**K-arm bandit: Recommendation systems**

Most service vendors acquire and maintain a large amount of content in their repository. It is generally difficult to model popularity and temporal changes based solely on content information. Contextual bandit algorithms have become popular for online recommendation systems such as Digg, Yahoo! Buzz, and news recommendation in general. Common practice is to create a simulator which simulates the online environment for the problem at hand and then run an algorithm against this simulator. However, creating simulator itself is often difficult and modelling bias is usually unavoidably introduced. Contextual bandits are a class of one-step reinforcement learning algorithms specifically designed for such treatment personalization problems where we would like to dynamically adjust traffic based on which treatment is working for whom. An important special case of the general contextual bandit problem is the well-known K-armed bandit (i) the arm set remains unchanged and contains K arms for all t, and (ii) the user is the same for all t. Since both the arm set and contexts are constant at every trial, they make no difference to a bandit algorithm, and so we will also refer to this type of bandit as a context-free bandit. Furthermore, in web services we often have access to user information which can be used to infer a user’s interest and to choose news articles that are probably most interesting to her. For example, it is much more likely for a male teenager to be interested in an article about iPod products rather than retirement plans. Therefore, we may “summarize” users and articles by a set of informative features that describe them compactly. By doing so, a bandit algorithm can generalize CTR information from one article/user to another, and learn to choose good articles more quickly, especially for new users and articles. Given asymptotic optimality and the strong regret bound of UCB methods for context-free bandit algorithms, it is tempting to devise similar algorithms for contextual bandit problems. Given some parametric form of payoff function, a number of methods exist to estimate from data the confidence interval of the parameters with which we can compute a UCB of the estimated arm payoff. Such an approach, however, is expensive in general.

Reference:  
<https://arxiv.org/pdf/1003.5956>

<https://arxiv.org/pdf/1003.0146>