How Hidden Aspects Can Improve Recommendation?

Conference Paper · November 2014

DOI: 10.1007/978-3-319-13734-6_19

4 authors:

Youssef Meguebli
France Télécom
17 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Mouna Kacimi
Free University of Bozen-Bolzano
45 PUBLICATIONS **492** CITATIONS
SEE PROFILE

Bich-Liên Doan
CentraleSupelec
88 PUBLICATIONS **239** CITATIONS
SEE PROFILE

Fabrice Popineau
Laboratoire de Recherche en Informatique
46 PUBLICATIONS **117** CITATIONS
SEE PROFILE

How Hidden Aspects Can Improve Recommendation?

Youssef Meguebli[1], Mouna Kacimi[2], Bich-liên Doan[1], and Fabrice Popineau[1]

[1] SUPELEC Systems Sciences (E3S), Gif sur Yvette, France
{youssef.meguebli,bich-lien.doan,fabrice.popineau}@supelec.fr
[2] Free University of Bozen-Bolzano, Italy
mouna.kacimi@unibz.it

Abstract. Nowadays, more and more people are using online news platforms as their main source of information about daily life events. Users of such platforms discuss around topics providing new insights and sometimes revealing hidden aspects about topics. The valuable information provided by users needs to be exploited to improve the accuracy of news recommendation and thus keep users always motivated to provide comments. However, exploiting user generated content is very challenging due its noisy nature. In this paper, we address this problem by proposing a novel news recommendation system that (1) enrich the profile of news article with user generated content, (2) deal with noisy contents by proposing a ranking model for users' comments, and (3) propose a diversification model for comments to remove redundancies and provide a wide coverage of topic aspects. The results show that our approach outperforms baseline approaches achieving high accuracy.

Keywords: News recommendation, Opinion mining, Diversification.

1 Introduction

News Media platforms play a crucial role in covering daily life topics ranging from social to political issues. Such platforms often allow users to publish their reactions to the published information and freely express their opinions. 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. By contrast, user generated content follows a bottom up approach where users start discussing some specific issues forming debates around a given topic. Consequently, they reveal hidden topic aspects which are not confined to any predefined plan and thus extend information by continuously bringing new insights. This calls for an effective strategy for news recommendation that would provide users 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 around the topic. For this reason, 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 comments for a more effective recommendation. User generated content is a free source of information which can be subject to a lot of noise. Thus, it is important to select only prominent comments using ranking strategy. Moreover, these comments have to be representative which require the application of diversification techniques to

L.M. Aiello and D. McFarland (Eds.): SocInfo 2014, LNCS 8851, pp. 269–278, 2014.

capture a wide set of aspects. Our proposed approach goes beyond existing techniques [1,20,26,4,26] that employ user generated content for search and recommendation in several ways. First, Ganesan et al., [4] use product reviews and assume that comments 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, comments on news sites usually contain a lot of noise, thus unlike the approach by Yee et al. [26], we do not use all comments to enrich the content of news article but we select only the topk comments. Additionally, we perform diversification on those comments to have a large coverage of new aspects. Our work aims at providing an effective news recommendation to facilitate the access of users to published news stories and more importantly, to motivate readers to comment on the news articles of interests and get involved in discussions with other users. We first propose an unsupervised technique for aspect extraction from user generated content and editorial content. Second, we propose a novel recommendation approach that (1) enriches the content of news articles with user generated content to improve the effectiveness of recommendation, (2) ranks user comments to select only prominent content and filter noise, and (3) proposes a comment diversification model based on authorities, semantic and sentiment diversification. Third, we test our approach on four datasets.

