Narrative Recommendations based on Natural Language Preference Elicitation for a Virtual Assistant for the Movie Domain

Alessandro Francesco Maria Martina[1], Cataldo Musto[1,*], Andrea Iovine[1], Marco de Gemmis[1], Fedelucio Narducci[2] and Giovanni Semeraro[1]

[1] *University of Bari, Italy*
[2] *Polytechnic University of Bari, Italy*

Abstract
In this paper[1], we present a strategy to introduce *natural language preference elicitation* in a virtual assistant for the movie domain. Our approach allows users to express preferences on *objective* movie features (*e.g.,* actors, directors, etc.) that are extracted from a structured knowledge base, as well as on *subjective* features that are collected by mining movie reviews. The effectiveness of the approach was evaluated in a user study (N=103), where our strategy was integrated in a virtual assistant that acquires users' preferences expressed in form of natural language statements and generates a suitable movie recommendation. Results showed that users experience some difficulties in expressing their preferences in terms of *subjective* features. However, when people succeed in expressing their preferences by also using *subjective* properties, this generally leads to better recommendations.

Keywords
Recommender Systems, Natural Language Processing, Opinion Mining, Dialogue, Preference Elicitation.

1. Introduction

The recent rise of Virtual Assistants (VAs) [2] has led to the diffusion of technologies such as Google Assistant, Siri, and Alexa. Even if these systems proved to be very effective in fulfilling a broad range of *informative needs*, ranging from booking flights and playing music [3] to health-related services [4], the development of *personalization* strategies for VAs is still in a preliminary stage, and a significant research effort is currently put in the development of VAs that provide recommendations [5]. In order to provide users with satisfying recommendations, *preference elicitation* [6] is one of the main issues to be addressed [7]. In this setting, *natural language elicitation* recently gained attention since it mimics user-to-user interaction and makes

[1] An extended version of this paper has been published at the 30th ACM Conference on User Modeling Adaptation and Personalization with the title "A Virtual Assistant for the Movie Domain Exploiting Natural Language Preference Elicitation Strategies", as a Late Breaking Result. [1]

Discussion Papers - 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), November 28 - December 2, 2022, University of Udine, Udine, Italy

*Corresponding author.
a.martina13@studenti.uniba.it (A. F. M. Martina); cataldo.musto@uniba.it (C. Musto); andrea.iovine@uniba.it (A. Iovine); marco.degemmis@uniba.it (M. de Gemmis); fedelucio.narducci@poliba.it (F. Narducci); giovanni.semeraro@uniba.it (G. Semeraro)

