Cross-Language Personalization through a Semantic Content-Based Recommender System

Pasquale Lops, Cataldo Musto, Fedelucio Narducci,
Marco de Gemmis, Pierpaolo Basile, and Giovanni Semeraro

Department of Computer Science,
University of Bari "Aldo Moro", Italy
{lops,musto,narducci,degemmis,basile,semeraro}@di.uniba.it
http://www.di.uniba.it/

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 the query is 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. Content-based filtering systems adapt their behavior to individual users by learning their preferences from documents that were already deemed relevant. The learning process aims to construct a profile of the user that can be later exploited in selecting/recommending relevant items. User profiles are generally represented using keywords in a specific language. For example, if a user likes movies whose plots are written in Italian, content-based filtering algorithms will learn a profile for that user which contains Italian words, thus movies whose plots are written in English will be not recommended, although they might be definitely interesting. 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 advanced language-independent representation based on word meanings. The proposed strategy relies on a knowledge-based word sense disambiguation technique that exploits MultiWordNet as sense inventory. As a consequence, content-based user profiles become language-independent and can be exploited for recommending 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.

Keywords: Cross-language Recommender System, Content-based Recommender System, Word Sense Disambiguation, MultiWordNet.

1 Introduction

Information Filtering (IF) systems are rapidly emerging as tools for overcoming information overload in the "digital era".

D. Dicheva and D. Dochev (Eds.): AIMSA 2010, LNAI 6304, pp. 52–60, 2010.

Specifically, content-based filtering systems [1] analyze a set of documents (mainly textual descriptions of items previously rated as relevant by an individual user) and build a model or profile of user interests based on the features (generally keywords) that describe the target objects. The profile is compared to the item descriptions to select relevant items.

Traditional keyword-based profiles are unable to capture the semantics of user interests because they suffer from problems of *polysemy*, the presence of multiple meanings for one word, and *synonymy*, multiple words with the same meaning. Another relevant problem related to keywords is the strict connection with the user language: an English user, for example, frequently interacts with information written in English, so her profile of interests mainly contains English terms. In order to receive suggestions of items whose textual description is available in a different language, she must explicitly give her preferences on items in that specific language, as well.

The main idea presented in this paper is the adoption of MultiWordNet [2] as a bridge between different languages. MultiWordNet associates a unique identifier to each possible sense (meaning) of a word, regardless the original language. In this way we can build user profiles based on MultiWordNet senses and we can exploit them in order to provide cross-language recommendations. The paper is organized as follows. Section 2 presents the architecture of the systems proposed in the paper, while Sections 3 and 4 describe the process of building language-independent documents and profiles. Experiments carried out in a movie recommendation scenario are described in Section 5. Section 6 analyzes related works in the area of cross-language filtering and retrieval, while conclusions and future work are drawn in the last section.

2 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*. 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*.

In this work, 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. 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 [8], and heavily relies on linguistic knowledge stored in lexical ontologies. In this work, the *Content Analyzer* relies on the *MultiWordNet* lexical ontology [2]. The generation of the cross-language user

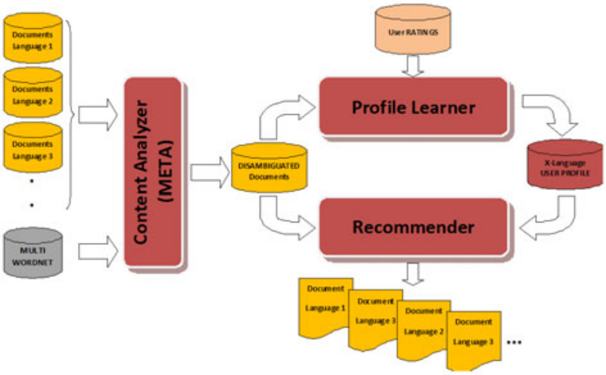


Fig. 1. General architecture of MARS

profile is performed by the *Profile Learner*, which infers the profile as a binary text classifier. 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).

3 Building Language-Independent Documents

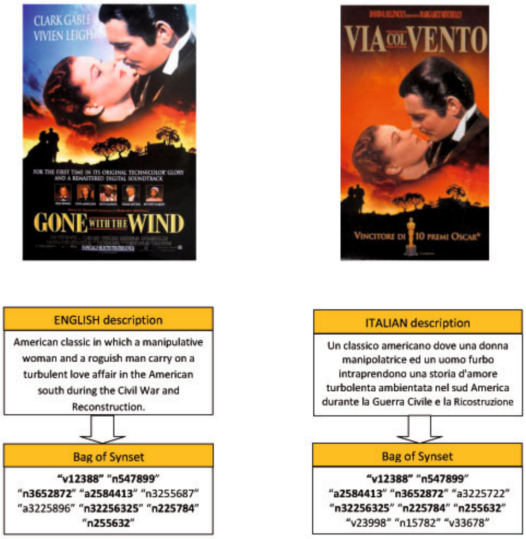
Semantic indexing of documents is performed by the Content Analyzer, which relies on META (Multi Language Text Analyzer) [9], a tool able to deal with documents in English and Italian. The goal of the semantic indexing step is to obtain a concept-based document representation. To this purpose, the text 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. In this work, the semantic indexing module exploits *MultiWordNet*[2] as sense-repository. MultiWordNet is a multilingual lexical database that supports the following languages: English, Italian, Spanish, Portuguese, Hebrew, Romanian and Latin.