2 Related Work

The emergence of Web 2.0 has led to a rapid growth of user generated content (UGC), such as product, movie, and hotel reviews, and comments on news stories. Due to its richness and insightfulness, user generated content was exploited by several studies [9,19,11,22,22,14,16,23,15,17] for different purposes including blog summarization [9], community detection for predicting the popularity of online content [19], spam detection [11], comments volume prediction [22], comments rating prediction [14], comments ranking [23,15], and identification of political orientation of users [17]. A key point for exploiting user generated content is to extract interesting and useful knowledge from it. Hence, some approaches [25,27] have focused on aspect extraction from annotated data. For instance, Wang et al., [25] identifies the main aspects of reviews by starting from few seed keywords which are fed into a bootstrapping-based algorithm. Most of these approaches are domain-specific, or usually highly dependent on the training data. In this paper, we employ an unsupervised approach to extract hidden aspects of news articles from their related users' comments. Another key point when exploiting user generated content is how to find the most useful or helpful information. To address this issue, several approaches have focused on ranking user reviews [8,12,24,3]. Danescu et al., [3] show, through extensive experiments, that exploiting relationships between reviews can significantly improve ranking quality. Litva et al., [15], propose to use PageRank to rank comments in news sites. In our work, we use this last technique to rank user comments due to its simplicity, domain-independence, and effectiveness. Directly related approaches to our work employ user generated content for search and recommendation [1,20,26,4,26,4]. Shmueli et al., [20] analyze the co-commenting patterns of users for recommending news articles to users who will likely comment them.

The closest works to ours are by Yee et al., [26,4] and Ganesan et al., [26,4] which exploit users' comments to enrich the content of documents. Yee et al., [26,4] prove that the potential of Youtube users' comments 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., [4] use the content of customer 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., [4] 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, comments on news sites usually contain a lot of noise, thus unlike the approach by Yee et al., [26] we do not use all comments to enrich the content of news articles but we select only the topk comments. Additionally, we perform diversification on those comments to have a larger coverage of new aspects.

3 Aspects Extraction

We describe here how aspects are extracted from user comments and news article content. Note that the same extraction method is used for both types of content, with the sole difference that the computation of aspects scores depends either on the corpus of comments or on one of the articles.

3.1 Generation of Candidate Aspects

To extract aspects from the comments of user u_i, we first identify the sentences[1] expressed in all his comments. Then, we rank their contained terms using $tf * idf$ scoring function. In our work, tf represents the term frequency in the set of sentences of user u_i, and idf represents the inverted document frequency in the set of sentences of all users in the platform. 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 platform. 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[2]. We check the redundancy of the generated candidates using Jaccard similarity [18]. 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 of a higher length.

[1] Using OpenNLP `http://opennlp.sourceforge.net/`

[2] In this work we have set k=500.

3.2 Selection of Promising Aspects

Generating n-grams that have high $tf * idf$ scores is not enough to identify the aspects discussed in users' comments 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 *pointwise mutual information* [21] (PMI) of words in n-grams based on its alignment to the narrow comments of each user (or article content). Formally, suppose $m_i = w_1...w_n$ is a generated n-grams. We define the $Score_n$ as follows:

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

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

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

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). 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. We used a modified PMI scoring [4] referred to as pmi' and is defined as:

$$pmi'(w_i, w_j) = \log_2 \frac{p(w_i, w_j) \cdot c(w_i, w_j)}{p(w_i) \cdot p(w_j)} \tag{3}$$

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 and $p(w_i, w_j)$ is the corresponding joint probability. The co-occurrence frequency, $c(w_i, w_j)$ 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 that moderately associated words with high co-occurrences have relatively high scores.

4 Comments Ranking

We adopt the opinion ranking approach proposed by Litva et. al., [15] because of its simplicity, domain-independence, and effectiveness. For each article a_j, we take all its related comments and build a graph where each node is a comment. An edge is created between two comments if their cosine similarity exceeds a given threshold[3]. Once we have the comments graph, we apply the PageRank algorithm to compute a score for each comment. The topk comments are then used to enrich the content of the news article a_j. We recall that the PageRank algorithm models use behavior in a hyperlink graph, where a random surfer visits a web page with a certain probability based on the page's PageRank. The probability that the random surfer clicks on one link is solely

[3] In our implementation we set the threshold to 0.5.

given by the number of links on that page. So, the probability for the random surfer reaching one page is the sum of probabilities for the random surfer following 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. By replacing pages by comments, and hyperlinks by similarity edges, we can directly apply PageRank to our comment graph.

