Crowdsourced Evaluation of Semantic Patterns for Recommendations

Valentina Maccatrozzo, Lora Aroyo and Willem Robert van Hage

The Network Institute
Department of Computer Science
VU University Amsterdam, The Netherlands
v.maccatrozzo@vu.nl, l.m.aroyo@vu.nl, w.r.van.hage@vu.nl

Abstract. In this paper we explore the use of semantics to improve diversity in recommendations. We use semantic patterns extracted from Linked Data sources to surface new connections between items to provide diverse recommendations to the end users. We evaluate this methodology by adopting a bottom-up approach, i.e. we ask users of a crowdsourcing platform to choose a movie recommendation from among five options. We evaluate the results in terms of a diversity measure based on the semantic distance of topics and genres of the result list. The results of the experiment indicate that there are features of semantic patterns that can be used as an indicator of its suitability for the recommendation process.

1 Introduction

Recommender systems help people cope with the amount of information available on the Internet. Widely used are collaborative filtering and content-based recommender systems. The first requires a high availability of ratings spread over the collection, otherwise it tends to suggest only rated items, preventing diversity. Content-based algorithms are based on the characteristics of the items, making less rated items more accessible, but still lacks diversity [4]. We extend the existing approaches with semantic patterns to improve diversity in recommendation results. Linked Data enables us to discover connections between items that otherwise would not surface. We use pattern frequency statistics in the linked datasets as indicators of the ability of patterns to produce recommendations. The goal of this experiment is to find the correlation between the objective statistical measures of patterns in linked data sources and the subjective user perception of their usefulness in order to define user-centered measures of relevance of the recommendations. We do this by performing the following steps: (1) identify relevant patterns in datasets, (2) define recommendation algorithms using these patterns, (3) evaluate with the crowd. This paper reports about the initial results on these contributions.

2 Related Work

Recommender systems developed upon Semantic Web Technologies were developed by Di Noia et al., who present a content-based recommender based only on Linked Data sources, showing its potentiality [5]. Their approach do not make use of content patterns. Oufaida and Nouali [10] propose a multi-view recommendation engine that integrates collaborative filtering with social and semantic recommendation. Our approach aligns more with the work of Aroyo et al. on a content-based semantic art recommender, where [1] explores a number of semantic relationships and patterns.

Semantic patterns as we define them share some similarities with the approach proposed by Sun et al. in [16] to define a path-based semantic similarity. However, our definition of patterns relies more on the work of Gangemi and Presutti [6], who introduce knowledge patterns to deal with the semantic heterogeneity of ontologies. Presutti et al. [12] used such patterns to analyze Linked Data, as a new level of abstraction. In this work, we define such semantic patterns for the purpose of diversity in recommendations.

The use of crowdsourcing for collecting users' contributions has been explored by different works. For instance, Kittur et al. [8] present an exploratory study to show how the experimental design influence the quality of the contributions, we follow their best practices. Crowdsourcing has been used also to build ground truth data by Aroyo and Welty in [2]. Also Sarasua et al. [13] make an interesting use of crowdsourcing for ontology alignment.

3 Semantic Patterns in Recommendations

In ontologies, patterns can emerge in the combination of data instances, the types of these instances, and the links created by the properties. A semantic pattern connects a source type T_1 with a target type T_{l+1} through steps consisting of property-type pairs. This can be formulated as an ordered set: $\{T_1, P_1, T_2, P_2, ..., T_l, P_l, T_{l+1}\}$. The length of the pattern is given by l. The type of the pattern depends on the instantiation of type T_2 to T_l, e.g. people pattern, etc. Patterns are called homogenous when T_2 to T_l are of the same type and heterogenous when the types are different. The workflow we define utilizes such patterns for recommendation purposes: we extract and select patterns suitable for recommendations, performing specific analysis, and we produce recommendations ranked by the diversity measure we define.

