Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models

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

Department of Computer Science,
University of Bari "Aldo Moro", Italy
{cataldomusto,narducci,basilepp,lops,degemmis,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 text retrieval and filtering systems. Anyway, since information exists in many languages, users could also consider as relevant documents written in different languages 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. *How could we represent user information needs or user preferences in a language-independent way?*

In this paper, we compared two content-based techniques able to provide users with cross-language recommendations: the first one relies on a knowledge-based word sense disambiguation technique that uses MultiWordNet as sense inventory, while the latter is based on a dimensionality reduction technique called Random Indexing and exploits the so-called *distributional hypothesis* in order to build language-independent user profiles.

Since the experiments conducted in a movie recommendation scenario show the effectiveness of both approaches, we tried also to underline strenghts and weaknesses of each approach in order to identify scenarios in which a specific technique fits better.

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

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 someway 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

R. Pirrone and F. Sorbello (Eds.): AI*IA 2011, LNAI 6934, pp. 250–261, 2011.

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 [18] 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). In this paper we investigated a simple research question: *how could we represent user profiles in order to create a mapping between preferences expressed in different languages and to provide cross-language recommendations with minimum costs?* We addressed this issue by comparing two different approaches: the first one exploits a Word-Sense Disambiguation technique based on MultiWordnet while the second one is based on an assumption typical of the so-called *distributional models*. It assumes that in every language each term often co-occurs with the same other terms (expressed in different languages, of course), thus by representing content-based user profiles in terms of the co-occurences of its terms, user preferences could become inerently independent from the language and this is sufficient to provide the user with cross-language recommendations. In this work we used a dimensionality reduction technique based on distributional hypothesis, called Random Indexing.

The paper is organized as follows. Section 2 analyzes related works in the area of cross-language filtering and retrieval. Recommendation models are presented

in section 3 and section 4, while the design of the experimental session carried out in a movie recommendation scenario is described in Section 5. Conclusions and future work are drawn in the last section.

2 Related Work

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

An attempt to define an effective multilingual information filtering system is proposed in [21]. The system is based on the fuzzy set theory. More specifically, the semantic content of multilingual documents is represented using a set of universal content-based topic profiles, encapsulating all feature variations among multiple languages. Using the co-occurrence statistics of a set of multilingual terms extracted from a parallel corpus (collection of documents containing identical text written in multiple languages), fuzzy clustering is applied to group semantically-related multilingual terms to form topic profiles.

Recently, the Multilingual Information Filtering task at CLEF 2009[1] has introduced the issues related to the cross-language representation in the area of Information Filtering. Damankesh et al. [5], 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 [17] gives a good overview of the approaches for cross-language retrieval. In [14] the authors propose an approach to build a model of the user's interests based on word senses rather that on simply words. The approach relyes on MultiWordNet to perform Word Domain Disambiguation and to create synset-based multilingual user profiles shown effective for news filtering. The most recent approaches to Cross-Language Retrieval mainly rely on the use of large corpora like Wikipedia. Potthast et al. [20] introduce CL-ESA, a new multilingual retrieval model for the analysis of cross-language similarity. The approach is based on Explicit Semantic Analysis (ESA) [8], extending the original model to cross-lingual retrieval settings. Furthermore, Juffinger et al.[1] recently presented the cross language retrieval system developed for the Robust WSD Task at CLEF 2008[2]. Finally, Gonzalo et al. [9] discuss ways in which EuroWordNet (EWN) [25] can be used in multilingual information retrieval activities, focusing on two approaches to Cross-Language Text Retrieval that exploit the EWN database as a large-scale multilingual semantic resource.