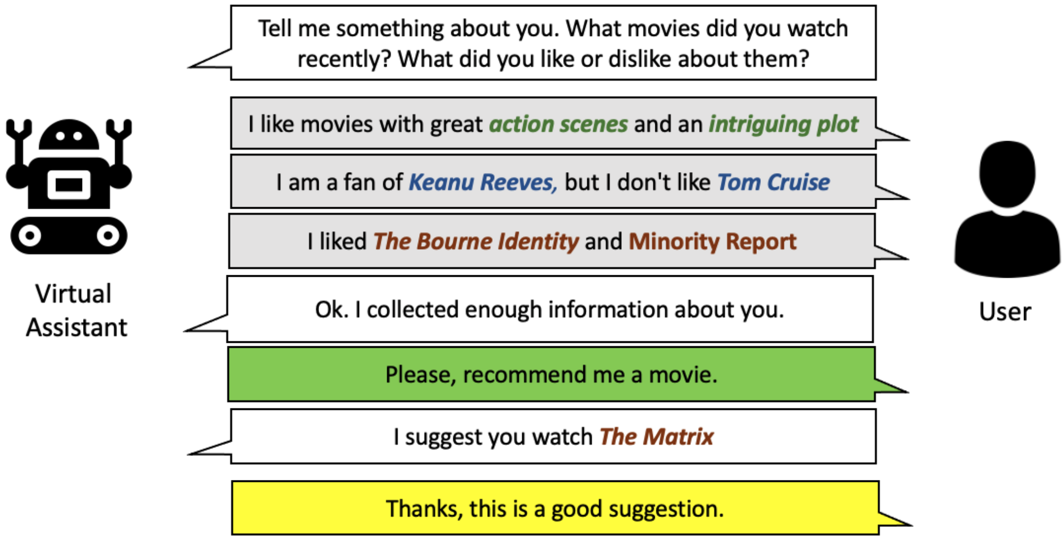

Figure 1: Interaction between a user and the VA

preference elicitation more *natural* and *satisfying* [8]. Based on these shreds of evidence, we present an approach to introduce *natural language preference elicitation* in a VA in the movie domain. Inspired by the paradigm of *narrative-driven recommendations* [9], the VA understands *narrative statements* expressing users' preferences and needs. An example of narrative request is: *"Movies with the genre 'Crime'.. Something like 'Nightcrawler'. And it is great if there is any form of plot twists"* [1]. By following the classification discussed in [10], people tend to express two types of preferences: *items* that they like or dislike (*"I like the Matrix"*), and *properties* that an item should or should not have. Properties can be either *objective* or *subjective*. Objective features concern non-controversial characteristics (*e.g.,* the actors of movie), while subjective features are based on opinions about the item. Accordingly, we designed a pipeline that allows the VA to recognize in natural language statements: *(i)* items liked by the users; *(ii)* objective characteristics of the items (*e.g., "I like movies with Keanu Reeves"*); *(iii)* subjective desired properties (*e.g., "I love amazing soundtracks"*). To evaluate our strategy, we carried out a user study (N=103) based on two variants of our preference elicitation strategy: one that allows users to talk only about *objective* properties, and one that also includes *subjective* properties.

The next sections are organized as follows: Section 2 describes the related work, Section 3 provides the details of the preference elicitation strategies and the general workflow. Section 4 describes the setup and results of the user experiments. Finally, Section 5 contains the conclusions and outlines future work.

[1]https://www.reddit.com/r/MovieSuggestions/comments/3fvycr/

2. Related Work

Preference Elicitation in Recommender Systems. Early attempts to gather users' interests relied on a *coarse-grained* preference elicitation based on item categories [11]. Next, research moved towards *explicit* preference elicitation. Popular approaches tackle this problem by asking users to *rate* or to *compare* a subset of items from the catalogue [12]. To reduce training time and maximize the information gathered from each answer, some strategies to automatically select the items to be rated have been also proposed [13]. The distinctive trait of our methodology in terms of *preference elicitation* lies in the fact that: *(1)* we designed a strategy which exploits *natural language statements* rather than classical explicit ratings; *(2)* we do not select the items to be rated. Conversely, we put the users *in control* of the elicitation process, and we allow them to freely express preferences and informative needs as natural language statements.

Natural Language Preference Elicitation. Early work [14] investigated how to acquire user preferences in the form of natural language statements. As previously stated, the distinction between *objective* and *subjective* features is inherited by [10]. However, differently from Kang et al., who just develop an interface to gather user statements and analyze how people use language, we applied this concept in the field. In particular: *(1)* we developed two different pipelines to extract subjective and objective features; *(2)* we designed a strategy to model users' interests based on the mentions to objective and subjective features in natural language statements; *(3)* we integrate our approach in a fully working VA for the movie domain. Another relevant piece of work is presented in [15], where the concept of *narrative-driven recommendations* (NDRs) is introduced. The effectiveness of NDRs is evaluated in [9, 16] in the books and movies domain. In both cases, authors carried out an *in-vitro* experiment. Differently from this work, we exploited *narrative statements* in the *preference elicitation* phase, since we introduce a strategy to recognize user preferences from natural language statements.

3. Description of the Methodology

Figure 2 describes the general workflow carried out by our strategy. The workflow can be split into two phases: *knowledge extraction*, whose goal is to extract descriptive features from structured and unstructured content, and *knowledge exploitation*, where the features are made available to the VA, which uses this information for *preference elicitation* and *recommendation*.

