Semantic Web Personalization in a Scientific Congress Scenario

Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile, Anna Lisa Gentile, Giuseppe Fraccalvieri

Dipartimento di Informatica
University of Bari
Via E. Orabona, 4 - 70125 Bari - Italy
E-mail: {semeraro, degemmis, lops, basilepp}@di.uniba.it, gentile.annalisa@gmail.com, giuseppefraccalvieri@yahoo.it

***Abstract*— Suppose you registered to a large scientific congress and you got from the Web site the conference program containing a long list of papers which will be presented. Which presentations do you choose to attend? Usually either you try to guess the most interesting talks from their titles and authors or you are forced to have a quick look at the conference proceedings. A recommender system able to learn your research interests from the latest papers you wrote or read, and use them to provide suggestions, might be of valuable help for you in this scenario. Content-based recommenders analyze documents previously rated by a target user, and build a profile exploited to recommend new interesting documents. One of the main limitations of traditional keyword-based approaches is that they are unable to capture the semantics of the user interests, due to the natural language ambiguity. We developed a semantic recommender system, called ITem Recommender[1], able to disambiguate documents before using them to learn the user profile. The Conference Participant Advisor service relies on the profiles learned by ITem Recommender to build a personalized conference program, in which relevant talks are highlighted according to the participant's interests.**

I. Introduction

Content-based recommenders typically require users to label documents by assigning a relevance score, and automatically infer user profiles exploited to rank suggested documents according to the user preferences. Traditional keyword-based approaches are unable to capture the semantics of the user interests. They are primarily driven by a string-matching operation: If a string, or some morphological variant, is found in both the profile and the document, that document is considered as 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 wrong documents could be deemed relevant due to polysemy. ITem Recommender (ITR) is a *semantic recommender* system able to learn accurate profiles which capture concepts expressing users' interests from relevant documents. These semantic profiles contain references to concepts defined in ontologies. The strategy implemented in ITR consists of two steps: the first is the semantic indexing of documents based on a word sense disambiguation technique that uses the WordNet lexical ontology to select, among all the possible meanings (senses) of a polysemous word, the correct one. In the second step, a naïve Bayes approach learns semantic sense-based user profiles as binary text classifiers (user-likes and user dislikes) from disambiguated documents. The system has been integrated into the *Conference Participant Advisor* service to recommend papers accepted at the "International Semantic Web Conference (ISWC) 2004". Test users provided training documents to ITR by rating papers presented during ISWC 2002 and ISWC 2003 events. After the training step, the system builds participant profiles and sends them the personalized ISWC 2004 programs by email.

[1]Available at http://193.204.187.223:8080/iswc_rebuild/

II. Related Work

Our research was mainly inspired by the following works. Syskill & Webert [1] learns user profiles as Bayesian classifiers able to recommend web pages. It adopts a document representation based on keywords. LIBRA [2] 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 keywords and are subdivided into slots, each one corresponding to a specific section of the document. Like Syskill & Webert, the main limitation of this work is that keywords are used to represent documents. SiteIF [3] exploits a sense-based representation to build the user profile as a semantic network whose nodes represent senses of the words in documents requested by the user. The semantic network is built by assigning each node with a score that is inversely proportional to its frequency over all the corpus, so that the score is higher for less frequent senses, avoiding that very common meanings become too prevailing in the user model. In our approach, a probability distribution of the senses found in the corpus of the documents rated by the user is learned and exploited to infer her profile. OntoSeek [4] is a system designed for content-based information retrieval from online yellow pages and product catalogs, which explored the role of linguistic ontologies in knowledge-retrieval systems. The approach has shown that structured content representations, coupled with linguistic ontologies, can increase both recall