Extraction & Selection of Patterns The sources where patterns can be discovered provide numerous candidates, hence it is critical to develop strategies to select relevant patterns. We perform a statistical analysis on the relation occurrences to select candidate patterns on the basis of their frequency, e.g. how many times the pattern is instantiated. Frequencies are calculated in two ways: considering only the properties involved in the pattern (*property frequency*), and considering also the types involved (*type frequency*). The property frequency is considered *global*, when calculated on the whole source, and *local*, when calculated in relation to an instance. For this experiment, we select patterns using different combinations of frequencies in order to test the correlation between frequencies and users' evaluations. We order the patterns on the basis of the property frequencies and we selected 6 patterns per frequency type: the two most frequent, the two less frequent and the two in the middle.

Diversity measure Diversity in recommendations is usually defined to be applied to list of items, aiming at reducing the number of similar items in the result set [14,17,7]. On the contrary, we designed a measure that does not require a list of recommendations because it is calculated with respect to the items in the user profile, hence, it can be applied also to single recommendations. This measure is defined upon the concept of semantic similarity, in a similar fashion of Middleton et al. [9] and Bogdanov et al. [3]. It allows us to suggest movies which are not exactly the users' favorites, but that are still related to them. We can consider all the metadata about a movie which consists of nouns (i.e. genre, topic, synopsis). Using relevance feedback we can identify the right value of diversity per metadata up to the right balance. Given two programs, p_1 and p_2, to calculate the measure we (1) extract genre and topic of p_1 and p_2; (2) calculate the semantic similarity between genres and topics; (3) calculate the diversity as one minus the semantic similarity; (4) calculate the diversity measure as the average of the previous ones. We use the Wu & Palmer measure [18], but other measures are possible as well.

$$Div(p_1, p_2) = \frac{(1 - sim(genre(p_1), genre(p_2))) + (1 - sim(topic(p_1), topic(p_2)))}{2}$$

4 Experimental Design

The experiment was performed on the platform CrowdFlower[1] to collect user feedback about recommendations generated using a selection of semantic patterns extracted from DBpedia[2]. We ask the users to select a match for a given movie from among five options, providing poster and synopsis. We proposed the following context: "*You are buying a movie for a friend and you want to get the"buy one, get two" promotion. Which of the following movies would you match with the starting one in order to surprise your friend with something not trivially related?*". Four options are defined with semantic patterns and ranked with IMDB ratings. We used IMDB to improve the probability of users knowing the movie to test different values of our diversity measure, as shown in Fig. 1. The fifth option is chosen from the Amazon[3] recommendations as a baseline to compare our performances. The options are in randomized order to avoid bias effects. We also ask the users to explain their choice, to obtain an indication on how they made it and to identify potential spammers. Additionally, we ask

[1] `http://crowdflower.com`

[2] `http://dbpedia.org`

[3] `http://amazon.com`

the users to type the third word in the synopsis of the movie they chose, as an additional spammer detecting question, following the best practices suggested by [8]. In particular spammers are supposed not to put any effort in the task, hence open questions are filled in with nonsense lists of characters. We use a bottom-up approach, i.e. instead of asking users to evaluate a recommendation, we ask them to choose it. In this way we try to be less intrusive as possible in affecting the users' choice.

Table 1: Generic example of options with related patterns.

Starting movie	Pattern	Selected Movie
The Devil Wears Prada	Amazon	Confessions of a Shopaholic
	Starring - Narrator	The Living Sea
	Writer	We Bought a Zoo
	Producer	Forrest Gump
	Set Location	The Bourne Ultimatum

Table 1 shows an example of the five options, starting with the Amazon recommendation, followed by the pattern *starring-narrator*, i.e. an actor in the starting movie performs as the narrator in the suggested movie. The last three options are movies that share the same properties with the starting movie: the writer, the producer, and the set location.

5 Results

We chose 12 movies of three different genres (thriller, history and crime), and selected 12 people patterns (i.e. patterns which involves only types "person") per movie. We built 36 tests and we collected 720 contributions (one contribution per user). 28 spammers were identified and eliminated from the results.

