Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation

Yongfeng Zhang
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
zhangyf07@gmail.com

ABSTRACT

Previous research on Recommender Systems (RS), especially the continuously popular approach of Collaborative Filtering (CF), has been mostly focusing on the information resource of explicit user numerical ratings or implicit (still numerical) feedbacks. However, the ever-growing availability of textual user reviews has become an important information resource, where a wealth of explicit product attributes/features and user attitudes/sentiments are expressed therein. This information rich resource of textual reviews have clearly exhibited brand-new approaches to solving many of the important problems that have been perplexing the research community for years, such as the paradox of cold-start, the explanation of recommendation, and the automatic generation of user or item profiles. However, it is only recently that the fundamental importance of textual reviews has gained wide recognition, perhaps mainly because of the difficulty in formatting, structuring and analyzing the free-texts. In this research, we stress the importance of incorporating textual reviews for recommendation through phrase-level sentiment analysis, and further investigate the role that the texts play in various important recommendation tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Filtering; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.3.5 [**Online Information Services**]: Web-based services

Keywords

Personalized Recommendation; Collaborative Filtering; Sentiment Analysis; Text Mining

1. INTRODUCTION

The continuous prospering of various Web2.0 online applications such as e-commerce and social networks has pushed users into the problem of information overwhelming [7]. The difficulty in accessing the desired online items further contributed to the emerging of Personalized Recommender Systems (PRS) [12], which attempt to make personalized and targeted item recommendations to users on various platforms and devices.

The research of personalized recommendation can be generally classified into Content-based [20], Collaborative Filtering (CF)-based [26] and Hybrid approaches [3]. The content-based approach attempts to construct user and item profiles, and thus to make recommendations through some meticulously designed matching strategies [14], while CF-based approaches attempt to learn the preferences of users automatically by considering the historical choices of other users [12]. Hybrid recommender system, on the other hand, aims to combine the advantages of various strategies to make more informed recommendations.

The CF-based approach has gained much attention from the research community, especially since the Netflix grand prize in the year of 2007 through 2009 [2], because the CF-approaches, especially those based on Matrix Factorization (MF) techniques [28, 8] on user-item numerical rating matrices, achieved important success in the task of rating prediction [12], and also exhibited great advantage in many other recommendation tasks [4].

However, the application of CF on numerical ratings has come across many difficulties in face of some key problems like data sparsity [35] and the explainability of numerical ratings [29, 32]. This further leads to some of the most concerned tasks in the research community, such as cold-start recommendation [36, 13], explainability of the recommendations [32], and automatic user/item profile generation.

The continuous growing of online textual user reviews, as another important information resource besides numerical ratings, has shed light on brand new solutions towards these issues. For example, although a user may have only made a single numerical rating towards a product in online shopping websites, she usually expresses more detailed opinions towards various product features/aspects in the corresponding piece of review text [33]. This is exposited in the sampled review in Figure 1, where the user expressed positive

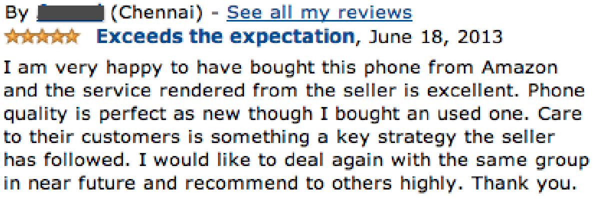

By [redacted] (Chennai) - See all my reviews
★★★★★ **Exceeds the expectation**, June 18, 2013
I am very happy to have bought this phone from Amazon and the service rendered from the seller is excellent. Phone quality is perfect as new though I bought an used one. Care to their customers is something a key strategy the seller has followed. I would like to deal again with the same group in near future and recommend to others highly. Thank you.

Figure 1: A piece of sampled user review towards the iPhone 5s product extracted from Amazon.com

WSDM'15, February 2–6, 2015, Shanghai, China.

ACM 978-1-4503-3317-7/15/02 ...$15.00.
http://dx.doi.org/10.1145/2684822.2697033.

opinions towards the features *service* and *phone quality* of the product, with the opinion words *excellent* and *perfect*, correspondingly, which composes into his overall numerical rating of five stars.

