Personalised PageRank for Making Recommendations in Digital Cultural Heritage Collections

Arantxa Otegi
IXA taldea
University of the Basque Country
Donostia, Basque Country
+34 943 015 110
arantza.otegi@ehu.es

Eneko Agirre
IXA taldea
University of the Basque Country
Donostia, Basque Country
+34 943 015 019
e.agirre@ehu.es

Paul Clough
Information School
University of Sheffield
Sheffield, UK
+44 114 2222664
p.d.clough@sheffield.ac.uk

ABSTRACT

In this paper we describe the use of Personalised PageRank (PPR) to generate recommendations from a large collection of cultural heritage items. Various methods for computing item-to-item similarities are investigated, together with representing the collection as a network over which random walks can be taken. The network can represent similarity between item metadata, item co-occurrences in search logs, and the similarity of items based on linking them to Wikipedia articles and categories. To evaluate the use of PPR, search logs from Europeana are used to simulate user interactions. PPR on each information source is compared to a standard retrieval-based baseline, resulting in higher performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Experimentation, Algorithms

Keywords

Recommender Systems, Random Walks, Cultural Heritage

1. INTRODUCTION

Recommender systems analyse user profiles, content items, and the connections between them to try and predict future user behaviour [1, 14]. They are commonly used to assist users with information discovery and help deal with the problem of information overload by suggesting specific content to users or groups of users. Recommender systems (and personalisation more generally) have been proven particularly useful when providing information access to cultural heritage content [2, 4], a domain in which the access to digital content by users with varying backgrounds and levels of expertise is growing at a rapid rate.

The EU-funded Personalised Access to Cultural Heritage project (http://www.paths-project.eu) investigated ways of assisting users with exploring a large collection of cultural heritage material selected from Europeana, the European aggregator for museums, archives, libraries, and galleries. The collection is represented as a graph-like structure, with nodes representing cultural heritage items (in this case Europeana records), and edges coming from diverse sources, including the similarity between metadata descriptions of items, implicit information derived from Europeana search logs, and the similarity between Wikipedia articles linked to items in the collection. This graph-based representation supports users' exploration through the collection. Recommendations are also being investigated as a means of further supporting exploration.

In prior work we have developed methods for producing generic, or non-personalised recommendations, of the kind *"users who viewed this item also viewed this item"* using combinations of collaborative and content-based filtering [7]. The focus of this paper, however, is personalised recommendation. Experiments using Personalised PageRank (PPR) [11] are conducted based on using different sources of information to weight links between nodes. Evaluation is performed using real search logs to simulate users and evaluate the success of PPR against a standard retrieval baseline for a relevant subset of Europeana. Results show that a PPR-based approach can successfully be used to generate recommendations. This approach is well suited to graph-based representations of digital collections and could be applied in domains other than cultural heritage.

This paper makes four contributions. The first is to show that Wikipedia provides a suitable additional information source for powering recommendations, with notable results for the graph built using the links between articles. The second is to show that PPR provides a good means to further exploit each information source. The third contribution is the evaluation methodology, which exploits search logs to automatically construct evaluation data to test several variant recommender systems cost-effectively. Finally, we perform a first evaluation of a recommender system over a relevant subset of Europeana.

2. RELATED WORK

Increasingly recommender systems are being used to assist users with information discovery by bringing relevant content to the user's attention [1]. 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 recipients of the recommendations (individual user or group recommendations). Personalisation and recommender systems are viewed as an important part of existing and future information services [14]. A further domain for adaptive IR techniques is cultural heritage [4]. A number of systems have been developed in recent years as a part of funded project work, including the CHIP (Cultural Heritage Information

978-1-4799-5569-5/14/$31.00 ©2014 IEEE

Personalisation) project that investigated the effects of transparency on trust and acceptance of user-adaptive systems of a content-based recommender system for artworks [8].

PageRank-based algorithms have been used for making recommendations on network structures. For instance, [9] used random-walks for collaborative recommendation. Alternatively, [10] presented a biased PageRank-like scoring algorithm named ItemRank, which can be used to rank products according to expected user preferences. Similarly, [17] proposed a content-based recommendation system that used random walk on an item similarity matrix. The experiments in [17] empirically showed that graphs are valuable to deal with the sparsity problem, which is also an issue in Europeana logs. In our case, we explore the use of both logs and item metadata similarity. In addition, we also propose to use the rich graph of information in Wikipedia links and categories. We show that random walks effectively use those information sources on a relevant subset of Europeana.

3. RECOMMENDER SYSTEM

The recommender system described in this paper utilises a session-based user model - the user profile is built up during an active search session from items viewed - and uses PPR to recommend further items of interest. Given one or many items, we use this graph-based algorithm to bias items that are more likely of interest to a user because they match the user's profile. Three scenarios are modelled:

No profile, item page: when the user is viewing an item, but there is no profile information because they have not logged into their account. Input information is the current item.