Knowledge Extraction. The knowledge extraction phase aims to: *(1)* collect descriptive features that characterize the items; *(2)* store the features into a knowledge base (KB). The knowledge extraction process is further split into two pipelines: the first focuses on *structured* knowledge sources, and aims to gather *objective* features from knowledge graphs (KG) such as Wikidata [17], while the second one runs opinion mining techniques [18] on *unstructured* knowledge sources (*i.e.,* user-generated content [19, 20] or users' reviews) to extract *subjective* features. In the first case, the procedure is carried out by mapping each *logical* entity that can be recommended, *i.e.,* items in the catalogue, with the corresponding *physical* entity in a KG. Mapping is obtained by matching item metadata (*i.e.,* title of a movie) with the URIs of the entities available in the KG. Once the mapping was completed, we extracted information about *actors, directors, genres.* An example of the output of this step is reported in the *left* box of the portion of KB presented in Figure 2.

On the other side, *subjective* properties are more related to the *perception* of the item, since they refer to characteristics that involve a degree of judgement [10]. To collect these properties, we designed an *opinion mining* pipeline inspired by [21]. In a nutshell, *subjective* features of an item are obtained by identifying *uni-grams* and *bi-grams* that are frequently mentioned with a *positive sentiment* in the reviews. To this end, for each item, we collect some reviews and we split them into *sentences*. Next, *sentiment analysis* is used [22] to determine the sentiment conveyed by each sentence. As in [23, 24], all the sentences expressing a negative or neutral sentiment are filtered out. After this step, all the lemmas mentioned in all the positive reviews are obtained. As for uni-grams, we maintain nouns and adjectives. As for bi-grams, we identify *noun-noun* and *adjective-noun* pairs. Our choices are based on previous research [25]. Next, we picked the top-100 uni-grams and bi-grams based on their TF-IDF score. To calculate TF-IDF, we considered all reviews for an item as an individual document. An example of the output is shown in the right box of the KB in Figure 2.

Knowledge Exploitation. This part of the pipeline aims to make the VA able to: *(1)* correctly catch the mentions to objective and subjective features in the preference elicitation process; *(2)* make the features available to the recommendation algorithm. As regards the first point (see the *dark grey* box in Figure 2), we provided the VA with some NLU capability, which is obtained through the combination of Intent Recognizer (IR), Named Entity Recognizer (NER) and Sentiment Analyzer. The goal of the IR is to correctly understand the *goals* or the *actions* the users have in mind when they interact with the VA. Operationally, the IR takes as input a message written by the user and classifies it into a fixed set of categories called *intents*. In our implementation, the IR classifies each message against three different intents *i.e., 'provide preferences'*, *'ask for recommendation'* and *'feedback on recommendation'*. These intents are directly inherited from similar approaches for CoRS [8]. As an example, by referring to Figure 1, the first three messages written by the user (highlighted in *grey*) can be classified as *'provide preferences'*, while the intent of the fourth and fifth messages (highlighted in *green* and *yellow*) are *'ask for recommendation'* and *'feedback on recommendation'*, respectively. Due to space limitations, we do not provide further details regarding the intent recognition process. For most of the general underlying concepts, the reader may refer to [8]. Next, in order to recognize mentions to entities contained in the text, we implemented a state-of-the-art NER module. The algorithms first exploits CRF to identify potential entities. Then, fuzzy string matching is used to map candidate entities to the elements in the knowledge base (*i.e.,* items properties). If a match is obtained, the entity is stored in the profile of the user as a *preference*. Finally, the VA should be also able to correctly understand the *sentiment* conveyed by the messages. In the second sentence highlighted in gray in Figure 1, the user mentions two entities in the same message, one with positive sentiment and one with negative sentiment. The goal of the Sentiment Analyzer is to process each sentence written by the user and to associate the correct sentiment to the fragment of text.