We see that the textual reviews contain both product-oriented information (i.e. product features) and user-oriented information (i.e. user opinions), and they usually exist in the form of pairs (user takes an opinion word to express his/her attitude towards a product feature). By conducting phrase-level sentiment analysis [15, 10, 5, 31] on the textual reviews, we are able to extract these feature-opinion word pairs, thus to gain more detailed information about the user's overall opinion towards a product (the overall rating), which helps to understand the item characterises and user needs in a wider range of dimension, and to alleviate the problem of data sparsity in the scenario of cold-start recommendation. The extracted features and opinions also help to make explanations about why or why not an item is recommended [32], and to construct user (or item) profiles automatically by estimating their preferences towards the features.

In this research, we aim to stress the importance of making further use of textual reviews in recommender systems. We focus on leveraging phrase-level sentiment analysis on the reviews to better solve the above mentioned cutting-edge research problems. In the following part, we review the related work in Section 2, and exposit some of the research topics, current research progress and the upcoming research plans on each topic in Section 3. Finally, we discuss, conclude and summarize the future directions in Section 4.

2. RELATED WORK

Collaborative Filtering (CF)-based techniques [26] have achieved great success in personalized recommender systems [12] due to their ability to take advantage of the wisdom of crowds, especially in the task of numerical rating prediction. With the remarkable performance on prediction accuracy, the Matrix Factorization (MF) [28] approaches have gained great popularity in both research community and the industry. Some of the commonly used matrix factorization algorithms include Singular Value Decomposition (SVD) [1, 24], Non-negative Matrix Factorization (NMF) [9], Probabilistic Matrix Factorization (PMF) [23, 22] and Max-Margin Matrix Factorization (MMMF) [25, 21].

However, the ratings made by each each is usually far less than the large volume of products in a typical system, which implies that the user-item rating matrices that CF algorithms attempt to tackle with are usually very sparse [35], as exampled in Figure 2, which shows the scattered small communities corresponding to the sparse submatrices on the Yelp rating dataset [35]. Besides, new users and items are continuously added to the online systems, which further worsens the sparsity [34]. All these factors lead to the important cold-start problem in recommender systems, where it is difficult to estimate the preferences or make recombinations to a user who rated only a few of the items [36, 13].

Fortunately, the ever growing availability of textual reviews has shed light on new approaches to alleviate the cold-start problem. The product features and user opinions included in the textual reviews can be extracted, formatted and summarized through Sentiment Analysis [11, 18] techniques. One of the core tasks in sentiment analysis is to determine the sentiment orientations that users express in reviews, sentences or on specific product features, corresponding to review(document)-level [19], sentence-level [30, 17] and phrase-level [31, 15, 5] sentiment analysis.

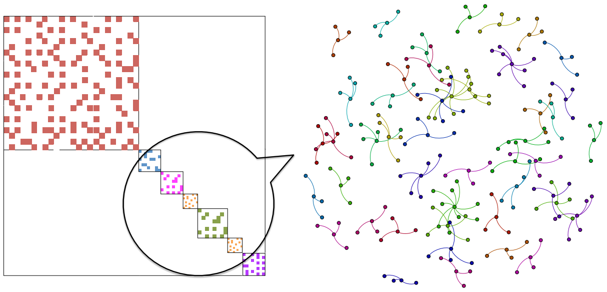

(a) Yelp dataset Matrix (b) Yelp dataset Graph

Figure 2: Structures of Yelp dataset. In the left is the exampled structure of the rating matrix, and in the right is the real structure of the scattered blocks.

Review- and sentence-level sentiment analysis attempt to label a review or sentence as one of some predefined sentiment polarities, which are typically *positive*, *negative* and sometimes *neutral* [11]. Phrase-level sentiment analysis aims to analyze the sentiment expressed by users in a finer-grained granularity. It considers the sentiment expressed on specific product features or aspects [6]. One of the most important tasks in phrase-level sentiment analysis is the construction of Sentiment Lexicon [27, 10, 5, 15], which is to extract feature-opinion word pairs and their corresponding sentiment polarities from these opinion rich user-generated free-texts.

