2013 IEEE 13th International Conference on Advanced Learning Technologies

Content-based Movie Recommendation within Learning Contexts.

Ricardo Kawase, Bernardo Pereira Nunes and Patrick Siehndel
Leibniz University of Hanover & L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
{*kawase, nunes, siehndel*}*@L3S.de*

***Abstract*—A good movie is like a good book. As a good book can serve entertaining and learning purposes, so does a movie. In addition to that, movies are in general more engaging and reach a wider audience. In this work, we present and evaluate a method that overcomes the challenge of generating recommendations among heterogeneous resources. In our case, we recommend movies in the context of a learning object. We evaluate our method with 60 participants that judged the relevance of the recommendations. Results show that, in over 74% of the cases the recommendations are in fact related to the given learning object, outperforming a text-based recommendation approach. The implications of our work can take learning outside the classroom and invoke it during the joy of watching a movie.**

I. Introduction

Due to the advances in technology, films and media play a greater role in our lives, reaching all audiences and supporting different goals. This work is motivated by two main circumstances. First, people devote significant part of their lives watching movies. Second, movies are indeed helpful in the learning process. In order to improve the learning experience inside and outside the classroom, our goal is to build a movie suggestion artifact. Our approach consists in a method that, given a learning object, is able to suggest contextualized movies that deal with the same topics.

Thus, the biggest challenge in our work is to overcome the barrier imposed by the different types of resources dealt with. On the one hand, we have Learning Objects (mostly textual) and, on the other hand we have movies that encompasses a short description. To accomplish our goals, we developed a method that creates semantic watermarks for objects. These watermarks are based on Wikipedia[1] main topic categories. Thus, our approach identifies and quantifies the relation of a given object to each category. By generating these watermarks, we are able to correlate learning objects and movies regarding their topic content.

II. Generating Watermarks Representation of Items

In order to compare learning objects to movies, we apply a method to generate watermarks for any text-based resource. A watermark is a histogram representation of a resource within a certain number of topics, in our case, the 23 main topic categories of Wikipedia. The process of creating watermarks is divided in a 3-step process chain: (a) entity extraction; (b) categorization and (c) profile aggregation.

Briefly, for any given object, our technique first recognizes its entities. After that, the entity categories are extracted and finally aggregated (following a weighting rule), creating the objects' watermark. With the watermarks of learning objects and movies, we are able to draw a comparison between these heterogeneous resources and, effectively generate recommendations to one or another.

During the first stage, *extraction*, entities are extracted from a given textual object. We first annotate the object to detect any mention of entities that can be linked to Wikipedia articles. For this purpose, we use the WikipediaMiner[1] service as an annotation tool. The process annotates a given document in the same way as a human would link a Wikipedia article. Our Wikipedia dataset contains over 4 million articles covering almost all knowledge domains.

In the second stage, *categorization*, we extract the categories of each entity that has been identified in the previous step. For each category, we follow the path of all parent categories, up to the root category. In some cases, this procedure results in the assignment of several top level categories to a single entity.

Following parent categories (which are closer the root category), we compute values of distance and siblings categories, resulting in each entity receiving 23 categories' scores. We used the Wikipedia category graph for relating one entity to the 23 main Wikipedia categories. The dataset we used contains 593.125 different categories. Each of these categories is linked to one or more of the main categories.

The graph walking algorithm for computing the relation of a category to the main categories follows a top-down approach that pre-computes main category weights for each entity (Wikipedia article). The relation of an article to the main categories is based on a depth first walk through the Wikipedia category graph: the algorithm remembers the distance from the root node, and follows only sub-category links of which the distance is larger or equal to the current distance.

Finally, in order to generate the final profile, the *aggregation* stage performs a linear aggregation over all of the scores of a given object. As a result of this profiling method, we have a 23 sized vector (number of Wikipedia

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

978-0-7695-5009-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ICALT.2013.53

main topic classifications), representing the watermark of a given object [5].

III. Evaluation

We perform two analogous user evaluations using a crowdsourcing platform to collect feedback. In one, we evaluate the outcomes of our method, while in the second, we evaluate a text-based approach.