and precision of content-based retrieval systems. By taking into account the lessons learned by the previously cited works, we conceived the ITR system as a text classifier able to deal with a sense-based document representation obtained by exploiting a linguistic ontology, as well as to learn a bayesian profile from documents subdivided into slots. The strategy we propose in order to shift from the classical keyword-based document representation to a sense-based one, is to integrate lexical knowledge in the indexing step of training documents. Several methods have been proposed to accomplish this task. In [5], the authors propose to include WordNet information at the feature level by expanding each word in the training set with all its synonyms available in WordNet in order to avoid a Word Sense Disambiguation (WSD) process. This approach has shown a decrease of effectiveness in the obtained classifier, mostly due to the word ambiguity problem, therefore it suggests that some kind of disambiguation is required. In [6], the authors experiment with three different settings for mapping words to senses: No WSD, most frequent sense as provided by WordNet, WSD based on context. They found positive results on the Reuters 25178[2], the OHSUMED[3] and the FAODOC[4] corpus. None of the previous approaches for embedding WSD in classification has taken into account the fact that WordNet is a hierarchical thesaurus. An interesting feature of ITR is that it adopts a similarity measure that takes into account the hierarchical structure of WordNet.

III. The ITem Recommender System in the Scientific Congress Scenario

The conceptual architecture of ITR is depicted in Figure 1. The *Content Analyzer* allows introducing semantics in the recommendation process by analyzing documents in order to identify relevant concepts representing the content. This process selects, among all the possible meanings (senses) of each polysemous word, the correct one according to the context in which the word appears. In this way, documents are represented using concepts instead of keywords, in an attempt to overcome the problems of the natural language ambiguity. The final outcome of the preprocessing step is a repository of disambiguated documents. This semantic indexing is strongly based on natural language processing techniques, such as Word Sense Disambiguation, and heavily relies on linguistic knowledge stored in the WordNet lexical ontology.

The *Profile Learner* implements a supervised learning technique for inferring a probabilistic model of the interests of a (target) user by learning from disambiguated documents rated according to her interests. This model represents the semantic profile, which includes those concepts that turn out to be most indicative of the user's preferences.

The *Recommender* exploits the user profile to suggest relevant documents, by matching concepts contained in the semantic profile against those contained in the documents to be recommended.

[2]http://about.reuters.com/researchandstandards/corpus/
[3]http://www.ltg.ed.ac.uk/disp/resources/
[4]http://www4.fao.org/faobib/index.html

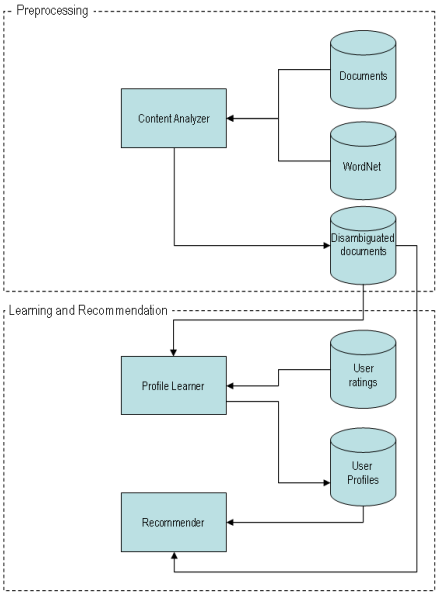

Fig. 1. The conceptual architecture of ITR

In the scientific congress scenario, the participant profile is learned from rated papers in the ISWC 2002 and 2003 paper repository. Then, the profile is matched against all ISWC 2004 accepted papers in order to identify the most relevant ones (which will be highlighted in the participant's personalized conference program).

The remainder of this paper describes the details of the process that leads to build semantic user profiles and the conference programs (properly) personalized for those profiles.

A. Semantic Document Indexing

The problem of learning user profiles can be cast as a binary text categorization task: Each document has to be classified as interesting or not on the ground of the user preferences. The set of categories is $C = \{c_+, c_-\}$, where c_+ is the positive class (*user-likes*) and c_- the negative one (*user-dislikes*). In our approach, a naïve Bayes algorithm learns *sense-based* user profiles as binary text classifiers (*user-likes* and *user-dislikes*) from disambiguated documents obtained by a *semantic indexing* phase performed by the *Content Analyzer*. The idea of learning user profiles from disambiguated documents was successfully introduced in [7]. In this work, we describe the positive effects of exploiting sense-based user profiles to obtain groups of users sharing the same interests in a new hybrid recommendation technique. The core of the *Content Analyzer* is a procedure for assigning senses to words. Here, *sense* is used as a synonym of *meaning*. This task is known as Word Sense Disambiguation and consists in determining which of the senses of an ambiguous word is invoked in a particular use of the word [8].