In [16], McAuley et al leveraged topic modelling to help extract the hidden topics from user reviews, thus to help improve the rating prediction accuracy. Textual reviews also help to construct intuitional explanations about why an item is recommended against the others. In [32], Zhang et al proposed a feature-level explainable recommendation strategy where the system persuades a user by telling him about his previously concerned product features in historical reviews.

As an integration of content- and CF-based recommendation strategies, the hybrid recommendation techniques [3] have achieved state-of-the-art performance in real-world applications [12]. However, the manual construction of user and item profiles requires a vast amount of domain knowledge, which is expensive and time consuming [20, 14]. Phrase-level sentiment analysis on textual reviews makes it possible to conduct automatic profile construction, by analyzing and structuring the reviews corresponding to a target user or product. In this work, we exposit the promising potentialities that textual reviews bring into recommender systems, state our current research achievements on the related topics, and pose some of the future research directions.

3. RESEARCH TOPICS

3.1 Cold-Start Recommendation

In CF-based recommendation algorithms, one of the most fundamental causes of cold-start comes from the absence of purchasing or rating information of new users or items. Although various CF techniques attempt to construct meticulously designed algorithms to estimate user preferences from a small number of ratings [13, 36], the performance remains limited as we know little about a user philosophically.

For example, many a user (about 49%) in the commonly used Yelp rating dataset[1] made only a single numerical rating, which makes it difficult to estimate his/her preference in a latent factorization space [36]. However, the corresponding piece of textual review plays a role of the user's explanation towards his/her rating, which may contain the user's opinions towards several features of the product that composes into his final rating. This information source could have been well used to tackle with the cold-start problem.

This procedure is exposited in Figure 4. We first integrate all the textual reviews in a product domain (mobile phone domain in this example) to construct a review corpus. By conducting phrase-level sentiment analysis on this corpus, a sentiment lexicon [27, 10, 5, 15] is constructed, which includes the product feature words (e.g. *screen, battery, price*) extracted from user reviews.

Next, we analyze the review(s) corresponding to each targeted user by extracting the features that he/she concerns. It is important to point out that this procedure can be conducted even though the user made only a single review, because there may also exist a bunch of mentioned features in the review text. Finally, the recommended items can be provided by selecting those products that perform well on those features that the user concerns, and the product performance on each feature can also be estimated from the reviews made by users towards that product.

3.2 Recommendation Explanation

An important problem of traditional CF-based recommendation algorithm in real-world application is the difficulty to explain the recommendation results. This is partially because of the fact that we do not know how a user composed his opinions from the many aspects into a single and simple numerical rating, and that CF algorithms (especially those based on matrix factorization techniques) only attempt to estimate these ratings in a latent (unknown) factorization space. These Latent Factor Models (LFM) makes it even more difficult to make the recommendations explainable, although the algorithm may achieve satisfactory rating prediction accuracies [29].

However, the existence of textual user reviews, as exposited in the previous section, provides a brand new information resource to help understand the user preferences and specific needs. By extracting the frequently mentioned product features from a user's historical reviews, we are able to get to know the product aspects that he/she concerns.

Different users may care about different produce features when making purchasing decisions. For example, a user may choose a mobile phone product given its large screen and good graphics performance, while another may make the same choice while considering its nice product design, although they may well give the same numerical rating of five stars. In many similar cases, the numerical ratings are insufficient to distinguish the preferences of different users, but the textual reviews tell us why a user made such a choice.

Preliminary studies on this research topics have been published in our paper [32], which attempts to improve the rating prediction accuracy and at the same time construct intuitional recommendation explanations. We will further investigate the explanations constructed from textual reviews by considering different explanation forms like product tags and word clouds, etc., as well as the scrutability, effectiveness and efficiency of explanations in recommendation. We will also attempt to bridge up explanations and its application in cold-start recommendation scenarios in the following research tasks.