5 Comments Diversification

In this section, we introduce the technique used to diversify comments on news sites which was inspired by the work in [10]. By diversifying comments, we aim to remove redundancies and thus to provide a wide coverage of topic aspects. We are given a set of comments $C = \{c_1, c_2, ..., c_n\}$ where $n \geq 2$. Our goal is to select a subset $L_k \subseteq C$ of comments that is diverse. We assume three main components that define the diversity of a set of comments : *authority*, *semantic diversity*, and *sentiment diversity*. Naturally, before discussing whether a set is diverse or not, it should first contains comments with high authority scores. Note that the *authority* of each comment is given by the PageRank score described in the previous section. To diversify a set of comments, we need to give more preference to dissimilar comments. We assume that two comments are dissimilar if (1) they discuss different 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 comments, where smaller the distance, the more similar the two comments are. The second one is a *sentiment distance* function $s : C \times C \rightarrow R^+$ between comments, where the smaller the distance, the closer in sentiments the two comments are. 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 comments such as the value of f is maximized. In other words, the goal is to find:

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

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. 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[4]. 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 authority scores are computed based on PageRank and the semantic distance is computed based on Jaccard similarity function. As for sentiment distance $s(a, b)$, it equals to 0 when a and b have the same sentiment, 1 otherwise. The sentiment orientation includes *positive*, *negative*, and *neutral* sentiments. The problem of diversifying search results is NP-hard [5,2]. However, there exist a well-known approximation algorithm to solve it [6], which works well in practice [10]. Gollapaudi et al. [6] show that their Max-sum diversification objective can be approached to a facility dispersion problem, known as the MaxSumDispersion problem [7,13]. 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. Thus, we define the distance function $d'(a, b)$ as follows:

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

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 as proposed in [7,13].

6 Experiments

We have crawled four real datasets based on the activities of 645 users on four news sites, namely CNN, The Telegraph, The Independent and Al-Jazeera.[5] The choice of these users was based on two key-properties: the number of users' comments and whether they follow the four news sites or not. More precisely, we start, by selecting the most active users on each news site based on the number of comments posted and then we choose users that have posted comments on the four news sites. 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,

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

[5] `http://www.cnn.com/`, `http://www.telegraph.co.uk/`, `http://www.independent.co.uk/` and `http://www.aljazeera.com/`

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. For each of those users, we have collected the details of his comments in the four news sites 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. Statistics about the number of commented articles and the number of comments for each dataset are shown in Table 1. To evaluate our approach, we have randomly selected 233 users among the most active users in the four news platforms described above. For each user we 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 comments 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.

Table 1. Dataset statistics

	Dataset1 (CNN Seed)		*Dataset2 (Telegraph Seed)*	
	#articles	**#comments**	**#articles**	**#comments**
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

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 we have missing information. 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. We have used two baseline approaches and tested several variations of our proposed technique. We have used the following strategies:(1) **NoEnrich** is the first baseline and its a simple content filtering approach based solely on the content of news articles. (2) **Yee** is the second baseline and its the closest works to ours which exploit all the set of user comments to enrich the content of documents (news articles in our case). (3)

Authority_k where we use our approach to enrich news articles with the topk authoritative comments related to it, selected as described in section 5. In our experiments we have used $k = 5$, $k = 10$, and $k = 20$. (4) **Diversity_k** where we use our approach to enrich news articles with the most diverse topk comments related to it, as described in section 6. In our experiments we have 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. The results of our experiments are shown in table 2. We can clearly see that our approach outperforms the baseline approaches by a significant margin. The improvement goes up to 17% in precision@5 compared to NoEnrich and 21% compared to Yee which is substantial. Having a closer look at the results, we can see that relying only on the content of news articles does not provide good performance. Even worse, when trying to enrich the content by all user comments, the precision decreases. By applying ranking, the precision improves but the gains are small ranging from 1% to 4%. However, when we apply diversification to the top 100 comments, the top5 and top10 diversified comments give the best results. These results meet our expectations since they perfectly reflect the role and the nature of comments in news platforms. Relying only on the content of articles does not perform well because user profiles built from comments focus on some aspects that might be different from the ones provided by the news article. Taking all comments into account is not a good idea either since comments are subject to noise and some of them might even deviate from the topic of interest, and thus this approach had the worst performance. Selecting the topk comments to be included in the article content is a good idea but due to redundancies this method loses its effect especially when k increases, which is the case of *Authority_20*. Finally, diversifying comments 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 comments are explicit in the news article, then their weight is increased, otherwise they are added which increase the chance of more users getting interested in the article. For example, the aspects extracted from the CNN 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 comment ranking and talk for example about *Australian Visa* and *Deportation*. Then, we see that diversification extracts more aspects such as *Australia tax* and *people contracts*.

