Implementing Recommendations in the PATHS System

Paul Clough[1], Arantxa Otegi[2], Eneko Agirre[2] and Mark Hall[1]

p.d.clough@sheffield.ac.uk, arantza.otegi@ehu.es,
e.agirre@ehu.es, m.mhall@sheffield.ac.uk

IXA NLP Group, University of the Basque Country
Information School, Sheffield University

Abstract. In this paper we describe the design and implementation of non-personalized recommendations in the PATHS system. This system allows users to explore items from Europeana in new ways. Recommendations of the type "people who viewed this item also viewed this item" are powered by pairs of viewed items mined from Europeana. However, due to limited usage data only 10.3% of items in the PATHS dataset have recommendations (4.3% of item pairs visited more than once). Therefore, "related items", a form of content-based recommendation, are offered to users based on identifying similar items. We discuss some of the problems with implementing recommendations and highlight areas for future work in the PATHS project.

Keywords: Digital libraries, recommendations, Europeana

1 Introduction

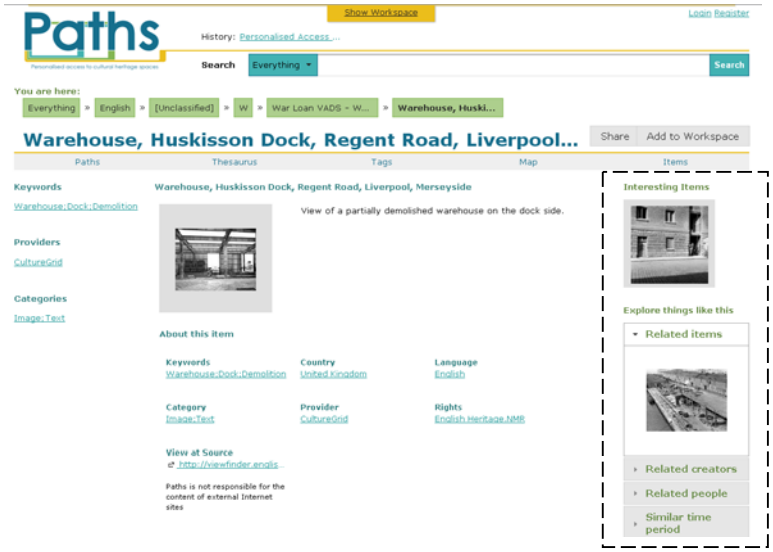
Increasingly recommender systems are being used to assist users with information discovery by bringing relevant content to users' attention. They are part of a wider set of techniques for providing personalization: the tailoring of systems or services to the specific needs of individual users or communities [1, 2]. Recommendation mechanisms provide advice on objects depending on the user context or profile. They can be broadly classified by the strategy they employ (content-based or collaborative filtering) and by the recipient of the recommendations (individual user or group recommendations). Recommender functionality (and personalization more generally) has been proven useful when providing information access to cultural heritage [3].

The EU-funded PATHS[1] (Personalized Access to Cultural Heritage) project [4, 5] is investigating ways of assisting users with exploring a large collection of cultural heritage material taken from Europeana[2], the European aggregator for museums, archives, libraries, and galleries. A prototype system has been developed that includes novel functionality for exploring the collection based on Google map-style interfaces,

[1] PATHS website: http://www.paths-project.eu/
[2] Europeana website: http://www.europeana.eu/portal/

data-driven taxonomies and supporting the manual creation of guided tours or paths. Another aspect being explored is the use of recommendations to promote information discovery. To date we have been exploring *non-personalized* recommendations based on item-to-item co-occurrences. These provide recommendations of the kind *"people who viewed this item also viewed this item."* Co-occurrence information (items that have been viewed consecutively in the same session) has been minded from a sample of Europeana logs to power the recommendations. Additionally, we provide links to "related items", a form of content-based recommendation, based on identifying 'similar' items and classifying the *type* of relation. In this paper we describe our recommendation work to date, difficulties in implementing recommendations and our plans for future work.

Fig 1. An example screenshot of the PATHS prototype[3] when viewing an item



2 The PATHS System

The current PATHS system interface is shown in Figure 1. At this point the user is viewing an item from the collection indexed in PATHS. This collection consists of approximately 540,000 items from items in Europeana that have English metadata. An additional 1.2 million Spanish items are in the process of being loaded. The prototype system as it stands supports vertical, top-down exploration through the provision of a thesaurus, a tag-cloud, topic map, and faceted search. However, to support the horizontal exploration at the level of individual items, only "paths",

[3] PATHS prototype system: http://prototype2.paths-project.eu

manually curated narratives through parts of the collection, are available. To further improve the horizontal exploration facilities, particularly in areas of the collection not covered by "paths", we are investigating non-personalized recommendations. This functionality is shown in Figure 1 in the dashed box.

2.1 Implementing *"people who viewed this also viewed this"*

We implemented a mechanism to automatically download transaction logs for the main Europeana portal on a daily basis. Currently we use a 6-months sample of logs (1 Jan to 30 June 2012), but have collected almost 2 years of data. We applied standard pre-processing, including the removal of lines not relating to user actions (e.g. cascading style sheets and images), removal of non-human actions (e.g. robots), session segmentation (based on a 30 min timeout between actions) and classification of requests (e.g. viewing an item). A 30 min timeout period of inactivity was selected based on previous research [6, 7], but we recognize that a fixed timeout period does have limitations for reliably detecting sessions and warrants further investigation [8].