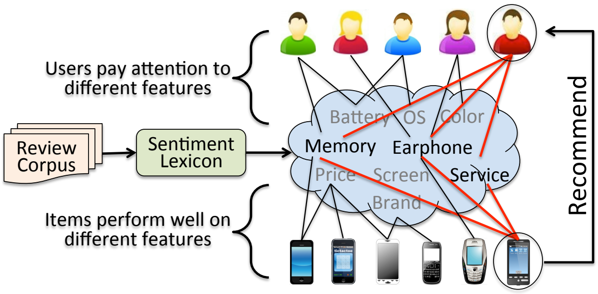


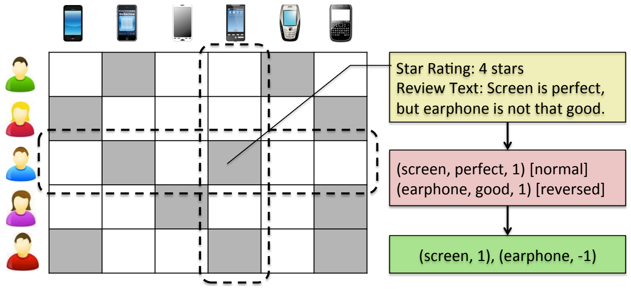
Figure 4: The product feature set of a domain (e.g. mobile phone) is extracted from a review corpus. The preference/profile of a user/item can be constructed by analyzing the corresponding review(s) of the user/item, thus to make cold-start recommendations when the ratings of a user is limited.

3.3 Automatic Profile Generation

The construction of accurate user/item profiles is the key factor in content-based recommendation, and they help to make more informed recommendations in hybrid recommender systems [3]. However, the profile construction process is time consuming, and usually requires the existence of domain knowledge, which is expensive in real-world scenarios.

However, the textual reviews in a specific product domain naturally serve as a kind of domain-dependent information, and it is possible to extract the user-generated domain knowledge with the help of the wisdom of crowds.

This can be intuitionally exposited in Figure 5. We first extract all the matched product features and their accompanying opinion words from each piece of review, and estimate the appropriate score that the user expressed towards each of the features, by considering the sentiment polarities of the opinion words.



Star Rating: 4 stars
Review Text: Screen is perfect, but earphone is not that good.

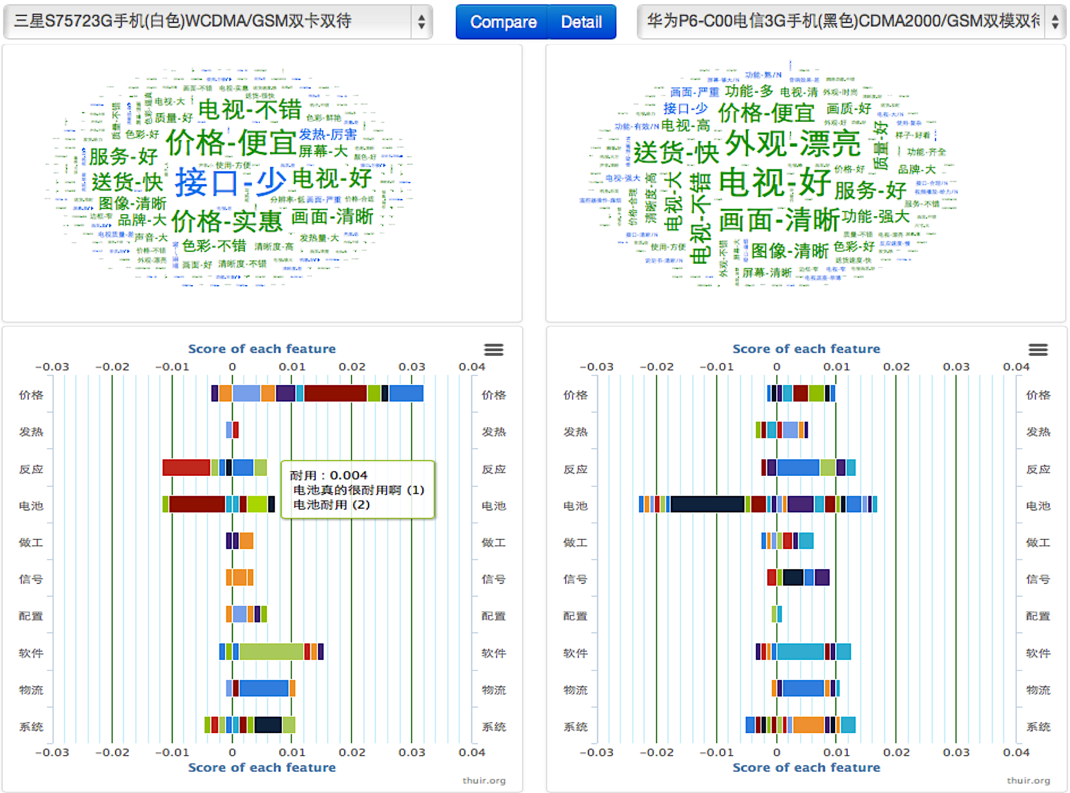
(screen, perfect, 1) [normal]
(earphone, good, 1) [reversed]