By comparing the results with the Amazon recommendation, all those suggestions that received an high number of votes (on average 27) are also reachable through semantic patterns, namely the starring pattern and the director patterns of length 2. The other Amazon recommendations received a low number of votes (on average 5.3) and performed clearly worse than the semantic patterns ones. This is an interesting result: our method can provide recommendations that can satisfy multiple needs. In order to evaluate the performances in these terms, we consider the explanation for the choices provided by the users. Although we asked the users to address diversity, the explanations show that this was not always what drove them. So, we clustered the choices on the basis of the users' comments into three categories: similar, different and not applicable (i.e. difficult to assess). Three patterns resulted peculiar for recommendations in the category '*different*': cinematographer-director, cinematographer-child-cinematographer and director-editing.

In Fig. 1, we can see the distribution of the diversity values over the movies used in the experiment. In the top right corner there are the movies that are more

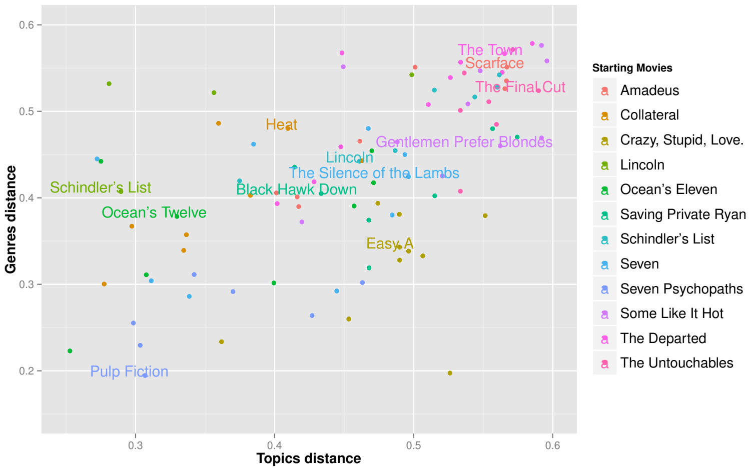

Fig. 1: For every starting movie, the diversity of recommendations is shown. In (0:0) there are the genre and topic of the starting movie. The labelled movies are the most chosen ones.

different from the starting one. Users that chose those movies did not always perceive this diversity, and they often disagree. For instance, the comments from two users who chose the pair Amadeus - Scarface were quite different. One user says: “Both movies are about the life of times of the lead characters.”, hinting at similarity. The other user says: “Pairing Scarface with Amadeus would be a surprise. Both films are American classics and contain amazing performances. Both films are biographical in nature as well. However, Amadeus is a ”period” film set in Austria and features classic works by Mozart. It's joyous and moving. Scarface is a crime drama focused on the dark underbelly of the drug cartels. It's big moments and shocks come not from musical masterpieces, but brutal violence.”, hinting at diversity. This suggests that the perception of the diversity is highly correlated with the users' knowledge of the movies, and attitude towards the task as well. However, this topic requires more investigation, which will be addressed in the future.

6 Analysis and Discussion

Our aim is to determine the most important features of a pattern to deliver meaningful recommendations. We consider local and global property frequencies, type frequencies, and length of the patterns. We perform correlation tests between the features and the users' feedback, using Spearman rank correlation test [15]. The results of this preliminary analysis show that there is a correlation between features of the semantic patterns and users' feedback. In particular, the global property frequency is positively correlated (0.32) to the users' feedback,

Table 2: Correlations between pattern features and users' feedback.