In total, the processed data consists of 14,164,379 requests (3,245,766 sessions), with 53.7% of requests for item views. We filter out those sessions without any request for items that map to the PATHS dataset. This results in 102,525 sessions (3.2% of the initial log) with 208,584 item requests. For each session we extract sequences of 2 viewed items (ignoring all other request types). For example for the action sequence $item_1 \rightarrow item_2 \rightarrow search_1 \rightarrow item_3$ we would extract the sequences $item_1 \rightarrow item_2$ and $item_2 \rightarrow item_3$. We ignored pairs containing repeated items (i.e. $item_1 = item_2$). This resulted in 55,521 different pairs of items and an average of 1.82 recommendations per item.

2.2 Implementing "Related Items"

For the "related items" functionality, the similarity between each pair of items is computed using a state of the art approach based on Latent Dirichlet Allocation over the text, allowing users to quickly find related items when browsing. An evaluation dataset was crowd-sourced to enable us to assess this approach [9]. In addition, a typed similarity approach is implemented to determine the 'type' of the relation, such as similar author, location, date, event, people involved or subject. With this extra functionality, users know *why* the system is making the suggestion, an aspect considered as important to recommender systems [10]. The approach is a combination of simple similarity heuristics, based on the appropriate metadata fields, and a lineal regression [11]. The latter method improved the results considerably, obtaining second position among several contenders at an open evaluation exercise[4].

[4] Semeval 2013: http://ixa2.si.ehu.es/sts/

3 Discussion

Like most cultural heritage systems the amount of interaction data generated by users of the PATHS system is insufficient for implementing "people who viewed this also viewed this" functionality, due to data sparseness. Therefore, we exploit usage information from a more widely used system (Europeana), but restrict the data to only those items we index. However, even using data from a more widely used system we can only make recommendations for 10.3% of items in the PATHS dataset.

A further issue is that only 2,407 pairs of items (4.3%) are viewed more than once which may be the threshold at which recommendations are acceptable. Therefore, we are also working on extracting more pairs between items based on transitivity (e.g. for the sequence $item_1 \rightarrow item_2 \rightarrow item_3$ we could also assume a relation exists between $item_1 \rightarrow item_3$) and duality (e.g. for the pair $item_1 \rightarrow item_2$ we could also extract $item_2 \rightarrow item_1$). Another approach to deal with data sparseness could be to map each item to a semantic category and then make recommendations at higher levels than item, i.e. suggest pairs of items for the same subject that are viewed consecutively.

One approach we adopt in the current prototype is utilization of additional content-based recommendations. These "related items" help to alleviate the problems of insufficient usage data. Combinations of approaches are commonly used to overcome the limitations in using collaborative filtering and content-based approaches independently [2]. Further work being planned includes evaluating recommendations in a controlled lab-based setting and field trials. Also, we are developing personalized recommendations based on a session-based user model (i.e. the user profile is built up during a session from items viewed) and using PageRank to identify items of interest.

4 Conclusions

This paper discusses the implementation of non-personalized recommendations at the item-level in the PATHS system, which assists users with exploring Europeana. Recommendations of the form "people who viewed this item also viewed this item" are powered by mining co-occurrences of items viewed in Europeana. To complement these recommendations, and alleviate some of the issues with data sparseness, we also implemented "related items" functionality. We discuss some of the issues with implementing non-personalized recommendations, in addition to avenues for further work on personalized recommendations in the PATHS system.

Acknowledgements

The research leading to these results was carried out as part of the PATHS project (http://paths-project.eu) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082.

References

1 Smeaton, A., and Callan, J. (2005) Personalisation and recommender systems in digital libraries, *International Journal on Digital Libraries*, vol. 5(4), pp. 299-308.

2 Adomavicius, G., and Tuzhilin, A. (2005) Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering*, vol. 17(6), pp. 734-749.

3 Ardissono, L, Kuflik, T., and Petrelli, D. (2012) Personalization in cultural heritage: the road travelled and the one ahead, *User Modeling and User-Adapted Interaction*, vol. 22 (1-2), pp. 73-99.

4 Agirre, E. et al. (2013) PATHS: A System for Accessing Cultural Heritage Collections, In *Proceedings of the 51st Annual Meeting of the Association for Computational* Linguistics (ACL'13), Sofia, Bulgaria, August 4-9 2013, pp. 151-156.

5 Fernie, K. et al. (2012) PATHS: Personalising access to cultural heritage spaces, In *Proceedings of 18th International Conference on Virtual Systems and Multimedia* (VSMM 2012), pp.469-474.

6 Tanasa, D., and Trousse, B. (2004) Advanced data preprocessing for intersites Web usage mining, *Intelligent Systems*, IEEE, 19(2), pp.59-65.

7 Catledge, L., and Pitkow, J. (1995) Characterizing browsing strategies in the world-wide web, In *Proceedings of the Third International World-Wide Web Conference on Technology, tools and applications*, Vol. 27.

8 Jones, R., and Klinkner, K. (2008) Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs, In *Proceedings of the 17th ACM conference on Information and knowledge management* (CIKM '08). ACM, New York, NY, USA, pp. 699-708.

9 Aletras, N., Stevenson, M., and Clough, P. (2013) Computing Similarity between Items in a Digital Library of Cultural Heritage, *Journal on Computing and Cultural Heritage*, vol. 5(4), Article 16.

10 Sinha, R. and Swearingen, K. (2002) The Role of Transparency in Recommender Systems, In *Proceedings of the Conference of Human Factors in Computing Systems*, 20-25 April, 2002, Minneapolis, MN, ACM, New York, NY, pp. 830-831.

11 Agirre, E. et al. (2013) UBC UOS-TYPED: Regression for typed-similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), Volume 1: Proceedings of the main conference and the shared task: Semantic Textual Similarity*. Atlanta Georgia, June 13-14 2013, pp. 132-137.