Review-driven Product Summarization through Stories

THESIS DESIGN

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF
MASTER OF SCIENCE

MARC VORNETRAN
11569565

MASTER INFORMATION STUDIES
HUMAN CENTERED MULTIMEDIA

FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

March 17, 2018

1st Supervisor
Dr. Stevan Rudinac
University of Amsterdam

Review-driven Product Summarization through Stories

Thesis Design

Marc Vornetran
University of Amsterdam
Amsterdam, The Netherlands
marc.vornetran@gmail.com

KEYWORDS

Product summarization, review analysis, feature extraction, storytelling, topic modeling

1 PROBLEM STATEMENT

The widespread adoption of e-commerce has established consumer generated reviews and ratings of products as a standard feature on merchant websites. These reviews provide guidance for potential new customers and help to meet their expectations after receiving an ordered product. Besides creating a knowledge database for consumers reviews contain valuable insights and measurements for both merchants and producers. While not every single customer authors a review especially popular products amass numerous reviews on large marketplaces during their life cycle. Reviews themselves differ in length, level of detail, quality of writing, structure, language and opinion features [6].

However, e-commerce platforms of producers, merchants and large scale retailers are by no means the only place for users to express their opinion about products. Review aggregation sites collect user reviews about products and services as a third party which does not offer or provide the reviewed subjects themselves. Well established household names include: Metacritic[1] (entertainment), Rotten Tomatoes[2] (movies, tv), IMDb[3] (movies, tv, actors) and TripAdvisor[4] (vacation, flights, restaurants). Aggregators act as independent platforms where users can find averaged reviews in a uniform fashion.

As a more specific instance with a much more narrow focus RateBeer[5] collects information about (craft) beer and primarily consumer generated reviews of beers and breweries. Established in 2000 it has since remained a popular exchange platform garnering more than five million reviews in total[6] with a rate of roughly one million reviews per year since 2016[7]. While more recent and exact figures remain hidden it is estimated that nearly 500,000 visitors view the site within one month[8]. Due to a number of factors RateBeer will be used as a testbed for this thesis project. First, the large quantity of historic reviews and constant stream of new content provides an abundance of data for processing and experimental testing. Second, according to RateBeer's quality assurance principles low quality and nonsense entries are swiftly removed from the platform through an extensive administration system. Furthermore, although Anheuser-Busch InBev acquired a minority stake of RateBeer[9] the site claims and emphasizes its mission to remain an independent platform which prohibits ratings by breweries and their affiliates. Third, reviews are of a semi-structured nature containing free text but also scaled ratings in five distinct categories: (1) aroma, (2) appearance, (3) taste, (4) palate, and (5) overall. Finally, being an international aggregator the site displays the location (city, country) of reviews when they were published by their authors. This allows to perform analysis scoped to specific cities, regions or countries.

With an abundance of reviews at hand readers are easily overwhelmed by the sheer number of expressed opinions, and thus extracting meaningful information becomes a challenging task. Existing solutions to this problem focus on different aspects in their summarization process. In feature-based summaries [5] product features are extracted and reviews are classified according to their sentiment towards these features. This approach aims to reduce opinion bias when only a subset of reviews is read. Similarly, this goal can also be achieved by ranking a set of questions about a product and matching reviews that are able to answer their respective question [8]. As other research has pointed out [1, 8] diversification is an important and sometimes overlooked characteristic of review summarization. Besides including representative reviews into the summary diverse sentiments and aspects should be included as well.

Bearing this in mind, the summarization problem can be approached with a wider scope taking into account the dominance of social multimedia [2]. The authors argue that summarization should be revisited in the light of modern multimedia to connect research areas which have been working in isolation from each other. In order to meet today's information needs of users tapping single media sources is not sufficient anymore. This allows to derive the following question which should lead this research project:
How to generate representative and diverse product story summarizations driven by user generated reviews?

As mentioned before RateBeer will act as a testbed for this experimental exploration. It provides user generated reviews for a large quantity of beers (products) and enables the possibility to apply the story generation to breweries and regions as well. The research will take other media sources into consideration such as further text or image content from social media or video advertisement which could be incorporated into the generated narrative.

