

Exploration / Exploitation Trade-Off in Mobile Context-Aware Recommender Systems

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski

Department of Computer Science, Télécom SudParis, UMR CNRS Samovar,
91011 Evry Cedex, France

{Djallel.Bouneffouf, Amel.Bouzeghoub,
Alda.Gancarski}@it-sudparis.eu

Abstract. The contextual bandit problem has been studied in the recommender system community, but without paying much attention to the contextual aspect of the recommendation. We introduce in this paper an algorithm that tackles this problem by modeling the Mobile Context-Aware Recommender Systems (MCRS) as a contextual bandit algorithm and it is based on dynamic exploration/exploitation. Within a deliberately designed offline simulation framework, we conduct extensive evaluations with real online event log data. The experimental results and detailed analysis demonstrate that our algorithm outperforms surveyed algorithms.

Keywords: recommender system, machine learning, exploration/exploitation dilemma, artificial intelligence.

1 Introduction

A considerable amount of research has been done in recommending interesting content for mobile users. Earlier techniques in Mobile Context-Aware Recommender Systems (MCRS) [3, 6, 12, 5, 22] are based solely on the computational behavior of the user to model his interests regarding his surrounding environment like location, time and near people (the user's situation).

The main limitation of such approaches is that they do not take into account the dynamicity of the user's content. This gives rise to another category of recommendation techniques that try to tackle this limitation by using collaborative, content-based or hybrid filtering techniques.

Collaborative filtering, by finding similarities through the users' history, gives an interesting recommendation only if the overlap between users' history is high and the user's content is static[18]. Content-based filtering, identify new documents which match with an existing user's profile, however, the recommended documents are always similar to the documents previously selected by the user [15].

Hybrid approaches have been developed by combining the two latest techniques; so that, the inability of collaborative filtering to recommend new documents is reduced by combining it with content-based filtering [13]. However, the user's content

in mobile undergoes frequent changes. These issues make content-based and collaborative filtering approaches difficult to apply [8].

The bandit algorithm is a well-known solution that has the advantage of following the evolution of the user's content and addresses this problem as a need for balancing exploration/exploitation (exr/exp) tradeoff.

One classical solution to the multi-armed bandit problem is the ϵ -greedy strategy [9]. With the probability $1-\epsilon$, this algorithm chooses the best documents based on current knowledge; and with the probability ϵ , it uniformly chooses any other documents uniformly.

The ϵ parameter controls essentially the exp/exr tradeoff between exploitation and exploration. One drawback of this algorithm is that it is difficult to decide the optimal value in advance. Instead, we extend the ϵ -greedy strategy with an update of the ϵ value by doing an exr/exp-tradeoff using a finite set of ϵ candidates.

The rest of the paper is organized as follows. Section 2 gives the key notions used throughout this paper. Section 3 reviews some related works. Section 4 presents our MCRS model and Section 5 describes the algorithms involved in the proposed approach. The experimental evaluation is illustrated in Section 6. The last section concludes the paper and points out possible directions for future work.

2 Key Notions

In this section, we sketch briefly the key notions that will be of use in the remainder of this paper.

2.1 The User's Model

The user's model is structured as a case based, which is composed of a set of situations with their corresponding user's preferences, denoted $U = \{(S^i; UP^i)\}$, where S^i is the user's situation and UP^i its corresponding user's preferences.

2.2 The User's Preferences

The user's preferences are deduced during the user's navigation activities. A navigation activity expresses the following sequence of events:

- The user's logs in the system and navigates across documents to get the desired information;
- The user expresses his preferences on the visited documents.

We assume that the preference is the information that we extract from the user's system interaction, for example the number of clicks on the visited documents or the time spent on a document.

Let UP be the preferences submitted by a specific user in the system at a given situation. Each document in UP is represented as a single vector $d=(c_1, \dots, c_n)$, where c_i ($i=1, \dots, n$) is the value of a component characterizing the preferences of d . We

consider the following components: the total number of clicks on d , the total time spent reading d and the number of times d was recommended.

