MARS: a MultilAnguage Recommender System

ABSTRACT

The exponential growth of the Web is the most influential factor that contributes to the increasing importance of cross-lingual text retrieval and filtering systems. Indeed, relevant information exists in different languages, thus users need to find documents in languages different from the one they are formulated in. In this context, an emerging requirement is to sift through the increasing flood of multilingual text: this poses a renewed challenge for designing effective multilingual Information Filtering systems.

In this paper, we propose a language-independent content-based recommender system, called MARS (MultilAnguage Recommender System), that builds cross-language user profiles, by shifting the traditional text representation based on keywords, to a more complex language-independent representation based on word meanings. As a consequence, the recommender system is able to suggest items represented in a language different from the one used in the content-based user profile.

Experiments conducted in a movie recommendation scenario show the effectiveness of the approach.

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

Keywords

Cross-language Recommender System, Content-based Recommender System, Word Sense Disambiguation, MultiWord-Net

1. INTRODUCTION

Nowadays the amount of information we have to deal with is usually greater than the amount of information we can process in an effective way. For this reason, user modeling and personalized information access are becoming essential to propose only (or firstly) the information that appear relevant or somehow related to the informative need of the target user.

Information Filtering (IF) systems are rapidly emerging in this context since they are helpful for carrying out this task in an effective way. These systems adapt their behavior to individual users by learning their preferences and storing them in a user profile. Filtering algorithms, exploiting the information stored in user profiles, perform a progressive removal of non-relevant content according to information about user interests, preferences or specific needs.

Specifically, the content-based filtering approach [1] analyzes a set of documents (usually textual descriptions of items previously rated as relevant by an individual user) and builds a model or profile of user interests based on the features (usually keywords) that describe the target objects. The profile is then exploited to recommend new relevant items.

If the profile accurately reflects user preferences, the information access process could be effective, since the profile could be used to filter search results, by deciding whether a user is interested in a specific item/document or not and, in the negative case, preventing it from being displayed. On the other side, these approaches have to deal with at least two kinds of problems: firstly, traditional keyword-based profiles are unable to capture the semantics of user interests because they are primarily driven by string matching operations. If a string, or some morphological variant of it, is found in both the profile and the document, a match is made and the document is considered as relevant. However, string matching suffers from problems of polysemy, the presence of multiple meanings for one word, and synonymy, multiple words with 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.
Another relevant problem related to string matching approaches is the strict connection with the user language: an English user, for example, frequently interacts with information written in English, so her (keyword-based) profile of interests mainly contains English terms. In order to receive suggestions of items whose textual description is in a different language, she must explicitly express her preferences on items in that specific language, as well. This means that the information already stored in the user profile cannot be exploited to provide suggestions for items whose description is provided in other languages, although they share some common features (e.g., an Italian and an English movie might share the same features but their plots could be written in two different languages).

On the other hand, collaborative filtering techniques are inherently multilingual, because they do not rely on the content of items, but they just exploit the users’ rating style (set of ratings given by users on the same items) for providing recommendations. The problem with this approach is that the similarity between users is only computable if they have common rated items. By the way, the number of ratings obtained from users is usually very small compared to the number of ratings that must be predicted. This problem is called sparsity, and has a negative effect on the predictions, because it affects the selection of neighbors: it is likely that the overlap of rated items between two users will be minimal in many cases. This problem is even more evident in the context of a multilingual environment: it is more likely that most users rated items whose description is available in their mother language.

The main idea presented in this paper is to exploit MultiWordNet [2] as a bridge between different languages in order to construct language-independent user profiles. MultiWordNet associates a unique identifier to each possible sense (meaning) of a word, regardless the original language. We describe a process for building user profiles based on MultiWordNet senses, which are exploited to provide cross-language recommendations. An interesting advantage of the proposed approach is related to the long tail phenomenon[3]. In [4] the authors demonstrated overwhelming evidence that the vast majority of users are a little bit eccentric, consuming niche products at least some times. These niche products are the so-called tail products in the long tail phenomenon. Tail products represent about 25-30% of sales in many cases. This problem is even more evident in the context of a multilingual environment: it is more likely that most users rated items whose description is available in their mother language.