The word-concept association is perform by META, which implements a Word Sense Disambiguation (WSD) algorithm, called JIGSAW.

The goal of a WSD algorithm is to associate a word w occurring in a document d with its appropriate meaning or sense s, selected from a predefined set of possibilities, usually known as *sense inventory*. JIGSAW takes as input a document encoded as a list of h words in order of their appearance, and returns a list of k MultiWordNet synsets ($k \leq h$), in which each synset s is obtained by disambiguating the target word w, by exploiting the *context* C in where w is found. The context C for w is defined as a set of words that precede and follow w. In the proposed algorithm, the sense inventory for w 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 [11], an adaptation of the Resnik algorithm has been used to disambiguate nouns [12], 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 adopted is published in [10].

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. 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.


CLARK GABLE
VIVIEN LEIGH
FOR THE FIRST TIME IN ITS ORIGINAL TECHNICOLOR GLORY AND A REMASTERED DIGITAL SOUNDTRACK
GONE WITH THE WIND
VIA COL VENTO
VINCITORE DI 10 PREMI OSCAR®

ENGLISH description

American classic in which a manipulative woman and a roguish man carry on a turbulent love affair in the American south during the Civil War and Reconstruction.

Bag of Synset

"v12388" "n547899"
"n3652872" "a2584413" "n3255687"
"a3225896" **"n32256325" "n225784"**
"n255632"

ITALIAN description

Un classico americano dove una donna manipolatrice ed un uomo furbo intraprendono una storia d'amore turbolenta ambientata nel sud America durante la Guerra Civile e la Ricostruzione

Bag of Synset

"v12388" "n547899"
"a2584413" "n3652872" "a3225722"
"n32256325" "n225784" "n255632"
"v23998" "n15782" "v33678"

Fig. 2. Example of synset-based document representation

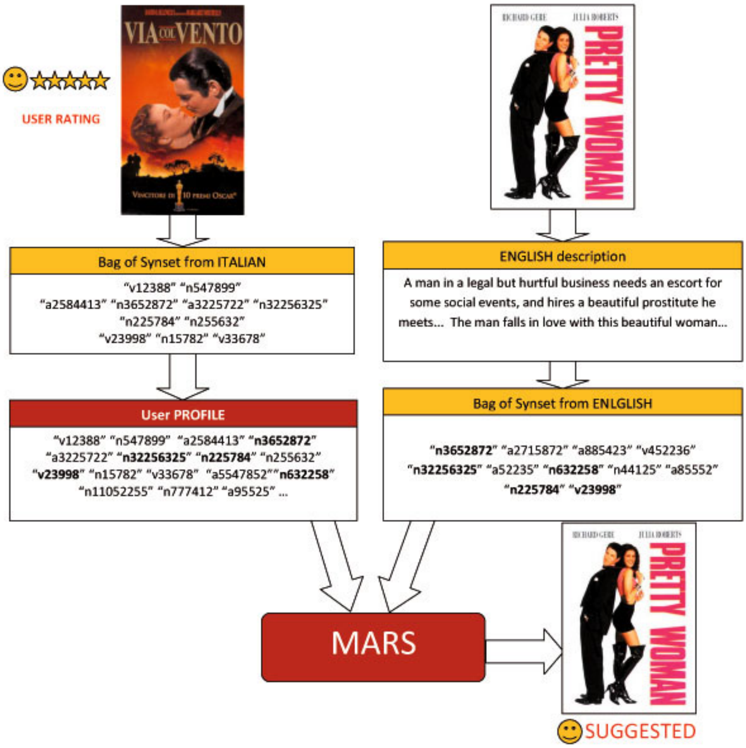
Figure 2 provides an example of representation for the movie by Victor Fleming *Gone with the wind*, corresponding to the Italian translation *Via col vento*. The textual description of the plot is provided both for the Italian and English version. The *Content Analyzer* produces the BOS containing the 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 none terms, while the adoption of the synset-based approach would allow a greater overlapping (seven shared synsets).

4 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. 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 Naïve Bayes text learning approach, widely used in content-based recommenders, which is not presented here because already described in [13]. 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_-.

Figure 3 provides an example of cross-language recommendations provided by MARS. The user likes a movie with an Italian plot and concepts extracted from the plot are stored in her synset-based profile. The recommendation step exploits this representation to suggest a movie whose plot is in English. The classical matching between keywords is replaced by a matching between synsets, allowing to suggest movies even if their descriptions do not contain shared terms.



