Enhanced Vector Space Models for Content-based Recommender Systems

Cataldo Musto
Dept. of Computer Science
University of Bari, Italy
cataldomusto@di.uniba.it

ABSTRACT

The use of Vector Space Models (VSM) in the area of Information Retrieval is an established practice within the scientific community. The reason is twofold: first, its very clean and solid formalism allows us to represent objects in a vector space and to perform calculations on them. On the other hand, as proved by many contributions, its simplicity does not hurt the effectiveness of the model. Although Information Retrieval and Information Filtering undoubtedly represent two related research areas, the use of VSM in Information Filtering is much less analzyed.

The goal of this work is to investigate the impact of vector space models in the Information Filtering area. Specifically, I will introduce two approaches: the first one, based on a technique called Random Indexing, reduces the impact of two classical VSM problems, this is to say its high dimensionality and the inability to manage the semantics of documents. The second extends the previous one by integrating a negation operator implemented in the Semantic Vectors[1] open-source package. The results emerged from an experimental evaluation performed on a large dataset and the applicative scenarios opened by these approaches confirmed the effectiveness of the model and induced to investigate more these techniques.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries, Indexing methods, Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

General Terms

Algorithms, Experimentation

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

RecSys2010, September 26–30, 2010, Barcelona, Spain.

Keywords

Content-based Recommender System, Information Filtering, Personalization, Vector Space Models

1. INTRODUCTION

The exponential growth of the available information, speeded up by the recent explosion of Social Web platforms, emphasize the need for systems able to effectively manage this surplus of information helping users in finding what they really need. In this scenario, Information Filtering systems are emerging as one of the most useful tools able to support users in this kind of activity. The goal of these systems, such as recommender systems, is to get information about a target user (what she knows, what she likes, the task to be accomplished, demographical or contextual informations and so on) and to exploit them in order to find the most relevant items for her, ranked according on a relevance criterion.

As underlined by Belkin and Croft [1], the models underlying Information Filtering (IF) present strong analogies with the ones of Information Retrieval (IR). As in Information Retrieval systems a query (describing short-term user needs) is submitted against a set of documents, in the same way in the Information Filtering model this role is played by a component called "user profile", that describes long-term user preferences and is used to filter the space from the irrelevant information for the target user. Despite these similarities, the impact of IR-based models in the area of IF has not yet been properly investigated. In area of the Information Retrieval the vector space model (VSM) emerged for almost three decades as one of the most effective approaches, thanks to its good compromise between expressivity, effectiveness and simplicity. However, VSM suffers from two important problems: the first one, related to the high-dimensionality of the vector space, makes impossible an incremental approach because the whole vector space has to be generated from scratch whenever a new item is added to the repository. Furthermore, Vector Space Models cannot manage the latent semantics and the position of the terms in a document. For example, given a document and a permutation of its terms, their representation in the VSM is absolutely the same, although the conveyed information could be likely different.

The main idea behind this work is to exploit the overlapping between IR and IF research areas with a twofold goal: first, I will evaluate the impact of IR-based models in the area of IF by comparing their performance with respect to other content-based filtering models. Furthermore, I will in-

vestigate models able to overcome classical VSM problems ensuring good efficiency, scalability and the ability of managing the latent semantics of documents in a more effective way. Specifically, in this work I compare two state-of-the-art Vector Space models. The first one, based on Random Indexing, implements a scalable and effective VSM-based approach. The second one tries to overcome the classical VSM problem that arise from the impossibility to manage the evidences about negative preferences by introducing a negation operator based on quantum mechanics.

This paper is organized as follows: first, related work are described in Section 2, while Section 3 focuses on the description of both filtering models. The results emerged from the experimental evaluation are described in Section 4. Finally, future directions of this research are sketched in Section 6.

2. RELATED WORK

Vector Space Model, introduced by Salton et al. [2] in 1975, is considered as one of most effective retrieval models in the IR research community. Some initial investigation about the use of VSM as content-based filtering model [3] has been proposed by Cohen and Hirsh [4] and by Nouali et Blache [5].

