Meaning is its use: towards the use of Distributional Semantics for Content-based Recommender Systems

Cataldo Musto

Department of Computer Science
University of Bari Aldo Moro, Italy
cataldo.musto@uniba.it

Abstract. The recent spread of collaborative platforms and social networks made the authoring of content easier and easier. However, due to this uncontrolled phenomenon, the amount of data continuously grows without a proper control in terms of quality and reliability of produced content. This makes the problem of Information Overload much more felt than the past.
As already demonstrated by many successful use cases such as Amazon [1], Netflix [2] and Pandora [3], Recommender Systems (RSs) are the tools that can cope the best with this issue, since they can help sifting this flow of information by providing users with personalized access to textual and multimedia information sources. Even if the most popular recommendation algorithms follow the collaborative approach, content-based recommender systems (CBRS) have proved to be effective in real-world scenarios, since they can face with some limitations of collaborative algorithms such as *scalability* and the *new item problem*.
Generally speaking, techniques for CBRS are based on the assumption that the relevance of an unseen item for a target user is usually predicted by matching the features stored in a user profile (inferred from the items previously considered as relevant) with those describing the new item. From this insight it immediately follows that CBRS pipelines need to model both items and user profiles with richer semantic representations, since the approaches based on simple string matching can not handle all the facets typical of natural languages and suffer from the typical issues of polysemy and synonymy.
Consequently, the research in both natural language processing and computational linguistics area is gaining more and more attention: beside the classical techniques based on the use of ontologies or the exploitation of word sense disambiguation algorithms, **Distributional Models (DM)** are recently emerging . These approaches got their name from distributional semantics and describe a set of techniques originally introduced in computational linguistics and cognitive sciences. These approaches are based on the assumption that the semantics of a term can be inferred by

[1] http://www.amazon.com
[2] http://www.netflix.com
[3] http://www.pandora.com

analyzing its use in large corpus of textual data. Specifically, these techniques rely on the *distributional hypothesis*, which states that *"Words that occur in the same contexts tend to have similar meanings"*.

By following the famous Wittgenstein's sentence (*"Meaning is its use"*) it is possible, as already demonstrated by Rubenstein and Goodenough in the mid-1960s [4], to infer the meaning of a term (such as *leash*) by analyzing the meaning of the other terms it co-occurs with (*dog, animal,* etc.). Similarly, the correlation between different terms (e.g., *leash* and *muzzle*) can be inferred by analyzing how similar are the contexts in which they are used. The use of this methodology provides a clear advantage, since a model to represent terms and documents in a lightweight semantic vector space, called WORD SPACE [1], can be built in a totally unsupervised way according to the *use* of the terms in a corpus of textual data.

The goal of this talk is to show how DM can be effective exploited to provide CBRS with a lightweight semantic representation of both items and use profiles. Specifically, a novel content-based recommendation framework that adopts DM as main building block, called **eVSM (enhanced Vector Space Model)**, is introduced [2]. In this framework the original VSM is extended by means of distributional models as well as a negation operator based on Quantum Logic and an incremental technique for dimensionality reduction, called Random Indexing.

In the experimental sessions the effectiveness of eVSM is evaluated in many offline and online settings, and it emerged that eVSM overcomes several state-of-the-art models in terms of goodness of recommendations and accuracy of the proposed ranking [3]. Specifically, it outperforms in a significant way LSI, the classical VSM and a Bayes text classifier in the task of providing users with recommendations about movies. The general outcomes of the offline evaluations are confirmed in the online ones as well, since eVSM showed its capability of providing good recommendations in two user studies carried out in the area of music recommendation.

References

1. Will Lowe. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 576–581, 2001.
2. Cataldo Musto. Enhanced vector space models for content-based recommender systems. In Xavier Amatriain, Marc Torrens, Paul Resnick, and Markus Zanker, editors, *RecSys*, pages 361–364. ACM, 2010.
3. Cataldo Musto, Giovanni Semeraro, Pasquale Lops, and Marco de Gemmis. Random indexing and negative user preferences for enhancing content-based recommender systems. In Christian Huemer and Thomas Setzer, editors, *EC-Web*, volume 85 of *Lecture Notes in Business Information Processing*, pages 270–281. Springer, 2011.
4. Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, 1965.