the representative sentence is focused). They then order the sentences using a given domain ontology by mapping sentences to the ontology nodes. The ontology basically encodes the key domain concepts and their relations. The sentences are ordered and organized into paragraphs following the tree such as they appear in a conceptually coherent fashion. Lu et al., [Lu10] also use online ontologies of entities and aspects to organize and summarize opinions. Their idea is closest to the above approach but it is different. They first select aspects that capture major opinions. The selection is done by frequency, opinion coverage (no redundancy), or conditional entropy. It then orders aspects and their corresponding sentences based on a coherence measure, which tries to optimize the ordering so that they best follow the sequences of aspect appearances in their original postings. Ku et al., [KLC06] perform blog opinion summarization, and produce two types of summaries: brief and detailed summaries, based on extracted topics (aspects) and sentiments on the topics. For a detailed summary, it lists positive-topical and negative-topical sentences with high sentiment degrees. For the brief summary, their method picks up the document/article with the largest number of positive or negative sentences and uses its headline to represent the overall summary of positive-topical or negative-topical sentences. Lerman et al. [LBGM09] define opinion summarization in a slightly different way. Given a set of documents D (e.g., reviews) that contains opinions about some entity of interest, the goal of an opinion summarization system is to generate a summary S of that entity that is representative of the average opinion and highlight its important aspects. They proposed three different models to perform summarization of reviews of a product. All these models choose some set of sentences from a review. The first model is called sentiment match (SM), which extracts sentences so that the average sentiment of the summary is as close as possible to the average sentiment rating of reviews of the entity. The second model, called sentiment match + aspect coverage (SMAC), builds a summary that trades-off between maximally covering important aspects and matching the overall sentiment of the entity. The third model, called sentiment-aspect match (SAM), not only attempts to cover important aspects, but cover them with appropriate sentiment. A comprehensive evaluation of human users was conducted to compare the three types of summaries. It was found that although the SAM model was the best, it is not significantly better than others. In [NHMK10], a more sophisticated summarization technique was proposed, which generates a traditional text summary by selecting and ordering

sentences taken from multiple reviews, considering both informativeness and readability of the final summary. The informativeness was defined as the sum of frequency of each aspect-sentiment pair. Readability was defined as the natural sequence of sentences, which was measured as the sum of the connectivity of all adjacent sentences in the sequence. The problem was then solved through optimization. Ganesan et al., [GZH10] give an abstractive summary of opinions using graphical model based method. Two main points make the difference between our proposed approach of extracting most important aspects during this thesis and the previous approaches (More details about our approach in chapter 5). Firstly, most of proposed approaches such as Ganesan et al., [GZV12] approach use product reviews which belong to an already known set of aspects. In our work, we are interested in aspects about daily life topics reported by news articles. These aspects are not classified but we extract them automatically using an unsupervised approach. Secondly, most of these approaches are domain-specific, or usually highly dependent on the training data. Unlike those approaches, we have proposed sentiment-independent technique to extract most important aspects which might be used on diverse domains without any training step.

2.3.3 Opinion ranking

Ranking opinions has received attention, in the past few years, driven by the need of automatic annotation of product reviews. The proposed approaches focus on how to find helpful product reviews [HLY$^+$12, KPCP06, LHAY08, TR09, DNMKKL09]. These approaches assign a helpfulness score to each review, based on past interactions in the system, and return to the user a ranked list of reviews. Different parameters have been exploited to rank reviews. Kim et. al., [KPCP06] exploit the multitude of user-rated reviews on Amazon.com, and train an SVM regression system to learn a helpfulness function. This helpfulness function is then applied to rank unlabeled reviews. Danescu et. al., [DNMKKL09] show, through extensive experiments, that social affect is a significant factor for measuring helpfulness. The social effect is based on the relationship of one user's opinion to the opinions expressed by others in the same setting. More precisely, the relationship of a reviews star rating to the star ratings of other reviews for the same product. Tsur et. al., [TR09] identifies a lexicon of dominant terms that constitutes the core of a virtual optimal review. This lexicon defines a feature vector representation. Reviews are then converted to this representation and ranked according to their distance from a "virtual core" review vector. Liu et. al., [LHAY08] show that the helpfulness of a review depends on three factors: the reviewer's expertise, the writing style of the review, and the timeliness of the review. Based on those features, they propose a nonlinear regression model for helpfulness prediction. Hong et. al., [HLY$^+$12] start from the assumption that user preferences are more explicit clues to infer the opinions of users on the review helpfulness. Thus, they use user-preferences based features including information need, credibility of the review, and mainstream opinions. The approaches described above use different features to define the helpfulness of a review ranging from its content and the expertise of its author to the preferences of users. However none of them takes into account the relationships between the reviews, meaning the debates that users engage into to discuss a given product. This is a very important aspect that has an impact on the helpfulness of a review. In our work, we take into account the relations between opinions and all the reactions

they got from users including nested opinions and explicit feedbacks. Then, we propagate the sentiments along those relations to compute the final score of an opinion. Additionally, unlike the approaches above described, we define user' expertise from the implicit ratings he/she gets for his actions.

2.4 Conclusion

In this chapter, we have introduced the fields of information retrieval, recommender systems, and opinion mining that we have used throughout this thesis. We have explained information retrieval techniques used on our different contributions. Then, we gave an overview about recommender systems which was classified into three categories: collaborative filtering, content-based and finally hybrid approaches. Finally, we present previous works on the area of opinion mining which has been exploited on our different contributions.

Part II

Research chapters

Chapter 3

Building Fine-grained User and Article profiles

3.1 Introduction

With growing online sources of news articles around the world, it becomes very challenging for news websites to help users find news articles that match with their interests. Personalized news services strive to adapt their services to individual users by making use of both content and user's information. Despite a few recent advances, this problem remains challenging for at least two reasons. Firstly, news service is featured with dynamically changing pools of content, making traditional collaborative filtering methods inapplicable. Secondly, the lack of accurate profile models used on describing both users' interests and the content of news articles make it difficult to match between news articles and users. In other words, it is difficult and even impossible to predict the appropriate news articles to recommend if the interests of a given user are not clearly described and/or the content of candidate news articles are not well defined. This calls for proposing specialized profiles on the domain of news in order to allow an accurate description of users' interests and news article contents. In this chapter, we propose a profile model for describing users' interests and news articles content. The proposed model is based on three components which are entities, aspects and sentiment. Entities represent well known concepts such as persons, location and organizations. Aspects represent concepts issued from a given ontology or entity properties. Sentiments represent the inclination of the content towards one or both of two components namely entities and/or aspects. The sentiment can be either positive, negative or neutral. In this thesis, we have employed two variations of our profile model namely sentiment-dependent and sentiment-independent profiles on three different applications. Firstly, we use a sentiment-dependent profile of our model to identify the political orientation of users. Secondly, we employ a sentiment-independent profile model on the context of classic news recommendation. Thirdly, a mixture of sentiment-dependent profile and sentiment-independent profile is used to diversify the list of recommended news articles and improve the information novelty in the results [CKC+08, CG98].

3.2 Profile Model

Several representations of either users' interests or news articles have been proposed in the context of news recommendation. The representations vary depending on the type of the available information used to describe them and on the intended use of the profile. In the following, we present the most used representations found in the literature:

- **Bag of Words.** It is a simple representation where keyword-based model [PMS09] is used. Even if such a representation is simple to build, it encounters problems such as polysemy, especially on the Web, where the chance of misinterpretation of word meaning is not negligible.

- **Vector.** Another approach to represent either users' interests or the content of news articles is to use a vector of weighted terms or concepts. It is even possible to use more than one vector to express more attitude or to compare different categories of terms or concepts if needed. The value of each vector element can be a boolean indicating for instance that a user has posted an opinion or has visited a news article regarding a given term or a concept. It can also be an integer value indicating achieved degree of knowledge about the concept or a term. Such representation is easy to implement and has been widely used in literature [BKA98, BP99, MKDP14a].

- **Taxonomy-based profiles.** Taxonomy-based model is a semantic representation where nodes represent concepts and edges hierarchical relationships between them like the work done in [AGHT11]. The disadvantage of such a model is that it is very difficult to build it automatically. However, once done, it can be easily shared with other adaptive applications, as it is strongly semantically grounded.

To describe users' interests or the content of news articles we have proposed a profile model which is based on three key components namely entities, aspects and sentiments.

- **Entity.** It represents well known concepts such as persons, organizations, locations, etc. For example, from the sentence *In US, the United States Congress and Obama don't have the authority to interfere on this issue.*, we can extract *US* (Location), *States Congress* (Organization), and *Obama* (Person) as entities. Entities can be considered as one of the main information used to describe events or viewpoints on either news articles or users' opinions. Some words might refer to the same entity. For instance, we have noticed that some users use the term UK while others use the term The United Kingdom and both of them refer to the same meaning.
  Figure 3.1 presents a manual example of some entities extracted from some users' opinions.

- **Aspect.** In product reviews, most of researchers have defined an aspect as a property of an entity that can be either a component or an attribute. For example, *screen*, and *battery* are aspects of the entity iPhone reflecting its components. By contrast, *size* is an aspect that is an attribute of the entity iPhone. In our case, we define aspects as concepts referring to a given ontology or entity properties. For instance, the aspects *Business* or *sport* can be

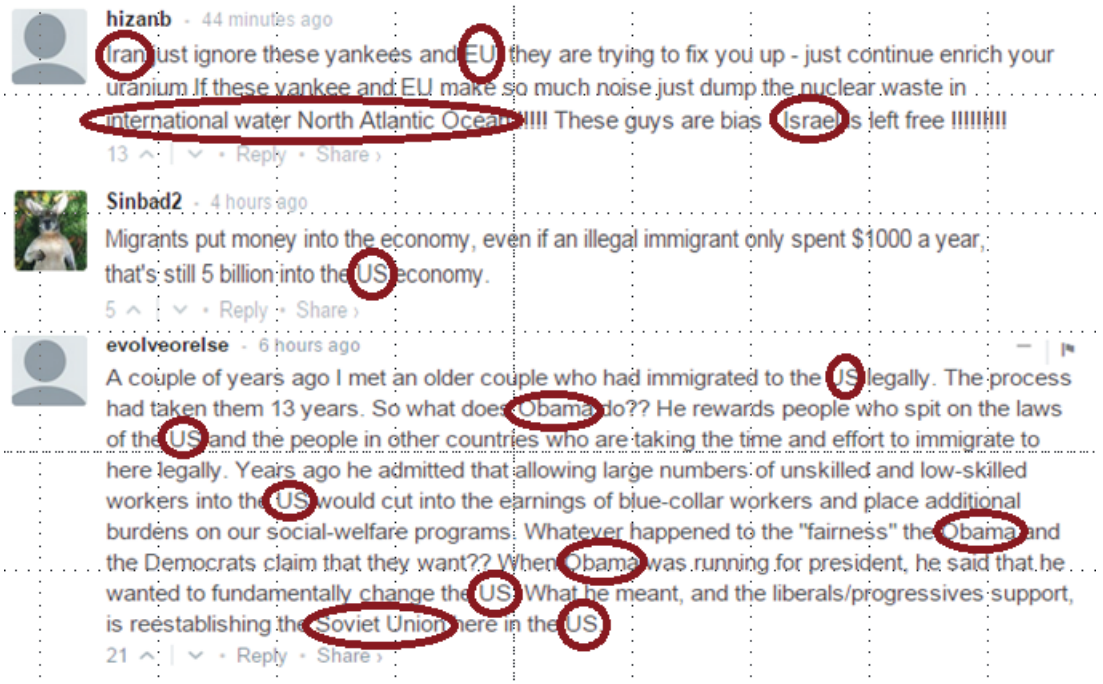

Figure 3.1: Some examples of extracted entities from opinions

defined as concepts, while *Tourism* or *Election* can be considered as entity properties for the entity France. Figure 3.2 shows a concrete example of some aspects extracted from a set of opinions.

- **Sentiment.** The last component of our model is sentiment which can be either positive, negative or even neutral towards one or both of the components entity and aspect. For example, we can find a user who has a positive sentiment towards the aspect *gun control* or a negative sentiment regarding the entity *François Hollande* (Person). We can have also sentiments that concern both components, for instance, we can have a user who has a positive sentiment regarding the aspect *Tourism* when it concerns the entity *France* and a negative sentiment about the same aspect *Tourism* when it concerns the entity *Iran*.

It is to note that we have used two approaches for extracting the list of entities and aspects during this thesis. First, an approach based on text mining techniques where both entities and aspects are extracted automatically. In this approach, we do not differentiate between the two components and we get as results from an input text content a set of entities and aspects. In the second approach, we have used a public web service named OpenCalais [Reu09] provided by Reuters [1] to extract the list of entities from each input of a text content. To extract the list of aspects, we have used the ODP taxonomy[2]. In the second approach, we differentiate between the extracted entities and aspects.

[1] http://www.reuters.com/
[2] www.dmoz.org

Figure 3.2: Some examples of extracted aspects from opinions

3.3 Sentiment-dependent profile

3.3.1 Motivation

Political views are freely and explicitly expressed through opinions in news websites. These opinions represent an interesting sample from political trends and orientations of users. Extracting such type of knowledge would allow news websites publishers to have an idea about the orientation of their commenters, the main issues related to each orientation, and the possible political persuasions and ideological viewpoints for all topics. The opinions expressed by users are not restrained by journalism values such as fairness or balance, and do not go through a formal editorial process. Moreover, the number of opinions about a given topic might continuously increase. The unstructured and the dynamic nature of opinions, provided in news websites, call for effective and efficient techniques for identifying political trends. The user classification might also be used to identify the political orientation of news articles like the work done in [ZRM11]. Several approaches have been proposed to classify political positions from texts. One line of work focused on using SVM with optimization of text feature selection [JA08] [OLK09] [YKD08] [HRG10], as well as complementing with sentiment analysis [DS06] [MM06] [MM07] [CGR+11]. Another line

of work used word frequencies, Bayesian statistical models, and topic models [LBC03] [MV08] [LC08b] [MCQ08] [SP08]. Most of these approaches use supervised techniques which can be expensive as they require training. Moreover, they mainly use semi-structured data to classify users. Examples include data extracted from twitter and microblogs which is characterized by short fragments (tweets, short messages), where each fragment covers a known and a unique aspect. By contrast, opinions in news articles cover almost always more than one aspect which are unknown. More specifically, approaches based on twitter samples use hashtags of controversial topics such as *USElection* or *Arabspring* and a set of stakeholders such as actors or politicians, to classify the stakeholders' opinions into pro or con categories for the respective topics [ARW12]. Opinions published in microblogs are frequently short and do not contain more than one aspect whereas, in news websites, users publish long opinions covering more than one aspect.
In this section, using our profile model, we propose an unsupervised technique for defining the political orientation of users based on their opinions in news websites. To the best of our knowledge, we are the first to propose an unsupervised approach on such unstructured and dynamic data. Our contribution is twofold (1) we generate user profile based on the entities, aspects and sentiments discussed or revealed in his opinions and (2) we construct a knowledge base of political orientations, using Wikipedia, to automatically classify users based on their profiles. We have conducted extensive experiments with US and Egypt user groups crawled from CNN and Al-Jazeera. The experiments showed that our approach provides high quality results to classify US users into Republican/Democrat leanings and Egypt users into secular/Islamist leanings.

3.3.2 Profile Generation

To define the political orientation of a given user U, we collect the opinions he has expressed, in a given news website, during a period of time T. Then, we analyze the opinions and extract from them all the entities and the aspects the user has discussed. For each aspect (or entity), we define the sentiment expressed by the user U. For example, a user can discuss the aspect of *abortion rights* and be *negative* about it. As a result, the user U is described by a set of aspects $\{a_1, ...a_n\}$, entities $\{e_1, ...e_n\}$ and their related sentiments $\{s_1, ...s_n\}$. Concretely, users' interests and article profiles are represented by two types of pairs, $< entity, sentiment >$ and $< aspect, sentiment >$. To this end, we proceed in three main steps.

Step1. Extraction of Opinionated Sentences.

We first identify the sentences using OpenNLP[3] expressed in all the opinions of user U. Second, we identify the sentiment of each sentence which might be positive, negative, or neutral. Third, we classify all sentences based on their sentiment to obtain three categories of sentences namely positive, negative and neutral. It is to note that we have used Alchemy Api[4] to compute the sentiment of each sentence.

[3]http://opennlp.sourceforge.net/
[4]http://www.alchemyapi.com/api/

Step2. Generation of Candidate Entities and Aspects.

We take all the opinionated sentences extracted from the previous step, and we rank their contained terms using $tf * idf$ scoring function. In our work, tf represents the term frequency in the set of opinionated sentences of user U, and idf represents the inverted document frequency in the set of opinionated sentences of all users. The idea is to select highly scored unigrams as a base for generating candidate aspects and entities. From these unigrams, we generate bi-grams, then we take the bi-grams as input and we build a set of n-grams by concatenating bi-grams that share an overlapping word. At each step we take the topk n-grams based on the score of their composed unigrams[5]. We check the redundancy of the generated candidates, using Jaccard similarity [RV96]. If two n-grams have a similarity higher than a defined threshold, we would discard one of them. In our work, we have set the maximum length of the n-grams to 5 since there were no meaningful n-grams of a higher length.

Step3. Selection of Promising Entities and Aspects.

Generating n-grams that have high $tf * idf$ scores is not enough to identify the entities and the aspects discussed in users' opinions. It is important for the words in the generated n-grams to be strongly associated within a sentence in the original text to avoid covering incorrect information. This property ensures that only a set of related words are used in the generated n-grams to avoid conveying incorrect information. To capture this association, we use *pointwise mutual information* [TC03] (PMI) of words in n-grams based on their alignment to the narrow opinions of each user. Formally, suppose $m_i = w_1...w_n$ is a generated n-grams. We define the $Score_n$ as follows:

$$S_{PMI}(w_1...w_n) = \frac{1}{n}\sum_{i=1}^{n} pmi_{local}(w_i) \tag{3.1}$$

where $pmi_{local}(w_i)$ is a local pointwise mutual information function defined as:

$$pmi_{local}(w_i) = \frac{1}{2C}\sum_{j=i=C}^{i+C} pmi'(w_i, w_j), i \neq j \tag{3.2}$$

where C is a contextual window size. The $pmi_{local}(w_i)$ measures the average strength of association of a word w_i with all its C neighboring words (on the left and on the right). For example, in *gun control law* phrase, assuming $C = 1$, for *gun* we would obtain the average PMI score of *gun* with *control* and for *control* we would obtain the average PMI of *control* with *gun* and *control* with *law*. When this is done for each $w_i \in m$, this would give a good estimate of how strongly associated the words are in m. To capture our second property, we used a modified PMI scoring [GZV12] referred to as pmi' where the pmi' between two words, w_i and w_j, is defined as:

$$pmi'(w_i, w_j) = \log_2 \frac{p(w_i, w_j) \cdot c(w_i, w_j)}{p(w_i) \cdot p(w_j)} \tag{3.3}$$

where $c(w_i, w_j)$ is the frequency of two words co-occurring in a sentence from the original text within the context window of C (in any direction) and $p(w_i, w_j)$ is the corresponding joint probability. The co-occurrence frequency, $c(w_i, w_j)$, which is not part of the original PMI formula,

[5]In this work we have set k=500

is integrated into our PMI scoring to reward frequently occurring words from the original text. By adding $c(w_i, w_j)$ into the PMI scoring, we ensure that low frequency words do not dominate and moderately associated words with high co-occurrences have relatively high scores.

3.3.3 Application: Defining Users' Political Orientations

An unsupervised technique of identifying the political orientation of users based on their profiles calls for the use of a knowledge base. To this end, we have created a knowledge base of political orientations taken from Wikipedia. For a given political orientation, we start from a Wikipedia seed page. We extract from text part of the page all outgoing links that point to other Wikipedia articles. Then, we consider the anchor text of these links as entities and aspects related to the political orientation. Each aspect or entity occurs in a sentence that have a sentiment orientation. Thus, in a similar way to user profile, we identify each political orientation by a set of aspects $\{a_1, ...a_m\}$, entities $\{e_1, ...e_m\}$ and their related sentiments $\{s_1, ...s_m\}$. Table 3.1 shows

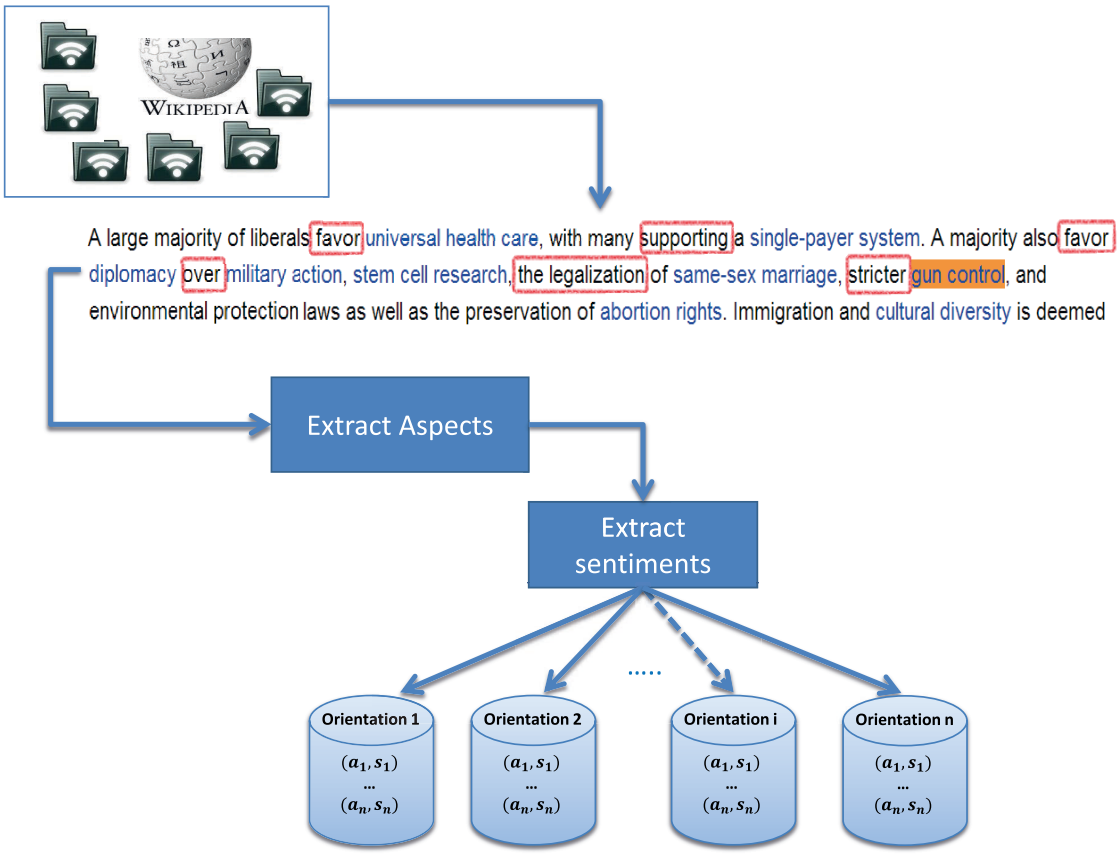

Figure 3.3: An illustration of how knowledge base of political orientations are created from Wikipedia

an example of Liberal and Conservative orientation and their relevant aspects extracted from Wikipedia. To cover more aspects and enrich the knowledge base, we also include the Wikipedia pages pointed by the seed page of the political orientation. For example, a seed page reports that liberals are in favor of universal health care. We take the Wikipedia page of universal health care and add its aspects to the favorite list of Liberals. The figure 3.3 presents a concrete example,

Orientation	**S**ome Aspects Extracted from Wikipedia	
Liberals	***f**avor*	universal health care, strict gun control, diplomacy, stem cell research, same-sex marriage, abortion rights
	***a**gainst*	increased military spending The Ten Commandments display in public buildings
Conservatives	***f**avor*	small government, low taxes, limited regulation free enterprise, school prayer, capital punishment
	***a**gainst*	same-sex marriage, abortion rights, multiculturalism

Table 3.1: The structure of the orientation knowledge base

using a paragraph from Wikipedia, of how the knowledge base of political orientations is created. To identify the political orientation of user U, we compute the similarity between its profile and the description of all the political orientations that exist in the knowledge base. The most similar description is assigned to the user as its political orientation.