In order to improve the effectiveness and the scalability of VSM, Berry et al. [6] pointed out the need to introduce some dimensionality reduction technique. LSA [7] and PLSI [8] are two of the most well-known techniques that implement this step. However, because of their computational complexity, these approaches are difficult to implement for real-world applications. In these scenarios effective techniques for dimensionality reduction such as Random Indexing [9] emerged. Effectiveness of this approach has already been demonstrated in [10] with an application for image and text data. The Semantic Vectors package, introduced by Widdows [11], extends the Random Indexing technique by introducing a negation operator based on Quantum Mechanics [12] . Some initial investigations about the effectiveness of the Semantic Vectors for retrieval tasks are reported in [13].

3. CURRENT STATE OF THE RESEARCH

In this section I will first introduce the main concepts behind the Random Indexing, because in both filtering approaches the vector space is built following this technique. Next, I will describe the Random Indexing-based model (RI) for Information Filtering and its extension through the introduction of a negation operator implemented in the Semantic Vectors (SV) package. The main difference between RI and SV approaches lies in the way they exploit the vector space to build user profiles.

In the next section we could refer to the *items to be filtered* and to the *user profiles* as *documents*. In fact, in a content-based filtering model they are considered synonyms because we assume that items to be filtered are described by means of some textual content. For example, in a movie recommendation scenario we can assume that an item (movie) will be represented its title, its cast, its plot and so on.

3.1 Indexing technique

Random Indexing is an efficient, scalable and incremental technique for dimensionality reduction. Following this approach we can represent terms and documents (in our scenario, *items* and *user profiles*) as points in a vector space with a considerable reduction of the features that describe them. To sum up, through this model we can obtain results comparable to other well-known methods (such as Singular Value Decomposition) but with a tremendous savings of computational resources. This approach is based on the so-called *distributional hypothesis*. According to that hypothesis, "words that occur in the same contexts tend to have similar meanings". Thus, the key concept behind this model is "context". In this case, we can think at the context as the set of words a term co-occurs with. Following the famous Wittgenstein sentence *"meaning is its use"*, Random Indexing 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. Specifically, the approach follows these steps:

1. *A context vector* is assigned for each term. This vector has a fixed dimension and it can contain only values in $\{-1, 0, 1\}$. Values are distributed in a random way but the number of non-zero elements is much smaller.
2. The Vector Space representation of a *term* (denoted by $\vec{t}$) is obtained by summing the context vectors of all the terms it co-occurs with.
3. The Vector Space representation of a *document* (denoted by $\vec{d}$) is obtained by summing the context vectors of all its terms.
4. The Vector Space representation of a *user profile* for user u (denoted by $\vec{p}_u$) is obtained by combining the context vectors of all the documents that user liked in the past. The unique difference between the filtering models proposed in this work is the way previously liked documents are combined.

Following that approach, given a set of documents, we can build a low-dimensional Vector Space that guarantees scalability, effectiveness and a better semantic modeling of the documents since each term is no longer represented in an atomic way, as in the classical keyword-based methods, but its position in the space depends on the terms it co-occurs with. The main idea behind our filtering model is to build a vector space for each user, where both items to be filtered and user profiles are represented as points in this space. Next, calculations based on similarity measures between vectors (such as the classical Cosine Similarity) allow us to efficiently 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.

3.2 Random Indexing-based model

This approach is 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 $d_1..d_n \in D$ be a set of already rated items, and $rate(u, d_i)$ $(i = 1..n)$ the rate given by the user u to the item d_i. We can describe the set of positive items for user u, denoted by I_u, as follows:

$$I_u = \{d \in D | rate(u, d) \geq \beta\} \quad (1)$$

As stated above, the *Random Indexing* is exploited to build the user profile in an incremental way, this is to say by

simply summing all the *document vectors* for each document in I_u. Let N be the cardinality of the set I_u and let $\vec{d_i}$ the vector space description of the document d_i, we can define the user profile $\vec{p}_u$ as follows:

$$\vec{p}_u = \sum_{i=1}^{N} \vec{d_i} \quad (2)$$

That is undoubtedly the simplest Random Indexing-based filtering model we could think at. In the experimental evaluation I will refer to this as RI. The main drawback of this method is that the user profile $\vec{p}_u$ is built without taking into account the rates provided by the target user for the items she liked. In other terms, it is *independent* from the rates provided by the target user (provided that they are above or below the threshold β).