The goal of a WSD algorithm is to associate the appropriate meaning (or sense) s to a word w_i in document d, by exploiting its *(window of) context* C, that is a set of words that precede

and follow w_i. The sense s is selected from a predefined set of possibilities, usually known as *sense inventory*. In our system, the sense inventory is obtained from WordNet (version 1.7.1)[5]. WordNet was designed to establish connections between four types of Parts of Speech (POS): Noun, verb, adjective, and adverb. The basic building block for WordNet is the SYNSET (SYNonym SET), which represents a specific meaning of a word. The specific meaning of one word under one type of POS is called a sense. Synsets are equivalent to senses, which are structures containing sets of words with synonymous meanings. Each synset has a gloss, a short textual description that defines the concept represented by the synset. For example, the words *night*, *nighttime* and *dark* constitute a single synset that has the following gloss: "the time after sunset and before sunrise while it is dark outside". Synsets are connected through a series of relations: Antonymy (opposites), hyponymy/hypernymy (IS-A), meronymy (PART-OF), etc.

B. The Word Sense Disambiguation Algorithm

JIGSAW is the WSD algorithm implemented by the *Content Analyzer*. It is based on the idea of combining three different strategies to disambiguate nouns, verbs, adjectives and adverbs. In this section we will describe the main idea behind the proposed approach. A more detailed description of the algorithm can be found in [9]. An adaptation of Lesk's dictionary-based WSD algorithm has been used to disambiguate adjectives and adverbs [10], an adaptation of the Resnik algorithm has been used to disambiguate nouns [11], while the algorithm we developed for disambiguating verbs exploits the nouns in the context of the verb and the nouns both in the glosses and in the phrases that WordNet utilizes to describe the usage of the verb. The algorithm disambiguates only words which belong to at least one synset.

The motivation behind our approach is that the performance of the WSD algorithms change in accordance to the POS tag of the word to be disambiguated. JIGSAW algorithm takes as input a document $d = \{w_1, w_2, \ldots, w_h\}$ and will output a list of WordNet synsets $X = \{s_1, s_2, \ldots, s_k\}$ $(k \leq h)$ in which each element s_i is obtained by disambiguating the *target word* w_i based on the information obtained from WordNet about a few immediately surrounding words. We define the *context* C of the target word to be a window of n words to the left and another n words to the right, for a total of $2n$ surrounding words. The algorithm is based on three different procedures for nouns, verbs, adverbs and adjectives, called $JIGSAW_{nouns}$, $JIGSAW_{verbs}$, $JIGSAW_{others}$, respectively. The POS tag of each word to be disambiguated is computed by the HMM-based tagger ACOPOST t3[6].

JIGSAW proceeds in several iterations by using the disambiguation results of the previous iteration to reduce the complexity of the next one. First, JIGSAW performs noun disambiguation by executing the $JIGSAW_{nouns}$ procedure. Then, verbs are disambiguated by $JIGSAW_{verbs}$ by exploiting the words already disambiguated by $JIGSAW_{nouns}$. Finally, the $JIGSAW_{others}$ procedure is executed. The WSD procedure is used to obtain a synset-based vector space representation that we called Bag-Of-Synsets (BOS), described in the next section.

C. Keyword-based and Synset-based Document Representation

In the Bag-Of-Synsets model (BOS), each document is represented by the vector of synsets recognized by the JIGSAW algorithm, rather than a vector of words, as in the classical Bag-Of-Words (BOW) model [12]. Another characteristic of the approach is that each document is represented by a set of *slots*. Each slot is a textual field corresponding to a specific feature of the document, in an attempt to take into account its structure. In our application scenario, in which documents are scientific papers, we selected three slots: *title*, *authors*, *abstract* (Figure 2). The text in each slot is represented according to the BOS model by counting separately the occurrences of a synset in the slots in which it appears.