When a sufficient number of preferences is collected by the VA, users can ask for a recommendation (*green* sentence in Figure 1). In this case, user preferences are passed to a recommendation algorithm, which in turn provides a suitable suggestion. Next, users can express feedback on the recommendation (*yellow* sentence in Figure 1) and whenever the suggestion is not liked, a new recommendation cycle starts. In our case, recommendations are provided by exploiting

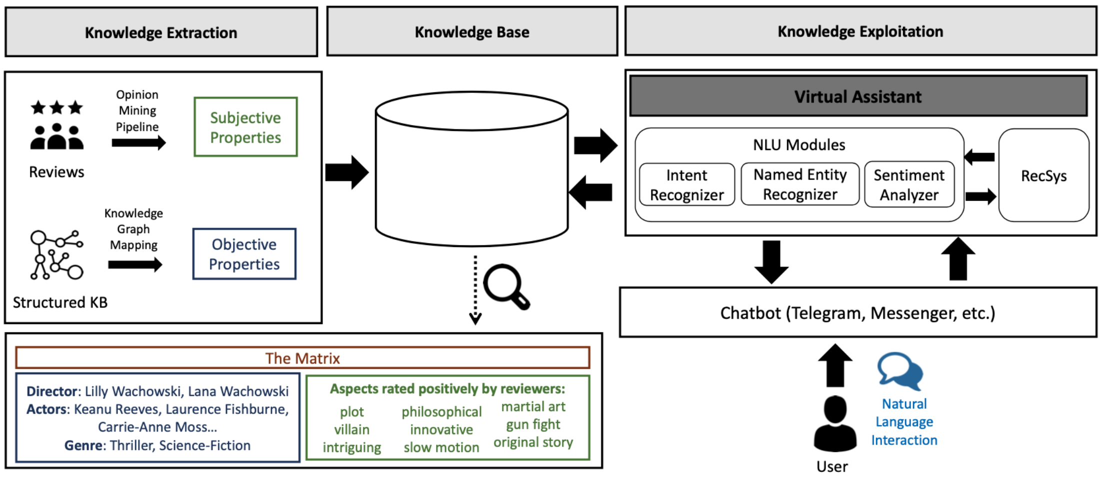

Figure 2: Workflow of our strategy for natural language preference elicitation

a content-based algorithm based on Doc2Vec[2] [26]. However, it is important to emphasize that the choice of the recommendation model is not crucial here, since in this work we aim to introduce a strategy to elicit user preferences through natural language. The analysis of further algorithms [27] is left as future work.

4. User experiment

In our user study (N=103), we asked users to interact with our VA by providing their preferences and by evaluating the recommendations they received.

In particular, we aim to answer to the following research questions: **(RQ1)** *How do people use natural language to express their preferences and needs?* We are interested in both quantitative (*e.g.*, amount of preferences expressed) and qualitative analyses (*e.g.*, the lexicon used by the users). Moreover, we want to assess to what extent users feel confident with our strategy; **(RQ2)** *How accurate are the recommendations generated by eliciting users' preferences through natural language statements? How do objective and subjective features affect the quality of the recommendations?*

Experimental Design. The user study involved 103 users (77.6% men, 100% aged 21-30, 61.8% already used a RS, 54.2% high interest in movies, 30% regularly or moderately used a VA), recruited by following the common *availability sampling* strategy. In order to investigate how the features impact the *quality of the recommendations*, we defined *two* experimental conditions: *(1)* **Objective**, where users can express their preferences by *only using* objective properties; *(2)* **Objective+Subjective**, where users can express their preferences for movies by using both the groups of properties. We discarded the configuration based *only* on subjective features because

[2]In our preliminary experiments, Doc2Vec was compared to other representation methods for word and sentence embeddings such as Word2Vec, BERT and so on. We preferred Doc2Vec based on the better ranking provided by the algorithm. Due to space reasons, we limit the discussion to one algorithm.

Table 1
HitRate results

	Objective	Objective + Subjective
HitRate@1	78%	**88.46%**
HitRate@2	96%	**98.11%**
HitRate@3	96%	**98.11%**

preliminary tests showed that users did not feel comfortable when they were forced to provide preferences by just using them.

