Knowledge Infusion into Content-based Recommender Systems

Giovanni Semeraro
Dept. of Computer Science
University of Bari "Aldo Moro",
Via E. Orabona, 4 - Bari, Italy
semeraro@di.uniba.it

Pasquale Lops
Dept. of Computer Science
University of Bari "Aldo Moro",
Via E. Orabona, 4 - Bari, Italy
lops@di.uniba.it

Pierpaolo Basile
Dept. of Computer Science
University of Bari "Aldo Moro",
Via E. Orabona, 4 - Bari, Italy
basilepp@di.uniba.it

Marco de Gemmis
Dept. of Computer Science
University of Bari "Aldo Moro",
Via E. Orabona, 4 - Bari, Italy
degemmis@di.uniba.it

ABSTRACT

Content-based recommender systems try to recommend items similar to those a given user has liked in the past. The basic process consists of matching up the attributes of a user profile, in which preferences and interests are stored, with the attributes of a content object (item).

Common-sense and domain-specific knowledge may be useful to give some meaning to the content of items, thus helping to generate more informative features than "plain" attributes. The process of learning user profiles could also benefit from the *infusion* of exogenous knowledge or open source knowledge, with respect to the classical use of endogenous knowledge (extracted from the items themselves).

The main contribution of this paper is a proposal for *knowledge infusion* into content-based recommender systems, which suggests a novel view of this type of systems, mostly oriented to content interpretation by way of the infused knowledge. The idea is to provide the system with the "linguistic" and "cultural" background knowledge that hopefully allows a more accurate content analysis than classic approaches based on words. A set of knowledge sources is modeled to create a memory of linguistic competencies and of more specific world "facts", that can be exploited to reason about content as well as to support the user profiling and recommendation processes. The modeled knowledge sources include a dictionary, Wikipedia, and content generated by users (i.e. tags provided on items), while the core of the reasoning component is a spreading activation algorithm.

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

Design

Keywords

Content-based Recommender Systems, Spreading Activation, Open source knowledge

1. INTRODUCTION

Systems implementing a content-based recommendation approach analyze a set of documents, usually textual descriptions of the items previously rated by an individual user, and build a model or profile of user interests based on the features of the objects rated by that user [7]. The profile is a structured representation of the user interests, adopted to recommend new interesting items. The recommendation process basically consists of matching up the attributes of the user profile against the attributes of a content object. The result of the matching process is a relevance judgment that represents the user's level of interest in that object. If a profile accurately reflects user preferences, it is of tremendous advantage for the effectiveness of an information access process. For instance, it could be used to filter search results, by deciding whether a user is interested in a specific Web page or not and, in the negative case, preventing it from being displayed.

One of the limitations of content-based techniques is that they cannot provide good suggestions if the content does not contain enough information to distinguish items the user likes from items the user does not like. Some representations capture only certain aspects of the content, but there are many others that would influence a user's experience. For instance, quite often there is not enough information in the word frequency to model the user interests in jokes or poems, while techniques for affective computing would be more appropriate [8] . To sum up, both automatic extraction and manually assignment of features to items could not

RecSys'09, October 23–25, 2009, New York, New York, USA.
Copyright 2009 ACM 978-1-60558-435-5/09/10 ...$10.00.

be sufficient to define distinguishing aspects of items that turn out to be necessary for the elicitation of user interests.

The aim of the approach proposed in this paper is to overcome this limitation.

The basic idea is to enhance a content-based recommender system by *knowledge infusion*, which is realized by:

- modeling the unstructured information stored in several (open) knowledge sources;
- exploiting the acquired knowledge in order to better understand the item descriptions and extract more meaningful features.

Modeling consists of processing the information provided by each knowledge source in order to discover relationships between words. The knowledge infusion process is completed by a reasoning mechanism, based on the theory of spreading activation [2], which allows to associate new features with an item or a user profile. In this way, an intelligent knowledge-based recommendation process may be realized. Given a set of features, called *clues* (e.g. a list containing the n most informative words of an item or a user profile), additional words related to the clues can be extracted from the modeled knowledge sources. Both clues and additional words are processed by the spreading algorithm to produce new features that can be exploited in several ways during the recommendation process. For example, if clues are features of an item, the new knowledge generated by the spreading algorithm may be used to discover further items related to the one clues refer to. If clues are features belonging to a user profile, new inferred features can be useful to find neighbors.

