Random Indexing for Content-based Recommender Systems

Cataldo Musto, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro

Department of Computer Science
University of Bari "Aldo Moro", Italy
{cataldomusto,lops,degemmis,semeraro}@di.uniba.it

Abstract. The use of Vector Space Models (VSM) in the area of Information Retrieval is an established practice, thanks to its very clean and solid formalism that allows us to easily represent objects in a vector space and to perform calculations on them. The goal of this work is to investigate the impact of VSM on Recommender Systems (RS) performance. Specifically, we will introduce two approaches: the first is based on a dimensionality reduction technique called Random Indexing, while the second extends the previous one by integrating a negation operator implemented in the Semantic Vectors open-source package. The results emerged from the experimental evaluation confirmed the predictive accuracy of the model. This work summarizes the results already presented in the RecSys 2010 Doctoral Consortium.

1 Introduction

Recommender Systems (RS) are emerging as one of the most useful tools able to support users to effectively manage the surplus of information they have to deal with. The goal of these systems is to get information about a target user and to exploit them in order to find the most relevant items for her. Although the models underlying Information Filtering (IF) present strong analogies with the Information Retrieval (IR) ones, the impact of IR-based models in the area of IF has not yet been properly investigated. Since 1975, the VSM [1] emerged as one of the most effective approaches in the area of IR, although it suffers from two important problems: the high-dimensionality of the vector space and the inability to manage negative preferences. The main idea behind this work is to investigate the impact of IR-based models on the area of IF by comparing their performance wrt other content-based filtering models. We introduced the definition of "enhanced vectors space models" (eVSM) to describe models able to overcome classical VSM problems. Specifically, we exploited Random Indexing, an incremental technique for dimensionality reduction, and a negation operator based on quantum mechanics to model negative user preferences. This paper is organized as follows: related work are described in Section 2, while Section 3 focuses on the description of both filtering models. Results emerged from the experimental evaluation are described in Section 4. Finally, future directions of this research are sketched in Section 5.

2 Related Work

Many dimensionality reduction approaches such as Latent Semantic Analysis (LSA) have been proposed in order to improve the effectiveness and the scalability of VSM. Recently, effective techniques for dimensionality reduction such as Random Indexing (RI) [2] emerged. The Semantic Vectors (SV) package [3] extends the RI technique by introducing a negation operator based on quantum mechanics.

3 eVSM for Content-based Recommender Systems

In our opinion, a VSM can be defined *enhanced* if the whole vector space is built in an *incremental way* and it is able to catch both the *semantics* of documents and the information coming from *negative evidences*.

In our approach we tackled the first two issues through the introduction of RI, while the last one is managed by exploiting SV. RI is an efficient, scalable and incremental technique for dimensionality reduction. Following this approach, we can represent terms and documents as points in a vector space with a considerable reduction of the features that describe them. RI is based on the so-called *distributional hypothesis*. According to that hypothesis, "words that occur in the same contexts tend to have similar meanings". RI builds the "meaning" of a term (its position in the vector space) in an incremental way, according to the other terms it co-occurs with. Further details about the dimensionality reduction process are contained in [4].

Through RI we can build low-dimensional vector spaces that maintain the original expressivity of the model because, as stated by Johnson and Lindestrauss in their lemma [5], the distance between points in the space is preserved. However, they still inherit a classic problem of VSM: the information coming from negative evidences is not managed. In order to tackle this issue we exploited the Semantic Vectors package[1] that introduces a negation operator based on quantum mechanics. While in SV it is used for retrieval tasks (i.e., to define queries that contain negative terms, such as *A not B*), in our recommendation model it is exploited to infer two vectors, one for positive preferences and one for negative ones. Specifically, the negation operator is used to identify the subspace that contains the items as close as possible to the positive preference vector and as far as possible to the negative one.

To sum up, the main idea behind our filtering models is to build a vector space where both items to be filtered and user profiles are represented as points in this space. Next, calculations based on similarity measures between vectors allow us to obtain the set of the most relevant items for the target user, this is to say, the points in the space that are nearest to her profile.