USER RATING
VIA COL VENTO
RICHARD GERE
JULIA ROBERTS
PRETTY WOMAN
Bag of Synset from ITALIAN
"v12388" "n547899"
"a2584413" "n3652872" "a3225722" "n32256325"
"n225784" "n255632"
"v23998" "n15782" "v33678"
ENGLISH description
User PROFILE
"v12388" "n547899" "a2584413" "n3652872"
"a3225722" "n32256325" "n225784" "n255632"
"v23998" "n15782" "v33678" "a5547852""n632258"
"n11052255" "n777412" "a95525" ...
Bag of Synset from ENLGLISH
"n3652872" "a2715872" "a885423" "v452236"
"n32256325" "a52235" "n632258" "n44125" "a85552"
"n225784" "v23998"
MARS
SUGGESTED

A man in a legal but hurtful business needs an escort for some social events, and hires a beautiful prostitute he meets... The man falls in love with this beautiful woman...

Fig. 3. Example of a cross-language recommendation

5 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; 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, and the languages adopted in the evaluation phase are English and Italian.

5.1 Users and Dataset

The experimental work has been carried out on the MovieLens dataset[1], containing 100,000 ratings provided by 943 different users on 1,628 movies. The original dataset does not contain any information about the content of the movies. The content information for each 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*.

5.2 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.* F_β measure, a combination of Precision and Recall, is also used to have an overall measure of predictive accuracy (β sets the relative degree of importance attributed to Pr and Re. In this work we set β as 0.5) 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.

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.

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

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.

5.3 Discussion of Results

Results of the experiments are reported in Table 1, averaged over all the users.

Table 1. Experimental Results

Experiment	Pr	Re	F_β
exp#1 – eng-ita	59.04	96.12	63.98
exp#2 – ita-eng	58.77	95.87	63.70
exp#3 – eng-eng	60.13	95.16	64.91
exp#4 – ita-ita	58.73	96.42	63.71

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.

6 Related Work

Up to our knowledge, the topic of Cross-Language and Multilanguage Information Filtering is still not properly investigated in literature.

Recently, the Multilingual Information Filtering task at CLEF 2009[2] has introduced the issues related to the cross-language representation in the area of Information Filtering. Damankesh et al. [3], propose the application of the theory of Human Plausible Reasoning (HPR) in the domain of filtering and cross language information retrieval. The system utilizes plausible inferences to infer new, unknown knowledge from existing knowledge to retrieve not only documents which are indexed by the query terms but also those which are plausibly relevant.

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 [4] gives a good overview of the approaches for cross-language retrieval.

Ballesteros et al. [5] 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.

The most recent approaches to Cross-Language Retrieval mainly rely on the use of large corpora like Wikipedia. Potthast et al. [6] introduce CL-ESA, a new multilingual retrieval model for the analysis of cross-language similarity. The approach is based on Explicit Semantic Analysis (ESA) [7], extending the original model to cross-lingual retrieval settings.

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 recommmendations is comparable to that of classical

[2] http://www.clef-campaign.org/2009.html

(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 are generally in different languages.

References

1. Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007)
2. Bentivogli, L., Pianta, E., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: First International Conference on Global WordNet, Mysore, India (2002)
3. Damankesh, A., Singh, J., Jahedpari, F., Shaalan, K., Oroumchian, F.: Using human plausible reasoning as a framework for multilingual information filtering. In: CLEF 2008: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, Corfu, Greece (2008)
4. Oard, D.W.: Alternative approaches for cross-language text retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. AAAI (1997)
5. Ballesteros, L., Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: SIGIR 1997: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 84–91. ACM, New York (1997)
6. Martin Potthast, B.S., Anderka, M.: A wikipedia-based multilingual retrieval model. In: Advances in Information Retrieval, pp. 522–530 (2008)
7. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Veloso, M.M. (ed.) IJCAI, pp. 1606–1611 (2007)
8. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. In: Text Categorization, ch. 16, pp. 575–608. MIT Press, Cambridge (1999)
9. Basile, P., de Gemmis, M., Gentile, A., Iaquinta, L., Lops, P., Semeraro, G.: META - MultilanguagE Text Analyzer. In: Proceedings of the Language and Speech Technnology Conference - LangTech 2008, Rome, Italy, February 28-29, 2008, pp. 137–140 (2008)
10. Basile, P., de Gemmis, M., Gentile, A., Lops, P., Semeraro, G.: UNIBA: JIGSAW algorithm for Word Sense Disambiguation. In: Proceedings of the 4th ACL 2007 International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic, June 23-24, 2007. Association for Computational Linguistics, pp. 398–401 (2007)
11. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 136–145. Springer, Heidelberg (2002)
12. Resnik, P.: Disambiguating noun groupings with respect to WordNet senses. In: Proceedings of the Third Workshop on Very Large Corpora. Association for Computational Linguistics, pp. 54–68 (1995)
13. de Gemmis, M., Lops, P., Semeraro, G., Basile, P.: Integrating Tags in a Semantic Content-based Recommender. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, pp. 163–170 (2008)