Opinion Retrieval Systems using Tweet-external Factors

Yoon-Sung Kim
Dept. of Computer Science
Korea University
Seoul, Korea
kys0205@korea.ac.kr

Young-In Song
Dept. of Computer Science
Korea University
Seoul, Korea
youngin.song@gmail.com

Hae-Chang Rim
Dept. of Computer Science
Korea University
Seoul, Korea
rim@korea.ac.kr

Abstract

Opinion mining is a natural language processing technique which extracts subjective information from natural language text. To estimate an opinion about a query in large data collection, an opinion retrieval system that retrieves subjective and relevant information about the query can be useful. We present an opinion retrieval system that retrieves subjective and query-relevant tweets from Twitter, which is a useful source of obtaining real-time opinions. Our system outperforms previous opinion retrieval systems, and it further provides subjective information about Twitter authors and hashtags to describe their subjective tendencies.

1 Introduction

Opinion mining is a natural language processing technique estimating an opinion from natural language text. This is useful for various users who report opinions in case of making reference to an opinion in various texts such as blogs, microblogs and forums. It can be used for identifying relevant opinions of customers about products or social issues. In addition, companies can utilize for analysis and establishing a marketing strategy using the analyzed results.

Analyzing sentiment for an entity from a large collection of data is costly. Therefore, an opinion retrieval system providing subjective and query-relevant documents can be useful. Especially, social network services such as Twitter are useful sources for estimating real-time public opinion. In case of social network services, there are limitations for retrieving subjective documents because of the limitations on the document length (Luo et al., 2012).

As a result of inherent document length limitations, several social search engine systems have been developed. The Sentiment140[1] system based on (Go et al., 2009) provides query-relevant and subjective documents in Twitter and the proportion of a collected and analysed sentiment. This system, however, does not perform sufficiently effective because it classifies documents using simple text features. Moreover, this system does not provide subjective information about Twitter authors and hashtags, although these features are useful for sentiment classification (Barbosa and Feng, 2010). Formerly, Twendz and Tweetfeel were the tools used for labeling Twitter polarity classification, however, these are no longer in service.[2]

We present a more powerful opinion retrieval system using tweet-external resources, which are used in state-of-the-art sentiment analysis approaches (Go et al., 2009; Luo et al., 2012; Luo et al., 2015). Our opinion retrieval system outperforms previous social network retrieval systems by adding features that are proposed in state-of-the-arts. In addition, our system provides information pursuant to subjectivity tendencies of Twitter authors and hashtags by showing sentiment statistics and tweet texts so that users can determine the opinion of the queries and subjective factors.

[1] http://www.sentiment140.com
[2] Twendz: http://twendz.waggeneredstrom.com, Tweetfeel: http://www.tweetfeel.com

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 126–130, Osaka, Japan, December 11-17 2016.

2 Opinion Retrieval System using Tweet-external Factors

Our opinion retrieval system retrieves subjective and query-relevant tweet documents. To provide the results, we use a learning-to-rank framework, utilizing several features that are helpful for opinion retrieval, as well as subjective information about each author and hashtag. Section 2.1 presents the opinion retrieval model, which forms the core of our system. Next, we describe how to use our opinion retrieval system in Section 2.2, and we show the architecture of our system in Section 2.3.

2.1 Opinion Retrieval Model using Learning-to-rank Framework

Our system performs opinion retrieval by re-ranking ad-hoc retrieval results. We re-rank our results using a learning-to-rank framework which is used in previous works (Luo et al., 2012). We use the features related to the document, author meta information, and Twitter-external information in our system.

Document features denote the characteristics that are observed in a tweet document. Several state-of-the-arts of polarity classification and opinion retrieval approaches use these features (Go et al., 2009; Barbosa and Feng, 2010; Luo et al., 2012; Luo et al., 2015). We use BM25 score, opinion word rate, and existence of link, hashtag, and mention.

Author-meta information is the information that we obtain from the author profile. It is useful for identifying subjectivity in Twitter (Luo et al., 2012). We use tweet number, follower number, friend number, and list number for author-meta features in our system.

Twitter-external information denotes the information that is related to Twitter-specific information such as the author of a tweet and hashtag in the tweet text. We aggregate the author-related tweet list written by the author wrote and we convert it into a document. In addition, we create a document based on the hashtag-related tweet list. which is the entire list of tweets using a particular hashtag (Kim et al., 2016). In this system, we use opinion word rate, retweet rate, pronoun rate, link rate, and average tweet length features in each aggregated document.