2.3 Context

A user's context C is a multi-ontology representation where each ontology corresponds to a context dimension $C=(O_{Location}, O_{Time}, O_{Social})$.

Each dimension models and manages a context information type. We focus on these three dimensions since they cover all needed information. These ontologies are described in [6] and are not developed in this paper.

2.4 Situation

A situation is an instantiation of the user's context. We consider a situation as a triple $S = (O_{Location}.x_i, O_{Time}.x_j, O_{Social}.x_k)$ where x_i , x_j and x_k are ontology concepts or instances. Suppose the following data are sensed from the user's mobile phone: the GPS shows the latitude and longitude of a point "48.8925349, 2.2367939"; the local time is "Mon Oct 3 12:10:00 2011" and the calendar states "meeting with Paul Gerard". The corresponding situation is:

$S=(O_{Location}. "48.89, 2.23", O_{Time}. "Mon_Oct_3_12:10:00_2011", O_{Social}. "Paul_Gerard")$.

To build a more abstracted situation, we interpret the user's behavior from this low-level multimodal sensor data using ontologies reasoning means. For example, from S , we obtain the following situation:

$MeetingAtRestaurant=(O_{Location}.Restaurant, O_{Time}.Work_day, O_{Social}.Financial_client)$.

For simplification reasons, we adopt in the rest of the paper the following notation: $S = (x_i, x_j, x_k)$. The previous example situation became thus:

$MeetingAtRestaurant=(Restaurant, Work_day, Financial_client)$.

Among the set of captured situations, some of them are characterized as high-level critical situations.

3 Related Work

We refer, in the following, recent recommendation techniques that tackle both making dynamic *ext/exp* (bandit algorithm) and considering the user's situation in recommendation.

3.1 Bandit Algorithms Overview (ϵ -greedy)

The (*ext/exp*) tradeoff was firstly studied in reinforcement learning in 1980's, and later flourished in other fields of machine learning [16, 19]. Very frequently used in reinforcement learning to study the (*ext/exp*) tradeoff, the multi-armed bandit problem was originally described by Robbins [16].

The ϵ -greedy is the one of the most used strategy to solve the bandit problem and was first described in [20]. The ϵ -greedy strategy choose a random document with epsilon-frequency (ϵ), and choose otherwise the document with the highest estimated

mean, the estimation is based on the rewards observed thus far. ε must be in the open interval $[0, 1]$ and its choice is left to the user.

The first variant of the ε -greedy strategy is what [9, 2] refer to as the ε -beginning strategy. This strategy makes exploration all at once at the beginning. For a given number $I \in \mathbb{N}$ of iterations, the documents are randomly pulled during the εI first iterations. During the remaining $(1-\varepsilon)I$ iterations, the document of highest estimated mean is pulled.

Another variant of the ε -greedy strategy is what Cesa-Bianchi and Fisher [2] call the ε -decreasing strategy. In this strategy, the document with the highest estimated mean is always pulled except when a random document is pulled instead with an ε_i frequency, where n is the index of the current round. The value of the decreasing ε_i is given by $\varepsilon_i = \{\varepsilon_0 / i\}$ where $\varepsilon_0 \in]0, 1]$.

Besides ε -decreasing, four other strategies are presented in [14]. Those strategies are not described here because the experiments done by [14] seem to show that, with carefully chosen parameters, ε -decreasing is always as good as the other strategies.

Compared to the standard multi-armed bandit problem with a fixed set of possible actions, in MCRS, old documents may expire and new documents may frequently emerge. Therefore it may not be desirable to perform the exploration all at once at the beginning as in [9] or to decrease monotonically the effort on exploration as the decreasing strategy in [2].

Few research works are dedicated to study the contextual bandit problem on Recommender System, where they consider user's behavior as the context of the bandit problem.