3.3.4 Experiments

Datasets

We have crawled 2 datasets from CNN and Al-Jazeera English news website. From CNN, we have extracted the activities of 11, 322 users. Indeed, for each user we have extract all their posted opinions. From Al-Jazeera, we have extracted the activities of 539 users, namely all their posted opinions. For each user, we have extracted all his opinions from *October 2009* to *September 2013*. More details about both datasets are shown in table 3.2. We have used on this application only opinions about politic.

Table 3.2: Datasets Statistics

	#Users	#Opinions	#News articles
CNN	11, 322	684, 058	15, 365
Al-Jazeera	539	24, 826	2, 773

Results

We have run our experiments on 500 users: 290 from US (CNN) and 210 from Egypt (Al-Jazeera). We have selected from the datasets only users that mention explicitly their nationality. We have shown the list of opinions of each user and asked human assessors, who were students not involved in this project, to analyze the users opinions and classify them into the following categories: *Democrat/Republican* for US users and *Secular/Islamist* for Egypt users. The result of the human assessment is the ground truth for our evaluation.
We applied our approach on the same 500 users selected before. The outcome of the classification was then compared to our golden standard. To measure the effectiveness of our approach, we have computed the accuracy which represents the fraction of users that were correctly classified. We have compared different variations of our approach. The first one uses the *top*100 unigrams, based on $tf * idf$ scoring function, to classify the user. The second approach uses n-grams of

Table 3.3: Accuracy of User classification

	US		**E**gypt	
	Democrats	**R**epublicans	**I**slamists	**S**eculars
Unigrams (tf*idf)	50%	16,66%	72,72%	25%
N-grams (tf*idf)	67,60%	50%	70,70%	51,56%
N-grams (PMI)	**9**5,07%	**7**9,41%	**8**5,85%	**8**4,37%

length between 1 and 5. The *top*100 n-grams, based on $tf * idf$ scoring function, are selected to classify the user. In the third approach, the *top*100 n-grams, based on PMI, are selected to classify the user. The results are shown in Table 3.3. We can see the impact of the different steps of our approach on the accuracy of our technique. Using only unigrams generates incomplete information about the aspects discussed by users and thus provides very inaccurate results. We can see that using n-grams improves the results in most cases, however they still have a low accuracy. Using PMI to select the aspects of opinions is the best providing an accuracy that goes up to $95,07\%$ To sump up, we have proposed a new technique for defining the political orientation of users based on their opinions about news articles. The proposed approach is promising as it provides means for dealing with unstructured source of information. Moreover, it is completely unsupervised which makes it flexible to be applied on any kind of dynamic knowledge such as opinions. As future work, we plan to extend the knowledge base to other types of orientations in other domains and propose a general approach for extracting the main aspects of daily life topics and their main trends.

3.4 Sentiment-Independent profile

3.4.1 Motivation

The accuracy of personalized recommendation depends mainly on how well user profiles are defined. Naturally, users' opinions represent a valuable information source since they reflect not only interesting entities for users but also more details about which entities and aspects they are interested in. Therefore, several past studies have exploited, in different ways, user-generated-content for news recommendation [AAYIM13, AGHT11, SKKL12, LHH$^+$10, CNN$^+$10, PMS09, MM10, HD10, WLJH10]. Most of these approaches use tweets [CNN$^+$10, AGHT11, MM10, HD10, WLJH10] and few others [AAYIM13, SKKL12, LHH$^+$10] exploit opinions on news websites. hmueli et. al., [SKKL12] restrict user profile to a set of tags extracted from related opinions. However, they do not take into account the difference between entities and aspects for defining the interests of users towards specific issues. Abbar et. al., [AAYIM13] build the profile of each user using a set of entities he has commented on with their related sentiments. While the proposed approach is interesting, it does not exploit all available information in users' opinions and thus it provides incomplete profiles. The reason is that a user can be interested in a specific entity when it is related to a given aspect and can be not interested in the same entity when it concerns another aspect. For instance, we can have a user who is interested by the entity *Tunisia* when it is related to the aspect *Tourism* and who is not interested in it when it is related to the aspect *Election*. In this section, we propose a personalized news recommendation approach that

pays particular attention to interesting aspects of each entity. To this end, we use our profile model to define users' interests and news articles using a set of tuples representing entities and their aspects. The idea is to have a fine-grained description of users and articles regarding general topics together with more specific issues. The profile of a user' interests is defined from the set of opinions he provides in the news website, and the article profile is extracted from its content and described by a set of tuples $(entity, aspect)$. We define each profile by the two main components *entities* and their related *aspects*. These profiles are then matched to recommend to each user the list of news articles that match with his interests. We evaluate our approach using four real datasets including The Independent, The Telegraph, CNN and Al-Jazeera. The experiments show that our approach outperforms baseline approaches with a large margin, in term of precision and NDCG.

3.4.2 Profile Generation

To define the profile of a given user u, we collect the opinions he has expressed, in all news websites, during a period of time T. Then, we analyze the opinions and extract from them a set of tuples $\{\{(e_1, a_{11}), (e_1, a_{12})..., (e_n, a_{nm})\}$ where e_i is an entity (e.g., Person, Location, Organization) and a_{ij} is the aspect related to each entity e_i. It is to note that we have used in this section the OpenCalais Api to extract the list of entities and the ODP taxonomy to extract the list of aspects related to each entity. For instance, from the opinion *"Obama is wrong to give work permits to young illegal immigrants"* we extract the entity Obama (Person) and their related aspects Work permit and illegal immigration. Practically, to build a user profile, we first identify all opinions expressed by the user u for a period of time T. We have used OpenNLP to identify all sentences from his opinions. Thus, for each sentence, we extract the different entities and their related aspects. Formally, the profile of a user u is defined by:

$$P(u) = \{(e_i, a_{ij}), w_u(e_i, a_{ij}) | e_i \in E, a_{ij} \in C, u \in U\} \tag{3.4}$$

Where C, E and U denote the set of entities, aspects and users respectively and $w_u(e_i, a_{ij})$ is the weight of each tuple (e_i, a_{ij}) computed using *tf*idf* technique . In our work, *tf* represents the tuple frequency in the set of opinions of user U, and *idf* represents the inverted document frequency in the set of opinions of all users. Similarly to user profile, we represent each news article by a set of tuples $< e_i, c_{ij} >$ extracted from its content. Practically, to build a news article profile, we first identify all sentences of its content using OpenNLP. Then, we extract the tuples corresponding to entities and the related aspects as described earlier. The weight of each tuple is defined through *tf*idf* technique where *tf* is the tuple frequency in the sentences of a given news article and *idf* is the inverted document frequency in all sentences of all news articles.

After extracting entities and aspects from opinions and articles contents, we can build the profile in three different ways

1. **E**ntity-centric Profile. It consists of a set of weighted entities as proposed in [AAYIM13]. This type of profile is suitable for recommending news articles that target mainly specific entities, like a location or a person, without addressing particular issues.

2. **A**spect-centric Profile. It consists of a set of weighted aspects. This type of profile is suitable for recommending news articles that address specific issues in a broad way meaning without focusing on a specific location, person, or organization.

3. **G**lobal Profile. It consists of a set of pairs tuples <entity, aspect>. This type of profile is suitable to give a more precise description of the view points expressed by the content of opinions and articles

The versatile profile model makes our approach applicable for different needs. It is important though to note that the Global profile is the most complete one and is the main contribution of our work.

3.4.3 Application: News Recommendation

Our goal is to propose a personalized news recommendation model tailored to users' interests. Typically, interests represent the conjunction between entities and their related aspects. In our setting, we identify the interests of a given user based on the opinions he has posted on the news websites. Using this information, the personalized news recommendation works as follows: Given a target user who is reading a seed article, we recommend a set of news articles that (1) are similar to the seed topic article for not deviating far away from user's interests and (2) match with specific issues that interest the user profile. The idea behind is to select, first, new articles that belong to the same topic than the seed article and then choose a subset that match with user interests. Formally, we define U as the set of users of a given news website, and A as the set of articles provided by the news website. Each user $u_i \in U$ provides a set of opinions C_i about a set of articles A' where $A' \subset A$. We assign to each user u_i a profile P_{u_i}, extracted from the set of his opinions C_i, which reflects his specific issues about what he reads in the past. Similarly, we assign to each article a_j a profile P_{a_j} extracted from its content. When user u_i is reading article a_j , we proceed as follows. First, we compute the similarity between the article profile P_{a_j} and the profiles of the set of articles A_t where $A_t \subset A$ and A_t corresponds to all the articles that were published in time interval t. By this way, we can restrict our search space to any time period specified by the user. The time interval can range from a few days to months depending on user needs. The set of articles A_t is then sorted from the most similar article to a_i to the least similar one resulting in list L_1. Second, we compute the similarity between the user profile P_{u_i} and the profiles of the articles contained in the set A_t, thus, providing another sorted list L_2 from the most similar article to user profile P_{u_i} to the least similar one. As a last step, we aggregate the two lists L_1 and L_2 to obtain the final list of sorted articles from which we recommend the topk articles to user u_i. It is to note that we have adopted cosine similarity to compute the similarity between profiles. This measure has been shown to be very effective in measuring similarity and detecting novelty between news articles [LMK+11]. In a standard search problem, a news article or user profile is represented by a vector of n dimensions where a term is assigned to each dimension and the value of the dimension represents the frequency of the term in the profile. In our setting we are interested in computing similarity between profiles described by a set of tuples, for this end we modify the vector representation as follows: each

profile is represented by one vector representing the set of tuples and the value of each dimension represents the frequency of the tuple on news article or user profile.

3.4.4 Experiments

Setup

We have crawled a dataset based on the activities of 164 users from The Independent news website. The choice of this site was based on the fact that it has a large number of active users that continuously post opinions on articles of various topics. Additionally and more importantly, users of The Independent follow also other news websites including The Telegraph, CNN and Al-Jazeera, so they have access to different types of articles covering different aspects for the same entity. For each one among those users, we have crawled his opinions in the four news websites mentioned earlier. Additionally, we have collected all the articles commented by each user from May 2010 to December 2013. Statistics about the number of opinions and articles from each news website are shown in Table 3.4 . To evaluate our approach, we have randomly selected 23 users. For each user we performed recommendation at different time points $t_1, t_2, ..t_n$. The reason behind time dependent evaluation is two fold: (1) to take into account profile updates since users continuously post opinions bringing new information about their interests, and (2) to use data before time point t_i for recommendation and data starting from time point t_i for assessment, as described later. The time points $t_1, t_2, ..t_n$ are chosen in such a way that between t_{i-1} and t_i, there is at least m opinions posted by the user. In our experiments, we have set $m = 100$ to have enough evidence that the user profile needs to be updated. This setting resulted in 189 rounds of recommendation. We have simulated the recommendation system in the following way. For each user and at each time point t_i, we build the user profile based on his opinions posted before t_i. Then, we choose as a seed article the first article that the user commented after time point t_i. We choose an article commented by the user to make sure that it matches user's interests. Based on the seed article and the user profile we return a set of articles that are similar to the seed article and at the same time have similar interests as the ones expressed in the user profile.

#Opinions	482, 073
#Independent articles	26, 096
#Telegraph articles	23, 154
#CNN articles	535
#Al-Jazeera articles	303

Table 3.4: Datasets Statistics

Figure 3.4 shows the distribution of articles by topic. We can see that most articles and opinions are related to the topic politic. Note that the list of the seed articles we have selected follows a very similar distribution to the overall set of articles. To assess the effectiveness of our approach we have used an automatic evaluation to avoid the subjectivity of manual assessments. We have considered the action of commenting on an article to be an indicator that the article fits the interests of the user. Based on this assumption, we check the list of recommended articles. The one that the user has commented on are considered relevant. Note that it is probable that

we systematically underestimate the interest of the user . A person might well be interested in an article even though he does not comment on it.

Results

We use two baselines strategies to assess our approach. The first one is based on aspect-centric profiles for both users and articles. The aspects were generated from users' opinions and news articles content using the ODP taxonomy as we have described earlier. The second strategy is based on entity-centric profiles for both users and articles. This strategy has been proposed in [AAYIM13] and it represents our second baseline. We compare both strategies to our contribution where we define a global profile for both users and articles. To compare the results of the different strategies, we use Precision and NDCG at k ($P@k$ and $NDCG@k$). The $P@k$ is the fraction of recommended articles that are relevant to the user considering only the top-k results. It is given by:

$$P@k = \frac{|Relevant\_Articles \cap topk\_Articles\_Results|}{k}$$

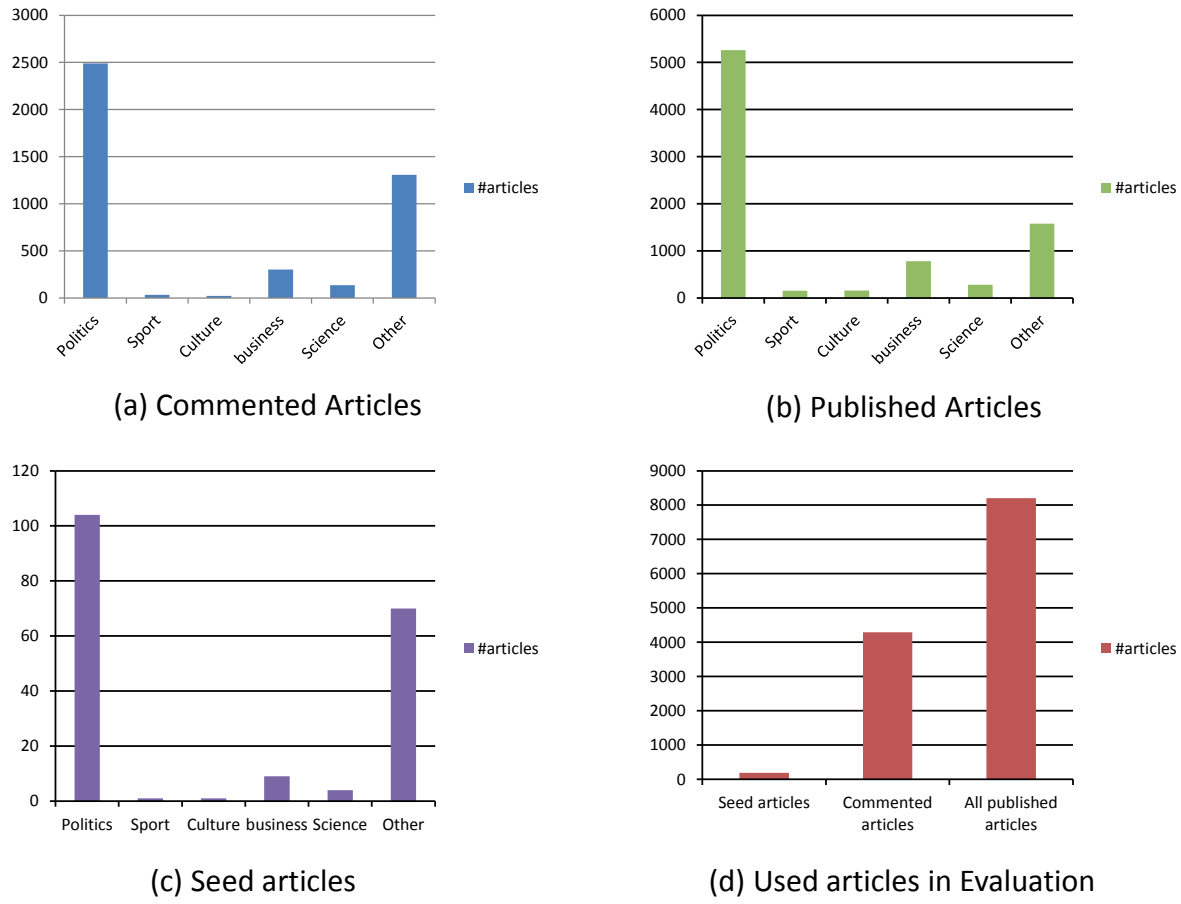

Figure 3.4: Statistics about categories of used articles in Evaluation

Additionally, we compute $NDCG$ to measure the usefulness (gain) of recommended articles based on their (geometrically weighted) positions in the result list. It is computed as follows:

$$NDCG(E,k) = \frac{1}{|E|}\sum_{j=1}^{|E|} Z_{kj} \sum_{i=1}^{k} \frac{2^{rel(j,i)}-1}{log_2(1+i)}$$

where Z_{kj} is a normalization factor calculated to make $NDCG$ at k equal to 1 in case of perfect ranking, and $rel(j,i)$ is the relevance score of a news article at rank i. In our setting, relevance scores $rel(j,i)$ have boolean values: 1(relevant) if the news article was commented by the user u, and 0(not relevant) if the news article was not commented by the user u. The precision and $NDCG$ results for the three strategies are shown in Table 3.5. We can clearly see that our

	P@5	**P**@10	**N**DCG @5	**N**DCG @10
Aspect-centric Profile	0.396	0.392	0.734	0.689
Entity-centric Profile [AAYIM13]	0.412	0.409	0.806	0.768
Global Profile	**0**.52	**0**.507	**0**.855	**0**.797

Table 3.5: Precision and NDCG values for all users

approach of using global profile outperforms the baseline approach with a gain of 10% in terms of precision and 5% in term of ranking at NDCG@5. We also observe that using only aspects to build user and article profiles performs worst. The reason is that most of the news articles do not address certain aspects without relating them to some entities. Thus, disregarding entities leads to worst results. Moreover, when viewpoints are expressed about entities, they usually refer to certain aspects of those entities. Thus, using only entities to build profiles decreases the performance. Consequently the combination of both entities and aspects give the best results. Note that real precision values must be higher than the one presented here. The reason is that opinions can tell us if a user is interested in an article or not but their absence does not mean the opposite. To sum up, we have proposed a new model for user and article profiles based on entities and their related aspects. We have performed experiments based on four news websites, namely The Independent, The Telegraph, CNN and Al-Jazeera. The results show that using both entities and aspects in the profile outperforms both entity-centric and aspect-centric approach with a minimum precision gain of 10% and 5% in term of ranking at $NDCG$@5.

3.5 Mixture Profile

3.5.1 Motivation

News recommendation services select often to users a list of news articles that match with their interests in the same form than retrieval results, i.e. as a ranked list. In most settings, the news recommendation service must recommend sets of news articles, rather than individual news article. Thus, a relevance score is computed for each news article to rank the list of recommended news articles. The accuracy of news recommendation service is very important facet of usefulness but it is not enough in practice. For instance, recommending redundant news articles leads to diminishing returns on utility, since users need to consume redundant information only once.

Thus, the recommended list of news articles should be well diversified. Diversification of the results list has recently been identified as a critical factor that significantly influences end-user satisfaction with a recommender system [VC11, LHCA10, HZ11]. To this end, we propose a 2 steps model: (i) First, we select only news articles that match with users' interests as done in previous section using a new similarity measure between user and news articles profiles.
(ii) Second, we diversify the list of selected news articles by applying a news articles diversification model based on two main components: (1) semantic diversification on the list of relevant news articles to avoid redundancy and to cover a diverse set of news articles presenting different arguments, and (2) sentiment diversification to cover different types of sentiments that can be positive, negative or neutral. We evaluate our approach using four real datasets including, The Independent, The Telegraph, CNN, and Al-Jazeera. The results show that users tend to comment diverse news articles and thus applying diversification help on improving the quality of recommendation.

3.5.2 Profile Generation

We define the content of news articles by a set of triplets $< e_i, a_{ij}, s_i >$ extracted from its content. Practically, to build a news article profile, we first identify all sentences of its content using OpenNLP. For each sentence, we define its sentiment orientation using the Alchemy Api. The sentiment orientation of a sentence can be positive, negative or neutral. Thus, for each news article we obtain three group of sentences corresponding to positive, negative and neutral profiles of the news article. For each group of sentences, we extract their tuples corresponding to entities and their related aspects. The weight of each tuple is defined through *tf*idf* technique where *tf* is the tuple frequency in the sentences of a given news article and *idf* is the inverted document frequency in all sentences of all candidate news articles. Combining entity elements, their related aspects and sentiment orientations, we define each news article as a set of triplets:

$$A = \{< e_i, a_{ij}, s >\} \tag{3.5}$$

To define users' interests, we have used the same technique described in previous section by a set of tuples (e_i, a_{ij}) representing entities and their aspects. The entities and their related aspects are extracted from the opinions of each user.