The approach is based on Explicit Semantic Analysis (ESA) for news filtering. The approach proposed in [10] exploits Latent Semantic Indexing for cross-language retrieval. The use of Singular Value Decomposition was investigated by Furnas et al. [11], who create a multidimensional indexing space for a parallel corpus of English documents and their French translations. Ballesteros et al. [12] underlined the importance of phrasal translation in cross-language retrieval and explored the role of phrases in query expansion via local context analysis and local feedback. Soergel [13] focused his attention on the use of thesauri and ontologies as knowledge bases in cross-language retrieval. In his work he provided a general introduction to thesaurus functions, structure, and construction with particular attention to the problems of multi-lingual thesauri. He proposed the creation of environments for distributed collaborative knowledge base development as a way to make large-scale knowledge-based systems more affordable.

The state of the art in the area of cross-language Information Retrieval is undoubtedly richer, and can certainly help in designing effective cross-language Information Filtering systems. Oard [7] gives a good overview of the approaches for cross-language retrieval. Moreover, Salton [8] showed that cross-language retrieval was nearly as effective as mono-lingual retrieval, when carefully constructed thesauri are used. Anyway, this study seems to be unrealistic, since for current standards it is impracticable to manually index large datasets. In [9] the authors propose to build a model of the user’s interests based on word senses rather than on simply words. The approach relies on MultiWordNet to perform Word Domain Disambiguation and to create synset-based multilingual user profiles shown effective for news filtering. The approach proposed in [10] exploits Latent Semantic Indexing for cross-language retrieval. The use of Singular Value Decomposition was investigated by Furnas et al. [11], who create a multidimensional indexing space for a parallel corpus of English documents and their French translations. Ballesteros et al. [12] underlined the importance of phrasal translation in cross-language retrieval and explored the role of phrases in query expansion via local context analysis and local feedback. Soergel [13] focused his attention on the use of thesauri and ontologies as knowledge bases in cross-language retrieval. In his work he provided a general introduction to thesaurus functions, structure, and construction with particular attention to the problems of multi-lingual thesauri. He proposed the creation of environments for distributed collaborative knowledge base development as a way to make large-scale knowledge-based systems more affordable.

The most recent approaches to Cross-Language Retrieval mainly rely on the use of large corpora like Wikipedia. Potthast et al. [14] introduce CL-ESA, a new multilingual retrieval model for the analysis of cross-language similarity. The approach is based on Explicit Semantic Analysis (ESA)
3. GENERAL ARCHITECTURE OF MARS

MARS (MultiLanguage Recommender System) is a system capable of generating recommendations, provided that descriptions of items are available in textual form. Item properties are represented in the form of textual slots, this is to say, a set of logical components in which the whole textual document can be split. For example, a movie can be described by slots title, genre, actors, summary. Figure 1 depicts the main components of the MARS general architecture: the Content Analyzer, the Profile Learner, and the Recommender.

The Content Analyzer is the main module involved in designing a language-independent content-based recommender system. It 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 occurs. In this way, documents are represented using concepts instead of keywords, in an attempt to overcome problems due to natural language ambiguity, and to the diversity of languages. The final outcome of the pre-processing step is a repository of disambiguated documents. This semantic indexing is strongly based on natural language processing techniques, such as Word Sense Disambiguation [19], and heavily relies on linguistic knowledge stored in lexical ontologies. The Content Analyzer relies on the MultiWordNet lexical ontology [2]. The generation of the cross-language user profile is performed by the Profile Learner, which infers the profile as a binary text classifier. The Profile Learner implements a supervised learning technique to infer a probabilistic model of user interests from disambiguated documents, rated in the training phase by that user.

Finally the Recommender exploits the cross-language user profile to suggest relevant items by matching concepts contained in the semantic profile against those contained in documents to be recommended (previously disambiguated). The user might receive recommendations in her own mother tongue, or in languages she knows. This is a decision of the specific application in which the recommender is integrated.

4. BUILDING LANGUAGE-INDEPENDENT DOCUMENTS

Semantic indexing of documents is performed by the Content Analyzer, which relies on META (Multi Language Text Analyzer) [20], a tool able to deal with documents in English and Italian. The goal of the semantic indexing step is to obtain a concept-based document representation. To this purpose, the text is first tokenized, then for each word possible lemmas (as well as their morpho-syntactic features) are collected. Part of speech (POS) ambiguities are solved before assigning the proper sense (concept) to each word. This step requires the identification of a repository for word senses and the design of an automated procedure for defining word-concept associations.