In [10], authors extend the ε -greedy strategy by updating the exploration value ε dynamically. At each iteration, they run a sampling procedure to select a new ε from a finite set of candidates. The probabilities associated to the candidates are uniformly initialized and updated with the Exponentiated Gradient (EG) [10]. This updating rule increases the probability of a candidate ε if it leads to a user's click. Compared to both ε -beginning and decreasing strategy, this technique improves the results.

In [21], authors model the recommendation as a contextual bandit problem. They propose an approach in which a learning algorithm selects sequentially documents to serve users based on contextual information about the users and the documents. To maximize the total number of user's clicks, this work proposes the LINUCB algorithm that is computationally efficient.

The authors in [2, 4, 9, 14, 21] describe a smart way to balance exploration and exploitation. However, none of them consider the user's situation during the recommendation.

3.2 Managing the User's Situation

Few research works are dedicated to manage the user's situation on recommendation. In [7, 17] the authors propose a method which consists of building a dynamic user's profile based on time and user's experience. The user's preferences in the user's profile are weighted according to the situation (time, location) and the user's behavior.

To model the evolution on the user's preferences according to his temporal situation in different periods, (like workday or vacations), the weighted association for the concepts in the user's profile is established for every new experience of the user. The

user's activity combined with the user's profile are used together to filter and recommend relevant content.

Another work [12] describes a MCRS operating on three dimensions of context that complement each other to get highly targeted. First, the MCRS analyzes information such as clients' address books to estimate the level of social affinity among the users. Second, it combines social affinity with the spatiotemporal dimensions and the user's history in order to improve the quality of the recommendations.

In [3], the authors present a technique to perform user-based collaborative filtering. Each user's mobile device stores all explicit ratings made by its owner as well as ratings received from other users. Only users in spatiotemporal proximity are able to exchange ratings and they show how this provides a natural filtering based on social contexts.

Each work cited above tries to recommend interesting information to users on contextual situation; however they do not consider the evolution of the user's content. As shown above, none of the mentioned works tackles both problems of exr/exp dynamicity and user's situation consideration in the exr/exp strategy. This is precisely what we intend to do with our approach.

We propose for exr/exp-tradeoff a calibration, which consists in updating the value of ε dynamically. At each iteration, we run the ε -decreasing algorithm to select a new ε from a finite set of candidates. The weight associated to the candidates increases if it leads to a user click.

4 MCRS Model

In our recommender system, the recommendation of documents is modeled as a contextual bandit problem including the user's situation information. Formally, a bandit algorithm proceeds in discrete trials $t = 1 \dots T$. For each trial t , the algorithm performs the following tasks:

4.1 Task 1

Let S^t be the current user's situation, and PS the set of past situations. The system compares S^t with the situations in PS in order to choose the most similar S^p using the following semantic similarity metric:

$$S^p = \arg \max_{S^p \in PS} \left(\sum_j \alpha_j \cdot \text{sim}_j(x_j^t, x_j^c) \right) \quad (1)$$

In Eq.1, sim_j is the similarity metric related to dimension j between two concepts x^t and x^c and α_j the weight associated to dimension j . This similarity depends on how closely x^t and x^c are related in the corresponding ontology (location, time or social). We use the same similarity measure as [24] defined by Eq. 2:

$$\text{sim}_j(x_j^t, x_j^c) = 2 * \frac{\text{deph}(\text{LCS})}{(\text{deph}(x_j^c) + \text{deph}(x_j^t))} \quad (2)$$

In Eq. 2, LCS is the Least Common Subsumer of x_j^t and x_j^c , and deph is the number of nodes in the path from the node to the ontology root.

4.2 Task 2

Let D be the document collection and $D_p \in D$ the set of documents recommended in situation S^p . After retrieving S^p , the system observes the user's behavior when reading each document $d_i \in D_p$. Based on observed rewards, the algorithm chooses the document d_p with the greater reward r_i .

4.3 Task 3

The algorithm improves its arm-selection strategy with the new observation: in situation S' , document d_p obtained a reward r_i .