Table 2. Overall performance of our approach

	P@1	**P@3**	**P@5**	**P@10**	**P@20**
NoEnrich	0.424	0.494	0.481	0.513	0.540
Yee [26]	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**

7 Conclusions

In this paper, we addressed the problem of recommendation in the context of news sites. In particular, we employed different ways to leverage user generated content on articles for refining the list of recommended news stories. Two approaches were proposed: (i) employing only relevant comments using comments ranking strategy, and (ii) using diverse comments. Our study on an extensive set of experiments showed that diverse comments achieve the best results compared to baseline approaches. As future work, we aim at exploring the impact of co-comments patterns. To this end, we plan to extend our model to a hybrid recommender model in which we employ collaborative filtering recommendation techniques.

References

1. Abbar, S., Amer-Yahia, S., Indyk, P., Mahabadi, S.: Real-time recommendation of diverse related articles. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, Republic and Canton of Geneva, Switzerland, pp. 1–12. International World Wide Web Conferences Steering Committee (2013)
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM, pp. 5–14 (2009)
3. Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., Lee, L.: 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 2009, pp. 141–150. ACM, New York (2009)
4. Ganesan, K., Zhai, C.: Opinion-based entity ranking. Inf. Retr. 15(2), 116–150 (2012)
5. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: WWW, pp. 381–390 (2009)
6. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 381–390. ACM, New York (2009)
7. Hassin, R., Rubinstein, S., Tamir, A.: Approximation algorithms for maximum dispersion. Operations Research Letters 21, 133–137 (1997)
8. Hong, Y., Lu, J., Yao, J., Zhu, Q., Zhou, G.: 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, pp. 495–504. ACM, New York (2012)
9. Hu, M., Sun, A., Lim, E.-P.: Comments-oriented document summarization: Understanding documents with readers' feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 291–298. ACM, New York (2008)
10. Kacimi, M., Gamper, J.: Diversifying search results of controversial queries. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 93–98. ACM, New York (2011)
11. Kant, R., Sengamedu, S.H., Kumar, K.S.: Comment spam detection by sequence mining. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 183–192. ACM, New York (2012)
12. Kim, S., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 423–430 (2006)

13. Korte, B., Hausmann, D.: An analysis of the greedy heuristic for independence systems. Annals of Discrete Mathematics 2, 65–74 (1978)
14. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 375–384. ACM, New York (2009)
15. Litvak, M., Matz, L.: Smartnews: Bringing order into comments chaos. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, KDIR, vol. 13 (2013)
16. Meguebli, Y., Kacimi, M., Doan, B.-L., Popineau, F.: Building rich user profiles for personalized news recommendation. In: Proceedings of 2nd International Workshop on News Recommendation and Analytics (2014)
17. Meguebli, Y., Kacimi, M., Doan, B.-L., Popineau, F.: Unsupervised approach for identifying users' political orientations. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 507–512. Springer, Heidelberg (2014)
18. Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic Biology 45(3), 380–385 (1996)
19. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2009, Arlington, Virginia, United States, pp. 452–461. AUAI Press (2009)
20. Shmueli, E., Kagian, A., Koren, Y., Lempel, R.: Care to comment?: Recommendations for commenting on news stories. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 429–438. ACM, New York (2012)
21. Terra, E., Clarke, C.L.A.: 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, NAACL 2003, pp. 165–172. Association for Computational Linguistics, Stroudsburg (2003)
22. Tsagkias, M., Weerkamp, W., de Rijke, M.: Predicting the volume of comments on online news stories. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1765–1768. ACM, New York (2009)
23. Tsagkias, M., Weerkamp, W., de Rijke, M.: News comments:Exploring, modeling, and online prediction. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 191–203. Springer, Heidelberg (2010)
24. Tsur, O., Rappoport, A.: Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In: International AAAI Conference on Weblogs and Social Media (2009)
25. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: A rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2010, pp. 783–792. ACM, New York (2010)
26. Yee, W.G., Yates, A., Liu, S., Frieder, O.: Are web user comments useful for search. In: Proc. LSDS-IR, pp. 63–70 (2009)
27. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie review mining and summarization. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, pp. 43–50. ACM, New York (2006)