Profile, landing page: when the user is at the general landing page and, as they have logged into their account, their profile information is available. Input information is the set of items in their profile.

Profile, item page: when the user is viewing an item and, as they have logged into their account, their profile information is available. Input information is the set of items in their profile and the current item.

Given a graph, PageRank ranks nodes in a graph according to their relative structural importance and can be viewed as the result of a random walk process across the graph [5]. Personalised PageRank (PPR) is an algorithm that, given a graph, ranks nodes according to their relative structural importance and the proximity from a set of focus nodes [11]. During initialisation of the random walk, stronger probabilities are assigned to the focus nodes of the graph, giving more importance to those and neighbouring nodes.

In summary, the process is the following: (1) Create a graph structure to represent the digital collection offline. Nodes represent items in the collection and weights are assigned to links between nodes. Additional nodes and links can be created for various sources (see Section 4.2). (2) Given the user model as a set of items and the focus item (defined according to each of the three scenarios), Personalised PageRank computes the weight of the items relative to the given items. (3) Items with the highest weights are then selected and returned as the recommendations.

4. EXPERIMENTS

4.1 Dataset

The recommender system described in this paper works over a collection of approximately 540,000 items with English metadata selected from Europeana. The collection is represented as a graph where nodes represent items and edges connect related/similar items (the weight reflecting the strength of the relationship). Various forms of semantic enrichment, including connecting similar items and mapping items to Wikipedia articles have already been carried out [3].

We also make use of search (or transaction) logs as a form of implicit evidence and for evaluating the PPR-based approach (Section 4.3). A 6-months sample of logs (1 Jan to 30 June 2012) from the main Europeana portal was collected and standard pre-processing applied [16], including the removal of lines not relating to user actions (e.g., CSS files 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., querying, viewing an item, etc.). In total, the processed data consists of 14,164,379 requests (3,245,766 sessions), with 53.7% of requests for item views. 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., $item_1 = item_2$.

4.2 Algorithms

We have implemented several algorithms making use of different sources of data, such as keywords in the metadata, item-to-item similarity, logs (Section 4.1), and links to Wikipedia and Wikipedia categories. Item-to-item similarity is computed using a state-of-the-art approach based on Latent Dirichlet Allocation [3], which is applied after analysing the metadata information of the items. As a result, each item in the dataset is linked to 25 most similar items, together with associated confidence values.

To link items with Wikipedia articles, the Wikipedia Miner API [13] is used over title, description and subject fields. Wikipedia Miner loads an instance of Wikipedia into a database, and then provides functions to detect and disambiguate relevant Wikipedia links from plain text. For each item, all candidate topic links identified by Wikipedia Miner are included. Based on those links to Wikipedia, items are linked to Wikipedia categories. We have used those sources of data and we also applied PPR to the graph produced by the respective information, as follows:

Keyword: uses Solr search engine to index item metadata (http://lucene.apache.org/solr/). Keywords from *dc:title* and *dc:creator* fields of the items corresponding to the chosen scenario are used as queries, and the retrieved items are returned as recommendations. For the "profile, item" scenario, the two retrieved ranks for "no profile, item" and "profile, landing" are combined summing the retrieval scores.

Similarity: based on the pre-computed similarity links between items [3] based on LDA. In the "no profile, item" scenario we return the list of similar items for the current item ordered by their weight. In the two "profile" scenarios, we aggregate the similar items for each item in the profile (and for the current item in the "profile, item"), summing their weights.

Logs: links between items when both items have been visited in succession in one session are extracted from Europeana search logs (see Section 4.1). The frequency of occurrence of pairs is used to weight the links, and applied to each scenario as described for similarity above.

PPR-wikilinks: applies PPR algorithm over a graph built from Wikipedia. In order to obtain the graph structure of Wikipedia, we

Table 1: Main results in the three scenarios. Results in bold with † and ‡ indicate, respectively, a higher result and statistically significant higher result than the respective baseline (keyword vs. PPR-wikilinks, similarity vs. PPR-sim, logs vs. PPR-logs)