In the field of document recommendation, when a document is presented to the user and this one selects it by a click, a reward of 1 is incurred; otherwise, the reward is 0. With this definition of reward, the expected reward of a document is precisely its Click Through Rate (CTR). The CTR is the average number of clicks on a recommended document, computed dividing the total number of clicks on it by the number of times it was recommended.

5 Exploration/Exploitation Tradeoff

In this section, we firstly present the ε -greedy algorithm (sub-section 5.1) and our proposed exr/exp-tradeoff algorithm named *decreasing- ε -greedy* (sub-section 5.2).

5.1 The ε -Greedy() Algorithm

The ε -greedy strategy is sketched in Algorithm 1. For a given user's situation, the algorithm recommends a predefined number of documents, specified by parameter N . In this algorithm, $UP = \{d_1, \dots, d_p\}$ is the set of documents corresponding to the current user's preferences; $D = \{d_1, \dots, d_N\}$ is the set of documents to recommend; getCTR (Alg. 1) is the function which estimates the CTR of a given document; Random (Alg. 1) is the function returning a random element from a given set; q is a random value uniformly distributed over $[0, 1]$ which defines the exploration/exploitation tradeoff; ε is the probability of recommending a random exploratory document.

Algorithm 1 The ε -greedy algorithm

Input: ε , UP , $D = \emptyset$, N

Output: D

For $i = 1$ **to** $i = N$ **do**

$q = \text{Random}(\{0, 1\})$

$$d_i = \begin{cases} \text{argmax}_{UP}(\text{getCTR}(d)) & \text{if } q > \varepsilon \\ \text{Random}(UP) & \text{otherwise} \end{cases}$$

$D = D \cup d_i$

Endfor

5.2 The Decreasing- ε -Greedy

To improve exploitation of the ε -greedy algorithm, we propose to extend it with setting out the optimal trade-off value ε . We iteratively update it by the method (*decreasing- ε -greedy* (Alg.2).

First we assume that we have a finite number of candidate values for ε , denoted by $H\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)$. Our goal is to select the optimal ε from $H\varepsilon$.

To this end, we apply the ε -decreasing strategy to propose an ε_i , and then we use a set of weights $w = (w_1, \dots, w_T)$ to keep track of the feedback of each ε_i , w_i is increased if we receive a number of clicks l_i from the user when we use ε_i .

Algorithm 2 decreasing- ε -greedy ()

Input: $H\varepsilon$, N , S^e , UP , n

Output: D

$\varepsilon\varepsilon = \emptyset$, $\tau = 0$, $w_i = 1$, $i = 1, \dots, T$

For $t = 1$ to n **do**

$\tau = 0.01 * t$

$q = \text{Random}([0, 1])$

$$\varepsilon_i = \begin{cases} \text{argmax}_{(i)}(w_i) & \text{if } q \leq \tau \\ \text{Random}(H\varepsilon - \varepsilon\varepsilon) & \text{otherwise} \end{cases}$$

$D = \varepsilon\text{-greedy}(\varepsilon_i, UP, D = \emptyset, N)$;

Receive a click feedback l_i from the user

$w_i = w_{i+} l_i$; $\varepsilon\varepsilon = \varepsilon\varepsilon \cup \varepsilon_i$;

Endfor

In Alg. 2. $\varepsilon\varepsilon$ is the set of ε that have been previously selected, n is the number of iteration of the learning algorithm, i is the identifier of ε and τ is the probability of proposing the $\text{argmax}_i(w_i)$, this parameter starts at low value and iteratively increases until the end of the learning ($\tau = 1$).

6 Experimental Evaluation

In order to evaluate empirically the performance of our approach, and in the absence of a standard evaluation framework, we propose an evaluation framework based on a diary set of study entries. The main objectives of the experimental evaluation are: to evaluate the performance of the proposed algorithms (Alg. 2).

In the following, we describe our experimental datasets and then present and discuss the obtained results.