A. Watermarks for Learning Objects

We based our experiments on a dataset sampled from the OpenScout project collection [2]. According to the Open Archives Initiative, the project gathers metadata information from learning resources located at different learning content repositories. The repository focus on business and management covering numerous topics, from *Management*, *Marketing* and *Economics* to *Human Resource* and *Law* among others.

For our evaluation, we selected a random set of documents whose language is English and had at least 500 characters in its description. In total, we collected 1,416 learning objects to be subject of our profiling method. The objects come from ten different online repositories[2]. General statistics of the learning objects dataset and the annotation process can be seen in Table I.

B. Watermarks for Movies

For the movies dataset, we used items from IMDb[3] collection. In total, the collection sums up to 2.3 million entries (movies, series, and so on).

Since our goal is to provide movies that can support the educational process, we selected only movies that are annotated with the genre *documentary*. Although a documentary provides a great deal of information, it does not impose the burden of a video lecture. The resulting set consisted of 31,991 documentaries. General statistics of the movies dataset and the annotation process can be seen in Table I.

The overall watermarks for learning objects and for movies are depicted in Figure 1. The distribution of topic coverage is highly influenced by the probabilities of each category in Wikipedia (there are much more articles related to the topic *Society* than to the topic *Agriculture*). Nevertheless, the graph shows the differences between the two sets where learning objects have a higher coverage on topics such as *Business*, *Science*, *Applied Sciences*, *Technology*, *Education* and *Life*.

C. Baseline Comparison

In order to generate recommendations of movies to learning objects, we used cosine similarity between the watermarks. Thus, given a learning object and its watermark,

[2]Please check the OpenScout Portal Web site for a detailed list of the repositories http://learn.openscout.net.

[3]http://www.imdb.com/

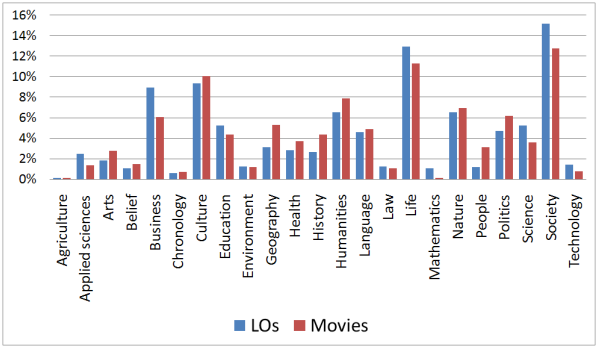

Figure 1. Overall watermarks for Learning Objecs(LOs) and IMDb movies(documentaries).

Table I
Datasets statistics.

	Learning Objects	Movies
Items	1,416	31,991
Avg. text length	878.77	615.7
Total entities found	46,211.00	234,491.00
Distinct entities	9,905.00	41,128.00
Avg. entities per object	32.66	7.6

Table II
User study results (1^{st} question) - Relevance of a movie for a LO.

Agreement	Watermark-based(%)	Text-based(%)
Strongly Agree	25.74	22.55
Agree	48.51	32.35
Undecided	3.96	13.73
Disagree	12.87	17.65
Strongly Disagree	8.91	13.73

we rank the movies according to their watermarks' cosine similarity. As a result, for each learning object, a ranked list of 'contextualized' movies is produced. With the purpose of comparison, we also generated rankings based solely on textual similarities.

To measure the textual similarity among the learning objects and movies, in our study, we used *MoreLikeThis*, a standard function provided by the Lucene search engine library[4]. *MoreLikeThis* calculates similarity of two documents by computing the number of overlapping words and giving them different weights based on TF-IDF [3]. *MoreLikeThis* runs over the fields we specified as relevant for the comparison - in our case the description of the learning objects and the movies' plots - and generates a term vector for each analyzed item (excluding stop-words). The ranking of the resulting items is based on Lucene's scoring function which is based on the Boolean model of Information Retrieval and the Vector Space Model of Information Retrieval [4].

D. User Study

Our user study consisted of a simple questionnaire to validate the quality of recommendations. Given the fact that there is no ground truth for learning objects/movies recommendations, we ran an online user evaluation. We

[4]http://lucene.apache.org/core/old_versioned_docs/versions/3_4_0/api/all/org/apache/lucene/search/similar/MoreLikeThis.html

Table III
USER STUDY RESULTS (2^{nd} QUESTION) - RELATEDNESS OF A MOVIE TO A LO.