3.5.3 Application: Diversification of Recommended News Articles

We propose a two stage recommendation model: In a first step, we select the topk[6] relevant news articles by computing cosine similarity between user profile and news articles profiles, where the unit item is a tuple (e_i, c_{ij}). This measure has been shown to be very effective in measuring similarity between documents [Sin01]. In a standard search problem, a document is represented by a vector of n dimensions where a term is assigned to each dimension and the value of the dimension represents the frequency of the term in the document. In our setting we are interested in computing similarity between tuples, so each profile is represented by a vector where the dimensions of each vector are assigned tuples and the value of each dimension represents the

[6] In this work we have empirically set k=200

*tf*idf* score of the tuple for the given profile. Formally the cosine similarity between a news article profile A and a user profile B is given by:

$$Similarity(A, B) = \frac{1}{3}(\frac{B.A^+}{||B||||A^+||} + \frac{B.A^-}{||B||||A^-||} + \frac{B.A^o}{||B||||A^o||})$$

where B is the vector corresponding to the user profile B, and A^+, A^-, and A^o are respectively the positive, negative, and neutral vectors corresponding to the news article profile A. We compute the cosine similarity between each type of vector and then we average the results to obtain the final similarity values. The more tuples an article profile and a user profile have in common, the more interesting is the article for the user. Note here that in this first stage we do not consider sentiments s to define news articles profiles. Thus, the profile of each news article is described like the user profile by a set of weighted tuples (e_i, a_{ij}). Consequently, we can formalize the relevance of each news article by a function $r : A \times U \rightarrow R^+$, where a higher value implies that the news article is more relevant to the user profile. In the second stage, we perform diversification of news articles. The technique used to diversify news articles was inspired by the works of Kacimi et al. in [KG11, KG12]. We are given a set of news articles $A = \{a_1, a_2, ..., a_n\}$ where $n \geq 2$. Our goal is to select a subset $L_k \subseteq A$ of news articles that is diverse. We assume that three main components define the diversity of a set of news articles : *relevance*, *semantic diversity*, and *sentiment diversity*. Naturally, before discussing whether a set is diverse or not, it should first contain relevant news articles. This is why it is important to include the relevance in diversification models [GS09a, AGHI09]. Note that the *relevance* of each news article is given by the cosine similarity score as described earlier. To diversify a set of news articles, we need to give more preference to dissimilar news articles. We assume that two news articles are dissimilar if (1) they contain different tuples of entities and/or aspects, and/or (2) they exhibit different sentiments about those tuples. To satisfy these two requirements, we define two distance functions. The first one is a *semantic distance* function $d : A \times A \rightarrow R^+$ between news articles, where the smaller the distance, the more similar two news articles are. This distance measures the *semantic diversity* of the set. The second one is a *sentiment distance* function $s : A \times A \rightarrow R^+$ between news articles, where the smaller the distance, the closest in sentiments two news articles are. The sentiment distance is used to compute the *sentiment diversity*. We formalize a set selection function $f : 2^A \times r \times d \times o \rightarrow R^+$, where we assign scores to all possible subsets of C, given a relevance function $r(.)$, a semantic distance function $d(.,.)$, a sentiment distance function $s(.,.)$, and a given integer $k \in Z^+(k \geq 2)$. The goal is to select a set $L_k \subseteq D$ of news articles such as the value of f is maximized. In other words, the goal is to find:

$$L_k^* = Max_{L_k \subseteq D, |L_k|=k} f(L_k, r(.), d(.,.), s(.,.))$$

where all arguments other than L_k are fixed inputs to the function. The purpose of this model is to maximize the sum of the relevance, the semantic dissimilarity, and the sentiment dissimilarity of the selected set. The function we aim at maximizing can be formalized as follows:

$$f(L) = \alpha(k-1)\sum_{a \in L} r(a) + 2\beta \sum_{a,b \in L} d(a,b)$$

$$+2\gamma \sum_{a,b \in L} s(a,b)$$

where $|L| = k$, and $\alpha, \beta, \gamma > 0$ are parameters specifying the trade-off between relevance, semantic diversity, and sentiment diversity[7]. The model allows to put more emphasis on relevance, on semantic diversity, on sentiment diversity, or on any mixture of these measures. Note that we need to scale up the three terms of the function. The reason is that there are $\frac{k(k-1)}{2}$ numbers in the semantic similarity sum, and $\frac{k(k-1)}{2}$ in the sentiment sum as opposed to k numbers in the relevance sum. The relevance scores are computed using cosine similarity and the semantic distance is computed using Jaccard similarity function. As for sentiment distance, we define it as follows:

$$s(a,b) = \begin{cases} 0, & \text{if the tuples have the same sentiment;} \\ 1, & \text{otherwise.} \end{cases}$$

where the sentiment orientation includes *positive*, *negative*, and *neutral* sentiments.

The problem of diversifying search results is NP-hard [GS09b, AGHI09]. However, there exist a well-known approximation algorithm to solve it [GS09b], which works well in practice [KG11, KG12]. Gollapaudi et al. [GS09b] show that their Max-sum diversification objective can be approached as a facility dispersion problem, known as the MaxSumDispersion problem [HRT97, KH78]. In our work, we follow the same principle and model our diversification problem as a MaxSumDispersion problem having the following objective function:

$$f'(L) = \sum_{a,b \in L} d'(a,b)$$

where d'(.,.) is a distance metric. We show in the following that f' is equivalent to our f function. To this end, we define the distance function $d'(a,b)$ as follows:

$$d'(a,b) = \begin{cases} 0, & \textit{if } a{=}b \\ r(a) + r(b)) + 2\beta d(a,b) & \\ +2\gamma s(a,b), & \textit{otherwise} \end{cases}$$

Considering the binary sentiment function, we claim that if d(.,.) is a metric then d'(.,.) is also a metric (proof skipped). We replace d'(.,.) by its definition in f'(L), disregarding pairwise distances between identical pairs, thus we obtain:

$$f'(L) = \alpha(k-1) \sum_{a \in L} r(a) + 2\beta \sum_{a,b \in L} d(a,b)$$

$$+2\gamma \sum_{a,b \in L} s(a,b)$$

we can easily see that each $r(a)$ is counted exactly $(k-1)$ times. Hence, the function f' is equivalent to our function f. Given this mapping, we can use a 2-approximation algorithm proposed in [HRT97, KH78] and illustrated by algorithm 1 to maximize our MaxSum objective *f*.

[7] In our implementation we have set $\alpha = \beta = \gamma = 1$

Algorithm 1 Algorithm for MaxSumDispersion

Input: News articles C, k
Output: Set $L(|L| = k)$ that maximizes $f(L)$
Initialize the set $L = \emptyset$
for $i \leftarrow 1$ **to** $\frac{k}{2}$ **do**
  $Find(a, b) = Max_{x,y \in D} d(x, y)$
  Set $L = L \cup \{a, b\}$
  Delete all edges from E that are incident to a or b
end for
If k is odd, add an arbitrary news article to L

3.5.4 Experiments

We have used the same data collection described in previous section 3.4.4. For the evaluation of our approach we have proceeded as follows: For each user, we performed recommendation after a time point t. We used data before time point t for creating the user profile and data starting from time point t for assessment. The time point t is chosen in such a way that there is at least 200 opinions posted by the user. We have used an automatic evaluation to avoid the subjectivity of manual assessments, where we consider the action of commenting on an article to be an indicator that the article fits the interests of the user. So, among the recommended articles, the ones commented by the user are considered relevant. Note that a person might well be interested in an article even though she/he does not comment on it but we did not consider that in our evaluation. As a baseline, we used the strategy proposed in [AAYIM13] where user profiles are represented by a set of entities and their related sentiments. Similarly, to the work done on [AAYIM13] we used the tool OpenCalais to extract entities from news articles content and users' opinions. To compare the results of the different strategies, we use Precision and NDCG at k ($P@k$ and $NDCG@k$). The $P@k$ is the fraction of recommended articles that are relevant to the user considering only the top-k results. It is given by:

$$P@k = \frac{|Relevant\_Articles \cap topk\_Articles\_Results|}{k}$$

Additionally, we compute $NDCG$ to measure the usefulness (gain) of recommended articles based on their (geometrically weighted) positions in the result list. It is computed as follows:

$$NDCG(E, k) = \frac{1}{|E|} \sum_{j=1}^{|E|} Z_{kj} \sum_{i=1}^{k} \frac{2^{rel(j,i)} - 1}{log_2(1 + i)}$$

where Z_{kj} is a normalization factor calculated to make $NDCG$ at k equal to 1 in case of perfect ranking, and $rel(j, i)$ is the relevance score of a news article at rank i.

In our setting, relevance scores $rel(j, i)$ have two different values: 1(relevant) if the news article was commented by the user u, and 0(not relevant) if the news article was not commented by the user u. The precision and $NDCG$ results for the three strategies are shown in Table 3.6. We can see in Table 3.6 that our approach of using global profile outperforms the baseline approach with a gain between 4 and 7 of % in term of precision and 5% in term of ranking at $NDCG@5$. The reason is that most of news articles do not address entities without relating them to some aspects. Moreover, when viewpoints are expressed about entities, they usually

	P@5	**P**@10	**N**DCG @5	**N**DCG @10
Entity-centric Profile [AAYIM13]	0.512	0.551	0.812	0.794
Global Profile	**0**.586	**0**.593	**0**.872	**0**.816

Table 3.6: Precision and NDCG values for all users

refer to certain aspects of those entities. Thus, using only entities to build profiles gives less room for diversification which penalizes the performance. Consequently the combination of both entities and aspects give the best results.

3.6 Conclusions

In this chapter we present our profile model used to describe either users' interests or the content of news articles. We have tested two variations of our profile model on three different applications. Firstly, we have used a sentiment-dependent profile on defining users' interests to identify their political orientation. Secondly, we have used a sentiment-dependent profiles for defining users' interests and the content of news articles on the context of news recommendation. Thirdly, we improve the quality of recommendation by applying a diversification model to reduce the redundancy of news articles using sentiment-dependent profile for defining the content of news articles and a sentiment-independent profile to define users' interests. We have observed that using only aspects to build user's and article profiles slightly performs. The reason is that most of the news articles do not address certain aspects without relating them to some entities. Further, viewpoints expressed about entities are usually referred to certain aspects of those entities. Thus, using only entities to build profiles decreases the performance. Consequently the combination of both entities and aspects give the best results on defining users' interests or even news articles. Sentiments might also be a valuable feature to increase the accuracy of news recommendation. In this thesis we have used sentiment feature to identify the political orientation of users and also as feature on a diversification model having as main goal to reduce the redundancy of the recommended news articles. During this thesis, we have used two techniques to extract entities and aspects. The first one is based on text mining techniques such as tf*idf and pointwise mutual information to define a list of n-grams presenting extracted entities and aspects. The main weakness of this approach that does not differentiate between entities and aspects. In fact, it is used mainly when we don't need to know whether the generated n-grams concern an entity or an aspect as was used to identify the political orientation of users. The second approach used to extract the list of entities and the list of aspects is based on OpenCalais and ODP taxonomy. This approach allows to make the difference between entities and aspects. However it is too time-demanding. Thus, it is not well appropriate for the case where we need to us it on real-time applications.

Chapter 4

Extraction of prominent opinions from news articles

4.1 Introduction

4.1.1 Motivation

News websites, like CNN and Al-Jazeera, provide the possibility to write opinions about any published article and engage in discussions with other users. Figures 4.1 and 4.2 present two examples of opinions space from two different news websites namely The Telegraph and Al-Jazeera. In these figures, we identify three different elements namely opinions, nested opinions and feedbacks. Opinions correspond to comments or reviews, given by users, for a news article to present often an agreement or disagreement about either the whole content of the news article or a specific aspects related to the main topic of the news article. Nested opinion is an opinion that replies to another opinion. A nested opinion has a sentiment orientation that can be positive, negative, or neutral. The set of nested opinions, related to a given opinion, forms a debate. Feedbacks are reactions to a given opinion or nested opinion. They can be one of the following types: like or dislike. Typically, opinions on news websites are unstructured making it hard to

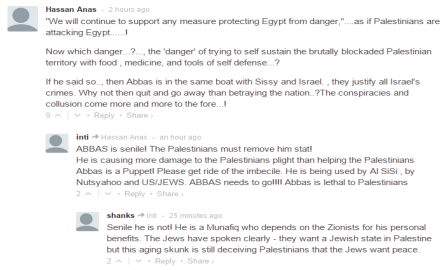

Figure 4.1: Opinions and nested opinions from Al-Jazeera news website

Figure 4.2: Opinions and nested opinions from the Telegraph news website

catch the flow of debates and to understand their main points of agreements and disagreements. Thus, there is a need for organizing opinions to (1) have a better understanding of the main

aspects related to each topic and to (2) facilitate the participation to debates and thus increase the chance of acquiring new opinions.

4.1.2 Contribution

This contribution have as main goal to organize user's opinions in news websites to facilitate their access, understand their trends, and provide a valuable source for enriching article contents by opinions presenting valuable information about the main topic of the news articles. The result of this work can be useful for many applications including news recommendation (such as our contribution in the context of personalized news recommendation which is more detailed in chapter 5), and the assessment of public opinion polls. However, this task can be very challenging since opinions are a free source of information which can be subject to a much of noise. Therefore, we focus on how to select high quality opinions about the different aspects of a given topic.
Our proposed approach goes beyond existing opinions ranking techniques [HLY+12, KPCP06, LHAY08, TR09, DNMKKL09, LM13a] in several ways.
Firstly, determining prominent opinions about daily life topics is much more complex than identifying helpful product reviews as suggested in prior work [HLY+12, KPCP06, LHAY08, TR09, DNMKKL09].
Secondly, the previous proposed approaches use different features to define the helpfulness of a review ranging from its content and the expertise of its author to the preferences of users. However none of them takes into account the relationships between the reviews, meaning the debates that users engage into to discuss a given product. In our work, we take into account the relationships between opinions and their replies, which we call nested opinions, propagating the sentiments along those relations to compute the final score of an opinion.
Thirdly, unlike existing approaches, we define user expertise not only based on explicit ratings, but also on implicit ratings the user gets from his actions. This is due to the fact that explicit ratings are impacted from different kind of bias [LCL+07] such as the winner circle bias, where opinions with many votes get more attention therefore accumulate votes in a disproportionate way, and the early-bird bias where the first opinion to be made tends to get more votes.
To sum up, the novel contribution by this work has the following salient properties:

1. We propose a novel scoring model for opinions based on their relevance to a given topic aspect and their prominence. We define the prominence of an opinion based on how much it is subject to replies and discussions, and the expertise of users reacting to it.

2. We model users' debates as a directed graph of opinions where links can either be positive or negative and represent agreements and disagreements between opinions.

3. We propose a new variation of the PageRank algorithm which handles both positive and negative links between graph nodes. The idea is to boost opinions scores along positive links and decrease them along negative links.

4. We test our approach by running experiments on three datasets crawled from, CNN, The Independent, and The Telegraph Web sites. The results show that our model achieves high

quality results, particularly for highly popular and highly controversial topics having a large amount of user debates. Thus, our model points out a very promising direction towards achieving our goal of finding valuable information despite the dramatic increase of the number of opinions and their noisy nature.

4.2 Debate-based Scoring Model

We consider a query $Q(u, q_1...q_n)$, issued by a query initiator u, as a set of keywords $q_1...q_n$ that describe one or several aspects related to a given news article. The goal is to retrieve high quality opinions that satisfy the user query. Result opinions should contain at least one of the query terms and be ranked according to a query-specific opinion score. Additionally, we propose to boost or decrease the score of an opinion based on the *reactions* of users to it. Users often start debates around a given opinion by providing feedbacks, supportive opinions, opposing opinions, or complementary ones. We capture the impact of these reactions around the opinion by introducing the concept of *prominence*. Both *r*elevance and *p*rominence scores are used to rank opinions that best match with the user query. Formally, we define the score of an opinion O about a news article, given a query Q, as follows:

$$Score(O, Q) = \alpha Rel(O, Q) + (1 - \alpha)Pro(O)$$

where $Rel(O, Q)$ reflects the relevance of opinion O to query Q, $Pro(O)$ reflects the prominence of opinion O, and α is a parameter used to balance the two components of the model.

4.2.1 Opinion Relevance

To compute the relevance of an opinion to user query about a news article A, we use BM25 (or Okapi) scoring function given by:

$$BM25(O, q_i) = IDF(q_i)\frac{f(q_i, O).(k_1 + 1)}{f(q_i, O) + k_1.(1 - b + b.\frac{|O|}{avgol})}$$

Where $f(q_i, O)$ is the count of term q_i in opinion O, $|O|$ is the length of opinion O, $avgol$ is the average opinion length in the collection of opinions about news article A, $k_1 = 1.2$ and $b = 0.75$. $IDF(q_i)$ is the inverse document frequency weight of the query term q_i which is computed as:

$$IDF(q_i) = log\frac{N_e - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N_e is the total number of opinions about a news article A, and $n(q_i)$ is the number of opinions about a news article A containing term q_i. Thus, the relevance score of an opinion is given by:

$$Rel(O, Q) = \sum_{i=1}^{n} BM25(O, q_i)$$

4.2.2 Opinion Prominence

An opinion might trigger reactions in the news platform and thus becomes the starting point of a debate. We call this kind of opinions seed opinions. A seed opinion can get replies from other users, then these replies get other replies and so on, and form a debate. We call an opinion replying to another opinion a nested opinion. Based on these patterns, we model the structure of a debate as a graph of opinions. More specifically, we use a directed tree as shown in figure 4.3 where the root represents a seed opinion. Each non root node is a nested opinion that replies

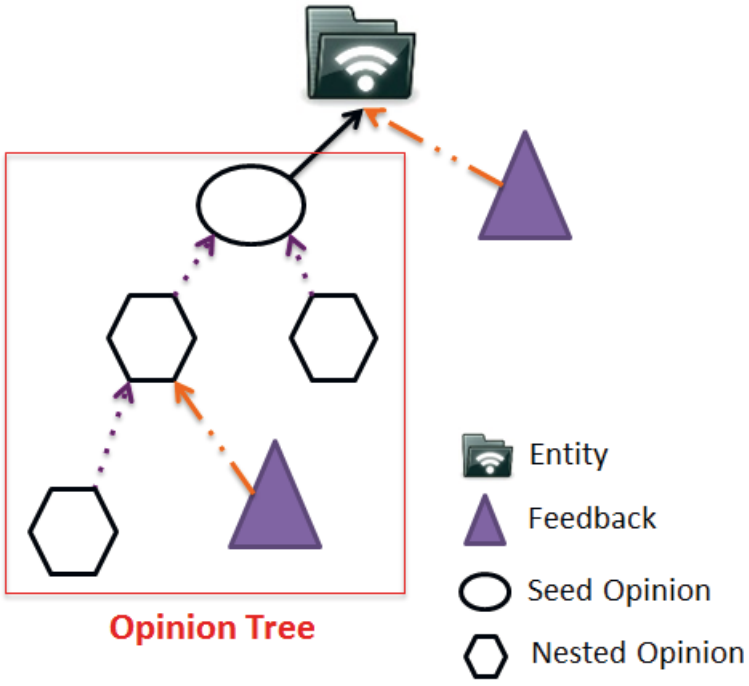

Figure 4.3: Content relations

to its parent. Leaf nodes are nested opinions that do not get any reply. Figure 4.4 shows an example of a debate structure. Edges are directed from children to parents where each link can be either positive or negative reflecting the sentiment the child expresses for its parent. Note that to get information about the sentiment orientation of nested opinions we have used Alchemy API. Using the debate graph, we compute the prominence of each opinion based on the number and quality of its incoming links. The underlying assumption is that prominent opinion are likely to receive many positive links from other opinions while less prominent ones are more likely (i) to receive more negative links or (ii) not to receive any reaction. To this end, we adopt a PageRank algorithm to compute the prominence scores of seed opinions as described in the next section. Note that nested opinions do not take part of query results because considering them as independent components risk to be meaningless. This means that nested opinions are not returned as results of the query. A nested opinion is answering another opinion, so getting it as a single result would be like looking at a part of a discussion without knowing why it started and what it is exactly about. Thus, we return to the user only seed opinions since they are certainly self-contained, and we use the nested opinions to compute the final score of their related seed opinions. When the user is interested in a seed opinion, then he/she can click on it to have access

to the debate that includes all related nested opinions.

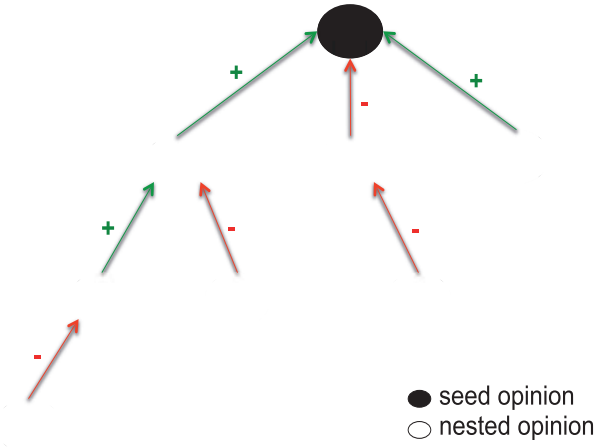

Figure 4.4: Debate Graph

4.3 OpinionRank Algorithm

OpinionRank adopts the same principle of PageRank Algorithm that models user behavior in a hyperlink graph, where a random surfer visits a web page with a certain probability based on the page PageRank. The probability that the random surfer clicks on one link is solely given by the number of links on that page. So, the probability that the random surfer reaches one page is the sum of probabilities of the random surfer to follow links to this page. It is assumed that the surfer does not click on an infinite number of links, but gets bored sometimes and jumps to another page at random. Besides its interpretation, the random jump is used to avoid dead-ends and spider traps in the graph.

