

Mining Semantically Indexed Documents for Intelligent User Profiling

Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile

Typically, personalized information recommendation services automatically infer a user profile, a structured model of the user interests, from documents the user already deemed as relevant. Traditional keyword-based approaches are unable to capture the semantics of the user interests. This work proposes a strategy consisting of two steps. The first one is a semantic indexing procedure based on a word sense disambiguation strategy which exploits the WordNet lexical database to select, among all the possible meanings (senses) of a polysemous word, the correct one. In the second step, semantically indexed documents are mined by a naïve Bayes learning algorithm that infer semantic, sense-based user profiles. Two experimental sessions were carried out to compare the performance of keyword-based profiles to that of sense-based profiles. We measured both the classification accuracy and the effectiveness of the ranking imposed by the two different kinds of profile on the documents to be recommended. The main outcome of both experiments is that the classification accuracy is improved without improving the ranking.

1 Introduction

Personalized systems adapt their behavior to individual users by learning their preferences during the interaction in order to construct a *user profile* that can be later exploited in the search process. Traditional keyword-based approaches are primarily driven by a string-matching operation: If a string, or some morphological variant, is found in both the profile and the document, a match is made and the document is considered relevant. String matching suffers from problems of *polysemy*, the presence of multiple meanings for one word, and *synonymy*, multiple words having the same meaning. The result is that, due to synonymy, relevant information can be missed if the profile does not contain the exact keywords in the documents while, due to polysemy, wrong documents could be deemed as relevant. These problems call for alternative methods able to learn more accurate profiles that capture concepts expressing users' interests from relevant documents. These *semantic profiles* will contain references to concepts defined in lexicons or, in a further step, ontologies. This paper describes an approach in which user profiles are obtained by machine learning techniques integrated with a word sense disambiguation (WSD) strategy based on the WordNet lexical database [10, 3]. The paper is organized as follows: After a brief discussion about the main works related to our research, we describe in Section 3 the first step of our semantic profile learning process, that is the WSD strategy we propose to represent documents by using WordNet. Section 4 presents the second step of the process: the naïve bayes text categorization method adopted to build *WordNet-based* user profiles. The method is implemented by our content-based profiling system ITem Recommender (ITR). Two experimental sessions were carried out in order to evaluate the proposed approach. The first experiment was performed on a content-based extension of the EachMovie dataset. The second one was performed on a cor-

pus of papers accepted to the International Semantic Web Conferences, held on 2002 and 2003, and rated by real users according to their preferences. The design and the main results of the experiments are presented in Section 5. Conclusions and future work are discussed in Section 6.

2 Related Work

Our research was mainly inspired the following works. *Syskill & Weibert* [12] learns user profiles as Bayesian classifiers able to recommend web pages, but represents documents by using keywords. *LIBRA* [11] adopts a Bayesian classifier to produce content-based book recommendations by exploiting product descriptions obtained from the Web pages of the Amazon on-line digital store. Documents are represented by using keywords and are subdivided into slots, each one corresponding to a specific section of the document. Like *Syskill & Weibert*, the main limitation of this work is that keywords are used to represent documents. *SitelF* [6] exploits a *sense-based* representation to build a user profile as a semantic network whose nodes represent senses of the words in documents requested by the user. In the modeling phase, *SitelF* considers the synsets (senses in WordNet) in the documents browsed during a user navigation session. Synsets are recognized by Word Domain Disambiguation (WDD), which is a variant of WSD where, for each word in a text, a domain label (Literature, Religion, ...) is chosen instead of a sense label. The WDD algorithm, for each *noun*, proposes the domain label appropriate for the word context. Then, the word synsets associated to the proposed domain are selected and added to the document representation. The system builds the semantic net by including in the nodes the synsets occurring in the browsed collection and by assigning to each node a score that is inversely proportional to its frequency over all the corpus. Arcs between nodes represent the co-