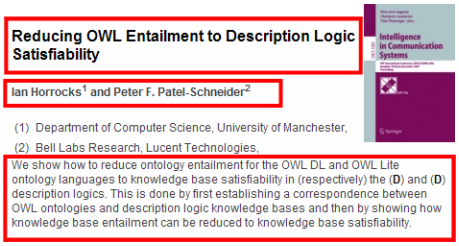

Fig. 2. The description of a paper structured in three slots

```
title: {document: 1; categorization: 1;
        classification: 1}
authors: {sam: 1; scott: 1}

abstract: {categorization: 2; learning: 1;
           classification: 1; AI: 1;
           artificial: 1; intelligence: 1}
```

Fig. 3. An example of document represented using the Bag-of-Words model. Each slot contains terms and their corresponding occurrences in the original text.

An example of BOW-represented document is depicted in Figure 3. The BOS-representation of the same document is presented in Figure 4.

Our hypothesis is that the BOS model helps to obtain profiles able to recommend documents semantically closer to the user interests. The difference with respect to keyword-based profiles is that synset unique identifiers are used instead of words. The next section describes the learning algorithm adopted to build semantic user profiles, starting from the BOS document representation.

[5]http://wordnet.princeton.edu
[6]http://acopost.sourceforge.net/

```
title: {06424377 [text file, document]
        ((computer science) a computer
        file that contains text
        (and possibly formatting
        instructions) using seven-bit
        ASCII characters): 1

        00998694} [categorization,
        categorisation, classification,
        compartmentalization,
        compartmentalisation, assortment]
         -- (the act of distributing
        things into classes or categories
        of the same type): 2}

authors: {}

abstract: {00998694 [categorization,
   categorisation, classification,
   compartmentalization,
   compartmentalisation, assortment]
   (the act of distributing things
   into classes or categories of
   the same type): 3

   06052624 [artificial intelligence,
   AI] (the branch of computer science
   that deal with writing computer
   programs that can solve problems
   creatively;
   "workers in AI hope to imitate
   or duplicate intelligence in
   computers and robots"): 2

   00590335 [learn, larn, acquire]
   (gain knowledge or skills;
   "She learned dancing from her
   sister"; "I learned Sanskrit";
   "Children acquire language at
   an amazing rate"): 1}
```

Fig. 4. An example of document represented using the Bag-of-Synsets model. Each slot contains the synsets associated by JIGSAW to the words in the original text. For the sake of readability, the synset descriptions (that are not included in the actual BOS representation) are also reported.

IV. Learning Semantic User Profiles

The Profile Learner module of ITR uses a Naïve Bayes text categorization algorithm to build profiles as binary classifiers (*user-likes* vs *user-dislikes*). The induced probabilistic model estimates the *a posteriori* probability, $P(c_j|d_i)$, of document d_i belonging to class c_j as follows:

$$P(c_j|d_i) = P(c_j) \prod_{w \in d_i} P(t_k|c_j)^{N(d_i,t_k)} \quad (1)$$

where $N(d_i, t_k)$ is the number of times token t_k appears in document d_i. Since each document is encoded as a vector of BOS (in the WordNet-based approach) or BOW (in the keyword-based approach), one for each slot, equation (1) becomes:

$$P(c_j|d_i) = \frac{P(c_j)}{P(d_i)} \prod_{m=1}^{|S|} \prod_{k=1}^{|b_{im}|} P(t_k|c_j, s_m)^{n_{kim}} \quad (2)$$

where S= $\{s_1, s_2, \ldots, s_{|S|}\}$ is the set of slots, b_{im} is the BOS or the BOW in the slot s_m of d_i, n_{kim} is the number of occurrences of token t_k in b_{im}. When the system is trained on BOW-represented documents, tokens t_k in b_{im} are words, and the induced categorization model relies on word frequencies. Conversely, when training is performed on BOS-represented documents, tokens are synsets and the induced model relies on synset frequencies. To calculate (2), the system has to estimate $P(c_j)$ and $P(t_k|c_j, s_m)$ in the training phase. The documents used to train the system are rated on a discrete scale from 1 to MAX, where MAX is the maximum rating that can be assigned to a document. According to an idea proposed in [2], each training document d_i is labeled with two scores, a "user-likes" score w^i_+ and a "user-dislikes" score w^i_-, obtained from the original rating r:

$$w^i_+ = \frac{r-1}{MAX-1}; \qquad w^i_- = 1 - w^i_+ \quad (3)$$

The scores in (3) are used for weighting the occurrences of tokens in the documents and to estimate their probabilities from the training set TR. The prior probabilities of the classes are computed according to the following equation:

$$\hat{P}(c_j) = \frac{\sum_{i=1}^{|TR|} w^i_j + 1}{|TR| + 2} \quad (4)$$

Witten-Bell smoothing [13] is adopted to compute $P(t_k|c_j, s_m)$, by taking into account that documents are structured into slots and that token occurrences are weighted using scores in equation (3):

$$\hat{P}(t_k|c_j, s_m) = \begin{cases} \frac{N(t_k,c_j,s_m)}{V_{c_j} + \sum_i N(t_i,c_j,s_m)} & \text{if } N(t_k, c_j, s_m) \neq 0 \\ \frac{V_{c_j}}{V_{c_j} + \sum_i N(t_i,c_j,s_m)} \frac{1}{V - V_{c_j}} & \text{otherwise} \end{cases} \quad (5)$$

where $N(t_k, c_j, s_m)$ is the count of the weighted occurrences of token t_k in the training data for class c_j in the slot s_m, V_{c_j} is the total number of unique tokens in class c_j, and V is the total number of unique tokens across all classes. $N(t_k, c_j, s_m)$ is computed as follows:

$$N(t_k, c_j, s_m) = \sum_{i=1}^{|TR|} w^i_j n_{kim} \quad (6)$$

In (6), n_{kim} is the number of occurrences of token t_k in slot s_m of token d_i. The sum of all $N(t_k, c_j, s_m)$ in the denominator of equation (5) denotes the total weighted length of the slot s_m in class c_j. In other words, $\hat{P}(t_k|c_j, s_m)$ is estimated as a ratio between the weighted occurrences of t_k in slot s_m of class c_j and the total weighted length of the slot. The final outcome of the learning process is a probabilistic model used to classify a new document in the class c_+ or c_-. The model can be used to build a personal profile that includes those tokens that turn out to be most indicative of the user's preferences, according to the value of the conditional probabilities in (5). When ITR is trained on documents disambiguated by JIGSAW, the tokens included in the user profiles are WordNet synsets.

V. The Conference Participant Advisor Service

The "Conference Participant Advisor" service is based on ITR and provides useful personalized support for conference participation planning. In the proposed scenario, the semantic profile of each test user registered to the service is exploited to suggest the most interesting talks to be attended at the conference by producing a one-to-one personalized conference program.

The prototype has been realized in the context of the "International Semantic Web Conference 2004", by adding to the conference homepage (a local copy of the official web site) a login/registration form to access the recommendation service (Figure 5). The user registers by providing an email address and can browse the whole document repository or search for papers presented during 2002 and 2003 ISWC events, in order to provide ratings. The search engine used to select the training examples relies on the BOS model in order to allow users to perform a *semantic* search and to reduce the overload in providing the system with appropriate positive and negative examples of documents the user is interested into. Let us suppose that the user now submits the query "categorization" to the paper retrieval system. The search engine analyzes the quey and shows the sense inventory corresponding to the keyword. Among all the possible senses listed, the user can choose one or more of them according to her wishes. In the proposed scenario, the user is interested in papers about "text categorization", which is the task of assigning documents to a list of predefined categories. Therefore, the most appropriate sense for the query is the third one in the sense inventory (Figure 6).

Each retrieved paper can be rated on a discrete rating scale, as shown in Figure 7. Notice that the word matching against the query, highlighted by the search engine, is different from the one in the query issued by the user. This is due to the fact that the two words are in the same synset, thus the system was able to realize a *semantic* matching by exploiting the synonymy relation in WordNet. This semantic search allows for a more accurate selection of training examples: the document retrieved in the aforementioned example would not have been retrieved by using a traditional keyword search.