Formally, the PageRank algorithm is given by:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)}\right)$$

where $PR(A)$ is the PageRank of page A, $PR(T_i)$ is the PageRank of pages T_i which links to page A, $C(T_i)$ is the number of outgoing links of page T_i, and d is a damping factor which can be set between 0 and 1. As we can see, the PageRank of page A is recursively defined by the PageRanks of pages which link to it. The PageRank of a page T is always weighted by the number of its outgoing links. This means that the more outbound links a page T has, the less will page A benefit from a link to it from page T. The weighted PageRank of pages T_i is then added up. Finally, the sum of the weighted PageRanks of all pages T_i is multiplied with a damping factor d which reflects the probability for the random surfer not stopping to click on links.

To achieve our goal, we propose the OpinionRank algorithm which adapts the PageRank algorithm to the requirements of our approach. Note that in this chapter, we consider *a*nswer relationships between opinions in the debate graph which results in a tree structure. Without loss of generality, we restrict the examples and the presentation of the idea to the tree for the

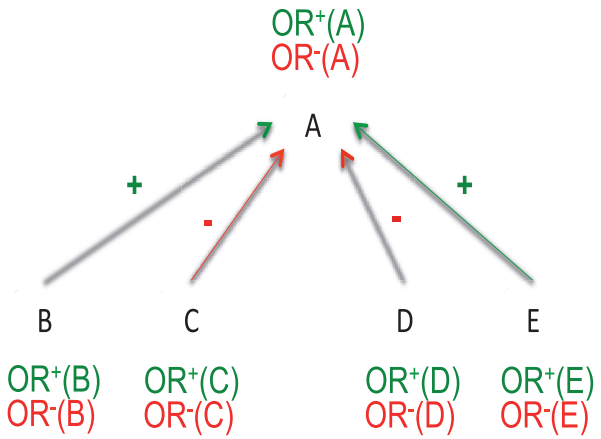

Figure 4.5: Example of OpinionRank

sake of simplicity. The algorithm can be applied to any other graph by simply increasing the number of outgoing links if needed. The reason is that a graph of opinions can be more general if we consider other types of relationships such as agreements and disagreements between reviews about a given product feature, or political tendencies where two opinions are linked if they agree or disagree on a given issue. Thus, our solution is general and can be applied to compute PageRank of nodes in any graph having positive and negative links.

Looking at the debate graph, shown in figure 4.4, we note that an opinion has only one outgoing link because it answers exactly one opinion. Thus, all $C(T_i)$, in the PageRank equation, should be set to 1. Additionally, links between opinions can reflect either positive or negative sentiments showing an agreement or a disagreement between opinions. Thus, a positive incoming link for page A should increase $A's$ PageRank, while a negative incoming link should decrease $A's$ PageRank. However, including subtractions will violate the properties of the probability distribution and give non trivial interpretation for the behavior of the random surfer. Thus, we propose to compute two OpinionRank scores for each opinion A: (1) a score that reflects the probability that the surfer reaches A with a positive sentiment and (2) a score that reflects the probability that the surfer reaches A with a negative sentiment. We can distinguish two cases.

If we have a node X that points to A with a positive link, meaning that X agrees with A, then if the surfer have a given sentiment for X (positive or negative) he will reach A with the same sentiment. By contrast, if X points to A with a negative link, then the surfer will reach A with an opposite sentiment to the one he/she has for X. This means that a positive link keeps the sentiment of the previous node and a negative link inverts it. Based on this principle, a direct application of the PageRank flow equation to compute the two OpinionRank scores of each node would result in the following:

$$OR^{+}(A) = (1-d) + d\left(\sum_{i=1}^{k} OR^{+}(P_i) + \sum_{j=1}^{m} OR^{-}(N_j)\right)$$

and

$$OR^-(A) = (1-d) + d\left(\sum_{i=1}^{k} OR^-(P_i) + \sum_{j=1}^{m} OR^+(N_j)\right)$$

where $OR^+(P_i)$ and $OR^-(P_i)$ are the OpinionRanks of opinions P_i which have a positive link to A.

Similarly, $OR^+(N_j)$ and $OR^-(N_j)$ are the OpinionRanks of opinions N_i which have a negative link to A. $OR^+(A)$ reflects the probability of reaching A with a positive sentiment and $OR^-(A)$ reflects the probability of reaching A with a negative sentiment.

As shown in figure 4.5, reaching A can be done via opinions B and E that agree with A or via opinions C and D that disagree with A. The intuition is that what agrees with B and E consequently agrees with A, and what disagrees with C and D consequently agrees with A. Thus, $OR^+(A)$ is computed as the sum of $OR^+(B)$, $OR^+(E)$, $OR^-(C)$, and $OR^-(D)$. Similarly, what disagrees with B and E consequently disagrees with A and what agrees with C and D consequently disagrees with A. Thus, $OR^-(A)$ is computed as the sum of $OR^-(B)$, $OR^-(E)$, $OR^+(C)$, and $OR^+(D)$.

To be able to compute the OpinionRank scores using the PageRank principle, each of the scores needs to satisfy the flow equation of PageRank given by:

$$r = M.r$$

r is a vector with an entry per page where r_i represents the importance score of page i. Each entry in the vector is computed by $r_j = \sum_{i \to j} \frac{r_i}{d_i}$, as described in the PageRank formula given earlier, which results in $\sum_i r_i = 1$. M is a Stochastic adjacency matrix. Considering page i has d_i outgoing links (in our case d_i is always 1), the entries of matrix M are given by:

$$M_{ij} = \begin{cases} \frac{1}{d_i} & \text{if } i \to j \\ 0 & \text{else} \end{cases}$$

From the matrix representation of the flow equation, it can be easily proven that r is an eigenvector of M and $M.r <= 1$, hence the power iteration method can be used to solve the problem efficiently. Additionally, the random walk interpretation considering the rank vector r as a probability distribution over pages would then lead to the stationary distribution.

It is clear, that in our case, the scores OR^+ and OR^- do not satisfy the flow equation of PageRank since they depend on each other and not only on their previously computed values due to the presence of negative links. This violates the properties of the rank vector r and the probability distribution, thus we cannot directly apply the PageRank algorithm. To solve this problem, we propose to convert the opinion graph with positive and negative links to an equivalent graph with only positive links. The resulting graph keeps the same properties of the original graph and allows the computation of both OpinionRank scores for each type of sentiment.

To convert the opinion graph into a graph with only positive links, we create a mirror node A^m for each opinion node A. The opinion node A would hold a score that reflects the probability that the random surfer reaches A with a positive sentiment, and its mirror node A^m would hold a score that reflects the probability that the random surfer reaches A with a negative sentiment. Further, for each node A, we check its outgoing link and we distinguish two cases.

Firstly, if A points its next node X with a positive link then we create a link from the mirror node A^m to the mirror node X^m. The reason is that, as described earlier, positive links keep the same sentiment from a node to its next node. Thus when the surfer arrives to A, this means that he has a positive sentiment for it, and when following its outgoing link, he will reach X with a positive sentiment.

By contrast, when the surfer reaches at node A^m, this means that he/she has a negative sentiment for A, consequently following its outgoing link he/she will end up in X^m meaning that he will have a negative sentiment for X. Second, if A points its next node X with a negative link, we change the outgoing link of A to point X^m instead of X. Then, we create a link from the mirror node A^m to the node X. Similarly, the reason is that negative links mean that a node has the opposite sentiment of its next node.

Consequently, when the surfer arrives to A meaning that he has a positive sentiment for it, he will have a negative sentiment for X which explains the link from A to X_m. By contrast, when the surfer arrives at node A^m meaning that he has a negative sentiment for A, he will end up in X meaning that he will have a positive sentiment for X. This graph would then allow the computation of OR^+ and OR^- without using negative links. In the following, both scores will be described by a single score OR since the sentiment is embedded in the nodes of the mirror graph.

Figure 4.6 shows an example of the graph transformation explained above. We can see that since node B has a positive link to A, it keeps the same link in the mirror graph while B^m points to A^m. In the case of D which has a negative link to B, the outgoing link is changed to point B^m and a link is created from D^m to B. The resulting graph connects nodes with positive links that can be considered as votes. Thus, we can use the PageRank algorithm on the mirror graph to compute the OpinionRank of each node. Recalling that each node in our case has only one outgoing edge, the OpinionRank Algorithm is given by:

$$OR(A) = (1 - d) + d\,(OR(O_1) + ... + OR(O_n))$$

where $OR(A)$ is the OpinionRank of opinion node A, where A can be an original or a mirror node, $OR(O_i)$ is the OpinionRank of opinion node O_i which links to the opinion node A in the mirror graph, and d is a damping factor which can be set between 0 and 1.

Typically, PageRank assumes a probability distribution between 0 and 1. Hence, the initial value for the score of each page is $\frac{1}{N}$ where N is the total number of pages in the graph. In our setting, we assume a non-uniform probability distribution where the initial score of each opinion is a function of the number of feedbacks it receives from users. The intuition is to boost opinions receiving positive feedbacks and lower those receiving negative feedbacks. In news websites, an opinion can receive two kinds of feedbacks: *l*ike and *d*islike. Thus, for each opinion O_i we set its OpinionRank score $OR(O_i) = \frac{L_i}{F}$ and the OpinionRank score for its mirror to $OR(O_i^m) = \frac{D_i}{F}$ where L_i and D_i are the number of likes and dislikes for opinion O_i respectively, and F is the total number of feedbacks for all opinions about the news article of interest.

When we compute for an opinion O the OpinionRank for positive sentiments and the OpinionRank for negative sentiments, reflected by the OpinionRanks of nodes O and O_m, its prominence score can be computed by:

$$Pro(O) = OR(O) - OR(O^m)$$

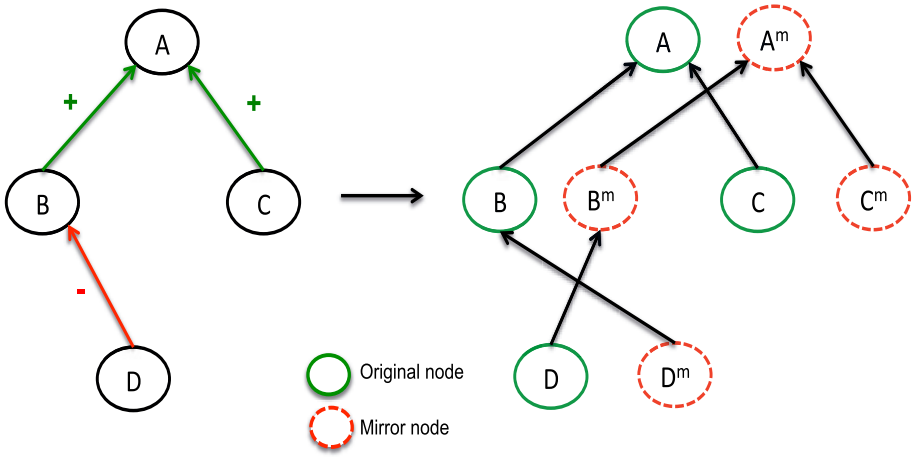

Figure 4.6: Example of a mirror graph

The prominence score gives more importance to opinions receiving many positive reactions and few negative reactions.

4.4 User-Senstive OpinionRank

We propose an extension to the OpinionRank Algorithm that weights the impact of an opinion based on the provider's confidence. For a given news article of topic T, the intuition is to give more importance to opinions provided by users with high confidence in topic T. To this end, at the initialization step, for each opinion O related to topic T we multiply its OpinionRank score by the confidence of its provider on topic T. The key question here is how to compute user

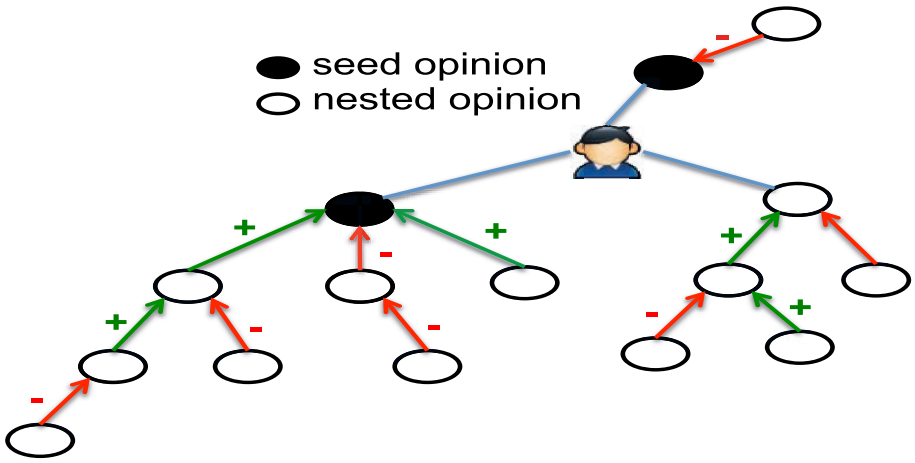

Figure 4.7: User Graph for a given Topic T

confidence. The confidence represents the expertise of a user on a given topic. Computing user's confidence in social networks have been addressed in many studies (e.g. [AAK11, ZAA07]) based on different measures including posts content, groups, and relationships between users. However these techniques have specific assumptions on their applications which make them hard to adapt

in our work. For this reason, we use a simple and intuitive way of computing user's confidence based on the reactions the user receives for his provided contents.

A user can provide different types of content including seed opinions, nested opinions, or feedbacks. Each of the seed and nested opinions belongs to a given topic and might receive reactions from other users. We represent the actions and the reactions a user performs within a given topic T by the graph shown in figure 4.7. To compute the user's confidence, we first compute the prominence score for each of its provided opinions and nested opinions. Then, the user's confidence is computed as the sum of the prominence scores of all the opinions he provided within topic T. Formally, user's confidence is given by:

$$C_{U,T} = \frac{\sum_{i=1}^{n} Pro(O_i)}{Max_{C_T}}$$

Note that user's confidence is normalized over the maximum user's confidence value Max_{C_T} in Topic T. The figure 4.8 presents an example of users' confidence distribution on the topic Politic for around 10000 users.

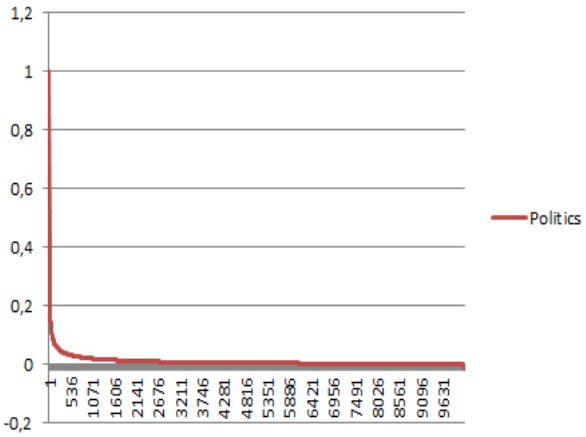

Figure 4.8: An Example of users' confidence distribution-Politic Topic

4.5 Experiments

4.5.1 Experimental Setup

Datasets. We have crawled three datasets of news websites from CNN, The Telegraph, and The Independent, which have a social service allowing users to communicate, discuss around topics, and perform a variety of rating actions. The choice of these datasets was based on their rich content of opinions and the possibility to get information about all actions and feedbacks of users allowing us to have a complete implementation of our model and validate our approach. Note that we could not implement our approach on any of the opinion datasets available in the literature, such as TREC datasets [GCC10, ZYM07, SHMO09], due to their lack of information about user reactions and the relations between seed opinions and nested opinions. We have crawled 40,334 articles from CNN, 40,136 articles from The Telegraph, and 10,408 from The

Independent. We have extracted all seed opinions related to these articles together with their nested opinions, and feedbacks. For each user who provided opinions, we have extracted his/her activities and the feedbacks he received for them. More statistics about these datasets are shown in Table 4.1.

Table 4.1: Datasets Statistics

	#News articles	#Users	#Seed opinions	#Nested opinions	#Feedbacks
CNN	40, 334	753, 185	12, 516, 409	23, 389, 867	80, 585, 030
Telegraph	40, 136	151, 813	7, 096, 741	11, 822, 323	122, 895, 681
Independent	10, 408	62, 171	747, 665	1, 411, 996	14, 445, 661

Baselines. We have used three baselines from the literature to assess the effectiveness of our approach.

BM25. We have chosen BM25 (or Okapi) scoring function to compute the relevance score of each opinion. Since BM25 does not incorporate any exploitation of opinions, it cannot be used for a fair comparison. So, the point of using BM25 as a baseline is to show the impact of the prominence score on the ranking and study the behavior of our model rather than assessing the effectiveness of our approach.

RevRank. We have chosen the RevRank technique [TR09], as a representative of the approaches using helpfulness to rank opinions. The proposed model can be applied on any type of opinions not only product reviews, which have motivated our choice. The idea of this work is to use the dominant terms as indicators for the key-concepts with respect to a specific news article, in order to compute a helpfulness score for each opinion. For example, the terms election or Obama are usually very frequent in the opinions about a news article on Obama presidential campaign. However their contribution to the helpfulness of an opinion is limited as they do not provide the user any new information or any new insights beyond the most trivial. On the other hand, terms like foreign policy and government are not very frequent but are potentially important, therefore the scoring algorithm should allow them to gain a dominance score. We have implemented the process of identifying dominant terms, as suggested by [TR09], in two stages. Firstly we compute, for each news article, the frequency of all terms that appear in its related opinions. Secondly, we re-rank the resulting terms by their frequency in the British National Corpus (BNC). To make the RevRank technique query-dependent, we have excluded the query terms from the process of defining dominant terms. In fact, for each news article a and query term q_i, we select the d most dominant terms to define a feature vector representation of c opinions containing term q_i. We refer to the feature vector having 1 in all of its coordinates as the core vector (CV_i) related to the query q_i. Each opinion O of news article e containing the query term q_i is mapped to V_O, a feature vector representation such that a coordinate k is 1 or 0 depending on whether or not the opinion O contains the k^{th} dominant term. Based on the feature vector representation of opinions and CV_i, we computed the helpfulness score of an opinion O with respect to the query term q_i as follows:

$$Help(O, q_i) = b(O, q_i) \times \frac{V_O.CV_i}{p(|O|) \times |O|}$$

Where $b(O, q_i)$ equals to 1 if the opinion O contains the query term q_i and 0 otherwise, $V_O.CV_i$ is the dot product of the representation vector of opinion O and CV_i, $|O|$ is the length of the opinion O, and $p(|r|)$ is a lowering factor equals to f[1] if $|O| < |\overline{O}|$ and 1 otherwise. The penalization factor f is needed to penalize opinions that are too short while the penalization for an excessive length is already given by the denominator $|O|$. Once the Helpfulness score is computed for each query term, we compute the RevRank score of an opinion O with respect to query $Q(q_1...q_n)$, as follows:

$$RevRank(O, Q) = \sum_{i=1}^{n} Help(O, q_i)$$

Smart. We have chosen SmartNews [LM13a] as our third baseline since this work is very close to ours. The idea of Smart is to rank opinions based on their relevance to a given paragraph which would be a query in our context. After relevance ranking, Smart proposes to boost or decrease the rank of opinions using PageRank scores. To construct the graph of opinions, the authors define each opinion by a vector of terms containing the $tf * idf$ score of each term. Then they define an outgoing link between two opinions O_i and O_j if their cosine similarity is higher than a given threshold [2]. We have implemented Smart in the very same way we have implemented our approach. The difference is that our opinion graph is different since it is based on users' reactions, moreover we additionally boost the scores with user's confidence.

Evaluation Metrics. To compare the results of the different methods, we use two quality measures: Precision at k ($P@k$) and Normalized Discounted Cumulative Gain (NDCG). The $P@k$ is the fraction of retrieved opinions that are relevant to the query considering only the top-k results. It is given by:

$$P@k = \frac{|Relevant\_opinions \cap topk\_opinions\_Results|}{k}$$

Additionally, we compute $NDCG$ to measure the usefulness (gain) of opinions based on their (geometrically weighted) positions in the result list.

$$NDCG(E, k) = Z_k \sum_{i=1}^{k} \frac{2^{rel(i)} - 1}{log_2(1 + i)}$$

where Z_k is a normalization factor calculated to make $NDCG$ at k equal to 1 in case of perfect ranking, and $rel(i)$ is the relevance score of an opinion at rank i. In our setting, relevance scores $rel(i)$ have three different values: 2(very relevant), 1(relevant), and 0(non relevant).

4.5.2 Strategies Under Comparison

We evaluate the effectiveness of our scoring model by using different strategies. For each news article, we rank its related opinions using the following strategies:

Rel. Results are ranked using the BM25 scoring described in section 4.2.1.

RevRank. Results are ranked on the basis of RevRank technique described in section 4.5.1.

Smart. Results are ranked based on Smart technique described in section 4.5.1.

[1] We have experimentally chosen $f = 3$

[2] we choose experimentally 0.5 as a threshold

Rel+Pro Results are ranked based on relevance and prominence computed using the opinionRank algorithm described in section 4.3.
Rel+Pro(Conf). Results are ranked based on relevance and prominence computed using the User-Sensitive opinionRank algorithm described in section 4.4 Finding a good set of queries is not an easy task since users might be interested in searching opinions on different aspects of a given news article, depending on their personal context and interests. Thus, we have conducted a user study with manual query selection and assessment. The task was carried out by 30 human

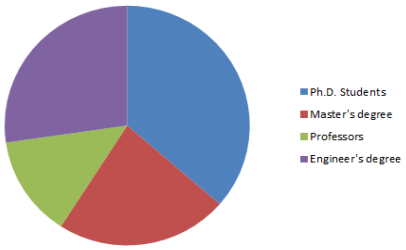

Figure 4.9: Education level of assessors

assessors who were researchers and students which are not involved in this project. More details about about those assessors are shown in figure 4.9. We have asked our human assessors to choose news articles of interests and suggest queries related to them according to their interests. This process resulted in 206 queries dealing on 206 topics from the three datasets. In Table 4.2, we have selected randomly some chosen queries classified by category. More precisely, we have tested 108 queries on `CNN`, 70 queries on `The Telegraph` and 28 queries on `The Independent`. For each query, we have applied the strategies described earlier and got the top 20 results for each strategy. We have shown the total results to our human assessors who evaluated them according to the following guidelines: (1) an opinion is considered as non relevant, and gets a score of 0, if it does not give an opinion related to the query, (2) an opinion is considered as relevant if it contains information about the query. In this case it gets a score of 1, (3) an opinion is considered very relevant if it is relevant to the query and provides additional information that was not given by the news article itself such as new view point, new arguments, or references to more information about the query topic. In this case it gets a score of 2. The assessment is done without having any idea about the adopted strategy.