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 rate provided by the user for it. More formally:

$$\vec{p}_u = \sum_{i=1}^{N} \vec{d_i} * rate(u, d_i) \quad (3)$$

In this way the model will increase the weight of the items the user liked more. In the future I will investigate about more complex weighting schemas. My initial goal has been to estabilish whether that simple weighting schema could improve the predictive accuracy of the filtering model.

3.3 Semantic Vectors-based model

Both models described in the previous section inherit a classic problem of Vector Space Models: user profiles are only modeled according to the positive preferences provided by the target user. The evidence about her negative preferences, this is to say the features describing the items whose rate is under the threshold, is not managed in any way. In order to overcome this important limit another model, called Semantic Vectors-based (SV), has been developed. This model is based on the Semantic Vectors open-source package, a set of libraries that implements a Random Indexing approach and extends it by introducing a negation operator based on quantum mechanics. Through this operator queries that contain negative terms, such as *A not B*, can be expressed, with the canonical semantics such that the answer set consists of all and only those documents that contain term A and not contain term B. From a theoretical point of view, this kind of query represents the projection of the vector A on the subspace orthogonal to those generated by the vector B.

The main idea behind SV model is to exploit this operator to represent in the user profile features that describe both positive and negative preferences, as it happens in the classical text classification approaches (e.g. Naïve Bayes, Support Vector Machines and so on). We can think at this model as an extension of the previously described RI model. Unlike RI, in which a single user profile $\vec{p}_u$ is build, in SV filtering model two user profiles, a positive and a negative one, are inferred. The set of positive items I_u^+ and the positive user profile $\vec{p}_{+u}$ are identical to the set of positive items I_u and the user profile $\vec{p}_u$ in RI, while the set of negative items, denoted by I_u^-, is defined as follows:

$$I_u^- = \{d \in D | rate(u, d_i) < \beta\} \quad (4)$$

The negative user profile, denoted by $\vec{p}_{-u}$, is built by summing the vector space representations of the items in I_u^-, following the same formula shown in (2). Thus, given the profiles $\vec{p}_{+u}$ and $\vec{p}_{-u}$ we can submit to the Semantic Vectors engine a query like $\vec{p}_{+u}$ *NOT* $\vec{p}_{-u}$, to find the items that contain as much as possible features that describe the documents in I_u^+ and as less as possible features in I_u^-. As RI, the SV model has its weighted counterpart, called W-SV. This model shares the same idea of the W-RI model and the same weighting schema described in (3), 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 target 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 term of predictive accuracy and effectiveness of the proposed ranking. Furthermore, we compared the behavior of these novel approaches with a bayesian filtering algorithm described in [14].

The experimental session has been carried out on a subset of the 100k MovieLens dataset[2], containing 40,717 ratings provided by 613 different users on 520 movies. Since content-based information were crawled from the English version of Wikipedia, we excluded from the original MovieLens dataset the movies without a Wikipedia entry. User profiles were learned by analyzing the ratings stored in the MovieLens dataset. Each rate was expressed as a numerical vote on a 5-point Likert scale, ranging from 1=strongly dislike to 5=strongly like. All the ratings above 2 were considered as positive, while the ratings under this threshold were considered as negative. The session was organized through a 5-fold cross validation: for each fold and for each user we built a vector space the user profile and the items to be filtered. By exploiting a simple cosine similarity measure we ranked the items, assuming the nearest ones as the most relevant. The metric used to evaluate the effectiveness of the approaches was the *Average Precision@n*, where n was set as 1,3,5,7 and 10. We preferred the Average Precision@n instead of the simple Precision@n because it takes into account also the position of the correctly classified items. 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 recommendation tools. As shown in Table 1, *W-SV* model gained the best results, increasing the Average Precision between 0.1% and 0.4%. Furthermore, Table 1 shows that the use of a weighting schema, even if in a naïve form, increases the precision of the system. In this case we have an average improvement that is higher for Random Indexing models (RI vs. W-RI) with respect to Semantic Vectors ones (SV vs. W-SV). This result suggests that different and more complex weighting schemas should be investigated, mostly for SV-based model. Finally, we can note that the introduction of the negation operator, the most novel feature of this approach, improves the Average