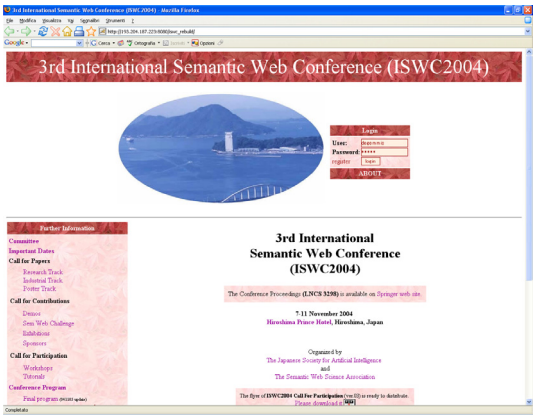

Fig. 5. ISWC 2004 Home page

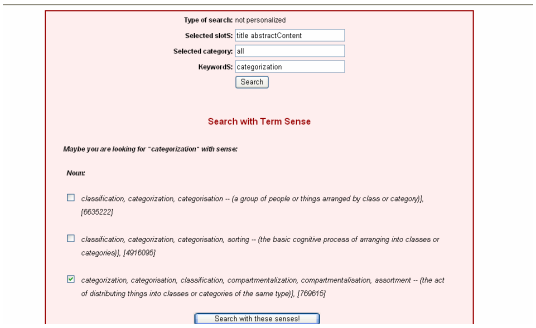

Fig. 6. The user selects the most appropriate sense for the keyword "categorization".

Given a sufficient number of ratings (at present the minimum number of training documents is set at 20), the system learns the semantic profile of the participant by exploiting the algorithm described in section IV. In the profile, *concepts* representing the participant's research interests are stored. ISWC 2004 accepted papers are classified using the learned profile to obtain a personalized list of recommended papers and talks, which is sent by email to the participant. Recommended talks are highlighted in the personalized electronic program (Fig. 8). In addition to the information contained in the program, the user is provided with further information to evaluate the effectiveness of the recommendations. In this version of the service, the abstracts of the recommended papers are also sent to the user in a separate file.

VI. Experimental Evaluation

The goal of the evaluation phase was to compare the performance of keyword-based profiles to that of synset-based profiles. Experiments were carried out on a collection of 100 papers (42 papers accepted to ISWC 2002, 58 papers accepted to ISWC 2003) rated by 11 real users, that we called *ISWC dataset*. Papers are rated on a 5-point scale mapped linearly to the interval [0,1]. Tokenization, stopword elimination and stemming have been applied to index the documents according to the BOW model. Documents have been processed by the

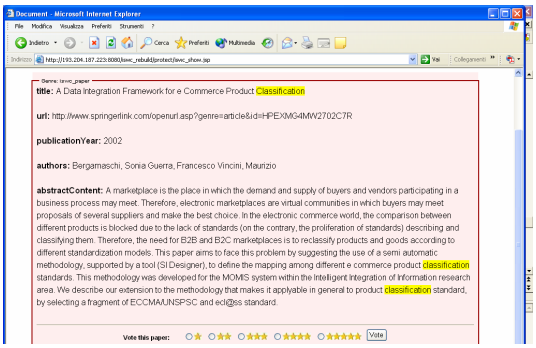

Fig. 7. A paper retrieved by the semantic search provided by ITR and the interface for rating it.

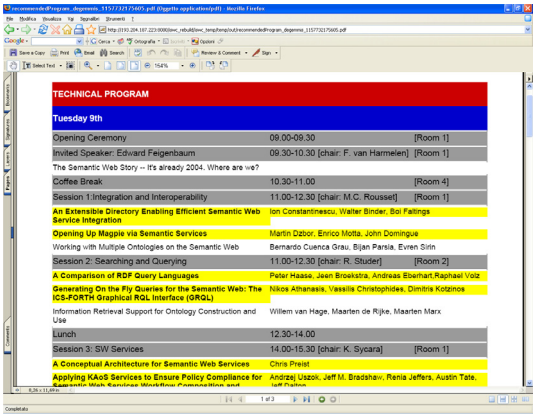

Fig. 8. The personalized program sent to the user

JIGSAW algorithm and indexed according to the BOS model, obtaining a 14% feature reduction ($20,016$ words vs. $18,627$ synsets). This is mainly due to the fact that synonym words are represented by the same synset. Keyword-based profiles were inferred by learning from BOW-represented documents, whilst synset-based profiles were obtained from BOS-represented documents. 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. Classification effectiveness was evaluated by the classical measures *precision*, *recall* [12]. We adopted the Normalized Distance-based Performance Measure (NDPM) [14] to measure the distance between the ranking imposed on papers by the user ratings and the ranking predicted by ITR, that ranks papers according to the a-posteriori probability of the class *likes*. Values range from 0 (agreement) to 1 (disagreement). In the experiments, a paper is considered *relevant* by a user if the rating is greater than 3, while ITR considers an item relevant if $P(c_+|d_i) > 0.5$, computed as in equation (2). We executed one experiment for each user. Each experiment consisted in:

1) selecting the papers and the corresponding ratings given by the user;
2) splitting the selected data into a training set *Tr* and a test set *Ts*;
3) using *Tr* for learning the user profile;
4) evaluating the predictive accuracy of the induced profile on *Ts*, using the aforementioned measures.