	No profile, item page					Profile, landing page					Profile, item page				
	#rec	#gs	Bpref	P@10	P_{cat}@10	#rec	#gs	Bpref	P@10	P_{cat}@10	#rec	#gs	bpref	P@10	P_{cat}@10
keyword	1000	232	0.2320	0.050	0.1349	1000	183	0.1830	0.028	0.0792	1000	250	0.2500	0.051	0.1320
similarity	875	54	0.0617	0.036	0.1144	958	56	0.0585	0.018	0.0792	980	86	0.0878	0.033	0.1085
logs	840	112	0.1333	0.105	0.1613	958	100	0.1044	0.068	0.1232	967	181	0.1872	0.138	0.2170
PPR-wikilinks	1000	199	0.1990	0.032	0.1085	1000	180	0.1800	0.017	**0.0909†**	1000	234	0.2340	0.024	**0.1408†**
PPR-sim	877	198	**0.2258‡**	0.017	0.1144	958	194	**0.2025‡**	0.008	**0.0850†**	980	235	**0.2398‡**	0.016	**0.1320†**
PPR-logs	857	483	**0.5636‡**	0.087	**0.1760†**	991	560	**0.5651‡**	0.050	**0.1525†**	992	587	**0.5917‡**	0.071	**0.2317†**

simply treat the articles as nodes and the links between articles as the edges. Items that link to Wikipedia articles are also added to the graph as nodes.

PPR-sim: exploits a graph created based on the similarity links. Each node in the graph corresponds to an item of the collection and an undirected edge connects two items if one item is in the list of similarity links of the other item.

PPR-logs: uses logs for creating the graph. The graph nodes are the items in the collection and there is an edge between two items if the pair exists in a log sequence.

In the "profile, item" scenario, profile and current items are considered as input information all together. But, in order to give more importance to the item viewed at the moment, when summing the weights in the first three baseline algorithms, the scores of the profile items are multiplied by 0.5. In the case of PPR algorithms, when initialising the random walk, the nodes corresponding to profile items are initialised to 0.5 and the current item to 1 (all other nodes are set to 0).

4.3 Evaluation Methodology

Various approaches have been proposed to evaluate recommender systems [12]. In this work we make use of a sample of Europeana search logs (described in Section 4.1) to create an evaluation dataset (gold standard) reflective of user interaction. A set of 1,000 sessions containing at least 5 viewed items was randomly selected from the logs[1]. Sequences of 5 viewed items (from this point referred to as '5-item log') were then extracted and used for evaluation. The output of the recommender systems were then judged on this gold standard for the following three scenarios:

No profile, item page: the fourth item of each 5-item log is used as the current item; the fifth item used as the gold standard recommendation.

Profile, landing page: the first three items of each 5-item log are used as profile items; the fifth item used as the gold standard recommendation.

Profile, item page: the first three items of each 5-item log are used as profile items; the fourth item as the current item; the fifth item used as the gold standard recommendation.

This automated evaluation methodology has the advantage of being efficient and cost-effective: allows testing and optimization of system variants without costly human involvement. We take the results based on this methodology to be indicative of performance differences between systems; however, more substantive user judgments would need to be taken based on user studies, for example as shown in [2]. Results are also an underestimation of the effectiveness of the method: the recommended items might be relevant, but we only count as relevant the one in the 5-item log. We also suspect that the results for the log and keyword methods could be overestimated, as the Europeana interface suggests items based on keywords and previous user interactions, and this may bias users. Given that we used logs from those users in the gold standard, our evaluation method might be biased in favour of them. Exploring whether this is the case is left for future work.

5. RESULTS

Table 1 shows the results on the three scenarios of the baselines and the three PPR variants. In the columns we report the number of items that obtain a recommendation (#rec), the number of items receiving recommendations including the gold standard item (#gs), and three evaluation measures (Bpref, P@10 and P_{cat}@10). Bpref [6] is a well-known evaluation measure used when relevance judgements are incomplete, as in our case. A precision score is used to indicate whether the relevant item (the 5thitem in the session) is present in the first 10 items shown to the user (P@10). A variant of P@10 is also used that scores a recommendation as correct if it is from the same Wikipedia category as the relevant item (P_{cat}@10). Our item to category mapping included 341 of the relevant items, that is, the evaluation figures are obtained on 341 items, not the full set of 1,000. Performance differences were judged significant if $p < 0.05$ [15].

Table 1 shows that using information derived from the search logs are the most effective for making recommendations amongst the three baselines. Using logs achieves the best results in all evaluation measures. Keyword search and PPR-wikilinks are also the only two methods that always return a recommendation, with similarity and logs trailing behind. The best results with P@10 are obtained using logs. None of the PPR methods attains better results than the respective baseline. The situation changes when taking into account categories (P_{cat}@10 results) or that the evaluations are incomplete (Bpref results), where the PPR-based methods, in general, beat their respective baselines, and where PPR-logs yields the best results in all three scenarios. Our interpretation is that the baseline methods are more precise in returning the item in the gold standard, but the Bpref and P_{cat}@10 results are strong indications that the recommendations of the PPR

[1] In order to allow fair evaluation, these 1,000 sessions were removed from the logs used by the recommendation algorithms (see Section 4.2).

methods are more useful in returning similar items in the same category.