2.2 System Usage

In this section, we describe how to use our opinion retrieval system. Our system is composed of the components shown in Figure 1:

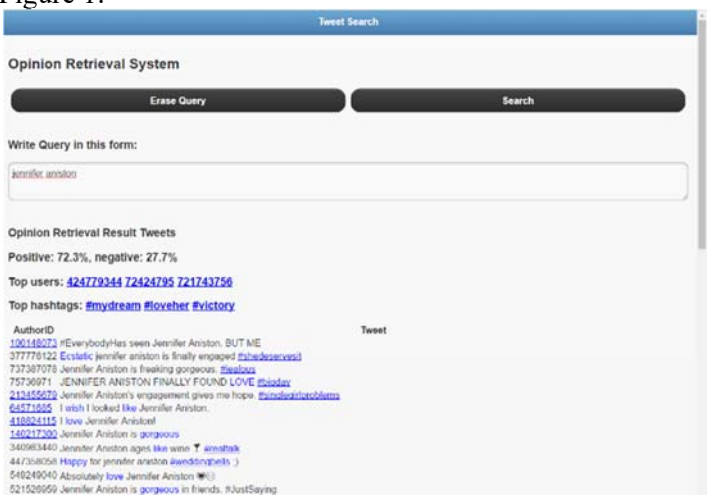

Figure 1. Screenshot of the main page of our proposed system

Figure 1 shows the screenshot of the main page of our site. On the main page, a user can write a query for retrieval. After writing such queries and deploying the search button, relevant and subjective results related to a query will be provided. As shown in Figure 1, the retrieved result consists of subjective information and the list of tweets. We provide polarity statistics, frequent authors and hashtags which have subjective information relevant to the authors' and hashtags' tendencies. Color marks in the tweet texts also give clues regarding the polarity (e.g., subjective lexicon).

The provided links for subjective author and subjective hashtags re-direct to a popup page. We provide the polarity information of the author and other subjective information and their tweet texts. Moreover, color marks of the texts are provided, which are same as the main tweet results.

2.3 System Architecture

In this section, we describe the architecture that comprises our system. Because of the absence of real-time Twitter API which provides hashtag text information, we use the English Twitter corpus using a Twitter public streaming API.

In our system, we first index the Twitter corpus for retrieval. We use Indri toolkit for indexing. Then, we extract features in the corpus for opinion retrieval, and create tweet lists for a tweet author, who wrote more than 100 tweets and hashtags which were used in more than 100 tweets. When a query is received, the retrieval system returns the top 500 tweets. We use BM25 algorithm for retrieving the queries. Then, we perform opinion retrieval by using the model described in Section 2.1. We use the ranking SVM algorithm from SVM Light toolkit. (Joachims, 1999). Finally, we provide the desired results and useful subjective information as described in Section 2.2. To provide sentiment statistics, we use a polarity classifier of (Barbosa and Feng, 2010). Finally, we utilize feature values for providing information of subjective authors and hashtags.

3 Experimental Result and Scenario

In this section, we describe the performance and effectiveness of the retrieval result of our system. In Section 3.1, we compare the performance to the previous opinion retrieval system, and in Section 3.2, we show an example of the retrieval result to explain the effectiveness of the retrieved results and the information of top users and hashtags.

3.1 Evaluation

To evaluate our system, we used the English Twitter corpus and the dataset in (Kim et al., 2016) The corpus was crawled during 1 month in 2012. The dataset was composed of tweets that are created by using Amazon Mechanical Turk. We utilized the corpus for composing the retrieval system, and then we evaluated the performance using the dataset. We used 10-fold cross validation for evaluation, and we used MAP metric. And, the Sentiment140 was used for our baseline system. We composed it using (Go et al., 2009). As shown in Table 1, our system significantly outperforms the baseline.

	Sentiment140	**Presented System**
MAP	0.2869	**0.3892**

Table 1. Performance of Sentiment140 and the presented system

3.2 Retrieval Scenario of our system

In this section, we show a scenario of retrieving a query to explain the effectiveness of our system and its usefulness in identifying the subjective tendency of the author and hashtag. Table 2 contains the top 10 retrieval results of our system when we search the query “Breaking Dawn”, the name of a movie.

Rank	**Tweet Text**	**Author ID**
1	Going to watch breaking dawn #Lonely /:	Iam_princess123
2	#30GoodMovies is Twilight 6x New moon 6x Eclipse 6x Breaking Dawn part 1 6x Breaking Dawn part 2 6x #loveall	Vira_thecoldone
3	#5FavouriteFilms twilight, new moon, eclipse, breaking dawn part 1 and breaking dawn part 2	TeamKristen
4	cant wait for breaking dawn part 2 #ashamed (:	DutchZaynsters
5	#6favMovies twilight,new moon,eclipse,breaking dawn part 1,breaking dawn part 2 and welcome to the Riley's #Krisbian #Robstener #Robsesed!x	Nic_in_twiland
6	My moms face during Breaking Dawn #priceless lmao	KissaNicole
7	Twilight, New Moon, Eclipse, Breaking Dawn Part 1..& when it comes out, I'm sure Breaking Dawn Part 2 will be my fifth. ♥ #5FavouriteFilms.	fearlesskristen
8	#MoviesThatMadeMeCry breaking dawn LOL	AbbyFrasca
9	@TeamKristen: #5FavouriteFilms twilight, new moon, eclipse, breaking dawn part 1 and breaking dawn part 2 HahahaHahahaHahahaHahaha	jaythom93
10	#ReplaceMovieTitleWithSabaw : The Breaking #Sabaw (The Breaking Dawn)	_iM_sOnNy_