Three different types of knowledge have been taken into account in the proposed knowledge infusion process:

- Linguistic knowledge: it is useful to recognize *general* concepts into item descriptions;
- Encyclopedic knowledge: it is useful to recognize *specific* domain-dependent concepts or named entities (usually not included in a dictionary). This type of knowledge is valuable especially for those systems (e.g. news recommenders) for which the adoption of domain ontologies is not feasible;
- Social knowledge: this type of knowledge is provided by users, therefore it is complementary with respect to the official item descriptions. The adoption of User Generated Content (UGC) is a way for taking into account evolving vocabularies besides traditional dictionaries.

The information coming from linguistic and encyclopedic knowledge sources might support a deeper (and hopefully more accurate) content analysis, while the knowledge hidden into UGC might help in associating “social” features to items, which integrate the features extracted from the “official” content. One of the forms of UGC that has drawn more attention from the research community is *folksonomy*, a taxonomy generated by users who collaboratively annotate and categorize resources of interests with freely chosen keywords called *tags*. We decided to consider tags as a valuable knowledge source since they might reveal some aspects of an item which are not covered by its official content description. Moreover, the adoption of tags in content-based recommendation techniques proved to be a promising approach [4].

The paper is structured as follows. Section 2 describes the *Knowledge Infusion* process, and provides details about both the modeling of different types of knowledge sources and the reasoning mechanism. Related work are briefly analyzed in Section 3 before drawing some final conclusions in the last section of the paper.

2. KNOWLEDGE INFUSION

We define *Knowledge Infusion* (KI) as the process of providing a system with the background knowledge which allows a “deeper understanding” of the items it deals with. In our vision, KI consists of two steps:

1. extracting and modeling relationships between words coming from several knowledge sources;
2. reasoning on the induced models in order to generate *new* knowledge, which can be useful for the recommendation step.

The name *Knowledge Infusion* arises from the fact that exogenous knowledge is introduced into the recommendation process, besides classical endogenous knowledge extracted from the items themselves.

The process is depicted in Figure 1. The core of the process is the knowledge base, which is built by modeling each type of knowledge source in order to make it usable by the reasoning module in an efficient and effective way. Since the adopted knowledge sources might have different characteristics, different heuristics should be used for building the model, and, at the same time, a uniform representation of that model is needed. We are interested in finding relationships existing between words, therefore the model of a knowledge source will be represented by the set of correlations existing between terms occurring in that specific source (a definition in a dictionary, a Wikipedia entry, etc). We decided to use a *term-term matrix* containing terms occurring in the modeled knowledge source; each cell of the matrix contains a weight representing the degree of correlation between the term on the row and the one on the column. The computation of the weights is different for each type of knowledge source and takes into account several parameters, as described in the following subsections.

KI starts from a set of given words, which we call *clues*, since they trigger the reasoning mechanism. The knowledge base is queried by using the clues, and a set of related words is obtained from the term-term model. Both clues and related words are passed to the spreading activation algorithm, which answers with a list of the most informative words related to the clues. For example, let us suppose that the clues are some keywords related to the movie “Star Wars”: *robot, alien, sword, space, battle*[1]. Some of the keywords in the list produced by the KI process are: *justice, extraterrestrial, fight*, which can be used to find related movies. The advantage of this approach is that the related movies can be selected by using keywords not necessarily included in the description of “Star Wars”, but inferred by a reasoning strategy.

[1] selected from the Internet Movie Database (IMDb) http://www.imdb.com

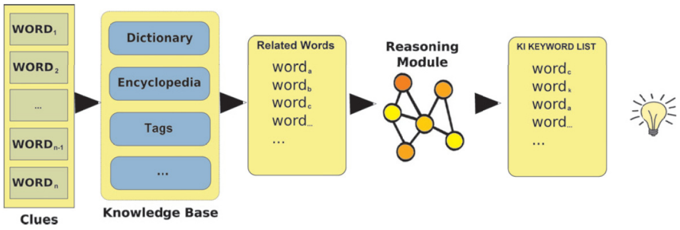

Figure 1: The Knowledge Infusion Process

2.1 Modeling a Dictionary

Our KI process exploits an Italian dictionary available on the Web, namely De Mauro Paravia[2], containing 160,000 lemmas. The model adopted to describe this type of knowledge source is a lemma-term matrix containing weights representing the relationship between a lemma and terms used to describe it in the dictionary. Because of the general lemma-definition organization of entries in the dictionary, we can fairly claim that the model is language-independent.