Before running the experiment, participants were informed about the goal of the experiment, were taught about the *lexicon* they could use to interact with the VA, (*i.e.,* only objective properties, or both objective and subjective), and about the statements (*i.e.,* the intents) the VA understands. Users assigned to configurations based on *objective* features were not aware of the opportunity of also expressing statements containing *subjective* characteristics of the items. To sum up, the experiment followed a *between-subjects* protocol: (1) Each participant was randomly assigned to one of the experimental conditions; (2) Each participant interacted with the VA and expressed her preferences. When the minimum amount of preference was collected (set to 5, according to a rough heuristic), the user could ask for a recommendation; (3) The system returned a recommendation, and the participant expressed a binary feedback. In case of positive feedback, the experiment ended. (4) Otherwise, participants were asked to provide new preferences, and they could ask for a new recommendation. The process ended in any case after three negative responses; (5) At the end of the experiment, each user answered a *post-usage questionnaire*.

Implementation Details. Our catalog of items was built by gathering all the elements belonging to the 'Movie' category from Wikidata. Next, objective properties were extracted from Wikidata by following the mapping procedure previously presented. Subjective properties were extracted by processing a subset of the Amazon Movie Reviews Data[3] by exploiting the opinion mining pipeline previously introduced. As shown in [10], sometimes objective and subjective properties may be overlapping. In order to simplify the analysis, we removed objective properties that refer to personal perceptions or emotions (*e.g.,* romantic movie). Table 2 depicts some statistics about the knowledge base[4] employed in the experiment. Intent Recognition relies on Google Dialogflow, while NER is implemented using a custom-trained model from CoreNLP. Finally, Sentiment Analyzer relies on the CoreNLP Sentiment Tagger. As for content-based recommendations, we exploited the Python implementation of Doc2Vec available in Gensim. Dimension of the vectors was set to 300, after parameter tuning.

Metrics and Questionnaire. We adopted Preference Count (PR) (*i.e.,* average number of preferences expressed by each user), and HitRate@K (HR@K), which is the number of satisfactory recommendations obtained within K turns of conversation. Moreover, we also asked the users to fill in a post-usage questionnaire based on the ResQue model [28], which aims to assess the effectiveness of the strategy. Answers are provided on a 5-point Likert Scale.

[3] https://snap.stanford.edu/data/web-Movies.html
[4] Link to the knowledge base on an anonymous repository: https://github.com/machetegrapefruit/fictional-waddle

Table 2
Knowledge Base statistics

Movies	17150				
Objective Properties	44404	**Avg. Objective Prop. per Item**	13.52 ± 10.61	**Avg. Item per Objective Prop.**	5.22 ± 36.67
Subjective Properties	12017	**Avg. Subjective Prop. per Item**	96.02 ± 11.74	**Avg. Item per Subjective Prop.**	42.39 ± 54.21

4.1. Discussion of the Results

In order to answer **RQ1**, we first analyze *how many* preferences were gathered, and *what characteristics* they have. Throughout the experiment, we collected 748 messages[5] (7.26 per user, on average).

By analyzing the 640 messages correctly recognized, we noted that the NLU algorithms recognized mentions to 664 entities (6.44 per user, on average). Next, by splitting the results based on preference type, we noted that 76.9% of the entities belong to *objective* properties (4.6 per user, on average), while only 23.1% refer to *subjective* properties. This is probably due to the fact that users are more familiar in expressing preferences in form *objective* properties, rather than indicating more articulated *subjective* characteristics. In order to deepen the analysis, we also analyzed the top-10 properties the users mentioned in their messages during the preference elicitation phase. In general, we noted that users mentioned *actors* and *movies* as objective features. Subjective features are used to refer to more specific characteristics, such as *soundtrack* or a *photography*.