As regards the first issue, in this work, the semantic indexing module exploits MultiWordNet as sense-repository. MultiWordNet is a multilingual lexical database that supports the following languages: English, Italian, Spanish, Portuguese, Hebrew, Romanian and Latin. Similarly to WordNet, the basic building block for MultiWordNet is the synset (SYNonym SET), a structure containing sets of words with synonymous meanings, which represents a specific meaning of a word. Some words have several different meanings, and some meanings can be expressed by several different word forms. Polysemy and synonymy can be viewed as complementary aspects of this mapping. In addition, MultiWordNet is organized by semantic relations between meanings, and since meanings are represented by synsets, semantic relations are pointers between synsets.

In MultiWordNet the Italian WordNet is strictly aligned with English WordNet 1.6 [21]. Italian synsets are attested in correspondence with English WordNet synsets, whenever possible, and semantic relations are imported and preserved.
Table 1: Mapping between an English and an Italian synset

<table>
<thead>
<tr>
<th>Language</th>
<th>Synset</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>world, human</td>
<td>all of the inhabitants of the earth</td>
</tr>
<tr>
<td></td>
<td>race, humanity, mankind, human beings, humans, mankind, man</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>mondo, umanità, uomo, genere umano, terra</td>
<td>insieme degli abitanti della terra, il complesso di tutti gli esseri umani</td>
</tr>
</tbody>
</table>

from the corresponding English synsets. Thus, when a relation occurs between two English WordNet synsets, the same relation occurs between the corresponding Italian synsets. However, some lexical idiosyncrasies between languages may exist, such as lexical gaps and denotation differences, but the multilingual hierarchy implemented in MultiWordNet is able to overcome these problems [2].

An example of mapping between an English and an Italian synset is reported in Table 1, in which the (synonymous) word forms used to define a specific meaning are reported, together with the gloss, that is the description of that specific meaning using free text.

As regards the issue for performing word-concept association, META implements a knowledge-based WSD algorithm, called JIGSAW [22], that takes as input a document encoded as a list of words in order of their appearance, and returns a list of \( k \) MultiWordNet synsets (\( k \leq h \)), in which each synset is obtained by disambiguating the target word \( w \). The goal of JIGSAW consists in assigning a word \( w \) occurring in a document \( d \) with its appropriate meaning or sense \( s \), by exploiting the context \( C \) in which \( w \) is found. The context \( C \) for \( w \) is defined as a set of words that precede and follow \( w \). The sense \( s \) is selected from a predefined set of possibilities, usually known as sense inventory. In the proposed algorithm, the sense inventory is obtained from MultiWordNet.

JIGSAW is based on the idea of combining three different strategies to disambiguate nouns, verbs, adjectives and adverbs. The main motivation behind our approach is that the effectiveness of a WSD algorithm is strongly influenced by the POS tag of the target word. An adaptation of Lesk dictionary-based WSD algorithm has been used to disambiguate adjectives and adverbs [23], an adaptation of the Resnik algorithm has been used to disambiguate nouns [24], while the algorithm we developed for disambiguating verbs exploits the nouns in the context of the verb as well as the nouns both in the glosses and in the phrases that MultiWordNet utilizes to describe the usage of a verb.

The complete description of the adopted WSD strategy is already published in [25, 22].

4.1 Document Representation

The WSD procedure implemented in the Content Analyzer allows to obtain a synset-based vector space representation, called bag-of-synsets (BOS), that is an extension of the classical bag-of-words (BOW) model. In the BOS model, a synset vector, rather than a word vector, corresponds to a document. MARS is able to suggest potentially relevant items to users, as long as an item having \( m \) properties can be represented in form of \( M \) textual slots. The text in each slot is represented by the BOS model by counting separately the occurrences of a synset in the slots in which it occurs. More formally, let \( s \) be the index of the slot, the document \( d \) is reduced to \( M \) bags of synsets, one for each slot:

\[
d^s = \langle t^s_1, t^s_2, \ldots, t^s_n \rangle \quad s = 1 \ldots M
\]

where \( t^s_i \) is the \( i \)th synset in slot \( s \) of the document, and \( n_s \) is the total number of synsets in slot \( s \) of document \( d \). For each bag of synsets \( d^s \), the corresponding synset-frequency vector is computed:

\[
f^s = \langle w^s_1, w^s_2, \ldots, w^s_n \rangle \quad s = 1 \ldots M
\]

\( w^s_i \) is the weight of the synset \( t^s_i \) in slot \( s \) of the document, and can be computed in different ways: it can be the frequency of synset \( t^s_i \) in \( s \) or a more complex feature weighting score. Note that the adoption of slots does not jeopardize the generality of the approach since the case of documents not structured into slots corresponds to have just a single slot in our document representation strategy. By invoking META on a text \( t \), we get \( \text{META}(t) = (\vec{x}, \vec{y}) \), where \( \vec{x} \) is the BOS containing the synsets obtained by applying JIGSAW on \( t \), and \( \vec{y} \) is the corresponding synset-frequency vector.