Each Web page describing a lemma has been preprocessed in order to extract the most relevant information useful for computing weights in the matrix. More specifically, each Web page contains the DEFINITION of the lemma, which describes its possible meanings and some example phrases that use that lemma (e.g. *intelligence* is ...), and the COLLOCATIONS, a group of words having a meaning not inferable from the meaning of the single words (e.g. *artificial intelligence*). The text of the Web page is processed in order to skip the HTML tags, even though the formatting information contribute to increase weights assigned to terms formatted using bold or italic font. Stopwords are eliminated and abbreviations used in the definition of the lemma are expanded. Weights are computed using a strategy based on a classic TF-IDF scheme [10], and normalized with respect to both the length of the definition in which the terms occur and the length of the entire dictionary.

2.2 Modeling Wikipedia

Exploiting Wikipedia as a knowledge source has several advantages, such as its constant development by the community, the availability in several languages, its high accuracy [5], and of course the availability of generic and specialized knowledge related to several domains. However, the process of modeling Wikipedia differs from the previous one, due to the huge amount of information to be processed. The resulting term-term matrix would result extremely large and sparse to be effectively used by the reasoning mechanism. Therefore, we adopted Semantic Vectors for modeling Wikipedia entries.

Semantic Vectors are based on the `WordSpace` model [9], which is a vector space whose points are used to represent semantic concepts, such as words and documents. The main idea behind Semantic Vectors is that words are represented by points in a mathematical space, in a manner that words with similar or related meanings are represented close in that space (geometric metaphor of meaning). One of the great virtues of semantic vectors is that they can be built using entirely unsupervised distributional analysis of free text. In addition, they make very few language-specific assumptions (tokenized text is just needed). We exploited the Semantic Vectors package [12], which relies on a technique called Random Indexing (RI) introduced by Kanerva in [6]. This allows building semantic vectors with no need for document-term or term-term matrix factorization, because vectors are inferred using an incremental strategy. This method faces efficiently the problem of dimensionality reduction, which is one of the key features used to uncover the "latent semantic dimensions" of a word distribution. The Semantic Vectors package supplies tools for indexing and retrieval of a collection of documents. This package relies on Apache Lucene[3] to create a basic term-document matrix, then it uses the Lucene API to create both a word-space and a document-space from the term document matrix, using Random Projection to perform dimensionality reduction without matrix factorization.

In order to exploit the Semantic Vectors model in our system, we downloaded the Italian Wikipedia XML dump and processed it by using Lucene for indexing. Then, the Semantic Vectors package was run on Lucene index. The corpus processed by the package consisted of Wikipedia pages. After these steps, we obtained a `WordSpace` model in which related words were near in that space. For example, if we search for terms near the word *pasta* we obtain: *spaghetti* 0.509 and *semola* (bram) 0.493. Each related term has a score which measures the correlation level.

2.3 Modeling Social Knowledge

In order to involve folksonomies in the recommendation process, tags are collected during the training step by letting users to express their preferences for items through a numerical rating, and to annotate rated items with free keywords. Well known systems, such as the movie recommender MovieLens[4], support member-contributed content in the form of tags. Tags are exploited either to create customized lists of items or simply to perform searching. For example, some tags associated with the movie "The Shining" by MovieLens users are: *Stephen King, Jack Nicholson, Stanley Kubrick, horror, hallucinatory, violent.* By clicking a tag in the tag cloud of the movie, a list of related movies is shown to the user. In our approach, given an item I, the set of tags provided by all the users who rated I is denoted as *SocialTags(I)*. *SocialTags(I)* can be viewed as an additional feature of an item, which is not *static* as the other features because tags evolve over time. Given a collection of items, *SocialTags(I)* is processed in order to count co-occurrences of tags. At the end of the process a tag-tag matrix is obtained, where the value associated with a pair of tags is the co-occurrence frequency in the collection.

2.4 The Reasoning Strategy

The first step of the KI process creates a memory of world facts and linguistic knowledge. The next step is to design an algorithm for retrieving the most appropriate *pieces of knowledge* associated with the clues. In order to solve this task, we adopted a *spreading activation model*, which has been used in other areas of Computer Science, such as Information Retrieval [3].

The pure spreading activation model consists of a network data structure upon which simple processing techniques are

[2] http://old.demauroparavia.it

[3] http://lucene.apache.org

[4] http://www.movielens.com

applied. The network consists of nodes interconnected by links. Links may be labeled and/or weighted, and usually have directions. The processing is initiated by labeling a set of *source nodes* with activation weights, and proceeds by iteratively propagating that activation to other nodes linked to the source nodes. For each iteration, a termination condition is checked in order to end the search process over the network.