4.5.3 Results and Analysis

The overall results on the three datasets are shown in Table 4.3. We can see that our approach almost always outperforms the most effective baseline approach by an increase up to 12% in terms of precision, and 3% in terms of NDCG. This shows that the prominence component of the model plays an important role in improving users' satisfaction. The datasets CNN, and The Telegraph have very similar performances while The Independent is slightly worse regarding $NDCG$ values. The reason is that The Independent dataset does not contain a lot of users' reactions, which makes the prominence component weaker. Additionally, the dataset having less queries makes it very sensitive to outliers. Average number of opinions, nested opinions and

Category	Generated queries
Business	Avoid inheritance tax, Titan French workers UK economy rating, European banker's bonuses
Media	Stop page3 Sun, Armageddon movie Jenna star news, Death Annette Funicello
Living	Infertility children couple, Horse meat scandal Antibiotics resistance, Women egg freezing
opinion	Solitary confinement for kids, Gun control laws Terrorist Boston attack, Dental patients tested for hepatitis
Politics	Obama Guantanamo plan, Obama Romney Job plans American president vote, Israel settlement

Table 4.2: Sample seed queries used to rank opinions

Table 4.3: Precision and NDCG values per DATASET

		P@10	**P**@20	**NDCG@10**	**NDCG@20**
CNN	*Rel*	0.610	0.624	0.816	0.798
	RevRank	0.708	0.645	**0.844**	0.797
	Smart	0.712	0.667	**0.844**	**0.833**
	Rel+ Pro	0.771	0.737	0.800	0.799
	Rel+ Pro (Conf)	**0.801**	**0.772**	0.836	0.832
Telegraph	*Rel*	0.665	0.674	0.811	0.804
	RevRank	0.789	0.704	0.862	0.844
	Smart	0.799	0.712	0.857	0.822
	Rel+ Pro	0.839	0.807	0.858	0.848
	Rel+ Pro (Conf)	**0.851**	**0.835**	**0.870**	**0.858**
Independent	*Rel*	0.694	0.652	0.843	0.832
	RevRank	0.773	0.710	**0.879**	**0.866**
	Smart	0.757	0.713	0.864	0.858
	Rel+ Pro	0.794	0.760	0.849	0.825
	Rel+ Pro (Conf)	**0.805**	**0.778**	0.854	0.831

feedbakcs per news article for the three datasets are shown in figure 4.10. It is also observed that User-sensitive opinionRank improves the effectiveness of opinionRank algorithm. Therefore, including user's confidence to compute prominence scores gives better results for opinion ranking. We also note that *S*mart is the technique with a performance close to our approach and in few cases slightly improves performance in terms of NDCG.

To have a more insightful analysis, we looked at the topic of news articles for 193 queries falling into 5 categories: Business (33 queries), Media (32 queries), Living (34 queries), Opinions (40 queries), and Politics (54 queries). Our results per category are shown in table 4.4. For all categories, our strategy almost always improves the two measures of Precision and NDCG. It is also noteworthy to say that the performance gain varies with the topic of the news article. For example, in the *Business* category the Precision@10 increased to 87.5% using our User-sensitive opinionRank model from 76.8% using Smart technique. By contrast, for the *Politics*, the absolute

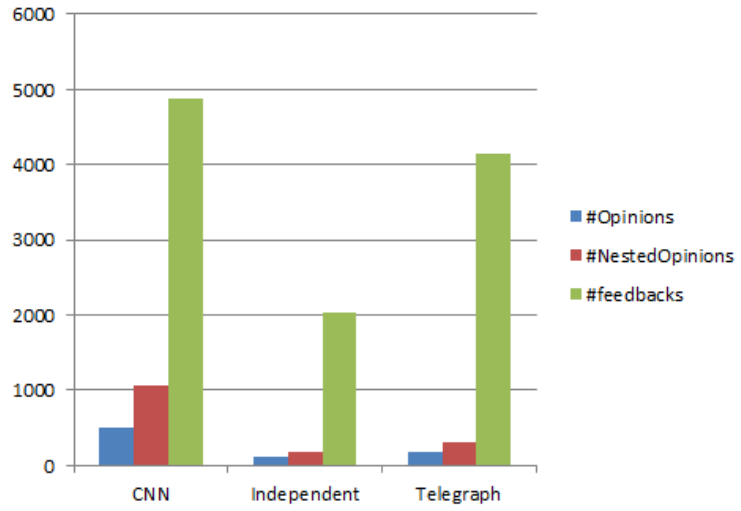

Figure 4.10: Average number of opinions, nested opinions and feedbacks for news articles per dataset

improvement is of 0.3% in Precision@10 which increases to 2% Precision@20. However, the rise

Table 4.4: Precision and NDCG values for Relevance-based Ranking per category

		Precision		**N**DCG	
		P@10	**P**@20	**N**DCG@10	**N**DCG@20
Business	*Rel*	0.85	0.625	**0.911**	0.790
	RevRank	0.768	0.656	0.907	0.890
	Smart	0.768	0.781	**0.911**	**0.906**
	Rel+ Pro	0.862	0.815	0.892	0.889
	Rel+Pro (Conf)	**0.875**	**0.837**	0.902	0.895
Media	*Rel*	0.587	0.556	0.857	0.813
	RevRank	0.668	0.571	0.849	0.828
	Smart	0.662	0.593	**0.882**	**0.865**
	Rel+ Pro	0.687	0.646	0.761	0.750
	Rel+ Pro (Conf)	**0.737**	**0.696**	0.832	0.813
Living	*Rel*	0.742	0.725	0.812	0.813
	RevRank	0.814	0.760	0.852	0.836
	Smart	0.742	0.732	0.835	0.826
	Rel+ Pro	0.814	0.817	0.858	0.845
	Rel+ Pro (Conf)	**0.821**	**0.835**	**0.863**	**0.861**
Opinion	*Rel*	0.621	0.683	0.797	0.794
	RevRank	0.757	0.681	0.859	0.830
	Smart	0.790	0.750	0.861	0.851
	Rel+ Pro	0.866	0.821	0.860	0.853
	Rel+ Pro (Conf)	**0.872**	**0.846**	**0.869**	**0.862**
Politics	*Rel*	0.667	0.645	0.789	0.772
	RevRank	0.739	0.611	0.854	0.747
	Smart	0.8	0.75	0.845	**0.844**
	Rel+ Pro	0.791	0.746	0.812	0.805
	Rel+ Pro (Conf)	**0.803**	**0.775**	**0.861**	0.803

in our approach can be much higher for many individual queries. Examples are shown in table 4.5. For instance, considering the query 'Osama bin laden death', the precision surged from 50% using RevRank and 65% using Smart to 100% using our User-sensitive opinionRank approach, and the NDCG from 73.9% and 67% to 98.1%. Similarly, we improve the precision of the query 'Gun control and suicide' from 60% and 70% to 100%, and the NDCG from 85.8% and 86% to 92.2%, giving high quality results.

The experimental results show that our model almost always outperforms the native rankings of opinions by a significant margin. In some cases, however, the gain improvements are small and generally depend on the category type. One explanation to this behavior is that topics of *Media* and *Opinion*, and *Politics* are usually very popular, gossip appealing, controversial, and about daily life subjects. Examples include, *Gun control and suicide* and *Kevin hart arrest*. Due to the large number of opinions and discussions in these categories, the prominence score improves the overall performance of our model. By contrast, other categories, such as *Living*, generally contain topics that are not very controversial and consequently generate less debates and discussions between users (e.g.,*School academies overspend*). Consequently, our model is less effective categories containing unpopular topics.

A closer look at Table 4.5 shows that we are performing particularly well for news articles about highly controversial and highly popular topics, which are subject to gossiping. For instance, we can clearly see the difference between the query about the *Gun control and suicide*, a highly controversial topic, and the *School academies overspend* where opinions are less diverse. The precision improves for the first query from 60% using RevRank and Smart techniques to 100% using User-sensitive opinionRank technique, while there is a slight improvement for the second query.

To sum up, our model works best for topics with very large number of opinions and debates. Figure 4.11 presents statistics about the average number of opinions, nested opinions and feedback per category for the queries extracted from CNN. It is clear that the number of reactions depends mainly on the topic of the news article. For instance, queries about the topic Business or Opinion gather more reactions than queries about topics Living or Culture. This is a very promising step towards our initial goal of selecting valuable information from the increasing amount of opinions in news sites. It is also clear that our model does not perform well with categories and news articles related to unpopular topics. To cope with that, one solution for selecting prominent opinions is to use the Smart technique which creates links between opinions based on their similarities. An opinion with many incoming links will thus be prominent even though there are no strong debates around it. Additionally, it would help filtering noisy content.

To have a concrete idea about the results of our approach, we take our motivation example of the news article "Boston Bombing" and we retrieve the top5 opinions about the topic by using (1) relevance only and (2) relevance and prominence. We can see, in Figure 4.12 that both result lists are relevant, however using the prominence score results in more opinions that bring new insights about the topic which is clearly an added value.

Table 4.5: Individual Precision and NDCG values for Relevance-based and Helpfulness-based Ranking

	P@20					**N**DCG@20				
	Rel	RevRank	Smart	Rel+ Pro	Rel+ Pro(Conf)	Rel	RevRank	Smart	Rel+ Pro	Rel+ Pro(Conf)
Titan workers problem	0.65	0.65	0.7	0.85	**0.95**	0.681	0.805	0.827	0.977	**0.988**
Italy election Europolitcs	0.6	0.7	0.75	0.9	**0.95**	0.816	0.939	0.934	0.979	**0.985**
School overspend	0.55	0.6	0.6	0.55	**0.65**	0.756	0.823	0.767	0.780	**0.882**
Kevin Hart arrest	0.4	0.45	0.45	0.75	**0.85**	0.808	0.825	0.825	0.746	**0.981**
Rennard sexual scandal	0.45	0.55	0.65	0.8	**0.85**	0.749	0.708	0.707	0.788	**0.785**
antibiotics resistance	0.75	0.8	0.8	0.9	**0.95**	0.876	0.961	0.881	0.955	**0.962**
Osama bin laden death	0.45	0.5	0.65	0.9	**1.0**	0.733	0.739	0.677	0.960	**0.981**
Gun control and suicide	0.6	0.6	0.7	0.7	**1.0**	0.779	0.858	0.863	0.637	**0.922**
Job plans US candidates	0.65	0.65	0.7	0.8	**0.85**	0.619	0.734	0.823	0.935	**0.942**
Russian Chechen War	0.6	0.6	0.65	0.7	**0.85**	0.695	0.673	0.776	0.767	**0.920**

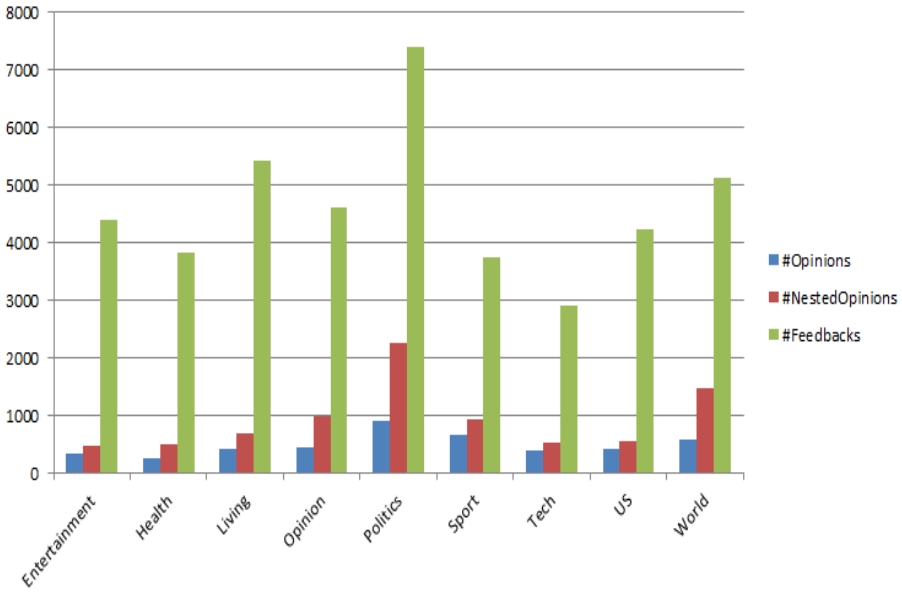

Figure 4.11: Average number of opinions, nested opinions and feedbacks for news articles per category

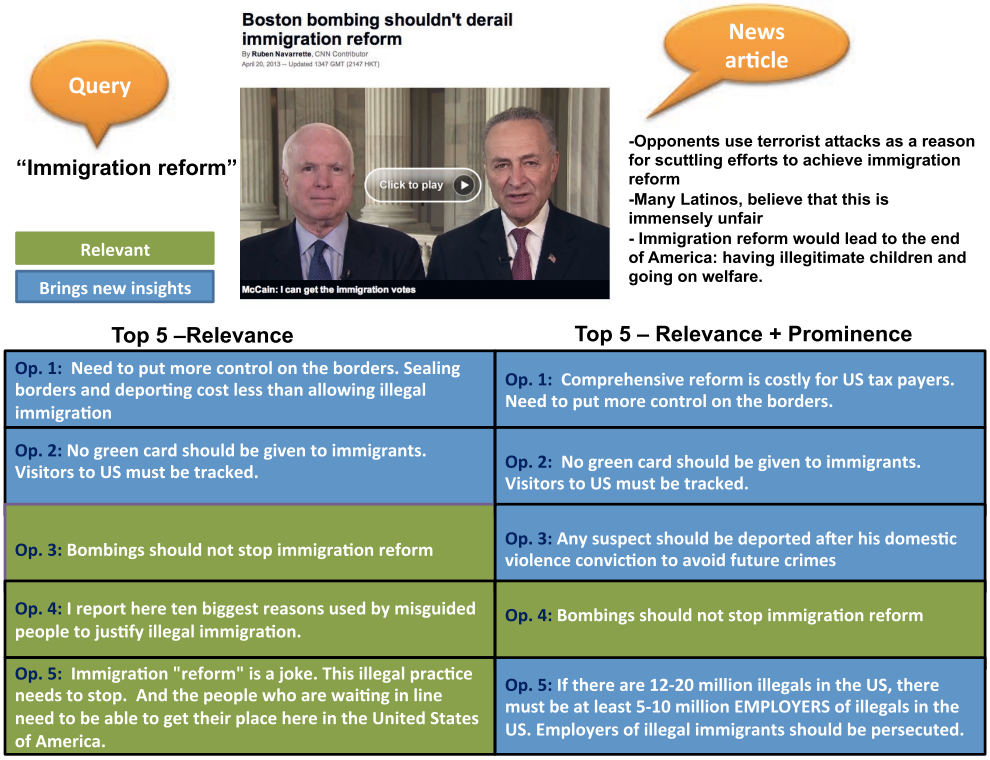

Figure 4.12: An Example of opinions Ranking for immigration reform

4.6 Conclusions

Retrieval and ranking of opinions in product reviews has received a great attention in the prior literature. In this chapter, we generalized this problem to retrieving and ranking opinions in news media, and paid particular attention to the exploitation of users' debates in such platforms to

retrieve the most prominent opinions. Our experiments showed that these debates, enhanced by explicit feedbacks, are definitely valuable and should be taken into account for ranking opinions. Thus, our proposed approach achieves its best performance when the query topic is highly controversial or popular. We have shown that our model can effectively exploit the large amount of debates and reactions to select the best opinions that are relevant to the user query. Further, using user's confidence to compute prominence scores gives better results for opinion ranking. An explanation is that opinions given by experts have more relevance and impact than opinions given by novice users It is also clear that our model does not perform well with categories and news articles related to unpopular topics.

Chapter 5

Enriching news articles using hidden aspects

5.1 Introduction

5.1.1 Motivation

News websites often allow users to express freely their opinions about the published information, except in cases where some users can go beyond the pale and consequently their opinions can be censured. The editorial content is generated using a top down approach where the provided information follows the publisher plan and target specific aspects that are made explicit in the editorial content through the headline of the news article and the article content. By contrast, opinions follows a bottom up approach where users start discussing some specific issues forming debates around a given topic. Consequently, they might reveal aspects that are not well described or even not presented in the content of news articles. We called, hidden aspects, such aspects which are not confined to any predefined plan and thus extend information by continuously bringing new insights. In other terms, we define hidden aspects as the set of aspects revealed on users' opinions which are related to the topic of news article, but which are not clearly described or even not presented in the editorial content . For example, Figure 5.1 shows a concrete example of extracting different hidden aspects from the users' opinions and which are not presented on the editorial content. This news article was published on CNN news websites and its title is *Boston bombing shouldn't derail immigration reform.* In this news article, the main content is about the consequences of Boston bombing event on the *Immigration reform.* By analyzing the content of users' opinions, we have discovered different aspects which are related to the main topic of this news article but which are not explicitly presented or described on the editorial content such as Green card, Border Control, and Employers Persecution. Table 5.1 shows more detailed examples of the explicit aspects provided by the editorial content extracted from the news article. It also shows the hidden aspects expressed by users on the same topic. Both explicit and hidden aspects presented in this table are extracted using a manual analysis of the article content and a list of selected opinions. The rich structure of news websites make them a valuable source for building knowledge bases about daily-life topics and their various sides. For this purpose, the knowledge extraction process should use editorial content as a seed for getting information, about

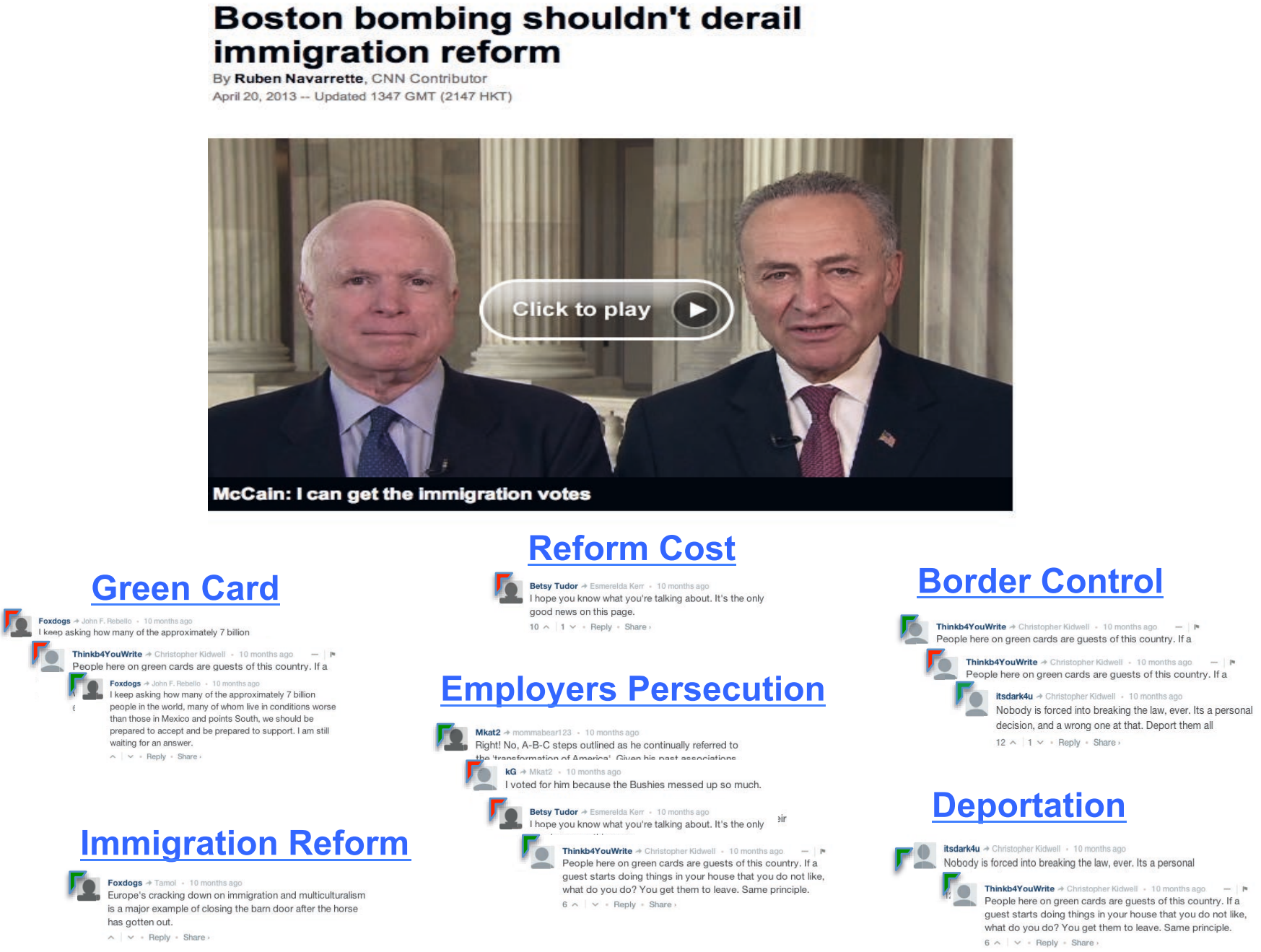

Figure 5.1: Example of revealed hidden aspects from a CNN news article

Explicit topic aspects from an editorial content
- Opponents use terrorist attacks as a reason for scuttling efforts to achieve immigration reform - Many Latinos believe that this is immensely unfair - Immigration reform would lead to the end of America
Hidden topic aspects from opinions
- Comprehensive reform is costly for US tax payers - Need to put more control on the borders - No green card should be given to immigrants - Visitors to US must be tracked - Bombings should not stop immigration reform - Employers of illegal immigrants should be persecuted - Sealing borders and deporting cost less than allowing illegal immigration - Any suspect should be deported after his domestic violence conviction to avoid future crimes

Table 5.1: Example of topic aspects of “Immigration reform"

any topic of interest, and then extend it with opinions to have a more complete picture. Thus, it is important to increase the number of opinions on published news articles and more importantly make them much apparent for users. This calls for an effective strategy for news recommendation that would provide users the news articles that match with their interests and on which they are willing to comment. The willingness to comment on a news article is driven by the kind of aspects discussed by users on the topic. Consequently, it is important to capture that information when recommending an article to a user. A straightforward way to achieve this goal is to enrich the content of news articles with user's opinions for a more effective recommendation. Typically, the editorial content is a reliable source of information since it is provided by professionals and is carefully reviewed before publication. However, opinions is a free source of information which can be subject to a lot of noise. Thus, it is important to select only prominent opinions using ranking strategy. Moreover, these opinions have to be representative which requires the application of diversification techniques to capture a wide set of aspects.