The methodology adopted for obtaining *Tr* and *Ts* was the 5-fold cross validation. The results of the comparison between the profiles obtained from documents represented using the two indexing approaches, namely BOW and BOS, are reported in Figure 9.

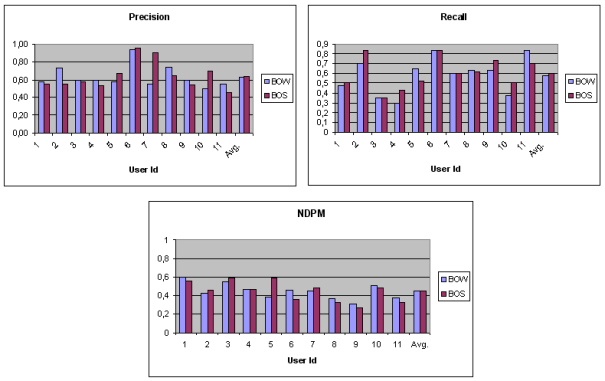

Fig. 9. Performance of the BOW - BOS profiles.

We can notice an improvement both in precision (+1%) and recall (+2%). In particular, precision improves for 4 users out of 11, while a more significant improvement (8 users out of 11) is obtained for recall. The BOS model outperforms the BOW model specifically for users 7 and 10, for whom we observe an increased precision, and in the worst case the same recall. The rating style of these users has been thoroughly analyzed, and we observed that they provided a well balanced number of positive and negative ratings (positive examples not exceeding 60% of the training set). Moreover, they had a very clean rating style, that is, they tend to assign the score 1 to not interesting papers, and the score 5 to interesting ones.

We also observed the effect of the WSD on the training set of these users. We interpreted this effect as follows: If a polysemous word occurs both in positive and negative examples, the system is unlikely to be able to detect the discriminatory power of that feature for the classification because the conditional probabilities of the word are almost the same for the two classes (likes and dislikes). On the other hand, once the system assigned the correct sense to the ambiguous word in each training example in which it occurred, it will be able to distinguish among the different meanings with which that word was differently used in positive and negative examples. Therefore, the occurrences of the *different* synsets assigned to the word will be heavily weighted due to the clean rating style of the users and this should result in more precise probability estimates that positively influenced the precision of the classification.

By the way, the main outcome is that it is difficult to reach a strong improvement both in precision and recall by using the BOS model: we observed a general improvement of both measures only on user 10. It could be noticed from the NDPM

TABLE I

A CASE IN WHICH CLASSIFICATION IS IMPROVED WITHOUT IMPROVING RANKING

Item	R_u	R_{BOS}	R_{BOW}
I1	6 (1)	0.65 (2)	0.65 (2)
I2	5 (2)	0.62 (3)	0.60 (3)
I3	5 (3)	0.75 (1)	0.70 (1)
I4	4 (4)	0.60 (4)	0.45 (5)
I5	4 (5)	0.43 (6)	0.42 (6)
I6	3 (6)	0.55 (5)	0.55 (4)
I7	3 (7)	0.40 (7)	0.40 (7)
I8	2 (8)	0.30 (8)	0.30 (8)
I9	1 (9)	0.25 (9)	0.25 (9)
I10	1 (10)	0.20 (10)	0.20 (10)