[1] http://code.google.com/p/semanticvectors/

3.1 Random Indexing (RI) and Weighted RI (W-RI) Models

These approaches are based on the assumption that the information coming from the items a user liked in the past can be a reliable source of information to build accurate user profiles. Therefore, let $d_1, d_2..d_n \in D$ be a set of already rated items, and $r(u, d_i)$ the rating given by the user u to the item d_i. We can define as I_u the set of the items for user u whose rating is over a fixed threshold. Intuitively, the user profile simply consists of the terms occurring in the documents she liked in the past. Formally, let $|I_u|$ be the cardinality of the set I_u and let $\boldsymbol{d}_i$ be the vector space representation of the document d_i, we can define the user profile $\boldsymbol{p}_u$ as follows:

$$\boldsymbol{p}_u = \sum_{i=1}^{|I_u|} \boldsymbol{d}_i \tag{1}$$

The main drawback of the RI method is that the user profile is built without taking into account the ratings provided by the target user for the items she liked. The second model, called *Weighted Random Indexing-based (**W-RI**)*, enriches the previous one by simply associating to each *document vector*, before combining it, a weight equal to the rating provided by the user for it.

3.2 Semantic Vectors (SV) and Weighted SV (W-SV) Models

In SV filtering model two user profile vectors, one for positive preferences and one for negative ones, are inferred. The set of positive items I_u^+ and the positive user profile vector $\boldsymbol{p}_{+u}$ are identical to the set of positive items I_u and the user profile $\boldsymbol{p}_u$ in RI, while the set of negative items, denoted by I_u^-, is defined as the set of the items whose rating is under the threshold. The negative user profile vector, denoted by $\boldsymbol{p}_{-u}$, is built by summing the vector space representations of the items in I_u^-. Given the profile vectors $\boldsymbol{p}_{+u}$ and $\boldsymbol{p}_{-u}$ we can instantiate the vector $\boldsymbol{p}_{+u}$ NOT $\boldsymbol{p}_{-u}$, that is exploited to find the items represented in the vector space that contain as much as possible features that occur in the documents in I_u^+ and as less as possible features from I_u^-. As RI, the SV model has its weighted counterpart, called **W-SV**. This model shares the same idea and the same weighting schema as the W-RI model, with the unique difference that in the negative profile I_u^- the items with a lower rate are given higher weights in order to exclude as much as possible the features disliked by the user.

4 Experimental Evaluation

The goal of the experimental evaluation was to measure the effectiveness of RI and SV models, as well as their weighted variants W-RI and W-SV, in terms of predictive accuracy. Furthermore, we compared the behavior of these novel approaches with a bayesian filtering algorithm described in [6]. The experimental

session has been carried out on a subset of the 100k MovieLens dataset. By exploiting a simple cosine similarity measure we ranked the items wrt the user profile, assuming the nearest ones as the most relevant. The metric used to evaluate the effectiveness of the approaches was the *Average Precision@n*.The results emerged from the experimental evaluation are presented in Table 1. We considered the results of the Bayesian classifier as baseline for our experiments, since this is the method currently implemented in our recommender system. As shown in Table 1, *W-SV* model gained the best results. A thorough description of the experimental session is contained in [4].

Table 1. Results of Average Precision@n on 100k MovieLens dataset

Metric	RI	W-RI	SV	**W-SV**	*Bayes*
AV-P@1	85,93	86,33	85,97	**86,78**	*86,39*
AV-P@5	85,75	86,10	85,99	**86,16**	*85,83*
AV-P@10	85,45	85,76	85,76	**85,85**	*85,75*

5 Conclusions and Future Directions

In this work we introduced the first results emerged from an initial investigation on the impact of eVSM, such as RI-based and SV-based ones, on Content-based Recommender Systems. The main outcome of the experimental evaluation was that this novel filtering model shows an accuracy comparable to the one obtained by other content-based filtering techniques such as Bayesian-based RSs. Furthermore, the introduction of a negation operator, a totally novel aspect for VSM, lets us manage the information about the disliked items and their features. The results obtained with the W-SV model represents a promising starting point for further investigations in this area.

References

1. G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.
2. M. Sahlgren, "An introduction to random indexing," in *Methods and Applications of Semantic Indexing Workshop, TKE 2005*, 2005.
3. D. Widdows, "Orthogonal negation in vector spaces for modelling word-meanings and document retrieval," in *ACL*, 2003, pp. 136–143.
4. C. Musto, "Enhanced vector space models for content-based recommender systems," in *Proceedings of the fourth ACM conference on Recommender systems*, pp. 361–364.
5. W. Johnson and J. Lindenstauss, "Extensions of Lipschitz maps into a Hilbert space," *Contemporary Mathematics*, 1984.
6. P. Lops, M. de Gemmis, G. Semeraro, C. Musto, F. Narducci, and M. Bux, "A semantic content-based recommender system integrating folksonomies for personalized access," in *Web Personalization in Intelligent Environment*, G. Castellano, L. C. Jain, and A. M. Fanelli, Eds. Springer (Berlin), 2009, pp. 27–47.