(screen, 1), (earphone, -1)

Figure 5: An example of user-item review matrix and automatic profile extraction. Each shaded block is a review made by a user towards an item; the entries included in the review are extracted, and further transformed to feature scores while considering the negation words. The features together with the user opinions extracted from the set of reviews corresponding to a user/item are further aggregated to generate a profile.

---

[1] `http://www.yelp.com/dataset_challenge`



耐用：0.004
电池真的很耐用啊 (1)
电池耐用 (2)

Figure 3: **Automatic product profile construction and comparison. The word-cloud based profile displays the Feature-Opinion word pairs extracted from the product's corresponding reviews, where the green ones are positive comments and blue ones are negative, giving an intuitional and first-sight view of the pros and cons of each product. The bar chart based profiles allow the users to examine the detailed scores and reviews on each specific product feature. This demo can be visited at http://mobile.thuir.org.**

After that, a simple and direct approach of profile generation for a user is to take all of his/her reviews and calculate the frequency of each feature therein, which intuitionally serve as an indicator of the extent that a user cares about a feature. Symmetrically, we can also consider all the reviews corresponding to a product, and integrate the opinions from different users to estimate its performance on each product feature, thus to construct the product profiles.

The product profiles can be displayed in different forms to help users better understand the pros and cons of a product, or to compare the profiles of two products so as to make more informed purchasing decisions. The reader might refer to our prototype demo systems on smart TV[2] and mobile phone[3] products, and this is also exampled in Figure 3, which displays and compares the profiles of two mobile phones, in the forms of both word cloud and bar charts.

Apart from exhibiting the profiles to users for product comparison, the user and product profiles can also be easily used for content-based or hybrid recommendation, because they model the users and items in the same and intuitional feature space, thus, for example, it would be very easy to calculate the similarity between two users, two products, or even a user and a product. We will further investigate the application of automatic profiling in personalized recommendation in the following work.

4. CONCLUSIONS AND RESEARCH PLAN

In this research proposal, we attempt to incorporate textual user reviews to tackle with some of the most cutting-edge topics in the research of personalized recommender systems. We particularly focus on the utilization of phrase-level sentient analysis technique to format, structure and summarize the free-text reviews, so as to take feature-level advantages of this information-rich resource.

We argue that the existence of textual reviews is of fundamental importance to personalized recommendation, even compared with the previously mostly concerned numerical ratings. However, the textual reviews have mostly been ignored for years in the previous research, perhaps due to the tremendous prosperity of collaborative filtering techniques based on user ratings. In the following work, we will continue to investigate the great potential of leveraging textual reviews in solving the problems of cold-start recommendation, recommendation explanation and automatic profile generation for hybrid recommender systems.