6.1 Evaluation Framework

We have conducted a diary study with the collaboration of the French software company Nomalys¹. This company provides a history application, which records the time, current location, social and navigation information of its users during their application use.

The diary study has taken 18 months and has generated 178 369 diary situation entries. Each diary situation entry represents the capture, of contextual time, location

¹ Nomalys is a company that provides a graphical application on Smartphones allowing users to access their company's data.

and social information. For each entry, the captured data are replaced with more abstracted information using time, spatial and social ontologies.

From the diary study, we have obtained a total of 2 759 283 entries concerning the user's navigation, expressed with an average of 15.47 entries per situation.

6.2 Experimental Results

In our experiments, we firstly have collected the 3000 situations with an occurrence greater than 100 to be statistically meaningful. Then, we have sampled 10000 documents that have been shown on any of these situations.

The testing step consists of evaluating the algorithms for each testing situation using the average CTR. The average CTR for a particular iteration is the ratio between the total number of clicks and the total number of displays. Then, we calculate the average CTR every 1000 iterations.

We have run the simulation until the number of iterations reaches 10000 and the number of documents (n) returned by the recommender system for each situation is 10.

In the first experiment, we have evaluated the two exr/exp algorithms described in section 5. In addition to a pure exploitation baseline, we have compared our algorithms (decreasing- ϵ -greedy) to the algorithms described in section 3: ϵ -greedy; ϵ -beginning, ϵ -decreasing and EG. In Fig. 1, the horizontal axis is the number of iterations and the vertical axis is the performance metric.

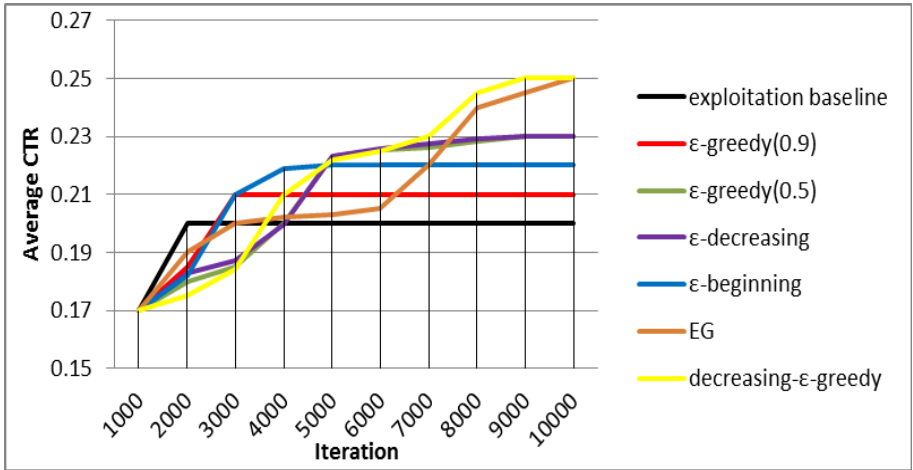


Fig. 1. Average CTR for different exr/exp algorithms

We have parametered the different algorithms as follows: For the ϵ -greedy algorithm, we experiment with two parameter values: 0.5 and 0.9. The ϵ -decreasing, EG, decreasing- ϵ -greedy share the same set of candidates $\{\epsilon_i = 1 - 0.01 * i, i = 1, \dots, 100\}$.

The ϵ -decreasing starts with the largest value and reduces it by 0.01 every 100 iterations until ϵ reaches the smallest value.

Overall exr/exp algorithms have better performance than the baseline. However, for the first 2000 iterations, with pure exploitation the exploitation baseline achieves a faster increase convergence. But in the long run, all exr/exp algorithms improve the average CTR at convergence.

We have several interesting observations regarding the behaviors of the different exr/exp algorithms. The *EG* algorithm has a similar convergence threshold than *decreasing- ϵ -greedy*.

