Explicit Semantic Analysis for Enriching Content-based User Profiles

Fedelucio Narducci, Giovanni Semeraro, Pasquale Lops, Marco de Gemmis

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

Abstract. A content-based recommender system suggests items similar to those previously liked by a user, therefore the recommendation process consists of matching up the features stored in a user profile with those of a content object (item). Usually a content-based user profile stores keywords that are more meaningful for that specific user. Common-sense knowledge could positively enrich that profile and content of items, thus helping to introduce more informative features than simple keywords. The idea of this work is to represent content objects, and consequentially user profiles, in terms of Wikipedia-based concepts.

1 Introduction

Content-based recommender systems analyze a set of documents (textual descriptions associated to items) and construct a profile of interests based on the features extracted from the items previously liked by the user [5]. The profile stores the user interests and then it is exploited in order to decide whether a new item is interesting or not for that specific user. Unfortunately, features extracted from content objects are often insufficient to compute similarity between two items or an item and a content-based profile, and this means that they are not effective for representing user preferences. The idea behind this work is to infuse knowledge in a content-based recommender system by: 1) modeling the unstructured information stored in Wikipedia; 2) exploiting the acquired knowledge in order to represent the content objects and the user profiles in terms of Wikipedia concepts (articles). In order to model the knowledge stored in Wikipedia we applied a technique proposed by Gabrilovich and Markovitch, called Explicit Semantic Analysis (ESA) [4]. By exploiting ESA, we can associate to each keyword in the text the most related Wikipedia-concepts. In this way, the classical Bag of Words (BOW) representation can be augmented with knowledge-based features.

This paper is structured as follows. Section 2 provides a general description of FIRSt, the content-based recommender system that is the starting point of this work; Section 3 focuses on the Knowledge Infusion process. Related work are briefly analyzed in Section 4 before drawing some final conclusions in the last section of the paper.

2 FIRSt: a content-based recommender system

FIRSt (**F**olksonomy-based **I**tem **R**ecommender sy**St**em) is a classic content-based recommender which can exploit static content (textual descriptions of items) and user-generated content (for example tags) in the recommendation step. Users were requested to express ratings for some of the items they liked. They can express their preferences on a Likert scale between a MIN and a MAX score. Since FIRSt is implemented as a text classifier, we need to split the dataset in two classes: *user-likes* and *user-dislikes*. Items whose ratings are greater than or equal to (MIN+MAX)/2 are supposed to be liked by the user and included in the positive training set, while items with lower ratings are included in the negative training set. User profiles are learned as binary text classifiers [6]. FIRSt applies a supervised learning technique for learning a probabilistic model of user tastes from textual descriptions, rated in the training phase by that user. After the learning step FIRSt is able to suggest relevant items by matching terms contained in the content-based profile against those contained in documents to be recommended. Item descriptions are pre-processed: the text is first tokenized, stopwords are eliminated, then for each word, the stem is obtained. After this step a document is represented by a BOW. The algorithm adopted for inferring user profiles is a Naïve Bayes text learning approach, widely used in content-based recommenders. Given a new document/item d_j, the *recommendation step* consists of computing the a-posteriori classification scores $P(c_{like}|d_j)$, used to produce a ranked list of potentially interesting items (belonging to the positive class *user-likes*).