5.1.2 Contribution

Several approaches employing opinions for purpose of search and recommendation have been proposed previously [AAYIM13, SKKL12, YYLF09, GZV12, YYLF09]. Shmueli et al., [SKKL12] analyze the co-commenting patterns of users for recommending news articles to users who will likely comment them. Abbar et al., [AAYIM13] extract from the content of users' opinions a set of features to be used for news articles diversification. The works which are close to ours are by Yee et al., [YYLF09] and Ganesan et al., [GZV12] which exploit users' opinions to enrich the content of documents. Yee et al., [YYLF09] prove that the potential of Youtube users' opinions in the search index yields up to a 15% improvement in search accuracy compared to user-supplied tags or video titles.
Similarly, Ganesan et al., [GZV12] use the content of customer's reviews to represent entities (hotels and cars) in the context of entity ranking. They measure the score of entities based on

how their reviews match with users' keyword preferences.
Two main points make the difference between our work and these approaches. First, Ganesan et al., [GZV12] use product reviews which belong to an already known set of aspects. In our work, we are interested in aspects about daily life topics reported by news articles. These aspects are not classified but we extract them automatically using an unsupervised approach. Second, opinions on news websites usually contain a lot of noise, thus unlike the approach by Yee et al., [YYLF09] we do not use all opinions to enrich the content of news articles but we select only the topk opinions. Additionally, we perform diversification on those opinions to have a larger coverage of new aspects. To the best of our knowledge, there is no existing work that exploits opinions to enrich news articles in the context of personalized news recommendation. Our work aims at providing an effective news recommendation to facilitate the access of users to published news articles. More importantly, it also aims at motivating users to comment on the news articles of interests and get involved in discussions with other users. These opinions are a valuable source of information that can bring new insights by revealing hidden aspects about topics and thus extending our knowledge about daily life topics. The novel contribution by this work has the following salient properties.

1. Firstly, we propose a novel recommendation approach that (1) enriches the content of news articles with opinions to improve the effectiveness of recommendation, (2) ranks news article opinions to select only prominent content and filter noise, and (3) offers an opinion diversification model based on authorities, semantic and sentiment diversification.

2. Secondly, we test our approach by running large experiments on four datasets crawled from CNN, The Independent, The Telegraph, and Al-Jazeera. The results show that our model achieves high quality results, particularly for diversified opinions which provide insightful aspects for enriching the content of news articles.

5.2 Problem formulation

Our goal is to propose a personalized news recommendation model tailored to users' interests. In our setting, we identify the profile of a given user from the type of articles he/she reads by exploiting the opinions he posts on the news website.
Formally, we define a news website to contain a set U of users and a set A of articles. Each user $u_i \in U$ provides a set of opinions C_i about a set of articles A^i where $A^i \subset A$. We assign a profile P_{u_i} for each user u_i extracted from the set of his opinions C_i.
Similarly, we assign a profile P_{a_j} for each article a_j extracted from its content. When user u_i connects to the news website, we compute the similarity between the user profile P_{u_i} and the profiles of the set of articles $A_t \subset A$ where A_t corresponds to all articles published in the time interval t. By this way, we can restrict our search space to any time period specified by the user, otherwise we set it to the latest period. The time interval can range from few days to months depending on user's preferences. The k most similar articles to user profile P_{u_i} are then recommended to user u_i.
We have adopted cosine similarity to compute the similarity between the user profile P_{u_i} and news articles profiles A_t. This measure has been shown to be very effective in measuring similarity

between documents [Sin01]. In a standard search problem, a document is represented by a vector of n dimensions where a term is assigned to each dimension and the value of the dimension represents the frequency of the term in the document. In our setting we are interested in computing similarity between n-grams, so each profile is represented by a vector where the dimensions of each vector are assigned aspects and the value of each dimension represents the average *tf*idf* score of the aspect terms for the given profile. Formally the cosine similarity between a news article profile A and a user profile P is given by:

$$Similarity(P, A) = \frac{P.A}{||P||||A||}$$

where P is the vector corresponding to the user profile u, and A is the vector corresponding to the news article profile a.
The key components of our model are user's and article profiles. To define the profile of a given user u_i, we collect the opinions C_i he/she has expressed during a period of time T. Then, we extract from them all the aspects user u_i has discussed. As a result, the profile P_{u_i} of user u_i is described by a set of aspects $\{as_1, ...as_m\}$. Similarly, we define the profile P_{a_j} of a given article a_j as a set of aspects $\{as_1, ...as_l\}$ extracted from its content.

A successful recommendation of an article a_j to user u_i would result in user u_i commenting on article a_j. Commenting an article goes beyond just finding its content interesting to read. It shows the involvement of the user in discussing, with other users, his/her opinions about the news article. These discussions can turn around aspects that are explicitly mentioned in the content of the news article or some hidden aspects brought up by users. In both cases the interesting aspects for the user (besides the general topic of the news articles) are the aspects discussed in the opinions of users. Thus, it will be valuable enriching the content of news articles with user opinions for a more effective recommendation. Naturally, the increasing number of opinions on news articles makes them more subject to noise and redundant information. To this end, we need to select the most prominent opinions to be added to the content of a news article. Moreover, the set of selected opinions should be diverse to avoid redundancy and increase the coverage of new topic aspects.
To sum up, we need two main components to perform news recommendation using our model. Firstly, a strategy for aspects extraction from users' opinions and the content of news articles. Secondly, an opinion diversification model. In the following we describe each of these components separately.

5.3 Aspects Extraction

In the following paragraph, we describe how aspects are extracted from users' opinions and news article content. Note that the same extraction method is used for both types of content. The only difference is the computation of aspects scores which depends either on the corpus of opinions or on the one of articles.

5.3.1 Generation of Candidate Aspects

To extract aspects from the opinions of user u_i, we first identify the sentences, using OpenNLP, expressed in all his/her opinions. We take all the extracted sentences, and we rank their contained terms using $tf * idf$ scoring function. In our work, tf represents the term frequency in the set of sentences of the user u_i, and idf represents the inverted document frequency in the set of sentences of all users in the news website. The idea is to select highly scored unigrams as a base for generating candidate aspects. Similarly, for a given article a_j, we use the same unigram extraction from its content however this time tf represents the term frequency in the set of sentences of article a_j and idf represents the inverted document frequency in the set of sentences of all news articles in the news website. From the selected unigrams, we generate bi-grams, then we take the bi-grams as input and we build a set of n-grams by concatenating bi-grams that share an overlapping word. At each step we take the topk n-grams based on the score of their composed unigrams[1]. We check the redundancy of the generated candidates using Jaccard similarity [RV96]. If two n-grams have a similarity higher than a defined threshold, we would discard one of them. In our work, we have set the maximum length of the n-grams to 3 since there were no meaningful n-grams with a higher length.

5.3.2 Selection of Promising Aspects.

Generating n-grams that have high $tf*idf$ scores is not enough for identifying the aspects discussed in users' opinions and articles content. It is important for the words in the generated n-grams to be strongly associated within a sentence in the original text to avoid covering incorrect information. To capture this association, we use the modified *pointwise mutual information* [TC03], already described in section 3.3.2, of words in n-grams based on its alignment to the narrow opinions of each user or the article content. Table 5.2 shows some examples of n-grams extracted from the content of a news article and its related opinions using our unsupervised approach.

N-grams	**Generated aspects**
Unigrams	Paris, Obama, Crisis, Settlement, Palestine Israel, Quds, Syria, Rape, Economy, European, Live, RealMarid
Bi-grams	Israel settlement, Obama president, Avoid tax, Titan workers UK economy, European banker's, London Live
Tri-grams	Solitary confinement kids, Stop page3 Sun, Gun control laws Terrorist Boston attack, Dental patients hepatitis

Table 5.2: Sample of generated aspects

5.4 Opinions Diversification

In this section, we introduce the technique used to diversify opinions on news sites which was inspired by the work in [KG11]. We aim by diversifying opinions to remove redundancies and

[1]In this work we have set k=500

thus provide a wide coverage of topic aspects. We are given a set of opinions $C = \{c_1, c_2, ..., c_n\}$ where $n \geq 2$. Our goal is to select a subset $L_k \subseteq C$ of opinions that is diverse. We assume three main components that define the diversity of a set of opinions : *authority*, *semantic diversity*, and *sentiment diversity*. Naturally, before discussing whether a set is diverse or not, it should first contain opinions with high authority scores. This is why the *relevance* is important to include in diversification models [GS09a, AGHI09]. Note that the *authority* of each opinion is given by applying our approach of opinion ranking described in section 4.3. To diversify a set of opinions, we need to give more preference to non similar opinions. We assume that two opinions are dissimilar if (1) they discuss different entities or aspects, and/or (2) they exhibit different sentiments about the news article topic, including positive, negative, and neutral sentiments. To satisfy these two requirements, we define two distance functions. The first one is a *semantic distance* function $d : C \times C \rightarrow R^+$ between opinions, where smaller the distance, the more similar the two opinions are. This distance measures the *semantic diversity* of the set. The second one is a *sentiment distance* function $s : C \times C \rightarrow R^+$ between opinions, where smaller the distance, the closest in sentiments the two opinions are. The sentiment distance is used to compute the *sentiment diversity*. Using the diversification framework proposed in [GS09a] and detailed in section 3.5.3, we formalize a set selection function $f : 2^C \times h \times d \times o \rightarrow R^+$, where we assign scores to all possible subsets of C, given an authority function $h(.)$, a semantic distance function $d(.,.)$, a sentiment distance function $s(.,.)$, and a given integer $k \in Z^+ (k \geq 2)$. The goal is to select a set $L_k \subseteq D$ of opinions such as the value of f is maximized. In other words, the objective is to find:

$$L_k^* = Max_{L_k \subseteq D, |L_k|=k} f(L_k, h(.), d(.,.), s(.,.))$$

where all arguments other than L_k are fixed inputs to the function. The goal of this model is to maximize the sum of the authority, the semantic dissimilarity, and the sentiment dissimilarity of the selected set of opinions. The function we aim at maximizing can be formalized as follows:

$$f(L) = \alpha(k-1) \sum_{a \in L} h(a) + 2\beta \sum_{a,b \in L} d(a,b) + 2\gamma \sum_{a,b \in L} s(a,b)$$

where $|L| = k$, and $\alpha, \beta, \gamma > 0$ are parameters specifying the trade-off between relevance, semantic diversity, and sentiment diversity[2]. The model allows to put more emphasis on relevance, on semantic diversity, on sentiment diversity, or on any mixture of these measures. The authority scores of opinions are computed based on opinion ranking strategy described in section 4.3 and the semantic distance is computed based on Jaccard similarity function. As for sentiment distance, we define it as follows:

$$s(a,b) = \begin{cases} 0, & \text{if a and b have the same sentiment orientation;} \\ 1, & \text{otherwise.} \end{cases}$$

where the sentiment orientation includes *positive*, *negative*, and *neutral* sentiments. We have used Alchemy Api to identify the sentiment of each opinion. As described in section 3.5.3, our diversification problem follow the same principle and model MaxSumDispersion problem which

[2]In our implementation we have set $\alpha = \beta = \gamma = 1$

have as objective to maximize the sum of all pairwise distances between points in a set S. We have used the 2-approximation algorithm proposed in [HRT97,KH78] and illustrated by algorithm 2 to maximize our MaxSum objective *f*.

Algorithm 2 Algorithm for MaxSumDispersion

Input: Opinions C, k
Output: Set $L(|L| = k)$ that maximizes $f(L)$
Initialize the set $L = \emptyset$
for $i \leftarrow 1$ **to** $\frac{k}{2}$ **do**
  $Find(a, b) = Max_{x,y \in D} d(x, y)$
  Set $L = L \cup \{a, b\}$
  Delete all edges from E that are incident to a or b
end for
If k is odd, add an arbitrary opinion to L

5.5 Experiments

5.5.1 Real-World Dataset

We have crawled four real datasets based on the activities of 645 users on four news websites namely CNN, The Telegraph, The Independent, and Al-Jazeera. The choice of these users was based on two key-properties: the number of users' opinions and whether they follow the four news websites or not. More precisely, we start, by selecting the most active users on each news website based on the number of posted opinions and then we choose principally users that have posted opinions on the four news websites. This process results in the selection of four datasets, the first one contains the activities of 150 users which are a subset of the most active users on CNN, the second dataset contains the activities of 180 users which are a subset of the most active users on The telegraph, the third dataset contains the activities of 164 users which are a subset of the most active users on The Independent and the last dataset contains the activities of 151 users which are a subset of the most active users on Al-Jazeera. Note that the four news websites mentioned earlier are using the same framework for commenting their news articles. It allows the extraction of the opinions of each user on the four news websites. For each of those users, we have collected the details of his opinions in the four news websites mentioned earlier (content, published time, etc.). Additionally, we have collected the details of all the commented news articles (e.g., news title, content, opinions, published time, etc.) from May 2010 to December 2013.

These news websites contain stories ranging from national and international events, sports, business to culture and entertainment. Statistics about the number of commented articles and the number of opinions for each dataset are shown in Table 5.3. It is clear from these statistics that most of users comment more than once opinion on news articles because in a lot of cases the number of opinions is largely higher than the number of commented news articles.

We can also see, in Figure 5.2, the distribution of users' opinions and commented news articles for each dataset.

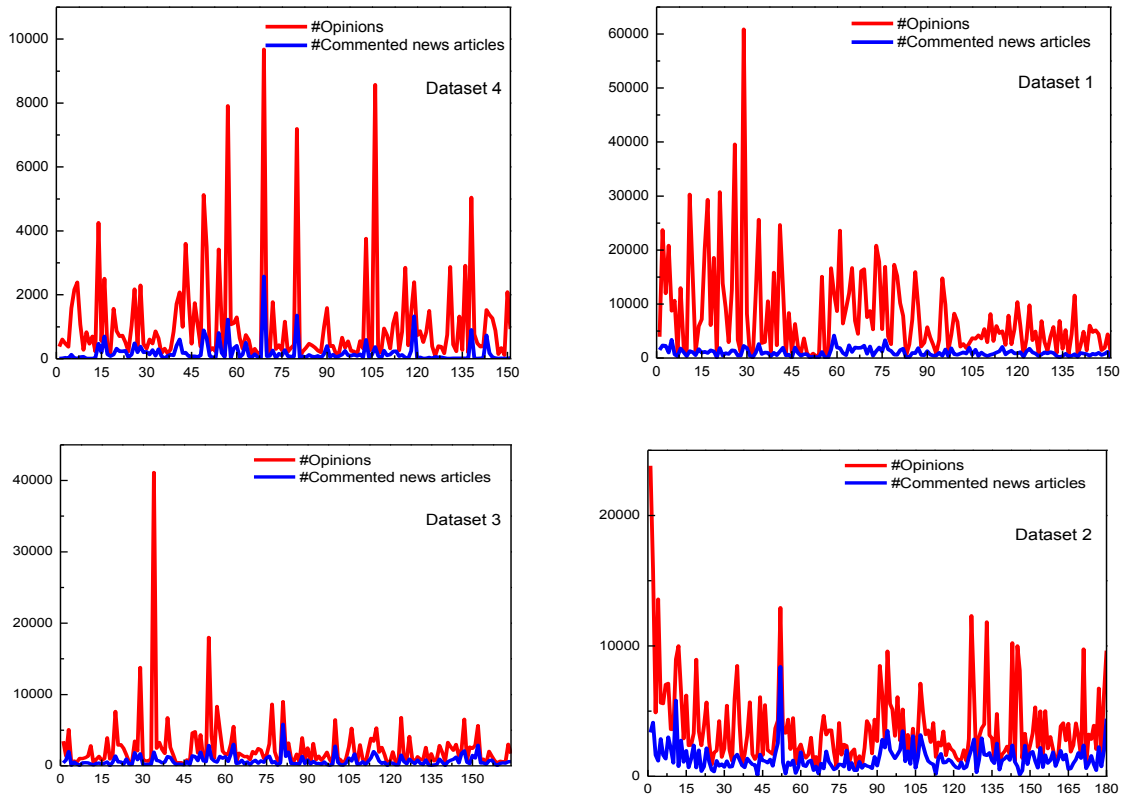

Figure 5.2: Users' Opinions and commented News articles Distribution per user for the four datasets

	Dataset1 (CNN Seed)		*Dataset2* (Telegraph Seed)	
	#articles	#opinions	#articles	#opinions
CNN	41, 245	12, 056, 789	665	874, 879
Telegraph	1, 908	1, 257, 645	56, 527	10, 704, 741
Independent	1, 412	987, 437	7, 999	1, 608, 665
Al-Jazeera	801	102, 254	451	62, 835
	Dataset3 (Independent Seed)		*Dataset4* (Al-Jazeera Seed)	
CNN	528	421, 542	2, 233	1, 652, 875
Telegraph	23, 272	6, 710, 580	1, 126	894, 710
Independent	27, 012	2, 985, 412	394	54, 760
Al-Jazeera	303	48, 058	9, 313	531, 452

Table 5.3: Dataset statistics

5.5.2 Setup

To evaluate our approach, we have randomly selected 233 users among the most active users in the 4 news websites described above. In fact, we have selected only users that have a continuous activity for at least 15 months and moreover they have commented at least 500 news articles. For each user we have performed recommendation at different time points $t_1, t_2, ..t_n$. The reason behind time dependent evaluation is twofold: (1) to take into account profile updates since users

continuously post opinions bringing new information about their interests, and (2) to use data before time point t_i for recommendation and data starting from time point t_i for assessment, as described later. The time points $t_1, t_2, ..t_n$ are chosen in such a way that between t_{i-1} and t_i, there is at least m news articles commented by the user. For each user u_i, we have chosen $m = \frac{N_i}{10}$ where N_i is the total number of commented news articles by the user u_i. This setting resulted in 2330 rounds of recommendation.

To assess the effectiveness of our approach we have used an automatic evaluation to avoid the subjectivity of manual assessments. We have considered the action of commenting on an article to be an indicator that the article fits the interests of the user. Based on this assumption, we check the list of recommended articles. The one that user has commented on are considered relevant. Note that it is probable that some information is missing. A person might well be interested in an article even though he does not comment on it. So, the actual results are most probably higher than our findings.

5.5.3 Strategies & Measures

We have used two baseline approaches and tested several variations of our proposed technique. The strategies that we have used are:
NoEnrich: The first baseline is a simple content filtering approach based solely on the content of news articles, meaning that no enrichment with opinions was performed.
Yee: The second baseline is the closest works to ours which exploit all the set of opinions related to each news article to enrich its content.
Authority_k: We use our approach to enrich news articles with the topk authoritative opinions related to it. The topk opinions are selected using the strategy described in section 4.3. In our experiments we used $k = 5$, $k = 10$, and $k = 20$
Diversity_k: We use our approach to enrich news articles with the most diverse topk opinions related to it. The diversification is performed as described in section 5.4. In our experiments we used $k = 5$, and $k = 10$

To compare the results of the different methods, we use Precision at k ($P@k$). The $P@k$ is the fraction of recommended articles that interest the user in question considering only the top-k results. It is given by:

$$P@k = \frac{|Relevant\_Articles \cap topk\_Articles\_Results|}{k}$$

5.5.4 Results

In a first step, we have compared the accuracy of news recommendation using only the news article content to the case when we enrich that content by all its related opinions as proposed in Yee's approach [YYLF09]. As shown in figure 5.3, it is clear that the precision has decreased while leveraging the news article contents by all its related opinions. For instance, the precision $P@1$ was dropped down from $42,4\%$ to $39,3\%$, and from $48,1\%$ to$44,5\%$ for the precision $P@5$.

In order to perceive the reason behind such a decrease, we have selected some articles and extracted the aspects representing the news article profile for both cases, namely (i) the case when we have used only the content of news article, and (ii) the case when we have used the

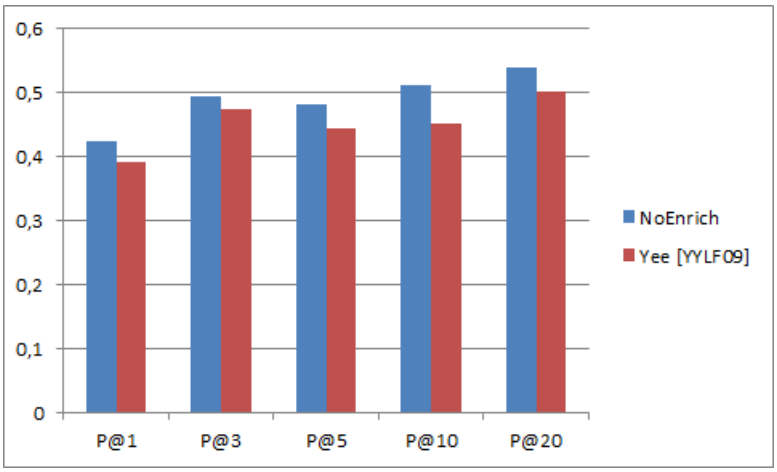

Figure 5.3: Impact of leveraging with all opinions

content of news article along with all its related opinions. In table 5.4, we randomly present some examples. By analyzing the extracted aspects for both cases, we have observed that the profile of news article, when it is leveraged with all its related opinions, contains some aspects that are far away from the main topic of news article. An explanation of such noise aspects might be referred to the presence of a lot of noise among the list of opinions which deviate the profile of news article from its main topic.

Once the quality of the generated news article profile is affected, the accuracy of news recommendation might decrease as the accuracy of news article profile is considered as one of the key-elements for revealing an accurate recommendation to users. To summarize, defining a news article profile using its content and all of its related opinions might likely be subject to raise an impact on the quality of news article profile, and hence the accuracy of recommendation. Ultimately, relying only on the news article content seems to exhibit a better performance.