The use of techniques for dimensionality reduction, such as Random Indexing [22], in the area of both monolingual and multilingual Information Filtering

[1] http://www.clef-campaign.org/2009.html
[2] http://www.clef-campaign.org/2008.html

is a relatively new topic. The effectiveness of this approaches has already been demonstrated in [4] with an application for image and text data. Recently the research about semantic vector space models gained more and more attention: S-Space[3] and Semantic Vectors (SV)[4] are the first packages developed in this area. The SV package was implemented by Widdows [26]: it implements a Random Indexing algorithm and defines a negation operator based on quantum mechanics. Some initial investigations about the effectiveness of the Semantic Vectors for retrieval and filtering tasks are reported in [3] and [16].

3 First Approach: Learning Profiles Based on MultiWordnet

In a classic content-based recommender system the item properties are represented in the form of *textual slots*. For example, a movie can be described by slots *title*, *genre*, *actors*, *summary*. In this approach we can imagine a general architecture composed by the three main components: a *Content Analyzer*, a *Profile Learner*, and a *Recommender*.

The *Content Analyzer*, which relies on META (Multi Language Text Analyzer) [2], a tool for the analysis and the processing of textual documents, 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. This step requires the identification of a repository for word senses and the design of an automated procedure for defining word-concept associations. For the first requirement we exploited the *MultiWordNet* lexical ontology [19]. Similary 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 MultiWordNet the Italian WordNet is strictly aligned with English WordNet 1.6 [15]. For the second requirement we implemented a Word Sense Disambiguation (WSD) procedure that, given some generical textual content represented through the classical bag-of-words (BOW), allows to obtain a richer synset-based vector space representation, called bag-of-synsets (BOS), where each word (or each *set* of words, for bigrams or trigrams) that occurs in the original BOW is mapped on the MultiWordnet concept it refers to. In the BOS model, a synset vector, rather than a word vector, corresponds to a document. Since each concept is represented through an unique *id* that is independent from the language, by shifting the representation from BOW to BOS

[3] `http://code.google.com/p/airhead-research/`
[4] `http://code.google.com/p/semanticvectors/`

we obtained a new and unified representation that is language-independent for both English and Italian documents.

The generation of the cross-language user profile is performed by the *Profile Learner*, which infers the profile as a binary text classifier [24] 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 Naïve Bayes text learning approach, widely used in content-based recommenders, which is not presented here because already described in [7]. 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. 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 Second Approach: Distributional Models

The second strategy used to represent items content in a semantic space relies on the distributional approach. This approach represents documents as vectors in a high dimensional space, such as `WordSpace` [23]. The core idea behind `WordSpace` is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space (geometric metaphor of meaning). Replacing words with documents results in a high dimensional space where similar documents are represented close. Therefore, semantic similarity between documents can be represented as proximity in a n-dimensional space. The main characteristic of the geometric metaphor of meaning is not that meanings are represented as locations in a semantic space, but rather that similarity between documents can be expressed in spatial terms, as proximity in a high-dimensional space. One of the great virtues of the distributional approach is that documents space can be built using entirely unsupervised analysis of free text. According to the *distributional hypothesis* [10], the meaning of a word is determined by the rules of its usage in the context of ordinary and concrete language behavior. This means that words are semantically similar to the extent that they share *contexts* (surrounding words). Co-occurrence is defined with respect to a context, for example a document. Hence, words are similar if they

have the same contexts, that is to say, they are similar if they occur in the same documents. It is important to underline here that a word is represented by a vector in a high dimensional space. Since these techniques are expected to handle efficiently high dimensional vectors, a common choice is to adopt *dimensionality reduction* that allows for representing high-dimensional data in a lower-dimensional space without losing information. *Latent Semantic Analysis (LSA)* [13] collects the text data in a co-occurrence matrix, which is then decomposed into smaller matrices with singular-value decomposition (SVD), by capturing latent semantic structures in the text data. The main drawback of SVD is scalability. Differently from LSA, *Random Indexing* (RI) [11] targets the problem of dimensionality reduction by removing the need for the matrix decomposition or factorization. RI incrementally accumulates context vectors, which can be later assembled into a new space, thus it offers a novel way of conceptualizing the construction of context vectors.