The good results for the PPR-based methods are especially strong for the "profile" scenarios, where the evidence from several items needs to be aggregated. The results on the "no profile, item" scenario are only strong for PPR-logs and PPR-sim, showing that PPR is especially effective when aggregating information from several items. It is interesting to note the good results of PPR-wikilinks, which compares favourably to the keyword-based recommendations on the "profile" scenarios on $P_{cat}@10$. This is remarkable given the fact that this method is based just on the Wikipedia structure and the mentions to Wikipedia articles that are automatically detected in the item metadata. Overall the results are encouraging and support the use of PPR as a suitable algorithm for providing recommendations when the collection is represented as a network or graph.

6. CONCLUSIONS

This paper investigated the use of Personalised PageRank for making personalised recommendations in a large cultural heritage collection. A number of information sources were explored to elicit recommendations: keywords from the metadata text, similar items according to item metadata, implicit information about consecutive items viewed within a session derived from Europeana search logs, and similar items as linked via Wikipedia. Results show that all methods are effective when used separately and provides insights into the operation of PPR with various sources of information.

This paper makes a number of contributions. Firstly, we show that Wikipedia provides a suitable additional information source for deriving recommendations, with notable results for the graph built based on links between articles. Secondly, we show that PPR provides a good means to further exploit each information source, improving results with respect to a standard retrieval baseline. Given the complementarity of each information source, we believe there is room for further improvements from combining all information into a single graph. Thirdly, our evaluation methodology exploits readily available log data to automatically construct evaluation data for testing several variants recommender systems cost-effectively. Finally, we evaluate personalised recommendations on a subset of the Europeana data. In future work we plan to experiment with combining the graphs created using each of the information sources, and perform user-based evaluations to confirm the validity of our evaluation methodology, similar to [2].

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270082.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, pp. 734-749, June 2005.

[2] M. Albanese et al., "A Novel Strategy for Recommending Multimedia Objects and its Application in the Cultural Heritage Domain," *Int. J. Multimed. Data Eng. Manag.,* vol. 2, no. 4, pp. 1-18, Oct. 2011.

[3] N. Aletras et al., "Computing similarity between items in a digital library of cultural heritage," *J. Comput. Cult. Herit.*, vol. 5, no. 4, pp. 16:1-16:19, Jan. 2013.

[4] L. Ardissono et al., "Personalization in cultural heritage: the road travelled and the one ahead," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 73-99, April 2012.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[6] C. Buckley and E. Voorhees, "Retrieval evaluation with incomplete information," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Research and development in information retrieval*, New York, NY, 2004, pp. 25-32.

[7] P. Clough et al., "Implementing Recommendations in the PATHS System," in *Proc. 2nd Int. Workshop on Supporting Users' Exploration of Digital Libraries*, 2013.

[8] H. Cramer et al., "The effects of transparency on trust and acceptance in interaction with a content-based art recommender," *User Modeling and User-Adapted Interaction*, vol. 18, no. 5, pp. 455-496, Nov. 2008.

[9] F. Fouss et al., "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, pp. 355–369, Mar. 2007.

[10] M. Gori and A. Pucci, "ItemRank: A random-walk based scoring algorithm for recommender engines," in *Proc. 20th Int. Joint Conf. Artifical Intelligence*, Hyderabad, India, 2007, pp. 2766–2771.

[11] T. Haveliwala, "Topic-sensitive PageRank," in *Proc. 11th Int. Conf. World Wide Web,* New York, NY, 2002, pp. 517-526.

[12] J. Herlocker et al., "Evaluating collaborative filtering recommender systems," *Trans. Inf. Sys.*, vol. 22, no. 1, pp. 5-53, Jan. 2004.

[13] D. Milne and I. H. Witten, "Learning to Link with Wikipedia," in *Proc. 17th ACM Conf. Information and knowledge management*, Napa Valley, California, USA, 2008, pp. 509–518.

[14] A. Smeaton and J. Callan, "Personalisation and recommender systems in digital libraries," *Int. J. Dig. Libr.,* vol. 5, no. 4, pp. 299-308, Aug. 2005.

[15] M. D. Smucker et al., "A comparison of statistical significance tests for information retrieval evaluation," in *Proc. 16th ACM Conf. information and knowledge management*, New York, NY, 2007, pp. 623–632.

[16] D. Tanasa and B. Trousse, "Advanced data preprocessing for intersites Web usage mining," *IEEE Intell. Syst*., vol. 19, pp. 59-65, Mar. 2004.

[17] H. Yildirim and M. S. Krishnamoorthy, "A random walk method for alleviating the sparsity problem in collaborative filtering," in *Proc. ACM Conf. Recommender Systems*, New York, NY, 2008, pp. 131–138.