[1]http://www.metacritic.com/
[2]https://www.rottentomatoes.com/
[3]http://www.imdb.com/
[4]https://www.tripadvisor.com/
[5]https://www.ratebeer.com/
[6]https://www.ratebeer.com/RateBeerBest/default_2013.asp, Accessed: 16-03-2018
[7]https://www.ratebeer.com/ratebeerbest/default_2016.asp, Accessed: 16-03-2018
[8]https://www.similarweb.com/website/ratebeer.com, Accessed: 16-03-2018
[9]https://www.nytimes.com/2017/06/18/business/media/anheuser-busch-inbev-ratebeer.html, Accessed: 16-03-2018

2 METHOD

In order to tackle the problem a system has to be devised which consists of four components building on each other. Firstly, as the foundation for further work access to the review dataset on RateBeer has to be established. This component collects historical reviews while also gathering new reviews being submitted to the site. Secondly, summarization techniques will be applied to the acquired beer reviews. Triage and evaluation of appropriate techniques will be supported by an initial literature study. The summarization and story component is responsible for performing text analysis on the corpora and outputting story generation data. Thirdly, the raw summarization and story output should be transformed in a visual representation which strikes a balance between factual and stylistic properties. Finally, the last component is used to measure the performance and effectiveness of the system.

2.1 Data Collection

As RateBeer does not publish official datasets of its reviews alternative ways have to be utilized. There are a few ways to approach this obstacle.

Using an open dataset[10] with 1,500,000 entries[11] can help to kick-start development of other system components while more reviews are collected in the meantime.

The official RateBeer API[12] is implemented in GraphQL[13] but is not directly suitable for collecting reviews on a large scale. It provides an endpoint for retrieving reviews of a beer given the respective ID number. Consequently, given a list of beer IDs this endpoint could potentially be used to retrieve reviews in a reliable, structured way. The default limitation of 5,000 requests per month and one request per second are an issue to keep in mind. Overcoming the API limitations could be possible by contacting RateBeer's operators and discussing if data can be made available more easily with exclusively research purposes in mind. In the past a dataset has been made available to researchers from Stanford University[14] which indicates a chance that a formal request could be successful.

Provided that the previously described approaches are unable to generate a suitable dataset scraping can be used to collect review data. This process involves programmatically visiting the website and extracting relevant data from the HTML response. In a similar fashion to crawling bots of search engines links have to be discovered to new pages. Essentially allowing the program to autonomously discover unknown beers and their respective reviews. However, it has to be ensured that the scraping process does not produce high load on the platform or floods it with requests in a rogue manner.

2.2 Summarization

Due to the high number of reviews performing summarization tasks on the full dataset may become unfeasible on consumer devices. As a support in this matter access has been granted to the Distributed ASCI Supercomputer (DAS) in its fourth generation [3]. Being designed for research areas with heavy demands for computational power and memory it is capable of handling the potentially large review dataset. Furthermore, it is a system specifically built for running algorithms in parallel and on distributed nodes with an easy to use interface.

The selection and evaluation of appropriate natural language processing (NLP) techniques for summarization will be one of the primary tasks during the research. Therefore, an initial literature study will be conducted to review currently used algorithms which can be deployed in this scenario. One approach is to cluster semantically similar reviews into groups before identifying latent topics [10]. This second step is called topic modeling which reveals underlying key concepts in text. Latent Dirichlet Allocation (LDA) [4] is regularly, successfully [10, 12] used to generate probability topic models. Notably, this approach can be combined with the MapReduce framework [10] which is popular for processing large amounts of data. Additionally, it is designed to be executed in parallel on distributed systems such as DAS-4. TextRank [9] is a different algorithm that will be evaluated in this project. It creates a graph-based ranking, inspired by web search page ranks, which can be used to extract keywords and summarize text.

2.3 Evaluation

Evaluating the performance and effectiveness of the proposed product summarization system is a challenging task. There is no direct ground truth available to compare the generated summary artifacts to. Nonetheless, official product descriptions by breweries and content from websites such as the *Brewers Association*[15] or *Brewers of Europe*[16] could potentially be used as comparisons. Popular scoring algorithms in the field of text summarization include ROUGE [7] and the Pyramid Method [11]. They have been successfully employed in comparable research projects [10, 12].

3 SCHEDULE

Table 1 lists weekly activities from April to July 2018. While the schedule extends into the month of July the main thesis work will take place between 02.04. and 29.06. Accordingly, the thesis document should be finalized before the end of June. The absence of an activity does not signify its exclusion from other weeks, for instance preparation for writing can begin early on in the process. In this regard, the proposed schedule can be interpreted as a roadmap with milestones which should be realized in their designated week. Sparsity and buffer times in June and July retain flexibility of the planning and allow activities to be stretched out if necessary.

Activities in an italic font do not directly translate to actual work items, namely *travel* and *buffer time*. They are included to keep an overview of where time bottlenecks could potentially arise or be resolved. Travel activities marked with an asterisk can be rescheduled if the circumstances require more time or presence in Amsterdam.