RI is based on the concept of Random Projection: the idea is that high dimensional vectors chosen randomly are "nearly orthogonal". This yields a result that is comparable to orthogonalization methods, such as Singular Value Decomposition [13], but saving computational resources.

Formally, given a $n \times m$ matrix A (in this scenario it represents the classical term/document matrix) and a $m \times k$ matrix R made up of k m-dimensional random vectors, we define a new $n \times k$ matrix B as follows:

$$A^{n,m} \cdot R^{m,k} = B^{n,k} \quad k << m \tag{1}$$

The new matrix B is a more compact representation of the original matrix A and it has the property to preserve the distance between points known as Johnson-Lindenstrauss lemma, if the distance between two any points of A is d, then the distance d_r between the corresponding points in B will satisfy the property that $d_r = c \cdot d$. A proof of that property is reported in [6].

Specifically, RI builds two spaces, namely WordSpace and DocumentSpace, by following three steps:

1. a context vector is assigned to each document. This vector is sparse, high-dimensional and ternary, which means that its elements can take values in {-1, 0, 1}. A context vector contains a small number of randomly distributed non-zero elements, and the structure of this vector follows the hypothesis behind the concept of Random Projection;
2. context vectors are accumulated by analyzing terms and documents in which terms occur. In particular, the semantic vector for a term is computed as the sum of the context vectors for the documents which contain that term. Context vectors are multiplied by term occurrences.
3. the semantic vector for a document is computed as the sum of the semantic vectors for the terms which occur in that document. Semantic vectors ot temrs are multiplied by term occurrences.

In this approach for each movie we extracted its plot (in English and Italian) and we built a multilingual space. The main difference between a multilingual

space and a monolingual one is that in this space each movie has two fields F_{L1} and F_{L2}, which store the same content but in two different languages (the plot in English and Italian). It is important to underline that not necessarily the content of F_{L2} is the perfect translation of F_{L1}. The power of distributional approaches is that two terms, in different languages, are similar because they share the same context. To build multilanguage space we need to generate four spaces: two *WordSpace* SW_{L1} and SW_{L2} and two *DocumentSpace* SD_{L1} and SD_{L2}. These spaces are built as follows:

1. a context vector is assigned to each movie (plot) as described in *RI* algorithm. We call this space *RB* (random base);
2. the semantic vector for a term in SW_{L1} is computed as the sum of the context vectors in *RB* for the movies (plots) which contain that term in the field F_{L1};
3. the semantic vector for a movie (plot) in DW_{L1} is computed as the sum of the semantic vectors for the terms in SW_{L1} which occur in that movie (plot) in the field F_{L1};
4. the semantic vector for a term in SW_{L2} is computed as the sum of the context vectors in *RB* for the movies (plots) which contain that term in the field F_{L2};
5. the semantic vector for a movie (plot) in DW_{L2} is computed as the sum of the semantic vectors for the terms in SW_{L2} which occur in that movie (plot).

Given a multilingual space built in that way, in order to provide recommendations we have also to build user profiles. In these work we compared two approaches called W-RI and W-SV, thoroughly described in [16]. In the first approach, given the set of the movies that a user liked in the past (namely, whose rating explicitly provided by the user is over a certain threshol) the user profile is computed as the sum of the semantic vectors for all of this movies in DW_{L1} or DW_{L2}. In the latter one the user profile is built in the same way, but the approach gives a bigger weight to the movies that the user liked more. The user profiles can be seen as a new element of the *DocumentSpace* and can be istantiated in the vector space.

Since all the four spaces share the same random base *RB*, this makes possible to compare elements that belong to different spaces. For example we can compute how a user profile in DW_{L1} (or, respectively, in DW_{L2}) is similar to a movie in DW_{L2} (or, respectively, in DW_{L1}) and we exploited this property in order to calculate items similarity and provide users with cross-lingua recommendation. Specifically, in this approach the user receives as recommendations the items whose similarity is the higher w.r.t. her profile.