5. ACKNOWLEDGEMENT

I would like to sincerely thank my supervisors Prof. Min Zhang, Yiqun Liu, Shaoping Ma and Prof. Yi Zhang from UCSC, as well as Prof. Tat-Seng Chua and Min-Yen Kan from NUS for their continuous mentoring and support during my research career. This work was supported by Tsinghua-Samsung Joint Lab, under the Chinese Natural Science Foun-

---

[2] `http://tv.thuir.org`
[3] `http://mobile.thuir.org`

dation (61073071) and National High Technology Research and Development (863) Program (2011AA01A205). Part of the work was conducted at the Tsinghua-NUS NExT Search Center supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

6. REFERENCES

[1] H. Abdi. Singular value decomposition (svd) and generalized singular value decomposition (gsvd). *Ency. of Measu. and Stat.*, pages 907–912, 2007.

[2] J. Bennett and S. Lanning. The Netflix Prize. *KDD Cup and Workshop*, 2007.

[3] R. Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[4] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. *RecSys*, 2010.

[5] X. Ding, B. Liu, and P. S. Yu. A Holistic Lexicon-Based Approach to Opinion Mining. *WSDM*, pages 231–239, 2008.

[6] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. *KDD*, pages 168–177, 2004.

[7] J. A. Konstan. Introduction to recommender systems: Algorithms and Evaluation. *ACM Transactions on Information Systems (TOIS)*, 22(1):1–4, 2004.

[8] Y. Koren, R. Bell, et al. Matrix factorization techniques for recommender systems. *Computer*, 2009.

[9] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. *NIPS*, pages 556–562, 2001.

[10] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. *WWW*, pages 342–351, 2005.

[11] B. Liu and L. Zhang. A Survey of Opinion Mining and Sentiment Analysis. *Jour. Mining Text Data*, pages 415–463, 2012.

[12] J. Liu, M. Chen, J. Chen, et al. Recent advances in personal recommender systems. *International Journal of Information and Systems Sciences*, 5(2):230–247, 2009.

[13] N. Liu, X. Meng, and C. Liu. Wisdom of the Better Few: Cold Start Recommendation via Representative based Rating Elicitation. *RecSys*, pages 37–44, 2011.

[14] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. *Recommender Systems Handbook*, pages 73–105, 2011.

[15] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. *WWW*, pages 347–356, 2011.

[16] J. McAuley and J. Leskovec. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. *RecSys*, pages 165–172, 2013.

[17] T. Nakagawa, K. Inui, and S. Kurohashi. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. *NAACL*, 2010.

[18] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[19] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. *EMNLP*, pages 79–86, 2002.

[20] M. J. Pazzani and D. Billsus. Content-based recommendation systems. *Adaptive Web LNCS*, 2007.

[21] J. Rennie et al. Fast maximum margin matrix factorization for collaborative prediction. *ICML*, 2005.

[22] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. *ICML*, 2008.

[23] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *NIPS*, 2008.

[24] N. Srebro and T. Jaakkola. Weighted low-rank approximations. *ICML*, 2003.

[25] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. *NIPS*, 2005.

[26] X. Su and T. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in AI.*, 2009.

[27] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguastics*, 37(2), 2011.

[28] G. Takacs, I. Pilaszy, B. Nemeth, and D. Tikk. Investigation of various matrix factorization methods for large recommender systems. *ICDM*, 2008.

[29] N. Tintarev and J. Masthoff. A Survey of Explanations in Recommender Systems. *ICDE*, 2007.

[30] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation (LREC)*, 2005.

[31] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *EMNLP*, pages 347–354, 2005.

[32] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. *SIGIR*, pages 83–92, 2014.

[33] Y. Zhang, H. Zhang, M. Zhang, Y. Liu, and S. Ma. Do Users Rate or Review? Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification. *SIGIR*, 2014.

[34] Y. Zhang, M. Zhang, Y. Liu, S. Ma, and S. Feng. Localized Matrix Factorization for Recommendation based on Matrix Block Diagonal Forms. *WWW*, 2013.

[35] Y. Zhang, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Understanding the Sparsity: Augmented Matrix Factorization with Sampled Constraints on Unobservables. *CIKM*, 2014.

[36] K. Zhou, S.-H. Yang, and H. Zha. Functional Matrix Factorizations for Cold-Start Recommendation. *SIGIR*, pages 315–324, 2011.