Feature	Correlation	p-value	Significance
Global property frequency	**0.32**	**7.921e-08**	**99% confidence level**
Local property frequency	0.19	0.001326	99% confidence level
Type frequency	0.23	0.0001292	99% confidence level
Length	**-0.35**	**1.616e-09**	**99% confidence level**
All features	-0.35	3.649e-09	99% confidence level
Global & Local property frequencies	-0.29	6.409e-07	99% confidence level
Global property & Type frequencies	-0.34	1.073e-08	99% confidence level
Global property frequency & Length	**-0.40**	**9.629e-12**	**99% confidence level**
Local property & Type frequencies	-0.20	0.0007238	99% confidence level
Local property frequency & Length	-0.36	1.08e-09	99% confidence level
Type frequency & Length	0.39	3.799e-11	99% confidence level
Global & Local property &Type frequencies	-0.28	1.977e-06	99% confidence level
Global & Local property frequencies & Length	-0.38	1.119e-10	99% confidence level
Local property & Type frequencies & Length	-0.30	3.171e-07	99% confidence level

i.e. the more frequent the pattern in the source, the more suitable it is for recommendations. The length of the pattern is, instead, negatively correlated (-0.35) to the users' feedback, i.e. longer patterns introduce too vague links between items, which seems not relevant for users. We performed the Principal Component Analysis [11] on the results to test different combination of the features. A combination of the global property frequency and the length of the pattern increased the correlation up to 0.40, confirming the prominence of these features in the prediction of the pattern usefulness in the recommendation process. These numbers represent a moderate correlation, however, given the limited size of the experiment, both in terms of patterns and users, and the fact that we do not take into consideration users' profile, these numbers are indicators for further research. In Table 2 we report the correlations, the p-value of the tests and their significance. In all cases we can reject the null hypothesis, i.e. all the correlation coefficients are significantly different from zero.

7 Future Work

We aim at improving our results by exploring other patterns features, as well as other sources, e.g. IMDB. We plan to perform larger scale experiments in order to compare general and domain specific vocabularies and analyze their differences in terms of patterns and coverage of items. Further, we will study the user perceived importance of each of the candidate patterns for the recommendation relevance and diversity, taking into consideration users' profiles.

Acknowledgments. This research is supported by the FP7 STREP "ViSTA-TV" project. We would also like to thank our colleague Chris Dijkshoorn for the valuable contribution for the design of the experiment.

References

1. L. Aroyo, N. Stash, Y. Wang, P. Gorgels, and L. Rutledge. CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. In *ISWC2007*, pages 879–886, 2007.
2. L. Aroyo and C. Welty. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci2013*. ACM, 2013.
3. D. Bogdanov, M. Haro, F. Fuhrmann, E. Gómez, and P. Herrera. Content-based music recommendation based on user preference examples. In *Womrad 2010*, 2010.
4. K. Bradley and B. Smyth. Improving Recommendation Diversity. In *The 12th Irish Conf. on Artificial Intelligence and Cognitive Science*, pages 85–94, 2001.
5. T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, and M. Zanker. Linked open data to support content-based recommender systems. In *I-SEMANTICS '12*, pages 1–8. ACM, 2012.
6. A. Gangemi and V. Presutti. Towards a pattern science for the Semantic Web. *Semantic Web - Interoperability Usability Applicability*, 1:61–68, 2010.
7. N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, 2011.
8. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *CHI*, pages 453–456. ACM, 2008.
9. S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, January 2004.
10. H. Oufaida and O. Nouali. Exploiting Semantic Web Technologies for Recommender Systems: A Multi View Recommendation Engine. In *ITWP 2009*, 2009.
11. K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11):559–572, 1901.
12. V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber. Extracting Core Knowledge from Linked Data. In *COLD2011*, 2011.
13. C. Sarasua, E. Simperl, and N. Noy. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *ISWC*, pages 525–541. Springer, 2012.
14. B. Smyth and P. McClave. Similarity vs. diversity. In *Case-Based Reasoning Research and Development*. Springer, 2001.
15. C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101, 1904.
16. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.
17. S. Vargas and P. Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys '11*, pages 109–116. ACM, 2011.
18. Wu, Z. and Palmer, M. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, 1994.