Abstract: Recommender systems have been widely adopted by electronic commerce and entertainment industries for individualized prediction and recommendation, which benefit consumers and improve business intelligence. In this article, we propose an innovative method, namely the recommendation engine of multilayers (REM), for tensor recommender systems. The proposed method utilizes the structure of a tensor response to integrate information from multiple modes, and creates an additional layer of nested latent factors to accommodate between-subjects dependency. One major advantage is that the proposed method is able to address the "cold-start" issue in the absence of information from new customers, new products or new contexts. Specifically, it provides more effective recommendations through sub-group information. To achieve scalable computation, we develop a new algorithm for the proposed method, which incorporates a maximum block improvement strategy into the cyclic block-wise-coordinate-descent algorithm. In theory, we investigate both algorithmic properties for global and local convergence, along with the asymptotic consistency of estimated parameters. Finally, the proposed method is applied in simulations and IRI marketing data with 116 million observations of product sales. Numerical studies demonstrate that the proposed method outperforms existing competitors in the literature. This is joint work with Xuan Bi and Xiaotong Shen.