

Improving Performance of Text Recommender Systems

メンドンサ, ドス, サントス, イスラエル

<https://hdl.handle.net/2324/2236256>

出版情報 : 九州大学, 2018, 博士 (工学), 課程博士
バージョン :
権利関係 :



(別紙様式 2)

氏 名 : イスラエル メンドンサ ドス サントス (Israel Mendonça dos Santos)

論 文 名 : Improving Performance of Text Recommender Systems
(テキスト・リコメンダー・システムの性能改善)

区 分 : 甲

論 文 内 容 の 要 旨

Recommender systems are built with the goal of helping common users to navigate through an ocean of data. Through suggestions, those systems try to find items that may be of the user interests, such as what to buy next, which movies to watch, or which documents he may be interested in. When making suggestions, a typical recommender system will employ the user's former data in order to decide what may or may not be of user's interests. However, in some occasions, such data may not be available, and in this case, the system needs to try to guess what to recommend by using mainly its knowledge of the items contained in its database. Recommender systems that use mainly the information of items on the database to make recommendations are called content-based recommender systems (CBRS).

In many CBRS, users need to make a prior evaluation of different items beforehand so that the system can generate accurate recommendations. Without this kind of initial user information, the system suffers from a problem called cold-start, in which the system has not yet acquired enough knowledge of which items different types of users are interested in.

There are a series of problems that can emerge when the recommended items are text documents. Usually, the more structured the data is, the more the system can analyze the parameters of the items, and recommend meaningful items. However, text documents have no structural constructing rules, and hence are hard to model in a single structure. Due to their unclear constitutional rules, when building a document recommendation system, we need to apply techniques to somehow represent those documents in a more structured way. Therefore, the problem of having an accurate model, that conveys all the important information is a non-trivial problem.

The most common ways to represent text documents are via bag of words (BOW), in which we represent a document as the set of its words, and term-frequency inverse document frequency (TF-IDF), in which we try to measure the importance of words from a document against the importance of the same words in a corpus. Yet these representations often fail to capture the full complexity that exists in the natural language text field including polysemous words (words with multiple meaning) and synonyms. Different authors attempted to solve the problem of representing documents in a more meaningful and structured way. One of the most promising

techniques is called Word2vec. The main objective of word2vec is to generate word embedding that carries the semantic meaning of words. Word2vec represent words as a high-dimensional vector.

The main contributions of this thesis are twofold: We tackle the problem of representing items in recommender systems, and also the problem of efficiently retrieving items from the database. Item Representation regards how we structure the items that the system uses for a recommendation. Different types of representation can have different performance results. Item Retrieval regards the techniques used when seeking items for recommendation. Depending on the search technique, it can affect the performance of the recommender system.

In relation to item representation, this thesis presents two new techniques that use the document's word vectors to represent documents: Column Average Representation (CAR), and Meta-Topics Representation (MTR). The Column Average Representation uses the average of each feature of the word vectors to represent the documents. The Meta-Topic Representation creates different meta-topics and uses the distance of words to those meta-topics to represent the documents. In short, both representations make transformations in the document's word vectors to be able to represent them. We build a text recommendation system and study how the performance, in terms of precision and recommendation time, is affected by both representations when compared to the classical document representation approaches. Experimental results show that the proposed method can achieve up to 50% more accurate recommendations when compared to existing document representation methods.

In relation to item retrieval. We tackle the problem of recommending items from a dense database. We evaluate the use of clustering algorithms to reduce recommendation time. We perform a study of how classic clustering algorithms behave when used together with our representation technique and study the pros and cons of using it. Experimental results show that recommender systems can greatly benefit from clustering, in which the algorithms were able to achieve partitions up to 60% similar to the original one, and improved recommendation time up to 5 times.

Our last contribution is an evaluation of the improvements achieved by combining both: our document representation techniques (CAR or MTR) and the classical clustering algorithms. We reformulate the recommender system workflow and manage to have recommendations up to 30 times faster than our baseline technique that uses no clustering.