The *decreasing- ϵ -greedy* method effectively learns the optimal ϵ and its convergence rate is close to the best setting. But for *EG* algorithm, it is only after the 6000 iterations that, the algorithm begins to show its advantage. For the *ϵ -decreasing* algorithm, the converged CTR increases as ϵ decreases.

Even after convergence, the *0.9-greedy* and *0.5-greedy* algorithms still give respectively 90% and 50% of the opportunities to documents with low average CTR, which decreases significantly their average CTR.

While the *ϵ -decreasing* algorithm converges to a higher CTR, its overall performance is not as good as *0.5-greedy*. Its CTR drops a lot at the early stage because of more exploration but does not converge faster.

The *decreasing- ϵ -greedy* algorithm has the best convergence rate, and increases CTR by a factor of 1.5 over the baseline and outperforms over all other exr/exp algorithms. The improvement comes from a dynamic tradeoff between exr/exp, controlled by reinforcement learning (ϵ -decreasing).

6.3 Size of Data

To compare the algorithms when data is sparse in our experiments, we reduce data sizes of 30%, 20%, 10%, 5%, and 1%, respectively. To better visualize the comparison results, figure 2 shows algorithms' average CTR graphs with the previous referred data sparseness levels.

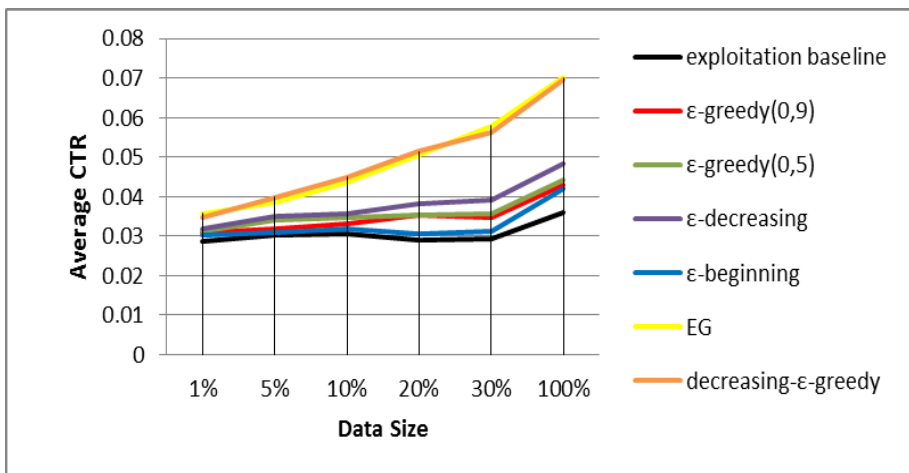


Fig. 2. Average CTR for different data size

Our main observation is that, Although very different from the EG algorithms, the Decreasing- ϵ -greedy tends to very similar results.

Except for exploitation baseline the ϵ -beginning seems to have a rather poor performance, its results are worse than any other strategy independently of the chosen parameters. The reason probably lies in the fact that the ϵ -beginning makes the exploration only at the beginning.

7 Conclusion

In this paper, we study the exploitation and exploration in mobile context-aware recommender systems and propose a novel approach that balances adaptively exr/exp.

We have evaluated our approach according to the proposed evaluation protocol and show that it is effective. In order to evaluate the performance of the proposed algorithm, we compare it with other standard exr/exp strategies.

The experimental results demonstrate that our algorithm performs better on CTR in various configurations. In the future, we plan to evaluate the scalability of the algorithm on-board a mobile device.