Table 2. Results of our system in case of retrieving “Breaking Dawn”

As shown in Table 2, we can figure out that the tweets containing opinion are in the top rank such as 1, 2, 3, 4, 5, 7, 8. The fact that the results with hashtags such as "#30GoodMovies", "#loveall", "#5FavouriteFilms" have higher ranks indicates the effectiveness of hashtag features because there are no clues about subjectivity except the hashtags.

In Table 3, we can see the lists of tweets for authors "TeamKristen" and "fearlesskristen" ranked 3 and 7 in Table 2, respectively. These tweets are about "Kristen Stewart", who is the actress of "Breaking Dawn", and the tweets are opinions about her and her movies. Table 4 shows the lists of tweets with hashtag "#30GoodMovies" and "#5FavouriteFilms" which rank 2 and 3 in Table 2, respectively. As shown in Table 4, there are subjective tweets about the movies for which these hashtags were used. Therefore, we can determine that the information provided by our subjective authors and relevant hashtags are useful for estimating subjectivity.

TeamKristen	**Fearlesskristen**
RT If you would go #LesbianForKristenStewart ;)	RT if your heart broke on 25th July 2012.
4 am and my love for Kristen is too big to let me sleep	"Ruperv." Lol omg, I love our fandom. :\')
I believe in Kristen. She's my role model for some reason.	#10PlacesIWantToGo: Kristen's house to tell her I love her. ♥ RT if you want to do that too.
IF YOU HATE KRISTEN STEWART UNFOLLOW ME, THANKS.	"You don't need to have the perfect face to be beautiful." - Kristen Jaymes Stewart. ♥
RT IF YOU STILL LOVE AND SUPPORT THE GIRL FROM MY ICON ♥	"Just follow your heart..you'll usually wind up where you want to be." - Kristen Stewart. ♥

Table 3. Examples of author-related tweets

#30GoodMovies	**#5FavouriteFilms**
I'm Legend #30GoodMovies	#5FavouriteFilms walk the line
Shaun the sheep #30GoodMovies	#5FavouriteFilms Batman: The Dark Knight
#30GoodMovies 6. Harry Potter and the half blood prince	#5FavouriteFilms the nutty professor, the original,,, jerry Lewis one
Harry Potter is one of #30GoodMovies	Oh no wait, I'm replacing Cool Runnings with Friends With Benefits #5FavouriteFilms
#30GoodMovies, anything and everything directed by Tim Burton.	#5FavouriteFilms inception - salt - 1984 - hunger games - beauty& the beast.

Table 4. Examples of hashtag-related tweets

4 Conclusion

We presented an opinion retrieval system for Twitter to find a subjective and query-relevant tweet related to a query. Our system outperformed previous opinion retrieval systems; it provides subjective information even about an individual user, which is not provided by other systems. In addition, our system can analyze the pros and cons of a product or service, which is certainly useful in the development of a marketing strategy. Unfortunately, we do not provide real-time tweet information due to the absence of a related Twitter API. If a Twitter is provided eventually, we will be able to provide real-time tweet service for real-time information of hashtag texts.

Acknowledgements

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Plannig (2012M3C4A7033344).

Reference

Barbosa L. and Feng J. 2010. *Robust Sentiment Detection on Twitter from Biased and Noisy Data*, Proceedings of the 23rd International Conference on Computational Linguistics, 36-44.

Go A., Bhayani R. and Huang L. 2009. *Twitter Sentiment Classification using Distant Supervision*, Technical Report, Stanford Digital Library

Joachim T. 1999. *Making large-Scale SVM Learning Practical*. Advances in Kernel Methods – Support Vector Learning, 169-184

Kim, Y., Song Y.-I., Rim H.-C. 2016. *Opinion Retrieval for Twitter Using Extrinsic Information*, Journal of Universal Computer Science, Volume 22, No 5, 608-629

Luo Z., Osborne M, and Wang T. 2012. *Opinion Retrieval in Twitter*, Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 507-510

Luo Z., Yu Y., Osborne M. and Wang T. 2015. *Structuring Tweets for improving Twitter search*, Journal of the Association for Information Science and Technology, Volume 66, Issue 12, 2522-2539