values that the relevant/not relevant classification is improved without improving the ranking. This situation can be explained by the example in Table I, in which each column reports the ratings of the items and the corresponding position in the ranking. Let R_u be the ranking imposed by the user *u* on a set of 10 items, let R_{BOS} and R_{BOW} be the ranking computed by both the BOS-generated and the BOW-generated profiles of *u* (ratings ranging between 1 and 6 - classification scores ranging between 0 and 1). An item is considered relevant if the rating is greater than 3 (symmetrically, the score is greater than 0.5). The BOS-generated profile has a better classification accuracy than the BOW-generated one (Recall=4/5, Precision=4/5 vs. Recall=3/5, Precision=3/4). NDPM is almost the same for both profiles because the two rankings are very similar. The difference is that I4 is ranked above I6 in R_{BOS} whilst I6 is ranked above I4 in R_{BOW}. The general conclusion is that the BOS model has improved the classification of items whose scores (and ratings) are close to the relevant/not relevant threshold, thus items for which the classification is highly uncertain.

A Wilcoxon signed ranked test, requiring a significance level $p < 0.05$, has been performed in order to validate these results. Each user is considered as a single trial for the test. The test confirmed that there is a statistically significant difference in favor of the BOS model only as regards recall.

VII. CONCLUSIONS AND FUTURE WORK

We presented a recommender system exploiting a Bayesian learning method to induce *semantic* user profiles from documents in which polysemous words are disambiguated by using a WordNet-based WSD procedure. Our hypothesis is that replacing words with synsets in the indexing phase produces a document representation that can be successfully used by learning algorithms to infer more accurate user profiles. As a consequence, more accurate recommendations are produced by using synset-based profiles. We evaluated this approach by designing a recommendation service for supporting users in the task of planning their attendance to a scientific conference. Experiments were conducted on a collection of papers in order to compare the performance of keyword-based profiles to that of WordNet-based profiles. The main outcome is that the integration of basic linguistic knowledge in the learning process improves the classification of documents whose classification score is close to the likes/dislikes threshold, that is, those items for which the classification is highly uncertain.

As a future work, we plan to integrate domain ontologies in the process of semantic representation and indexing of documents.

ACKNOWLEDGMENTS

This research was partially funded by the European Commission under the 6^{th} Framework Programme IST Integrated Project VIKEF No. 507173, Priority 2.3.1.7 Semantic Based Knowledge Systems. The authors would like to thank Giovanni Giancaspro for his enthusiasm and dedication in designing and developing the Conference Participant Advisor service.

REFERENCES

[1] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," *Machine Learning*, vol. 27, no. 3, pp. 313–331, 1997.

[2] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the 5^{th} ACM Conference on Digital Libraries*. San Antonio, US: ACM Press, New York, US, 2000, pp. 195–204.

[3] B. Magnini and C. Strapparava, "Improving user modelling with content-based techniques," in *Proc. 8th Int. Conf. User Modeling*. Springer, 2001, pp. 74–83.

[4] N. Guarino, C. Masolo, and G. Vetere, "Content-based access to the web," *IEEE Intelligent Systems*, vol. 14, no. 3, pp. 70–80, 1999.

[5] S. Scott and S. Matwin, "Text classification using wordnet hypernyms," in *COLING-ACL Workshop on usage of WordNet in NLP Systems*, 1998, pp. 45–51.

[6] S. Bloedhorn and A. Hotho, "Boosting for text classification with semantic features," in *Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Mining for and from the Semantic Web Workshop*, 2004, pp. 70–87.

[7] M. Degemmis, P. Lops, and G. Semeraro, "A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation," *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (to appear)*, 2007.

[8] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, US: The MIT Press, 1999, ch. 16: Text Categorization, pp. 575–608.

[9] G. Semeraro, M. Degemmis, P. Lops, and P. Basile, "Combining learning and word sense disambiguation for intelligent user profiling," in *Twentieth International Joint Conference on Artificial Intelligence, January 6-12, 2007, Hyderabad, India (to appear)*, 2007.

[10] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. London, UK: Springer-Verlag, 2002, pp. 136–145.

[11] P. Resnik, "Disambiguating noun groupings with respect to WordNet senses," in *Proceedings of the Third Workshop on Very Large Corpora*. Association for Computational Linguistics, 1995, pp. 54–68.

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, 2002.

[13] I. Witten and T. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, vol. 37, no. 4, 1991.

[14] Y. Y. Yao, "Measuring retrieval effectiveness based on user preference of documents," *Journal of the American Society for Information Science*, vol. 46, no. 2, pp. 133–145, 1995.