In a second step, we have compared the recommendation of news articles between NoEnrich

Article Title	**NoEnrich**	**Yee [YYLF09]**
British couple to be deported from Astralia for living in wrong suburb	Australia Living, British Australia, couple Live, Australia life	Immigrants life, couple Law, Australian goverment, Australian suburb
Barack-Obama a dithering controlling risk averse	Obama policies, Clinton foreign policy, American policies	Obama foreign policy, president budget, world policy, Obama election
Cameron needs to capture some of Boris Johnson sunshine	Boris Johnson, Cameron Johnson, Cameron party	Cameron party, Boris Johnson, Cameron voters, Election

Table 5.4: Example of aspects extracted from news article profiles

and Authority. In other terms, we have compared the accuracy of recommendation by using two different news article profiles. Firstly, with profiles defined using only the content of news article and profiles defined using the news article content and a set of *topk* opinions ranked using our strategy of opinion ranking.

We have obtained a slight improvement compared to the approach where only the content of

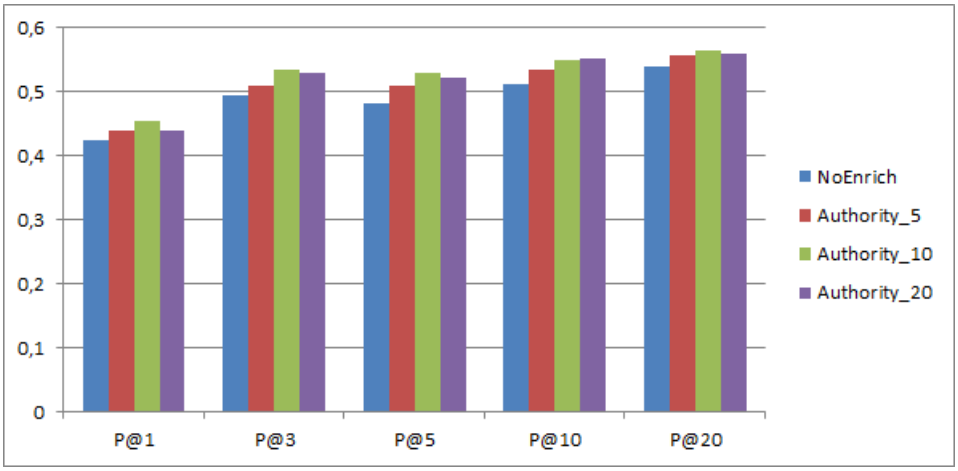

Figure 5.4: Impact of leveraging topk opinions

news article is used. For instance, the precision $P@10$ was increased to $53,4\%$ from $51,3\%$ with *top*5 opinions and to $56,5\%$ from 54% with *top*10 opinions.

Subsequently, we have used three primitives, news articles profiles with *top*5, *top*10, and *top*20

Article Title	**NoEnrich**	**Authority_10**
Argentina pulls out of falklands talks	Islanf falkland, Island Argentina, referendum	Island referundum, Argentina regime, British colonialism, Argentina Island
Barack-Obama a dithering controlling risk averse	Obama policies, Clinton foreign policy, American policies	Foreign policy, foreign budget, Obama foreign policy, World policy
Cameron needs to capture some of Boris Johnson sunshine	Boris Johnson, Cameron Johnson, Cameron party	Cameron policies, economy, Cameron finance policies, Economic leaders

Table 5.5: Example of aspects extracted from news articles profiles

relevant opinions. Thus, the aspects describing news articles profiles are defined in a second case using the content of news article and the content of *topk* opinions. We have obtained a slight improvement as compared to the approach where only the content of news article is employed. For instance, the precision $P@10$ has increased to $53,4\%$ from $51,3\%$ with *top*5 opinions and to $56,5\%$ from 54% with *top*10 opinions.

For a deep comprehension of the reasons leading to this improvement, we present below the aspects defining the news articles profiles without leveraging opinions, as well as with leveraging the *top*10 opinions. We have also observed a slight increase of the accuracy of news recommendation between leveraging *top*10 and *top*20 opinions. This might be due to the redundancy of opinions, and thus lowering the quality of news article profile.

As a last comparison, we have compared the results of NoEnrich strategy with Diverse strategy. In this case, we have compared the recommendation of news articles between the case when we have used only the content of news article and the case where we apply the ranking and diversification

approaches described earlier on opinions related to each news article. Thus, the news articles profiles are described using their content and a list of prominent and diverse opinions.

In this case, we have observed a significant improvement in terms of Precision either when

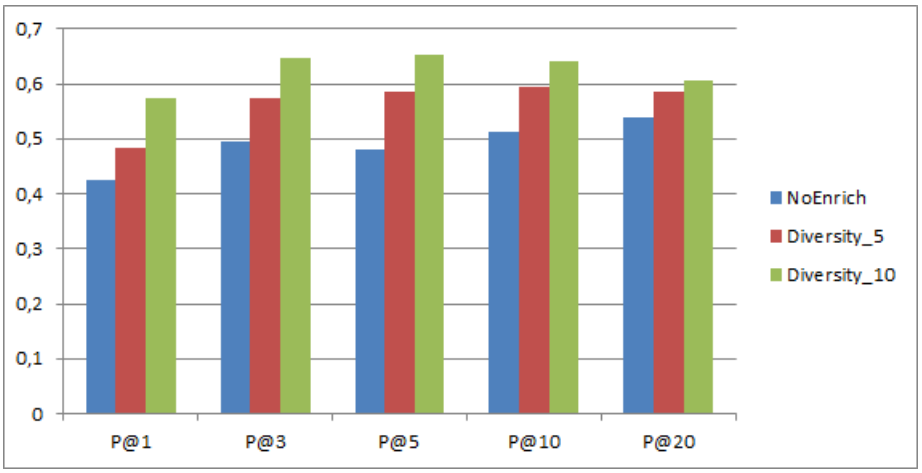

Figure 5.5: Impact of leveraging relevant and diverse users' opinions

opinions. For instance, the precision $P@5$ was raised to $58,7\%$ from $48,1\%$ for $Diversity_5$ and to $65,4\%$ from $48,1\%$. The profile of news article profiles are more relevant to the topic of news article and contain diverse aspects than other profiles. It is also clear that most of used aspects to describe the news article are well attached to the topic of news article.

To sum up, we can clearly observe that leveraging all opinions to define the news article contents

Article Title	**NoEnrich**	**Diversity_10**
Argentina pulls out of falklands talks	Islanf falkland, Island Argentina, referendum.	Falkland living, Falkland people, Colonialism, Referendum, Argentina government.
Barack-Obama a dithering controlling risk averse	Obama policies, Clinton foreign policy, American policies.	Afghanistan policy, American foreign policies, Clinton policy, Foreign policy budget.
British couple to be deported from Australia for living in wrong suburb	Australia living, British Australia, Couple live, Australia life.	Australian Tax, Australian people, Deportation, Australian Visa, Contract people, Live rules, living condition.

Table 5.6: Example of aspects extracted from news articles profiles

can affect the accuracy of news recommendation. Moreover, leveraging opinions by applying only opinion ranking strategy can improve the accuracy. The recommendation has also some limits due mainly to the redundancy of selected opinions. Thus, ranking opinions and use them to enrich the news article contents is not enough. Enriching the content of news articles with a subset of diverse opinions give a better results on the accuracy of recommendation. This is might be explained by the fact that when we apply this strategy we can deal with two key-weaknesses of leveraging opinions which are noise and redundancy of opinions. The detailed results of our experiments are depicted in table 5.7. To sum up, we can clearly see that our approach

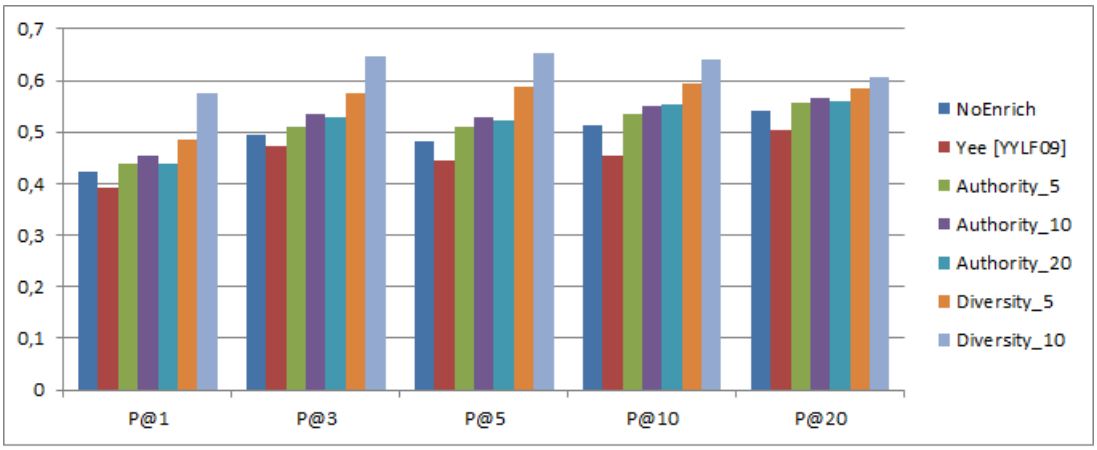

Figure 5.6: All results

	P@1	**P**@3	**P**@5	**P**@10	**P**@20
NoEnrich	0.424	0.494	0.481	0.513	0.540
Yee [YYLF09]	0.393	0.474	0.445	0.453	0.503
Authority_5	0.439	0.510	0.509	0.534	0.558
Authority_10	0.454	0.535	0.530	0.550	0.565
Authority_20	0.439	0.530	0.521	0.553	0.559
Diversity_5	0.484	0.575	0.587	0.595	0.586
Diversity_10	**0**.575	**0**.646	**0**.654	**0**.640	**0**.607

Table 5.7: Overall performance of our approach

outperforms the baseline approaches by a significant margin. The improvement goes up to 17% in terms of *precision*@5 compared to NoEnrich and 21% compared to Yee, which is substantial. Having a closer look at the results, we notice see that relying only on the content of news articles does not provide a good performance. Furthermore, even when trying to enrich the content by all users' opinions, the precision decreases. By applying ranking, the precision improves but the performance gain is small, ranging from 1% to 4%. However, when we apply diversification to ranked opinions opinions, the *top*5 and *top*10 diversified opinions give the best results.

These results meet our expectations since they perfectly reflect the role and the nature of opinions in news websites. Relying only on the content of articles does not perform well because user profiles built from opinions focus on some aspects that might be different from the one provided by the news article. Regarding the enrichment of news article with opinions there are different observations. First, taking all opinions into account is not a good idea since opinions are subject to noise and some of them might even deviate from the topic of interest. Yet, this approach had the worst performance. Second, selecting the *topk* opinions to be included in the article content is a good idea. However, due to redundancies, this method loses its effect especially when k increases, which is the case of $Authority_2$0.

Finally, diversifying opinions before enriching the content of articles provides a high gain in precision. This is because of the wider coverage of aspects. If the aspects discussed in the opinions are explicit in the news article, then their weight is increased, otherwise they are added

which increases the chance of having more users getting interested in the article. We can see for example that the aspects extracted for the news article British couple to be deported from Australia for living in wrong suburb are too generic with *NoEnrich* and Yee strategies. They are mainly about Australian Live. By contrast, the aspects become more focused with opinion ranking and talk for example about Australian Visa and Deportation. Then, we can see that diversification extracts more aspects such as Australia tax and people contracts.

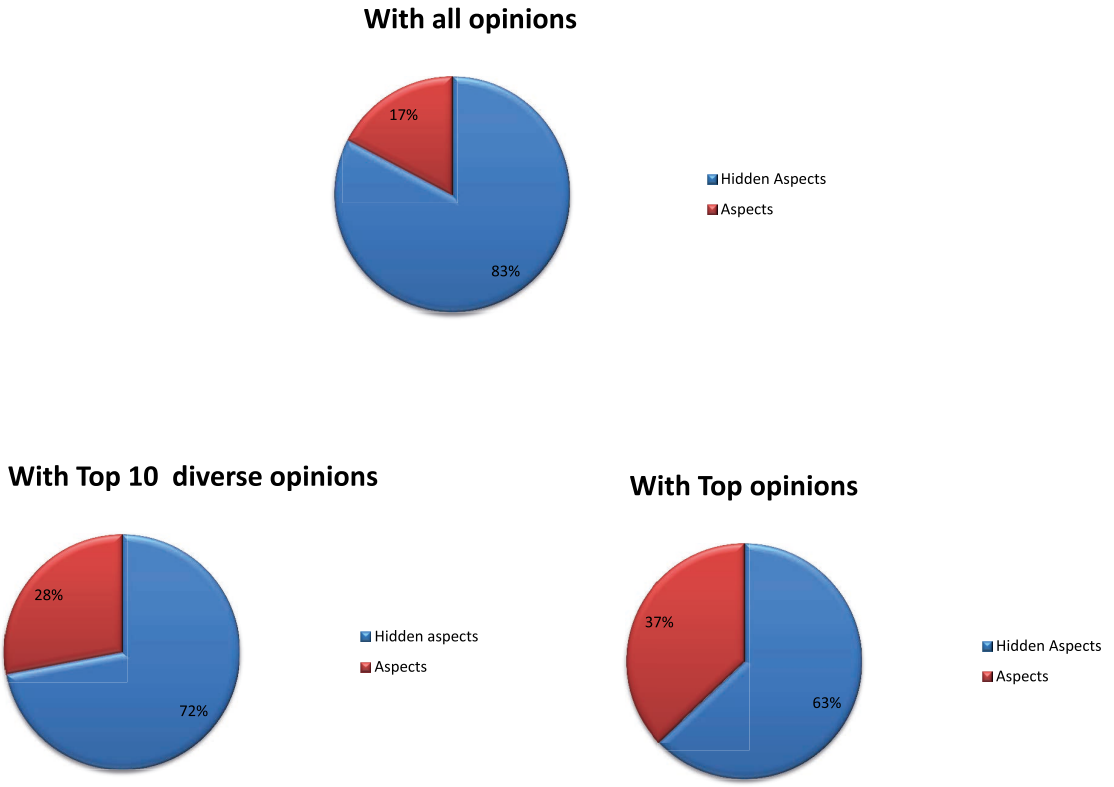

Figure 5.7: Proportions of Aspects and Hidden Aspects in news articles profiles

Figure 5.7 shows the proportions of aspects extracted from news articles and hidden aspects extracted from users' opinions. With all opinions, the description of the news article is mainly based on users' opinions, the top k reduce that dramatically but result a very small contribution. Then diversification covers more hidden aspects without biasing the content of news articles which explains its good performance.

5.6 Conclusions

In this chapter, we investigated three ways of leveraging opinions on the content of news articles for refining the list of recommended news articles. Two approaches outperform the baseline where only news article content is used: (i) leveraging the news article contents with only relevant opinions using opinions ranking strategy, and (ii) leveraging the news article contents with diverse opinions using an opinion diversification model. Our study, which was conducted through an

extensive set of experiments, showed that diverse opinions achieve the best results compared to baseline approaches. This result shows that opinions is a valuable source of information that can capture well users' interests and contribute to the extension of knowledge about daily life topics.

Part III

Conclusions and Outlooks

Chapter 6

Conclusions and Future Work

The main motivation of this thesis was to propose a personalized news recommender that facilitates access of users to the news articles that match with their interests. We have demonstrated that opinions can be exploited to boost the quality of personalized news recommendation. During this thesis, we have addressed this problem by proposing three main contributions. Firstly, we have proposed a profile model that accurately describes both users' interests and news article contents. The profile model was tested on three different applications ranging from identifying the political orientation of users to the context of news recommendation and the diversification of the list of recommended news articles. Results show that our profile model give much better results compared to state-of-the-art models. Secondly, we have investigated the problem of noise on opinions and how we can retrieve only relevant opinions in response to a given query. The proposed opinion ranking strategy is based on users' debates features. We have used a variation of PageRank technique to define the score of each opinion. Results show that our approach outperforms two recent proposed opinions ranking strategies, namely Smart [LM13b] and RevRank [TR09], particularly for controversial topics. Thirdly, we have investigated different ways of leveraging opinions on news article contents including all opinions, topk opinions based on opinion ranking strategy, and a set of diverse opinion. To extract a list of diverse opinions, we have employed a variation of an existing opinion diversification model. Results show that diverse opinions give the best performance over other leveraging strategies. In section 6.1, we present our answers to research questions that were raised in the first chapter of this thesis. Then, in section 6.2, we list the future research directions following from this thesis.

6.1 Answers to Research Questions

Research question 1. *How should users' interests and news articles be described ?*
In chapter 3, we have proposed a profile model that accurately describes users' interests and news article contents. The proposed model is based on three main components which are respectively entities, aspects and sentiments. Indeed, by analyzing the content of a large number of opinions and news article contents, we have observed that their content is often composed by one or a set of following three components which are entities, aspects, and sentiment. Entities represent well known concepts such as persons, locations, and organizations. The aspects might be a list of concepts extracted from a given ontology or entities' properties. The last component of our

profile model is the sentiment which represents the feeling towards one or both of the first two components and it can be either positive, negative or neutral.
To assess the effectiveness of our profile model, we have tested two versions of our profile model which are sentiment-independent and sentiment-dependent. The first version, sentiment-independent, is a set of tuples (entity, aspects) while the second version is represented by two different tuples (aspect, sentiment) and (entity, sentiment). We have tested our model on two different applications.
In the first application, we have used our profile model to propose a new technique for identifying the political orientation of users based on their opinions around news articles. We define users based on a set of tuples (entity, sentiment) and (aspect, sentiment). The proposed approach is promising as it is completely unsupervised which makes it flexible to be applied on any kind of dynamic knowledge such as opinions. Moreover, it provides a means for dealing with an unstructured source of information. The proposed approach is an unsupervised technique because we have created a knowledge base of political orientations based on Wikipedia. For each political orientation, we start from a Wikipedia seed pages that describe this political leaning. We extract from the textual part of the pages all outgoing links that point to other Wikipedia articles. Then, we select the anchor text of these links which correspond either to a set of entities or aspects related to the political orientation. Each entity or aspect occurring in a sentence that has a sentiment orientation is extracted. Thus, in a similar way to the user profile, we describe each political orientation by a set of entities $\{e_1, ...e_n\}$, aspects $\{a_1, ...a_m\}$, and their related sentiments $\{s_1, ...s_m\}$. We have conducted experiments with US and Egypt user groups crawled from CNN and Al-Jazeera English news websites. We have run our experiments on 500 users: 290 from US (CNN) and 210 from Egypt (Al-Jazeera). The results show that through our profile model the results in most cases are improved by providing the best accuracy over the state-of-the art methods and goes up to $95,07\%$.
In the second application, we have used a sentiment-independent version of our profile model in the context of classic personalized news recommendation. To this end, we introduce a new approach that models the profile of users and articles based on a set of tuples representing entities and their aspects. The idea is to have a fine-grained description of users and articles regarding general topics together with more specific issues. The profile of a user is extracted from the set of opinions he provides in the news websites, and the article profile is extracted from its content and described by a set of tuples (entity; aspect). These profiles are then matched to recommend to each user the list of articles that match with user profile interests and the current article he is reading. We evaluate our approach using four real datasets including The Independent, The Telegraph, CNN, and Al-Jazeera. The experiments show that our approach outperforms baseline approaches with a large margin, in term of precision and NDCG.
In the third application, we have used a combination of sentiment-dependent and sentiment-independent profiles on the context of personalized news recommendation. The main contribution in this application was the adapting of a diversification model to diversify the list of recommended news articles. We have used a sentiment-independent profile to describe users' interests based on a set of weighted tuples <entity,aspect>. We define the news articles profiles based on sentiment-dependent profile namely a list of triplets corresponding to the three components of

the profile model namely entity, aspect and sentiment. We have applied a diversification model on the list of recommended news articles to deal with problem of redundancy. The model is based on two components: (1) semantic diversification to avoid redundancy and to cover a diverse set of news articles presenting different arguments, and (2) sentiment diversification to cover different types of sentiments that can be positive, negative or neutral. The results show that diversification of news articles can have a positive impact on the accuracy of recommended news articles.

Research question 2. *How to deal with the problem of noise on opinions ?*

To deal with the problem of noise on opinions and get only relevant opinions in response to a given query, we have proposed in chapter 4 an opinion ranking model. This work aims at the organization of opinions on news websites to facilitate their access, understand their trends, and provide a valuable source for enriching article contents. The result of this work can be useful for many applications including news recommendation, and the assessment of public opinion polls. Our approach goes beyond existing approaches for three main reasons.

Firstly, determining prominent opinions about daily life topics is much more complex than identifying helpful product reviews as suggested in prior work.

Secondly, unlike existing approaches, we define user expertise not only based on explicit ratings, but also on implicit ratings the user gets for his actions. This is due to the fact that explicit ratings suffer different kinds of bias [LCL+07] such as the winner circle bias, where opinions with many votes get more attention and therefore, accumulate votes disproportionately, and the early bird bias where the first opinion to be published tends to get more votes.

Thirdly, none of the existing approaches takes into account implicit ratings provided by users' debates and exchange of opinions. In our work, we take into account the relationships between opinions and their replies, which we call nested opinions, propagating the sentiments along those relations to compute the final score of an opinion. We model users' debates as a directed graph of opinions where links can either be positive or negative representing agreements and disagreements between opinions. Then, we propose a new variation of the PageRank algorithm which handles both positive and negative links between graph nodes. The idea is to boost opinions scores along positive links and decrease them along negative links. We test our approach by running experiments on three datasets crawled from CNN, The Independent, and The Telegraph news websites. The results show that our model achieves high quality results, particularly for highly popular and highly controversial topics having a large amount of user debates.

Research question 3. *How to improve the effectiveness of personalized news recommendation with opinions ?*

In chapter 5, we have investigated the impact of leveraging opinions on the content of news articles. Two main points highlight the difference between our work and previous approaches.

Firstly, approaches that use product reviews to enrich the description of their products assume that these reviews belong to an already known set of aspects. In our work, we are interested in aspects about daily life topics reported by news articles. These aspects are not classified but we extract them automatically using an unsupervised approach.