[2]http://www.grouplens.org/node/73

Table 1: Average Precision

Metric	RI	W-RI	SV	**W-SV**	*Bayes*
AV-P @1	85,93	86,33	85,97	**86,78**	*86,39*
AV-1 @3	85,78	85,97	86,19	**86,33**	*85,97*
AV-P @5	85,75	86,10	85,99	**86,16**	*85,83*
AV-P @7	85,61	85,92	85,88	**85,97**	*85,77*
AV-P @10	85,45	85,76	85,76	**85,85**	*85,75*

Precision of the filtering model. The paired comparison RI vs. SV and W-RI vs. W-SV shows an improvement of about 0.4% for top-three items in favor of models that exploit the negation operator.

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this work I introduced the first results emerged from an initial investigation on the impact of enhanced VSM, such as Random Indexing-based and Semantic Vectors-based ones, on Content-based Recommender Systems. The main outcome of the experimental evaluation was that, even in this first prototype and even with naive weighting schemas, the filtering model shows an accuracy comparable to the one obtained by other content-based filtering techniques such as the Bayesian classifier. Furthermore, the introduction of a negation operator, a totally novel aspect for VSM, lets us manage also 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. In the future, I will introduce other weighting schemas and I will compare the results with those obtained by a VSM based on the classical term/document matrix, in order to establish whether the dimensionality reduction someway sacrifices the predictive accuracy. Finally, future directions of the research could be summarized in two points. (1) Given a good filtering model, I will study how the information produced in Social Web applications such as Facebook and LinkedIn could be exploited in order to skip the classical training step of the filtering process and to build accurate user profiles. The information coming from these social networks is useful because it merges the precision of the explicit feedback methods with the low intrusiveness of the implicit ones. (2) The recent phenomenon of Linked Data gives a good cue for researchers in the Information Filtering area, because the representation shift from classical keyword-based user profiles to *Linked Data* user profiles (based on microformats such as RDFa) allows to view user profiles no longer as isolated pieces of information, but as part of more complex structures in which relationships are explicitly coded and can be exploited for recommendation tasks.

6. ACKNOWLEDGMENTS

Thanks to my tutor, prof. Giovanni Semeraro, for the fruitful discussions and the continuous support which made this work possible.

7. REFERENCES

[1] N. Belkin and B. Croft, *Information Filtering and Information Retrieval*. Comm. ACM, vol. 35, no. 12, pp. 29,37, 1992.

[2] 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

[3] D. Billsus and M. J. Pazzani, *User Modeling for Adaptive News Access*. UMUAI, vol. 10, no. 2-3, pp. 147,180, 2000

[4] W. W. Cohen and H. Hirsh, *Joins that generalize: Text classification using WHIRL*. in KDD, 1998, pp.169,173

[5] O. Nouali and P. Blache, *A semantic vector space and features-based approach for automatic information filtering*. Expert Syst. Appl., vol. 26, no. 2, pp. 171,179, 2004

[6] M. W. Berry, Z. Drmac and E. R. Jessup, *Matrices, Vector Spaces and Information Retrieval*. SIAM Review, vol. 41, no. 2, pp. 335,362, 1999.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*. JASIS, vol. 41, no. 6, pp. 391,407, 1990.

[8] T.Hofmann, *Probabilistic latent semantic indexing*. in Proceedings of the 22nd Annual International SIGIR Conference, 1999

[9] M.Sahlgren, *An introduction to random indexing*. in Methods and Applications of Semantic Indexing Workshop, TKE 2005, 2005.

[10] E. Bingham and H. Mannila, *Random projection in dimensionality reduction: applications to image and text data*. in KDD '01. ACM, 2001, pp. 245,250

[11] D.Widdows, *Orthogonal negation in vector spaces for modelling word-meanings and document retrieval*. in ACL, 2003, pp. 136,143.

[12] C. J. van Rijsbergen, *The Geometry of Information Retrieval*. Cambridge, UK: Cambridge University Press, 2004

[13] P. Basile, A. Caputo, and G. Semeraro, *Semantic vectors: an information retrieval scenario*. in Proceedings of the First Italian Information Retrieval Workshop (IIR-2010).

[14] 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.