Next, in Table 3 we report the results obtained by the post-usage questionnaire. Results show that the use of subjective features led to mixed outcome in terms of *interface adequacy*: the users found it easier to express their preferences using objective properties alone, but they found that the system was able to better understand what they were talking about when all properties are available. A small decrease was observed for *control* and a tiny increase was noted in terms of *ease of use*, thus it is likely that users had difficulty expressing their preferences, but this did not adversely affect the usability. To conclude, this part of the study confirmed that the combination of subjective and objective features provides users with more opportunities to express their preferences. However, the fact that users are not particularly familiar with the lexicon to use is of hindrance for their complete exploitation. In order to answer **RQ2**, we analyze the average HitRate. First, we can state that our preference elicitation strategy allows users to get good recommendations since, more than 80% of the users obtain a good recommendation after the first turn, and the value further increases between 96% and 98% at turn three (see Table 1). Moreover, results showed that the injection of subjective features led to more accurate recommendations, especially for HitRate@1. This means that the presence of subjective features allows our content-based algorithm to generate more precise recommendations, and this happens regardless the difficulties that the users experienced in the preference elicitation step.

[5]This value is based on the messages whose intent recognized by the IR is *'provide preference'*. Other kind of messages are not of interest here.

Table 3
Average answers collected in the post-usage questionnaire. Best-performing configuration is reported in bold.

Construct	Question	Objective	Obj + Subj
Ease of Use	*I became familiar with the recommender system very quickly*	4.24	**4.26**
Control	*I feel in control of expressing my actual preferences*	**4.28**	4.19
Int. Adequacy	*I found it easy to tell the system what I like/dislike*	**4.28**	4.15
Int. Adequacy	*It is easy for the recommender to understand what I said*	4.1	**4.25**

5. Conclusion

The main contribution of this paper is a strategy to elicit natural language user preferences in a VA for the movie domain. Our approach is based on a *knowledge extraction* pipeline, where both objective and subjective features are obtained from structured and unstructured knowledge sources, which is followed by a *knowledge exploitation* pipeline, where the information previously extracted are exploited by NLU modules and recommendation algorithm. The proposed pipeline was then evaluated in a user study whose results showed that the approach allows users to express their preference and to receive accurate recommendations. Another outcome was that people tend to express their preferences in terms of *objective* features, and discard *subjective* features. However, when the users are able to use subjective features, better recommendations are usually generated. As future work, we aim to further increase the external validity by evaluating our approach in a different domain [29], and we will introduce better techniques for recommendation and NLU modules. Moreover, techniques for implicit preference modeling based on holistic user profiles will be investigated as well [30].

References

[1] A. F. M. Martina, C. Musto, A. Iovine, M. de Gemmis, F. Narducci, G. Semeraro, A virtual assistant for the movie domain exploiting natural language preference elicitation strategies, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 7–12.

[2] A. S. Tulshan, S. N. Dhage, Survey on virtual assistant: Google assistant, siri, cortana, alexa, in: International symposium on signal processing and intelligent recognition systems, Springer, 2018, pp. 190–201.

[3] C. Musto, F. Narducci, M. Polignano, M. De Gemmis, P. Lops, G. Semeraro, Myrrorbot: A digital assistant based on holistic user models for personalized access to online services, ACM Transactions on Information Systems (TOIS) 39 (2021) 1–34.

[4] M. Polignano, F. Narducci, A. Iovine, C. Musto, M. De Gemmis, G. Semeraro, Healthassistantbot: a personal health assistant for the italian language, IEEE Access 8 (2020) 107479–107497.

[5] D. Rafailidis, Y. Manolopoulos, Can Virtual Assistants Produce Recommendations?, in: Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, 2019, pp. 1–6.