5 Experimental Evaluation

The goal of the experimental evaluation was to measure the predictive accuracy of both content-based multilingual recommendation approaches, by comparing the language-independent (cross-language) user profiles represented through

BOSs with the W-SV and W-RI approaches based on distributional hypothesis and Random Indexing. 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 a cross-language recommender system is comparable to that of a monolingual one and 3) whether a specific approach gets a significative improvement w.r.t. the other ones, becoming preferable in some specific recommendation scenario.

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

5.1 Users and Dataset

The experimental work has been performed on a subset of the MovieLens dataset[5], 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*. For both approaches the text in each slot has been tokenized, stemmed and the stopwords have been removed. For the model based on the bayesian classifier the POS tag has been identified before running the WSD algorithm.

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.

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. The effectiveness of the recommendation approaches has been evaluated by means of *Precision@n*, where n has been set as 5 and 10. In the experiment, an item is considered *relevant* for a user if the rating is greater than or equal to 3. The dataset used in the experiment is really unbalanced in terms of positive and negative ratings (83% positive, 17% 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.

[5] `http://www.grouplens.org`

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) or 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) or 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 [12].

5.3 Discussion of Results

Results of the experiments are reported in Table 1 and 2, averaged over all the users.

Table 1. Precision @5

Experiment	W-SV	W-RI	Bayes
exp#1 – eng-ita	84,65	84,65	85,61
exp#2 – ita-eng	84,85	84,63	85,20
exp#3 – eng-eng	85,23	85,29	85,23
exp#4 – ita-ita	85,27	84,84	85,71

By summing up, in this experimental session we tried to compare two very different approaches. The first one, based on a classical Bayes classifier, exploits external linguistic knowledge and relies on the assumption that the BOS can be

Table 2. Precision @10

Experiment	W-SV	W-RI	Bayes
EXP#1 – ENG-ITA	84,73	84,43	84,60
EXP#2 – ITA-ENG	84,77	84,54	84,56
EXP#3 – ENG-ENG	85,10	84,86	84,89
EXP#4 – ITA-ITA	85,11	84,86	84,93

an effective bridge to represent user preferences expressed in different languages. The second one, based on Random Indexing, does not require any linguistic pre-processing and is totally based on the distributional hypothesis. It assumes that the similar distribution of the terms, even in different languages, makes the preferences independent from the language and a simple projection of the user profile built in one language into the space built in another one is sufficient to provide the user with cross-language recommendations. In general, the main outcome of the experimental session is that the strategy implemented for providing cross-language recommendations is quite effective for both approaches. There is no significative difference by comparing the accuracy of the models previously presented. 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. Specifically, the approach based on the bayesian classifier gained the best results in the *Precision@5*. This means that this model has an higher capacity to rank the best items at the top of the recommendations list. Furthermore, it is worth to note that the comparison of the results of the Exp#1 with the results of the Exp#3 shows that the cross-lingua recommendations based on the profiles learned in English improve the Precision with respect to the monolingual one. This is due to the better accuracy of the WSD process for english contents, for which the disambiguation process introduces less noise when the BOS are built.

However the approach based on the bayesian classifier and MultiWordNet might seem too elaborate, because of the several operations needed to represent documents as bag-of-synsets. On the other side, the absence of a linguistic pre-processing is one of the strongest point of the approaches based on the distributional model and the results gained by the W-SV and W-RI models in the Precision@10 further underlined the effectiveness of this approach. Indeed, in all of the experiments, the Precision of the W-SV model is higher with respect to the W-RI model and the Bayesian one. The first result, that confirms the results already presented in [16], shows the importance of modeling negative user preferences and the goodness of the negation operator based on Quantum logic.

In conclusion, both approaches gained good results. Even in most of the experiments the cross-lingua recommendation approaches get worse results w.r.t. the mono-lingual ones, the difference in the predictive accuracy does not appear statistically significative. In general the bayesian approach fits better in scenarios where the number of items to be represented is not too high, and this can justify the application of the pre-processing steps required for building BOSs, while the

distributional models, thanks to their simplicity and effectiveness, fit better in scenarios where real-time recommendations that ensure a good accuracy need to be provided.