References

1. Alexandre, S., Moira Norrie, C., Grossniklaus, M., Signer, B.: Spatio-Temporal Proximity as a Basis for Collaborative Filtering in Mobile Environments. In: CAiSE (2006)
2. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite Time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 235–256 (2002)
3. Bader, R., Neufeld, E., Woerndl, W., Prinz, V.: Context-aware POI recommendations in an automotive scenario using multi-criteria decision making methods. In: Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation, USA, pp. 23–30 (2011)
4. Baltrunas, L., Ludwig, B., Peer, S., Ricci, F.: Context relevance assessment and exploitation in mobile recommender systems. *Personal and Ubiquitous Computing* 16(5) (2011)
5. Bellotti, V., Begole, B., Chi, E.H., Ducheneaut, N., Fang, J., Isaacs, E.: Activity-Based Serendipitous Recommendations. In: Proceedings on the Move, pp. 1157–1166. ACM, USA (2008)
6. Bouneffouf, D., Bouzeghoub, A., Gançarski, A.L.: Following the User's Interests in Mobile Context-Aware recommender systems. In: AINA Workshops, pp. 657–662. IEEE Computer Society, USA (2012)
7. Adomavicius, G., Mobasher, B., Ricci, F., Tuzhilin, A.: Context-Aware Recommender Systems. *AI Magazine* 32(3), 67–80 (2011)
8. Chu, W., Park, S.: Personalized recommendation on dynamic content using predictive bilinear models. In: Proc. of World Wide Web, pp. 691–700. ACM, Spain (2009)
9. Even-Dar, E., Mannor, S., Mansour, Y.: PAC Bounds for Multi-armed Bandit and Markov Decision Processes. In: Kivinen, J., Sloan, R.H. (eds.) COLT 2002. LNCS (LNAI), vol. 2375, pp. 255–270. Springer, Heidelberg (2002)
10. Kivinen, J., Manfred, K.W.: Exponentiated gradient versus gradient descent for linear predictors. In: Information and Computation, vol. 132, pp. 132–163. Academic Press, USA (1997)

11. Langford, J., Zhang, T.: The epoch-greedy algorithm for contextual multi-armed bandits. In: *Advances in Neural Information Processing Systems*, vol. 20, pp. 1096–1103. Curran Associates, Canada (2008)
12. Lakshmish, R., Deepak, P., Ramana, P., Kutila, G., Garg, D., Karthik, V., Shivkumar, K.: A Mobile context-aware, Social Recommender System for Low-End Mobile Device. In: *Mobile Data Management: Systems, Services and Middleware CAESAR*, pp. 338–347. IEEE Computer Society, USA (2009)
13. Lihong, L., Wei, C., Langford, J., Schapire, E.: A Contextual-Bandit Approach to Personalized News Document Recommendation. CoRR, Presented at the Nineteenth International Conference on World Wide Web, Raleigh, pp. 661–670. ACM, USA (2010)
14. Mannor, S., Tsitsiklis, J.: The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Computational Learning Theory* 5, 623–648 (2004)
15. Mladenic, D.: Text-learning and related intelligent agents: A survey. *IEEE Intelligent Agents* 14, 44–54 (1999)
16. Robbins, H.: Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* (1952)
17. Samaras, G., Panayiotou, C.: Personalized portals for the wireless user based on mobile agents. In: *Proc. 2nd Int'l Workshop on Mobile Commerce*, pp. 70–74. ACM, USA (2002)
18. Schafer, J.B., Konstan, J., Riedi, J.: Recommender systems in e-commerce. In: *Proc. of the 1st ACM Conf. on Electronic Commerce*, pp. 158–166. ACM, USA (1999)
19. Sutton, R., Barto, A.: *Reinforcement Learning: An Introduction*. MIT Press (1998)
20. Watkins, C.J.: *Learning from Delayed Rewards*. Ph.D. thesis. Cambridge University (1989)
21. Wei, L., Wang, X., Zhang, R., Cui, Y., Mao, J., Jin, R.: Exploitation and Exploration in a Performance based Contextual Advertising System. In: *KDD*, pp. 133–138. ACM, USA (2010)
22. Woerndl, W., Schulze, F.: Capturing, Analyzing and Utilizing Context-Based Information About User Activities on Smartphones. In: *Activity Context Representation Techniques and Languages: Papers from the 2011 AAAI Workshop* (2011)
23. Wu, Z., Palmer, M.: Verb Semantics and Lexical Selection. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138 (1994)