[6] M. d. Gemmis, L. Iaquinta, P. Lops, C. Musto, F. Narducci, G. Semeraro, Learning preference models in recommender systems, in: Preference Learning, Springer, 2010, pp. 387–407.
[7] B. Priyogi, Preference Elicitation Strategy for Conversational Recommender System, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 824–825.
[8] D. Jannach, A. Manzoor, W. Cai, L. Chen, A Survey on Conversational Recommender Systems, arXiv preprint arXiv:2004.00646 (2020).
[9] T. Bogers, M. Koolen, "I'm looking for something like ...": Combining Narratives and Example Items for Narrative-driven Book Recommendation (2018) 10.
[10] J. Kang, K. Condiff, S. Chang, J. A. Konstan, L. Terveen, F. M. Harper, Understanding How People Use Natural Language to Ask for Recommendations, in: Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17, ACM Press, Como, Italy, 2017, pp. 229–237. URL: http://dl.acm.org/citation.cfm?doid=3109859.3109873. doi:10.1145/3109859.3109873.
[11] H. Nguyen, P. Haddawy, The decision-theoretic interactive video advisor, in: In Proceedings of the Fifteenth conference on Uncertainty in artificial, 1999, pp. 494––501.
[12] B. Carterette, P. N. Bennett, D. M. Chickering, S. T. Dumais, Here or there, in: European Conference on Information Retrieval, Springer, 2008, pp. 16–27.
[13] M. Elahi, F. Ricci, N. Rubens, A survey of active learning in collaborative filtering recommender systems, Computer Science Review 20 (2016) 29–50. Publisher: Elsevier.
[14] I. Nunes, S. D. Barbosa, D. Cowan, S. Miles, M. Luck, C. J. de Lucena, Natural language-based representation of user preferences, Interacting with Computers 27 (2015) 133–158.
[15] T. Bogers, M. Koolen, Defining and Supporting Narrative-driven Recommendation, in: Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17, ACM Press, Como, Italy, 2017, pp. 238–242. URL: http://dl.acm.org/citation.cfm?doid=3109859.3109893. doi:10.1145/3109859.3109893.
[16] L. Eberhard, S. Walk, L. Posch, D. Helic, Evaluating narrative-driven movie recommendations on reddit, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 1–11.
[17] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85. Publisher: ACM New York, NY, USA.
[18] L. Chen, P. Pu, Survey of preference elicitation methods, Technical Report, 2004.
[19] P. Lops, M. d. Gemmis, G. Semeraro, C. Musto, F. Narducci, M. Bux, A semantic content-based recommender system integrating folksonomies for personalized access, in: Web Personalization in Intelligent Environments, Springer, 2009, pp. 27–47.
[20] C. Musto, F. Narducci, M. De Gemmis, P. Lops, G. Semeraro, Star: a social tag recommender system, Proceedings of the ECML/PKDD Discovery Challenge (2009) 215–227.
[21] R. Dong, M. Schaal, M. P. O'Mahony, B. Smyth, Topic extraction from online reviews for classification and recommendation, in: Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI 13), Beijing, China, 3-9 August 2013, AAAI, 2013, pp. 1310–1316.
[22] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intelligent systems 28 (2013) 15–21.
[23] C. Musto, P. Lops, M. de Gemmis, G. Semeraro, Justifying recommendations through

aspect-based sentiment analysis of users reviews, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 4–12.

[24] C. Musto, M. de Gemmis, P. Lops, G. Semeraro, Generating post hoc review-based natural language justifications for recommender systems, User Modeling and User-Adapted Interaction (2020) 1–45.

[25] H. Nakagawa, T. Mori, A simple but powerful automatic term extraction method, in: COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14, Association for Computational Linguistics, 2002, pp. 1–7.

[26] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[27] C. Musto, T. Franza, G. Semeraro, M. de Gemmis, P. Lops, Deep content-based recommender systems exploiting recurrent neural networks and linked open data, in: Adjunct Publication of the 26th conference on user modeling, adaptation and personalization, 2018, pp. 239–244.

[28] P. Pu, L. Chen, A User-Centric Evaluation Framework of Recommender Systems 612 (2010) 8.

[29] A. Buccella, A. Cechich, D. Gendarmi, F. Lanubile, G. Semeraro, A. Colagrossi, Building a global normalized ontology for integrating geographic data sources, Computers & Geosciences 37 (2011) 893–916.

[30] C. Musto, M. Polignano, G. Semeraro, M. de Gemmis, P. Lops, Myrror: a platform for holistic user modeling, User Modeling and User-Adapted Interaction 30 (2020) 477–511.