Secondly, opinions on news sites usually contain a lot of noise, and thus, unlike the approach by Yee et al., [YYLF09] applied on YouTube videos, we do not use all opinions to enrich the

content of news articles but we select only the top-k opinions. Additionally, we have used a diversification model to diversify the list of opinions to have a large coverage of new aspects. To the best of our knowledge, there is no existing work that exploits opinions to enrich news articles in the context of personalized news recommendation. Our work aims at providing an effective news recommendation to facilitate the access of users to published news articles. More importantly, it motivates readers to comment on the news articles of interests and get involved in discussions with other users. For this reason, we have proposed a novel recommendation approach that enriches the content of news articles with opinions to improve the effectiveness of recommendation. We have adopted three strategies to enrich the content of news articles. We have applied our strategies on real data activities of users extracted from four datasets crawled from CNN, The Independent, The Telegraph, and Al-Jazeera. Firstly, by leveraging the content of news articles with all its related opinions. In this case, we got worst results than using only the content of news articles. By analyzing the extracted aspects from news articles profiles, we have observed that it contains some aspects that are far away from the main topic of news articles. An explanation of the presence of such aspects might be referred to the presence of a lot of noise among the list of opinions, which points out aspects that are far away from the main topic of the news article.
Secondly, by selecting only prominent opinions. In this case, we have obtained a slight improvement compared to the approach where only the content of news article is used. For instance, the precision $P@10$was improved from $51,3\%$ to $53,4\%$ with $top5$ opinions and from 54% to $56,5\%$ with $top10$ opinions.
Thirdly, by leveraging the content of news articles using a list of diverse opinions. To diversify opinion on news websites, we have employed an opinion diversification model based on three components namely authorities, semantic and sentiment component. In which case, we have observed a significant improvement in terms of Precision. For instance, the precision $P@5$ was raised from $48,1\%$ to $58,7\%$ for $Diversity\_5$ and from $48,1\%$ to $65,4\%$.
In the rest of this chapter, we will outline our plans for future work, and discuss possible research topics beyond what have been addressed in this thesis.

6.2 Future Work

There are many potential directions for future work to improve the accuracy of personalized news recommendation. Some research directions that can be envisioned from the current status of this work are sketched in the following.

- **Time-sensitive user's profile** This thesis presents a user profile model that shows within several empirical studies that it achieves better results compared to previous works. However, this thesis does not explore the impact of time on defining users' interests and does not analyze the evolution of users' interests over time. This is useful to give insights on the exploration of users' commenting and reading behaviors. Further, in the proposed model we do not take into account viewpoint changes through time which can be a problem if there are many contradicting pairs in the profile.

- **Opinion ranking** For pragmatic reasons, our experiments included news datasets that have very similar structures. However, exploring other datasets of different types of entities, of users, and kinds of opinions is worthwhile in order to show the wide applicability of our model. To this end, we are planning to assess the effectiveness of our approach using a dataset crawled from Youtube, which is more subject to noise.

- **Opinion Summarization** Presenting the list of relevant opinions in response to a given query is not always enough. The reason is that the user has to go through all opinions to understand the most important aspects revealed on opinions. Thus an improvement might be an approach that summarize the most important aspects and sentiments revealed for each given aspects.

- **News recommendation** Lastly, we aim at deploying the proposed approach on live traffic data which will bring us new challenges and opportunities on the context of news recommendation.

Bibliography

[AAK11] Sahin Albayrak Akram Al-Kouz, Ernesto William De Luca. Latent semantic social graph model for expert discovery in facebook. In *11th International Conference on Innovative Internet Community Systems*, page 269. GI Edition, 2011.

[AAYIM13] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, and Sepideh Mahabadi. Real-time recommendation of diverse related articles. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1–12, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[AB06] Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, 2006.

[AC10] Deepak Agarwal and Bee-Chung Chen. flda: Matrix factorization through latent dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 91–100, New York, NY, USA, 2010. ACM.

[AGHI09] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[AGHT11] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, UMAP'11, pages 1–12, Berlin, Heidelberg, 2011. Springer-Verlag.

[ALM] YOANN ALMERAS. Offshore outsourcing of journalism.

[AM03] Sarabjot Singh Anand and Bamshad Mobasher. Intelligent techniques for web personalization. In *Proceedings of the 2003 international conference on Intelligent Techniques for Web Personalization*, pages 1–36. Springer-Verlag, 2003.

[ARW12] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. Harmony and dissonance: organizing the people's voices on political controversies. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 523–532, New York, NY, USA, 2012. ACM.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.

[BC09] Michael Bendersky and W. Bruce Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 8–14, New York, NY, USA, 2009. ACM.

[BHK98] John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[BKA98] Krishna Bharat, Tomonari Kamba, and Michael Albers. Personalized, interactive news on the web. *Multimedia Syst.*, 6(5):349–358, September 1998.

[BMP11] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[BOHG13] J. Bobadilla, F. Ortega, A. Hernando, and A. GutiéRrez. Recommender systems survey. *Know.-Based Syst.*, 46:109–132, July 2013.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[BP99] Daniel Billsus and Michael J. Pazzani. A personal news agent that talks, learns and explains. In *Proceedings of the Third Annual Conference on Autonomous Agents*, AGENTS '99, pages 268–275, New York, NY, USA, 1999. ACM.

[Bur02] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[Bur05] Robin Burke. Hybrid systems for personalized recommendations. In *Proceedings of the 2003 International Conference on Intelligent Techniques for Web Personalization*, ITWP'03, pages 133–152, Berlin, Heidelberg, 2005. Springer-Verlag.

[Bur07] Robin Burke. The adaptive web. chapter Hybrid Web Recommender Systems, pages 377–408. Springer-Verlag, Berlin, Heidelberg, 2007.

[BWC+12] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 185–194, New York, NY, USA, 2012. ACM.

[BYRN+99a] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[BYRN99b] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[CC09] Ben Carterette and Praveen Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM*, pages 1287–1296, 2009.

[CCCT09] Elica Campochiaro, Riccardo Casatta, Paolo Cremonesi, and Roberto Turrin. Do metrics make recommender algorithms? In *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*, pages 648–653. IEEE, 2009.

[CFMH12] Michel Capelle, Flavius Frasincar, Marnix Moerland, and Frederik Hogenboom. Semantics-based news recommendation. In *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 27:1–27:9, New York, NY, USA, 2012. ACM.

[CG98] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[CGM+99] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR workshop on recommender systems*, volume 60. Citeseer, 1999.

[CGN+11] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part III*, INTERACT'11, pages 152–168, Berlin, Heidelberg, 2011. Springer-Verlag.

[CGR+11] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 192–199. IEEE, 2011.

[CGT12] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Trans. Interact. Intell. Syst.*, 2(2):11:1–11:41, June 2012.

[CKC+08] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[CLA+03] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 585–592, New York, NY, USA, 2003. ACM.

[CMO14] Matteo Catena, Craig Macdonald, and Iadh Ounis. On inverted index compression for search engine efficiency. In *Advances in Information Retrieval*, pages 359–371. Springer, 2014.

[CMS10] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[CNN+10] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: Experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.

[CTFLH12] Sergio Cleger-Tamayo, Juan M. Fernández-Luna, and Juan F. Huete. Top-n news recommendations in digital newspapers. *Know.-Based Syst.*, 27:180–189, March 2012.

[CTLM08] Paolo Cremonesi, Roberto Turrin, Eugenio Lentini, and Matteo Matteucci. An evaluation methodology for collaborative recommender systems. In *Proceedings of the 2008 International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution*, AXMEDIS '08, pages 224–231, Washington, DC, USA, 2008. IEEE Computer Society.

[DCCG12] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[DGM08] Souvik Debnath, Niloy Ganguly, and Pabitra Mitra. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1041–1042, New York, NY, USA, 2008. ACM.

[DLB09] Mariam Daoud, Lynda-Tamine Lechani, and Mohand Boughanem. Towards a graph-based user profile modeling for a session-based personalized search. *Knowl. Inf. Syst.*, 21(3):365–398, November 2009.

[DLP03] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW*, pages 519–528, 2003.

[DLY08] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM*, pages 231–240, 2008.

[DNMKKL09] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 141–150, New York, NY, USA, 2009. ACM.

[DS06] Kathleen T Durant and Michael D Smith. Mining sentiment classification from political web logs. In *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*, 2006.

[DTB10] Mariam Daoud, Lynda Tamine, and Mohand Boughanem. A personalized graph-based document ranking model using a semantic user profile. In *User Modeling, Adaptation, and Personalization*, pages 171–182. Springer, 2010.

[Eli75] Peter Elias. Universal codeword sets and representations of the integers. *Information Theory, IEEE Transactions on*, 21(2):194–203, 1975.

[FBY92] William B Frakes and Ricardo Baeza-Yates. Information retrieval: data structures and algorithms. 1992.

[GAC+13] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 515–526, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[GACOR05] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric K. Ringger. Pulse: Mining customer opinions from free text. In *IDA*, pages 121–132, 2005.

[GCC10] Shima Gerani, Mark James Carman, and Fabio Crestani. Proximity-based opinion retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 403–410, New York, NY, USA, 2010. ACM.

[GCP03] Susan Gauch, Jason Chaffee, and Alexander Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1:1–3, 2003.

[GDH04] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 482–490, New York, NY, USA, 2004. ACM.

[GM09] Asela Gunawardana and Christopher Meek. A unified approach to building hybrid recommender systems. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 117–124, New York, NY, USA, 2009. ACM.

[GNOT92] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, December 1992.

[GS09a] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *WWW*, pages 381–390, 2009.

[GS09b] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 381–390, New York, NY, USA, 2009. ACM.

[GZH10] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 340–348, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[GZV12] Kavita Ganesan, ChengXiang Zhai, and Evelyne Viegas. Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 869–878, New York, NY, USA, 2012. ACM.

[HCH04] Chun-Nan Hsu, Hao-Hsiang Chung, and Han-Shen Huang. Mining skewed and sparse transaction data for personalized shopping recommendation. *Mach. Learn.*, 57(1-2):35–59, October 2004.

[HD10] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[Hea92] Marti A. Hearst. *Direction-based text interpretation as an information access refinement*, pages 257–274. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1992.

[HL04a] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[HL04b] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[HLY+12] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou. What reviews are satisfactory: novel features for automatic helpfulness voting. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 495–504. ACM, 2012.

[Hof04] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):89–115, January 2004.

[HRG10] Graeme Hirst, Yaroslav Riabinin, and Jory Graham. Party status as a confound in the automatic classification of political speech by ideology. 2010.

[HRT97] Refael Hassin, Shlomi Rubinstein, and Arie Tamir. Approximation algorithms for maximum dispersion. *Operations Research Letters*, 21:133–137, 1997.

[HSDL12] Jonathon S Hare, Sina Samangooei, David P Dupplaw, and Paul H Lewis. Imageterrier: an extensible platform for scalable high-performance image retrieval. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 40. ACM, 2012.

[HW00] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING*, pages 299–305, 2000.

[HZ11] Neil Hurley and Mi Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, March 2011.

[JA08] Maojin Jiang and Shlomo Argamon. Political leaning categorization by exploring subjectivities in political blogs. In *DMIN*, pages 647–653. Citeseer, 2008.

[JBE+13] Arjan JP Jeckmans, Michael Beye, Zekeriya Erkin, Pieter Hartel, Reginald L Lagendijk, and Qiang Tang. Privacy in recommender systems. In *Social Media Retrieval*, pages 263–281. Springer, 2013.

[JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[KG11] Mouna Kacimi and Johann Gamper. Diversifying search results of controversial queries. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 93–98, New York, NY, USA, 2011. ACM.

[KG12] Mouna Kacimi and Johann Gamper. Mouna: Mining opinions to unveil neglected arguments. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2722–2724, New York, NY, USA, 2012. ACM.

[KH78] Bernhard Korte and Dirk Hausmann. An analysis of the greedy heuristic for independence systems. *Annals of Discrete Mathematics*, 2:65–74, 1978.

[KH06] Soo-Min Kim and Eduard H. Hovy. Identifying and analyzing judgment opinions. In *HLT-NAACL*, 2006.

[KLC06] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 100107, 2006.

[KMM+97] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):77–87, March 1997.

[Kob94] Alfred Kobsa. User modeling and user-adapted interaction. In *Conference companion on Human factors in computing systems*, pages 415–416. ACM, 1994.

[KPCP06] S.M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, 2006.

[KR12] Joseph A. Konstan and John Riedl. Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, April 2012.

[LBC03] Michael Laver, Kenneth Benoit, and Trinity College. Extracting policy positions from political texts using words as data. pages 311–331, 2003.

[LBGM09] Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 514–522, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[LC08a] George Lekakos and Petros Caravelas. A hybrid approach for movie recommendation. *Multimedia tools and applications*, 36(1-2):55–70, 2008.

[LC08b] Frank Lin and William W Cohen. The multirank bootstrap algorithm: Self-supervised political blog classification and ranking using semi-supervised link classification. In *ICWSM*, 2008.

[LCL+07] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, pages 334–342, 2007.

[LDP10] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 31–40, New York, NY, USA, 2010. ACM.

[LHAY08] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 443–452. IEEE, 2008.

[LHC05] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[LHCA10] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 210–217, New York, NY, USA, 2010. ACM.

[LHH+10] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 653–661, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[LHO05] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, volume 5, pages 17–24, 2005.

[LM13a] Marina Litvak and Leon Matz. Smartnews: Bringing order into comments chaos. In *KDIR*, 2013.

[LM13b] Marina Litvak and Leon Matz. Smartnews: Bringing order into comments chaos. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, KDIR*, volume 13, 2013.

[LMK+11] Yuanhua Lv, Taesup Moon, Pranam Kolari, Zhaohui Zheng, Xuanhui Wang, and Yi Chang. Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 57–66, New York, NY, USA, 2011. ACM.

[Lov68] Julie B Lovins. *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[LP07] Hong Joo Lee and Sung Joo Park. Moners: A news recommender for the mobile web. *Expert Systems with Applications*, 32(1):143–150, 2007.

[LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003.

[Lu10] Bin Lu. Identifying opinion holders and targets with dependency parser in chinese news texts. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, HLT-SRWS '10, pages 46–51, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Luh57] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317, October 1957.

[LWL+11] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. Scene: A scalable two-stage personalized news recommendation system. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 125–134, New York, NY, USA, 2011. ACM.

[LXL+12] Chen Lin, Runquan Xie, Lei Li, Zhenhua Huang, and Tao Li. Premise: Personalized news recommendation via implicit social experts. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1607–1611, New York, NY, USA, 2012. ACM.

[LY09] Wu-Jun Li and Dit-Yan Yeung. Relation regularized matrix factorization. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1126–1131, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[LZS09] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 131–140, New York, NY, USA, 2009. ACM.

[McA95] Melinda McAdams. Inventing an online newspaper. *Interpersonal Computing and Technology*, 3(3):64–90, 1995.

[MCQ08] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. volume 16, pages 372–403. SPM-PMSAPSA, 2008.

[MK10] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[MKDP14a] B-LD Youssef Meguebli, Mouna Kacimi, Bich-liên Doan, and Fabrice Popineau. Building rich user profiles for personalized news recommendation. In *Proceedings of 2nd International Workshop on News Recommendation and Analytics*, 2014.

[MKDP14b] Youssef Meguebli, Mouna Kacimi, Bich-Liên Doan, and Fabrice Popineau. Unsupervised approach for identifying users' political orientations. In *Advances in Information Retrieval*, pages 507–512. Springer, 2014.

[MKL09] Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 203–210, New York, NY, USA, 2009. ACM.

[MLK09] Hao Ma, Michael R. Lyu, and Irwin King. Learning to recommend with trust and distrust relationships. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 189–196, New York, NY, USA, 2009. ACM.

[MM06] Tony Mullen and Robert Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 159–162, 2006.

[MM07] Robert Malouf and Tony Mullen. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web*, 2007.

[MM09] Frank McSherry and Ilya Mironov. Differentially private recommender systems: Building privacy into the net. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 627–636, New York, NY, USA, 2009. ACM.

[MM10] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[MMN02] Prem Melville, Raymod J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Eighteenth National Conference on Artificial Intelligence*, pages 187–192, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[MR09] Tariq Mahmood and Francesco Ricci. Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 73–82, New York, NY, USA, 2009. ACM.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[MV08] Lanny W Martin and Georg Vanberg. A robust transformation procedure for interpreting political text. volume 16, pages 93–100. SPM-PMSAPSA, 2008.

[NHMK10] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. Optimizing informativeness and readability for sentiment summarization. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 325–330. Association for Computational Linguistics, 2010.

[OLK09] Alice H Oh, Hyun-Jong Lee, and Young-Min Kim. User evaluation of a system for classifying and displaying political viewpoints of weblogs. In *ICWSM*, 2009.

[Paz99] Michael J. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artif. Intell. Rev.*, 13(5-6):393–408, December 1999.

[PE05] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *HLT/EMNLP*, 2005.

[PHLG00] David M. Pennock, Eric Horvitz, Steve Lawrence, and C. Lee Giles. Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, pages 473–480, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[PKCK12] Deuk Hee Park, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11):10059–10072, 2012.

[PL05a] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, 2005.

[PL05b] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[PLV02a] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *CoRR*, cs.CL/0205070, 2002.

[PLV02b] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[PMS09] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[Por80] Martin F Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[PPM+06] Seung-Taek Park, David Pennock, Omid Madani, Nathan Good, and Dennis DeCoste. Naïve filterbots for robust cold-start recommendations. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 699–705, New York, NY, USA, 2006. ACM.

[PTLMHV12] C. Porcel, A. Tejeda-Lorente, M. A. Martínez, and E. Herrera-Viedma. A hybrid recommender system for the selective dissemination of research resources in a technology transfer office. *Inf. Sci.*, 184(1):1–19, February 2012.

[PUPL01] Alexandrin Popescul, Lyle H. Ungar, David M. Pennock, and Steve Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 437–444, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[RD06] Filip Radlinski and Susan T. Dumais. Improving personalized web search using result diversification. In *SIGIR*, pages 691–692, 2006.

[Reu09] T Reuters. Opencalais, 2009.

[RF13] Hongda Ren and Wei Feng. Concert: A concept-centric web news recommendation system. In *Proceedings of the 14th International Conference on Web-Age Information Management*, WAIM'13, pages 796–798, Berlin, Heidelberg, 2013. Springer-Verlag.

[RIS+94] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[RSJ88] Stephen E. Robertson and Karen Sparck Jones. Document retrieval systems. chapter Relevance Weighting of Search Terms, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.

[RV96] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. volume 45, pages 380–385. Oxford University Press, 1996.

[RV97] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, March 1997.

[RW03] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Sal71] Gerard Salton. The smart retrieval system—experiments in automatic document processing. 1971.

[SAyMY08] Julia Stoyanovich, Sihem Amer-yahia, Cameron Marlow, and Cong Yu. Leveraging tagging to model user interests in del.icio.us. In *In AAAI SIP*, 2008.

[SB07] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, pages 300–307, 2007.

[SCTB+12] David Sontag, Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Susan Dumais, and Bodo Billerbeck. Probabilistic models for personalizing web search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 433–442, New York, NY, USA, 2012. ACM.

[SFHS07] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[SHMO09] Rodrygo L. T. Santos, Ben He, Craig Macdonald, and Iadh Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 325–336, Berlin, Heidelberg, 2009. Springer-Verlag.

[Sin01] Amit Singhal. Modern information retrieval: a brief overview. *BULLETIN OF THE IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING*, 24:2001, 2001.

[SK09] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.

[SKKL12] Erez Shmueli, Amit Kagian, Yehuda Koren, and Ronny Lempel. Care to comment?: Recommendations for commenting on news stories. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 429–438, New York, NY, USA, 2012. ACM.

[SKKR01] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[SKR99] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, pages 158–166, New York, NY, USA, 1999. ACM.

[SKR01] J. Ben Schafer, Joseph A. Konstan, and John Riedl. E-commerce recommendation applications. *Data Min. Knowl. Discov.*, 5(1-2):115–153, January 2001.

[SM83] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1983.

[SMO10] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Selectively diversifying web search results. In *CIKM*, pages 1179–1188, 2010.

[SP08] Jonathan B Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. volume 52, pages 705–722. Wiley Online Library, 2008.

[SPUP01] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Generative models for cold-start recommendations. In *Proceedings of the 2001 SIGIR Workshop on Recommender Systems*, volume 6. Citeseer, 2001.

[SPUP02] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[TC03] Egidio Terra and C. L. A. Clarke. Frequency estimates for statistical word similarity measures. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 165–172, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[TDE10] Swati Tata and Barbara Di Eugenio. Generating fine-grained reviews of songs from album reviews. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1376–1385. Association for Computational Linguistics, 2010.

[TLZ12] Bin Tan, Yuanhua Lv, and ChengXiang Zhai. Mining long-lasting exploratory user interests from search history. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1477–1481, New York, NY, USA, 2012. ACM.

[TM08] Ivan Titov and Ryan T McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer, 2008.

[TR09] O. Tsur and A. Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[Tur02a] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424, 2002.

[Tur02b] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[VC11] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 109–116, New York, NY, USA, 2011. ACM.

[VH+05] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[VH06] Ellen M Voorhees and Donna Harman. Common evaluation measures. In *The Twelfth Text Retrieval Conference (TREC 2003)*, pages 500–255, 2006.

[VJN13] Jan Vosecky, Di Jiang, and Wilfred Ng. Limosa: A system for geographic user interest analysis in twitter. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 709–712, New York, NY, USA, 2013. ACM.

[WLJH10] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[WWH04] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *AAAI*, pages 761–769, 2004.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[WZ09] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *SIGIR*, pages 115–122, 2009.

[YGK+08] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *Trans. Audio, Speech and Lang. Proc.*, 16(2):435–447, February 2008.

[YKD08] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. volume 5, pages 33–48. Taylor & Francis, 2008.

[YYLF09] Wai Gen Yee, Andrew Yates, Shizhu Liu, and Ophir Frieder. Are web user comments useful for search. *Proc. LSDS-IR*, pages 63–70, 2009.

[ZAA07] Jun Zhang, Mark S. Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 221–230, New York, NY, USA, 2007. ACM.

[ZKL+10] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[ZL06] ChengXiang Zhai and John D. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.

[ZM06] Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6, 2006.

[ZRM11] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *ICWSM*, 2011.

[ZYM07] Wei Zhang, Clement Yu, and Weiyi Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 831–840, New York, NY, USA, 2007. ACM.

[ZYZ11] Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. Functional matrix factorizations for cold-start recommendation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 315–324, New York, NY, USA, 2011. ACM.