3 Exploiting ESA for Knowledge Infusion in FIRSt

The *Knowledge Infusion* (KI) process can be defined as the procedure of providing a system with external heterogeneous knowledge which allows a knowledge-rich textual description and feature space. The KI process consists of two steps: 1) extracting and modeling relationships between words coming from a knowledge source; 2) reasoning on the induced models in order to generate *new* knowledge, which can be useful for the recommendation step. The first step can be performed by exploiting the Explicit Semantic Analysis method. ESA uses knowledge concepts explicitly defined and manipulated by humans. The knowledge source of ESA is Wikipedia. In order to describe how ESA works, we assume that each article in Wikipedia is considered as a concept. We have $\langle C_1, C_2, C_3, ..., C_n \rangle$ representing the vector of basic concepts, where n is the total number of Wikipedia articles. The idea behind ESA is very simple: given a set of concepts $\{C_1, C_2, C_3, ..., C_n\}$ and a set of associated documents (the Wikipedia articles themselves) $\{d_1, d_2, d_3, ..., d_n\}$ we constructs a sparse matrix T where each of the n columns corresponds to a concept (identified by the title of the Wikipedia article), and each of the rows corresponds to a word that occurs in not less than three documents. So, an entry $T[i, j]$ represents the $TF - IDF$ value of term t_i in document d_j. Several heuristics were applied in order to filter poorly relevant concepts. ESA was already used to augment the BOW with knowledge-based features in the text categorization task [4]. Given a document to be classified, ESA allows to represent it in terms of Wikipedia concepts. Furthermore in [3], the authors demonstrated that it is better to enrich the BOW rather than replace it with the generated concepts. The experimental results demonstrated that the Wikipedia-based classifier is significantly superior to the baseline (pure text), in particular when few training examples are available. This situation is very common in the recommendation scenario in which users usually rate very few items. In this work we are investigating the application of ESA to recommendation task, implementing a knowledge infusion mechanism in FIRSt. In this way, we would improve the effectiveness of FIRSt in those situations in which the Bayesian classifier fails, for example when there is a poor overlap among textual descriptions (documents share a little number of words). In this case the KI process could facilitate the matching between a user profile and the textual descriptions, by increasing the number of shared terms. FIRSt applies a simple criterion to weigh the new generated features. It puts together all the related concepts (the vector extracted by ESA) to each term in the BOW and assigns them a score considering the frequency of the original feature (already in the BOW) and the $TF-IDF$ value that relates it to the new generated concept; the k most related concepts are then selected and added to the original BOW (k is a parameter whose value needs to be tuned). More formally:

$$w_c = \sum_{i=1}^{n} (f_i * TF - IDF_{ic})$$

where w_c is the weight for the generated feature c (Wikipedia concept), f_i is the normalized frequency of the term i (already in the original BOW), $TF-IDF_{ic}$ is the $TF-IDF$ value that relates the term i to the concept c in the ESA matrix, and n is the number of terms in the BOW related to the concept k. For example, given the ESA matrix in table 1 and the $BOW = \{volley(0.2), football(0.8), ...\}$ (the normalized frequency in the brackets), we can assign to the new Wikipedia concept *Sport* the value $w = (0.2 * 0.5) + (0.8 * 0.7)$. This step will be performed for each concept related to *football* and *volley* and all of them will be put together in a new set; after that, the k concepts in the set with the highest score, will be added to the BOW.

Term	**Sport**	**Wikipedia Concept 2**	**...**	**Wikipedia Concept n**
football	0.7	$TF-IDF$	...	$TF-IDF$
volley	0.5	$TF-IDF$	...	$TF-IDF$
...	$TF-IDF$	$TF-IDF$	...	$TF-IDF$
$term_z$	$TF-IDF$	$TF-IDF$	...	$TF-IDF$

Table 1. ESA matrix

4 Related Work

Some works about *knowledge-based* (KB) recommender systems are described in [2]. This kind of recommender systems suggests items based on inferences about needs and preferences of the active user. They exploit *functional* knowledge, so they are able to reason about relationships that link a need with an item. Our approach is different because it does not require knowledge engineering efforts, and because it exploits open source common-sense knowledge for a deeper understanding of the content description of items. As regards the feature generation process, in [1] it is presented an approach to augment Reuters-21578 documents with WordNet sysnsets. However, WordNet has some drawbacks when used as a knowledge base for text classification: a fairly small coverage, limited information about synsets, too many distinct senses associated to common words [4].

5 Ongoing and Future Work

This work is an ongoing project, however we carried out some preliminary experiments. We performed the preprocessing step on Wikipedia pages in order to construct the ESA matrix. We used the English Wikipedia dump released on March 12th, 2010 containing 4,909,224 articles. We exploited the Apache Lucene[1] search engine library to create an index for that dump. Preliminary evaluation sessions carried out on the MovieLens dataset [2] demonstrated the effectiveness of our approach. We observed that in general the knowledge infusion process produces improvements in term of classification accuracy (in particular we obtained an highest *Precision* by exploiting the Wikipedia-based BOW). We also applied some feature selection techniques in order to reduce the noise in the BOW. Actually we are investigating the implementation of different future generation approaches and we are also building the ESA matrix for the Italian language.

References

1. S. Bloehdorn and A. Hotho. Boosting for text classification with semantic features. In *In Proceedings of the MSW 2004 Workshop at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 70–87, 2004.
2. R. Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
3. E. Gabrilovich and S. Markovitch. Feature Generation for Text Categorization Using World Knowledge. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1048–1053. Professional Book Center, 2005.
4. E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Intell. Res. (JAIR)*, 34:443–498, 2009.
5. D. Mladenic. Text-learning and related intelligent agents: a survey. *IEEE Intelligent Systems*, 14(4):44–54, 1999.
6. F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[1] http://lucene.apache.org/

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