Finally, weekly meetings have been scheduled with the supervisor which are not explicitly denoted in the schedule. At the time of writing they will take place on Thursdays at 10 am. They involve all

[10] https://s3.amazonaws.com/demo-datasets/beer_reviews.tar.gz
[11] https://github.com/awesomedata/awesome-public-datasets/pull/339, Accessed: 17-03-2018
[12] https://www.ratebeer.com/api-documentation.asp, Accessed: 17-03-2018
[13] https://graphql.org/
[14] https://snap.stanford.edu/data/web-RateBeer.html, Accessed: 17-03-2018
[15] https://www.brewersassociation.org/
[16] http://brewersofeurope.org/

other students working on the same overarching review analytics system.

Week	*Activities*
#14, 02.04.-08.04.	*Travel*, building scraping system, DAS-4 introduction
#15, 09.04.-15.04.	Finish scraping system, literature study, data collection & exploration
#16, 16.04.-22.04.	Evaluate appropriate NLP and ML techniques, related work
#17, 23.04.-29.04.	Build first summarization system
#18, 30.04.-06.05.	Writing, iterate on summarization system
#19, 07.05.-13.05.	Writing, iterate on summarization system, create first frontend, **mid-term presentation**
#20, 14.05.-20.05.	Writing, iterate on summarization system & frontend, begin employing scoring algorithms , *travel**
#21, 21.05.-27.05.	Writing, further work on the full system, evaluation of system performance *travel*
#22, 28.05.-03.06.	Writing, finalize work on summarization system
#23, 04.06.-10.06.	Writing
#24, 11.06.-17.06.	Writing, *buffer time*
#25, 18.06.-24.06.	Writing, *buffer time*, *travel*
#26, 25.06.-01.07.	**Final submission**
#27, 02.07.-08.07.	Prepare thesis defense, *travel**
#28, 09.07.-15.07.	Prepare thesis defense
#29, 16.07.-22.07.	**Thesis defense (Friday, 20.07.)**
#30, 23.07.-29.07.	*Buffer time*

Table 1: Proposed weekly activity schedule

REFERENCES

[1] Mohammed Al-Dhelaan and Abeer Al-Suhaim. 2017. Sentiment diversification for short review summarization. *Proceedings of the International Conference on Web Intelligence - WI '17* (2017), 723–729. https://doi.org/10.1145/3106426.3106525

[2] Stevan Rudinac B, Tat-seng Chua, Nicolas Diaz-ferreyra, Gerald Friedland, Tatjana Gornostaja, Benoit Huet, Rianne Kaptein, and Krister Lind. 2018. MultiMedia Modeling. 10704 (2018), 632–644. https://doi.org/10.1007/978-3-319-73603-7

[3] Henri Bal, Dick Epema, Cees De Laat, Rob Van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. 2016. A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term. *Computer* 49, 5 (2016), 54–63. https://doi.org/10.1109/MC.2016.127

[4] David M Blei, Blei@cs Berkeley Edu, Andrew Y Ng, Ang@cs Stanford Edu, Michael I Jordan, and Jordan@cs Berkeley Edu. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993 arXiv:1111.6189v1

[5] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (2004), 168. https://doi.org/10.1145/1014052.1014073

[6] Minqing Hu and Bing Liu. 2004. Mining Opinion Features in Customer Reviews. *19th national conference on Artifical intelligence* (2004), 755–760. https://doi.org/10.1145/1014052.1014073 arXiv:cs/0112017

[7] C Y Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on text summarization branches out (WAS 2004)* 1 (2004), 25–26.

[8] Mengwen Liu, Yi Fang, Alexander G. Choulos, Dae Hoon Park, and Xiaohua Hu. 2017. *Product review summarization through question retrieval and diversification.* Vol. 20. Springer Netherlands. 575–605 pages. https://doi.org/10.1007/s10791-017-9311-0

[9] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. *Proceedings of EMNLP* 85 (2004), 404–411. https://doi.org/10.3115/1219044.1219064 arXiv:arXiv:1011.1669v3

[10] N. K. Nagwani. 2015. Summarizing large text collection using topic modeling and clustering based on MapReduce framework. *Journal of Big Data* 2, 1 (2015), 1–18. https://doi.org/10.1186/s40537-015-0020-5

[11] a Nenkova and R Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *Proceedings of HLT-NAACL* 2004 (2004), 145–152. papers2://publication/uuid/DC675E84-0A45-48B7-A26C-F08B4B9398D3

[12] Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. 2013. Personalized time-aware tweets summarization. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* c (2013), 513. https://doi.org/10.1145/2484028.2484052