Relatedness		Watermark-based(%)	Text-based(%)
Related	5	14.85	15.69
	4	29.70	28.43
	3	19.80	17.65
	2	15.84	16.67
Unrelated	1	19.80	21.57

set up our evaluation on CrowdFlower[5], a crowdsourcing platform. With CrowdFlower we are able to reach a broader, unbiased audience to judge our outcomes.

The task posted for the participants consisted in the evaluation of relevance and relatedness between a learning object and movie. Each participant was presented with the description of a learning object and the description of the top ranked recommended movie (same descriptions used for the annotation process). After reading the descriptions, participants were asked the following two questions

- *Q1: Do you think that the suggested movie is relevant for the learning object?*
- *Q2: In which degree the movie is related to the main topic of the learning object?*

The responses were registered in a 5-point Likert scale model. The first question aims at measuring the quality of the movie recommendations under the perspective of an extracurricular activity. The second one aims at uncovering the real relatedness of the movie and the learning object. The answers are not necessarily dependent. A movie may not be relevant for a learning object, and yet topic-wise related.

E. Results

In total, we had 60 participants in our evaluation. These participants evaluated 606 pairs (a learning object and a recommended movie). The responses were evenly distributed between watermarks and text-based approach (303 judgments for each).

In general, for the watermark based strategy, 74% of the participants *Agreed* or *Strongly Agreed* with the recommendations. In contrast, the positive agreement results for the text-based strategy sums up to only 55% (see Table II).

Regarding the relatedness between learning objects and movies, results turned out to be quite similar. Both strategies produced around 44% related (>3) recommendations. While the watermarks approach produced 44.5% of related suggestions, the text-based produced 44.1%.

To extend our analysis, we calculated the Pearson's coefficient of correlation between the first and the second question, resulting in 0.52 for the watermarks strategy and 0.80 for the text-based. In both cases we see a high correlation, specially in the text-based approach. The main reason is that the text-based approach is unable to capture different aspects other them explicit terms in the description. Thus, if it produces a relevant result, most probably it will also be related. On the other hand, watermarks identify relevance without relatedness. In fact, results show that for the watermark approach, in 13.9% of the judged pairs (learning object - movie), the participants stated that the movies were relevant (*Agree* or *Strong Agree*) but not related (relatedness 1 or 2). For the opposite case, where movies were related (relatedness 4 or 5) but not relevant (*Disagree* or *Strong Disagree*), it only happened in 1.3% of the judgments. Respectively, the numbers for the text-based approach are 7.2% and 1.3%. Even though a movie is unrelated to the main topic of a learning object, in some cases, it might still be relevant for the learning process.

In the end, the results show that the watermark approach produces significant($p<0.05$) better recommendations of movies as an extracurricular activity. Our *watermarking* approach is able to identify the context of a learning object on a higher level of abstraction. In contrast, a text-based approach is not able to identify these general topics relying solely in the term-to-term identification. In general, text-based approaches fail to identify latent topics in rather short descriptions.

[5]https://www.crowdflower.com/

IV. CONCLUSION

In this work, we presented a strategy to recommending movies to learning objects. The main challenge of this task consists in making the two types of resources comparable. In order to overcome such challenge, we used a watermark approach that identifies the main topics of each item, independent of its type.

In total, we had 60 participants in our evaluation that generated 606 judgments. The amount of positive agreements of movies suggested by the watermark reaches 74%, outperforming a text-based approach. We believe that, our method can be especially useful for teachers and tutors that are looking for alternatives to enrich their lessons.

REFERENCES

[1] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[2] K. Niemann, U. Schwertel, M. Kalz, A. Mikroyannidis, M. Fisichella, M. Friedrich, M. Dicerto, K.-H. Ha, P. Holtkamp, R. Kawase, and R. Kawase. Skill-based scouting of open management content. In *EC-TEL*, pages 632–637, 2010.

[3] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[4] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[5] P. Siehndel and R. Kawase. Twikime! - user profiles that make sense. In *International Semantic Web Conference (Posters & Demos)*, 2012.