Figure 2 provides an example of representation for the movie by Stanley Kubrick titled Clockwork Orange, corresponding to the Italian translation Arancia Meccanica. The textual description of the plot is provided both for the Italian and English version. The Content Analyzer produces the BOS containing the concepts extracted from the plot. The MultiWordNet-based document representation creates
a bridge between the two languages: in a classical keyword-based approach the two plots would share only the term Beethoven, while the adoption of the synset-based approach would allow a greater overlapping (six shared synsets).

The following section describes how BOS-indexed documents in different languages are used in a content-based information filtering scenario for learning accurate language-independent user profiles.

5. PROFILE LEARNER

The generation of the cross-language user profile is performed by the Profile Learner, which infers the profile as a binary text classifier [26] since each document has to be classified as interesting or not with respect to the user preferences. Therefore, the set of categories is restricted to $c_+$, the positive class (user-likes), and $c_-$ the negative one (user-dislikes).

The induced probabilistic model is used to estimate the a posteriori probability, $P(c|d_j)$, of document $d_j$ belonging to class $c$. The algorithm adopted for inferring user profiles is a Naive Bayes text learning approach, widely used in content-based recommenders, which is not presented here because already described in [27]. What we would like to point out here is that the final outcome of the learning process is a probabilistic model used to classify a new document, written in any language, in the class $c_+$ or $c_-$. Given a new document $d_j$, the model computes the a-posteriori classification scores $P(c_+,|d_j)$ and $P(c_-|d_j)$ by using probabilities of synsets contained in the user profile and estimated in the training phase by exploiting both ratings provided by users and synset frequencies. Each user provided ratings on items using a 5-point Likert scale ranging from 1 (MIN - strongly dislikes) to 5 (MAX - strongly likes). Items whose ratings are greater than $(\text{MIN}+\text{MAX})/2$ 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. In this case, the threshold value is set as 3, so all the items whose ratings were equals to 4 or 5 have been labelled as positive.

The profile contains the user identifier and the a-priori probabilities of liking or disliking an item ($P(c_+), P(c_-)$), apart from its content. Moreover, the profile is structured in two main parts: profile_like contains features describing the concepts able to deem items relevant, while features in profile_dislike should help in filtering out not relevant items. Each profile is structured in slots, resembling the same representation strategy adopted for documents. Each slot reports the features (MULTIWORDNET synset identifiers) occurring in the training examples, with corresponding frequencies computed in the training step. Frequencies are used by the Bayesian learning algorithm to induce the classification model (i.e. the user profile) exploited to suggest relevant items in the recommendation phase. The profile learning process for user $u$ starts by selecting all items (disambiguated documents) and corresponding ratings provided by $u$. Each item falls into either the positive or the negative training set depending on the user rating, in the same way as previously described in this section. Therefore, given a new document (previously disambiguated) $d_j$, the recommendation step consists in computing the a-posteriori classification scores $P(c_+,|d_j)$, used to produce a ranked list of potentially interesting items, from which items to be recommended can be selected.

6. EXPERIMENTAL EVALUATION

The goal of the experimental evaluation was to measure the predictive accuracy of language-independent (cross language) user profiles built using the BOS model. More specifically, we would like to test 1) whether user profiles learned using examples in a specific language can be effectively exploited for recommending items in a different language, and 2) whether the accuracy of the cross-language recommender system is comparable to that of the monolingual one.

Experiments were carried out in a movie recommendation scenario in which the languages adopted in the evaluation phase are English and Italian.

We do not provide experiments for evaluating whether synset-based profiles outperform keyword-based ones, because we have already shown this result in the context of a movie recommendation scenario [28].

6.1 Users and Dataset

The experimental work has been carried out on the MovieLens dataset, containing 100,000 ratings provided by 943 different users on 1,628 movies. The dataset was quite unbalanced because 57% of the ratings was labelled as positive (this is to say, with a rating equal or greater than 4). The original dataset does not contain any information about the content of the movies. The content information for each

3http://www.grouplens.org
movie was crawled from both the English and Italian version of Wikipedia. In particular the crawler gathers the Title of the movie, the name of the Director, the Starring and the Plot.