This model was adopted as reasoning mechanism of the KI process. In the network for KI, nodes represent words, while links denote associations between words, obtained from the knowledge sources processed as described in the previous sections. The spreading activation is triggered by words given as clues. The activation of clues causes words with related meaning (as modeled in the knowledge sources) to become active. At the end of the weight propagation process, the most "active" words are included in the list provided as output by KI and might be exploited in the recommendation process. For example, let us suppose that the clues are the following five IMDb keywords related to the movie "The Shining": *axe, murder, paranormal, hotel, winter*. The output of the KI process is a list of words with related meaning, such as *perceptions, killer, psychiatry*. By searching in IMDb using some of those keywords, the following movies are retrieved: 1) "Carrie" 2) "The silence of the lambs", which seem good recommendations for users who liked "The Shining".

The spreading algorithm is thoroughly described in [11], where it was successfully used in the context of a linguistic game. Obviously, the design of an experiment is needed in order to validate the proposed approach when used in content-based recommender systems.

3. RELATED WORK

The process of infusing knowledge into a content-based recommender systems has some overlapping with research in *knowledge-based* (KB) recommender systems, which attempt to suggest items based on inferences about needs and preferences of the active user. The peculiarity of that type of recommender systems is that they have *functional* knowledge about how a particular item meets a particular user need, thus they are able to reason about the relationships between a need and a possible recommendation [1].

Our KI process is different, compared to inevitable knowledge engineering efforts required by KB systems. We do not model *specific* knowledge which guides the recommendation process, but general open source knowledge is exploited to understand the content description of items. For example, a KB restaurant recommender might use knowledge of cuisines to infer similarity between restaurants, while a travel recommender needs knowledge of locations. In our approach, knowledge is modeled and infused once (but can be extended incrementally by adding new knowledge sources) and can be exploited by different types of recommender systems.

As regards the reasoning strategy, spreading activation techniques have been successfully applied in information retrieval to discover associations among documents [3]. The distinctive features of our approach are: 1) nodes in the network correspond to terms coming from the knowledge sources, 2) weights on the links between nodes are computed according to the frequency of terms in the knowledge sources. Therefore, the reasoning mechanism is able to discover general associations between keywords, and can be consequently applied in different contexts.

4. CONCLUSIONS AND ONGOING WORK

The hypothesis discussed in this paper is that content-based techniques could benefit from the *infusion* of exogenous knowledge, with respect to the classical use of endogenous knowledge. The basic idea is that the information stored in knowledge sources might support feature extraction techniques in the content analysis step in order to build features more meaningful than simple words. Different types of knowledge sources have been taken into account in designing the knowledge infusion strategy: a dictionay (for linguistic knowledge), Wikipedia (for world knowledge), folksonomies (for social knowledge provided by users). The information stored in the knowledge sources is modeled in order to extract relationships between words, which are exploited by a spreading activation algorithm able to reason about content of items. We are convinced that this strategy might result is an effective (content+knowledge) hybrid recommender system. Currently, we are integrating all the knowledge sources described in the paper in order to perform a complete experiment on a movie dataset (content crawled from imdb.com) and to validate our working hypothesis.

5. REFERENCES

[1] R. Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[2] A. M. Collins and E. F. Loftus. A Spreading Activation Theory of Semantic Processing. *Psychological Review*, 82(6):407–428, 1975.

[3] F. Crestani. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence*, 11(6):453–582, 1997.

[4] M. de Gemmis, P. Lops, G. Semeraro, and P. Basile. Integrating Tags in a Semantic Content-based Recommender. In *Proc. of the 2008 ACM Conference on Recommender Systems*, pages 163–170, 2008.

[5] J. Giles. Internet Encyclopaedias Go Head to Head. *Nature*, 438:900–901, 2005.

[6] P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

[7] D. Mladenic. Text-learning and Related Intelligent Agents: A Survey. *IEEE Intelligent Systems*, 14(4):44–54, 1999.

[8] R. W. Picard. *Affective Computing*. MIT Press, 2000.

[9] M. Sahlgren. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces*. PhD thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics, 2006.

[10] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[11] G. Semeraro, P. Lops, P. Basile, and M. de Gemmis. On the Tip of my Thought: Playing the Guillotine Game. In *IJCAI 2009, Proc. of the 21st International Joint Conference on Artificial Intelligence*, pages 1543–1548. Morgan Kaufmann, 2009.

[12] D. Widdows and K. Ferraro. Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.