6 Conclusions and Future Work

In this paper we presented a comparison between two content-based approaches for providing cross-language recommendations. The key idea behind the first one 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, while the second one relies on a totally unsupervised learning method based on the distributional hypothesis. 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 (monolingual) content-based recommendations for both approaches. In the future, we are planning to investigate the effectiveness of both models 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 available in different languages.

References

1. Andreas Juffinger, R.K., Granitzer, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Evaluating Systems for Multilingual and Multimodal Information Access, pp. 155–162 (2009)
2. 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, pp. 137–140 (2008)
3. Basile, P., Caputo, A., Semeraro, G.: Semantic vectors: an information retrieval scenario. In: Melucci, M., Mizzaro, S., Pasi, G. (eds.) IIR 2010 - Proceedings of the First Italian Information Retrieval Workshop, Padua, Italy, January 27-28, pp. 1–5 (2010)
4. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: KDD 2001, pp. 245–250. ACM, New York (2001)
5. Damankesh, A., Singh, J., Jahedpari, F., Shaalan, K., Oroumchian, F.: Using Human Plausible Reasoning as a Framework for Multilingual Information Filtering. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Penas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241. Springer, Heidelberg (2010)
6. Dasgupta, S., Gupta, A.: An elementary proof of the Johnson-Lindenstrauss lemma. Tech. rep., Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA (1999)
7. de Gemmis, M., Lops, P., Semeraro, G., Basile, P.: Integrating Tags in a Semantic Content-based Recommender. In: Proc. of the 2008 ACM Conf. on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, pp. 163–170 (2008)

8. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In: Veloso, M.M. (ed.) IJCAI, pp. 1606–1611 (2007)
9. Gonzalo, J., Verdejo, F., Peters, C., Calzolari, N.: Applying EuroWordNet to Cross-Language Text Retrieval, vol. 32, pp. 185–207. Springer, Netherlands (1998)
10. Harris, Z.: Mathematical Structures of Language. Interscience, New York (1968)
11. Kanerva, P.: Sparse Distributed Memory. MIT Press, Cambridge (1988)
12. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proc. of IJCAI 1995, pp. 1137–1145 (1995)
13. Landauer, T.K., Dumais, S.T.: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review 104(2), 211–240 (1997)
14. Magnini, B., Strapparava, C.: Improving user modelling with content-based techniques. In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds.) UM 2001. LNCS (LNAI), vol. 2109, pp. 74–83. Springer, Heidelberg (2001)
15. Miller, G.: WordNet: An On-Line Lexical Database. International Journal of Lexicography 3(4) (1990) (Special Issue)
16. Musto, C.: Enhanced vector space models for content-based recommender systems. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 361–364. ACM, New York (2010), http://doi.acm.org/10.1145/1864708.1864791
17. Oard, D.W.: Alternative Approaches for Cross-Language Text Retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, pp. 154–162 (1997)
18. 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) iSBN 978-3-540-72078-2
19. Pianta, E., Bentivogli, L., Girardi, C.: MultiwordNet: developing an aligned multilingual database. In: Proc. of the 1st Int. WordNet Conference, Mysore, India, pp. 293–302 (2002)
20. Potthast, M., Stein, B., Anderka, M.: A wikipedia-based multilingual retrieval model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008)
21. Chau, R., Yeh, C.-H.: Fuzzy multilingual information filtering. In: 12th IEEE International Conference on Fuzzy Systems, FUZZ 2003, pp. 767–771 (2003)
22. Sahlgren, M.: An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop, TKE 2005 (2005)
23. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis, Stockholm University, Department of Linguistics (2006)
24. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1) (2002)
25. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities 32(2-3), 73–89 (1998)
26. Widdows, D.: Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In: ACL 2003: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp. 136–143. Association for Computational Linguistics, Morristown (2003)