It is worth to note that English and Italian movie descriptions are not a mere translation from one language to another: different contributors have provided the English and Italian movie plots, thus different sets of terms have been used (see Table 2). This means that a simple approach in which user profiles are learned on items in a specific language, and recommendations are provided by suggesting items properly translated in another language could be not effective. It is the same for the approach in which profiles are learned on items properly translated in the language for which recommendations should be provided.

The text in each slot has been tokenized and stemmed, and the POS tag has been identified for running the WSD algorithm. A named entity recognition procedure has been performed and stopwords have been eliminated just from the plots. Table 2 reports the number of features used to represent movies. We can notice that the WSD procedure allows a reduction of the number of features used to represent movies (roughly 9% for English, 16% for Italian).

<table>
<thead>
<tr>
<th>Slot</th>
<th>#WORDS</th>
<th>#SYNSETS</th>
<th>English</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>875</td>
<td>1,023</td>
<td>763</td>
<td>913</td>
</tr>
<tr>
<td>Director</td>
<td>642</td>
<td>642</td>
<td>525</td>
<td>498</td>
</tr>
<tr>
<td>Starring</td>
<td>2,535</td>
<td>2,535</td>
<td>2,490</td>
<td>2,234</td>
</tr>
<tr>
<td>Plot</td>
<td>25,846</td>
<td>21,932</td>
<td>23,330</td>
<td>18,268</td>
</tr>
<tr>
<td>Total</td>
<td>28,898</td>
<td>26,132</td>
<td>27,108</td>
<td>21,913</td>
</tr>
</tbody>
</table>

In order to learn accurate user profiles, we have not performed the evaluation for those users who provided less than 20 ratings. Moreover, we selected all the movies for which both the English and Italian description is available. To sum up, the dataset after this processing contained 40,717 ratings provided by 613 different users on 520 movies.

6.2 Design of the Experiment

User profiles are 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.

MARS is conceived as a text classifier, thus its effectiveness was evaluated by classification accuracy measures, namely Precision and Recall [26], where Precision (Pr) is defined as the number of relevant selected items divided by the number of selected items, and Recall (Re) is defined as the number of relevant selected items divided by the total number of relevant items. $P_\beta$ measure, a combination of Precision and Recall, is also used to have an overall measure of predictive accuracy ($\beta$ sets the relative degree of importance attributed to Pr and Re) [29]. Since users should trust the recommender, it is important to reduce false positives. It is also desirable to provide users with a short list of relevant items (even if not all the possible relevant items are suggested), rather than a long list containing a greater number of relevant items mixed-up with not relevant ones. Therefore, we set $\beta = 0.5$. These specific measures were adopted because we are interested in measuring how relevant a set of recommendations is for a user.

In the experiment, an item is considered relevant for a user if the rating is greater than or equal to 4, while MARS considers an item relevant for a user if the a-posteriori probability of the class likes is greater than 0.5.

The dataset used in the experiment is balanced in terms of positive and negative ratings (57% positive, 43% negative).

We designed two different experiments, depending on 1) the language of items used for learning profiles, and 2) the language of items to be recommended:

- **Exp#1 – ENG-ITA**: profiles learned on movies with English description and recommendations provided on movies with Italian description;
- **Exp#2 – ITA-ENG**: profiles learned on movies with Italian description and recommendations produced on movies with English description. We compared the results against the accuracy of classical monolanguage content-based recommender systems:
- **Exp#3 – ENG-ENG**: profiles learned on movies with English description and recommendations produced on movies with English description;
- **Exp#4 – ITA-ITA**: profiles learned on movies with Italian description and recommendations produced on movies with Italian description.

We executed one experiment for each user in the dataset. The ratings of each specific user and the content of the rated movies have been used for learning the user profile and measuring its predictive accuracy, using the aforementioned measures. Each experiment consisted of:

1. selecting ratings of the user and the description (English or Italian) of the movies rated by that user;
2. splitting the selected data into a training set $Tr$ and a test set $Ts$;
3. using $Tr$ for learning the corresponding user profile by exploiting the:
   - English movie descriptions (Exp#1);
   - Italian movie descriptions (Exp#2);
4. evaluating the predictive accuracy of the induced profile on $Ts$, using the aforementioned measures, by exploiting the:
   - Italian movie descriptions (Exp#1);
   - English movie descriptions (Exp#2);

In the same way, a single run for each user has been performed for computing the accuracy of monolingual recommender systems, but the process of learning user profiles from $Tr$ and evaluating the predictive accuracy on $Ts$ has been carried out using descriptions of movies in the same language, English or Italian.

The methodology adopted for obtaining $Tr$ and $Ts$ was the 5-fold cross validation [30].
Table 3: Experimental Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Pr</th>
<th>Re</th>
<th>$F_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP#1 – ENG-ITA</td>
<td>59.04</td>
<td>96.12</td>
<td>63.98</td>
</tr>
<tr>
<td>EXP#2 – ITA-ENG</td>
<td>58.77</td>
<td>95.87</td>
<td>63.70</td>
</tr>
<tr>
<td>EXP#3 – ENG-ENG</td>
<td>60.13</td>
<td>95.16</td>
<td>64.91</td>
</tr>
<tr>
<td>EXP#4 – ITA-ITA</td>
<td>58.73</td>
<td>96.42</td>
<td>63.71</td>
</tr>
</tbody>
</table>

6.3 Discussion of Results

Results of the experiments are reported in Table 3, averaged over all the users. The main outcome of the experimental session is that the strategy implemented for providing cross-language recommendations is quite effective. More specifically, user profiles learned using examples in a specific language, can be effectively exploited for recommending items in a different language, and the accuracy of the approach is comparable to those in which the learning and recommendation phase are performed on the same language. This means that the goal of the experiment has been reached.

The best result was obtained by running Exp#3, that is a classical monolanguage recommender system using English content both for learning user profiles and providing recommendations. This result was expected, due to the highest accuracy of the JIGSAW WSD algorithm for English with respect to Italian. This means that the error introduced in the disambiguation step for representing documents as bag-of-synsets hurts the performance of the Profile Learner. It is worth to note that the result of Exp#4 related to movies whose description is in Italian is quite satisfactory. The result of the second experiment, in which Italian movie descriptions are used for learning profiles that are then exploited for recommending English movies is also satisfactory.

This confirms the goodness of the approach designed for providing cross-language personalization.

The approach proposed in the paper might seem too elaborate, because of the several operations needed to represent documents as bag-of-synsets. Translating every term in the documents in other languages might appear to be the most straightforward solution to cross-language personalization. However, polysemous terms can have the effectiveness of the translation process and as a consequence the filtering performance. Furthermore, due to linguistic and cultural differences, not all terms have direct translations in all languages, even if conceptually related terms with slightly different meanings are always available. Translating every term in the user profile could be a solution even easier, but the translation process could be not effective due to the lack of context of those terms. However, a syntactic approach based on the translation of documents in different languages can fail in producing effective cross-language recommendations because documents to be recommended might be represented using keywords different from those contained in the user profile. It is the case of our experiments. To sum up, a mere syntactic approach to cross-language information filtering is hardly feasible, and semantic approaches for representing the text seems to be necessary. This is more evident from the results of the experiments, in which the cross-language recommendations are accurate even though movie descriptions are not a direct translation of English to Italian or vice versa.

In our opinion, the strongest aspects of the approach presented in this work are the following:

1. the process of learning content-based user profiles does not depend on the specific language in which training examples are represented. Indeed, the Profile Learner is able to process documents represented as bag-of-synsets, independently of the language used;

2. the Content Analyzer, able to represent documents as bag-of-synsets, relies on a WSD algorithm in which a lexical resource such as MultiWordNet is exclusively needed. That resource must contain the association between word forms and word meanings, and the alignment between senses in different languages. This alignment allows representing documents in different languages, without any additional process;

3. the WSD algorithm used by the Content Analyzer is totally unsupervised, thus it does not need any parallel corpus for associating senses to words in different languages;

4. the BOS model allows a semantic representation of user profiles that captures concepts concerning user interests, independently from the words and the language used for representing them. This means that a semantic profile is sufficient to represent user interests in several languages, while a syntactic approach would need a different profile for each different language.

7. CONCLUSIONS AND FUTURE WORK

This paper presented a semantic content-based recommender system for providing cross-language recommendations. The key idea is to provide a bridge among different languages by exploiting a language-independent representation of documents and user profiles based on word meanings, called bag-of-synsets. The assignment of the right meaning to words is based on a WSD algorithm that exploits MultiWordNet as sense repository. Experiments were carried out in a movie recommendation scenario, and the main outcome is that the accuracy of cross-language recommendations is comparable to that of classical (monolingual) content-based recommendations. In the future, we are planning to investigate the effectiveness of MARS on different domains and datasets. More specifically, we are working to extract cross-language profiles by gathering information from social networks, such as Facebook, LinkedIn, Twitter, etc., in which information is generally available in different languages.

8. REFERENCES


