Keystone IC1302

STSM Report

STSM Reference: COST-STSMIC1302-33354

Title: ***Defining research framework for automatic recommendation of new PhD publications and personalization of ranking search results***

Host: Dr. Dragan Ivanović, University of Novi Sad, Serbia

Beneficiary: Dr. Joel Azzopardi, University of Malta, Malta

Duration: 3rd April, 2016 – 9th April, 2016

Purpose:

Researching and implementing a personalised recommendation system for the University of Novi Sad's PhD Dissertations Digital Library (PHD UNS).

Description of the work:

The University of Novi Sad has been maintaining a Digital library of PhD theses (PHD UNS) since 2012 (http://dosird.uns.ac.rs/phd-uns-digital-library-phd-dissertations). This digital library consists of the full text of the dissertations, and along with some other metadata. Currently, the PHD UNS users can search digital library and manually browse the list of new PhD dissertations added to PHD UNS.

The main aim of this short term scientific mission is to develop a personalized recommendation system for users of this PHD UNS digital library.

The first objective of this STSM was to review existing literature about personalised recommendation systems. The algorithms behind automatic recommendation systems are categorized into 2 main types, namely:

- **Content-based techniques** – whereby recommendation is performed on the basis of similarity between documents on the basis of their content; and
- **Collaborative techniques** – whereby recommendation is performed on the basis of what other 'similar' users have found useful. Collaborative recommendation may also be performed on the basis of document similarity

where similarity in this case is calculated on the basis of the overlap of users accessing that document.

For this research, it was decided to implement a number of content-based and collaborative recommendation systems. Evaluation would then be performed to determine which system produces the best result.

The required data to perform collaborative recommendation consisted of the download logs for the PHD UNS digital library. In preparation for this STSM, Dragan Ivanović had extended the PHD UNS software functionality in order to log the users' interactions in terms of downloads. The available download logs covered the period 20/2/2016 13:14:26 till 1/4/2016. 08:50:53. These logs covered 22409 downloads of 753 different dissertations from 13882 different users. These download logs were pre-processed in order to create a user-by-document matrix where each cell contains the number of downloads that a user has performed from a particular dissertation. Collaborative recommendations were calculated on the basis of this matrix.

Content-based recommendations require analyses of the dissertations' content. The PHD UNS digital library contains 1897 different dissertations. Apart from a very few exceptions, the dissertations are in Serbian. A number of the dissertations are written using the Cyrillic alphabet, and the others are written using the Latin alphabet. Our initial target was to convert these dissertations into the bag-of-words representation. Dragan provided a tool that performs all the required pre-processing, namely: it converts the Cyrillic characters into their equivalents from the roman alphabet; it performs case folding, stop word removal and stemming; and it converts non-ascii characters to corresponding characters from the ascii character set. Given this tool, we were able to convert all the dissertations from pdf to text, pre-process them using this tool, and eventually construct a term-by-document matrix where the cells contain the TF.IDF weight of each stemmed term within each document (dissertation).

The following approaches were implemented:

1. Content-based recommendation – where a user model is built by calculating the average document vector from the document vectors within the term-by-document matrix that correspond to the dissertations downloaded by the user. This mean document vector is then used to identify other similar dissertations that have not yet been downloaded by the user.

2. Content-based recommendation with LSA – this is similar to the previous approach apart except that Latent Semantic Analysis (LSA) techniques are used by decomposing the term-by-document matrix, and then using only a sub-set of the original dimensions.

3. Collaborative recommendation based on similarity between the different users – the user-by-document matrix is used to identify users similar to the current user. These 'similar' users are then used to identify other unseen dissertations that these similar users have downloaded.

4. Collaborative recommendation based on similarity between the different users using LSA. This is similar to the previous approach but the user-by-document matrix is decomposed beforehand, and only a sub-set of the original dimensions are utilised.

5. Collaborative recommendation based on 'user' similarity to the downloaded documents. In this approach the user-by-document matrix is used to identify unseen dissertations that are similar to the dissertations that have been downloaded by the user. The similarity between different dissertations is based on the overlap of users that have downloaded these dissertations.

6. Collaborative recommendation based on 'user' similarity to the downloaded documents using LSA. This is similar to the previous approach but the user-by-document matrix is decomposed beforehand, and only a sub-set of the original dimensions are utilised.

Latent Semantic Analysis have been used quite often in the case of content based analysis and recommendation. However, it has rarely been used for collaborative recommendation. Our idea for using LSA in conjunction with collaborative techniques is based on the fact that collaborative techniques suffer from the sparsity problem – i.e. having a sparse user-by-item matrix. LSA could help solve this issue.

Apart from these recommendation techniques, we implemented also a system that is extracts association rules from the user-by-document matrix using the Apriori algorithm. The extracted association rules contain only a single antecedent, and are filtered based on support count and confidence. These association rules can be used when a user has downloaded a dissertation to inform him/her about what other dissertations users tend to downloaded as well along with that dissertation.

Results Obtained:

Besides implementing the approaches described previously, we evaluated the 6 recommendation techniques listed in the previous section. We have not yet evaluated the performance of the association rules.

Evaluation was performed using a new set of download logs – these described 7122 downloads that occurred between 2/4/2016 18:47:18 and 8/4/2016 10:49:59. The

evaluation performed measured what fraction of downloads in the new set of download logs from existing users of dissertations that were not already downloaded by them were predicted by our algorithms. One has to note that this evaluation is not optimal, especially when considering that a user may not be aware of 'interesting' dissertations that may be recommended by our algorithms. However, we considered this evaluation to be a suitable initial indicator of how the different algorithms compare with each other.

For this evaluation, we had each of our system generate 20 recommendations. We calculated recall on the basis of these 20 recommendations. As a baseline, we had a system that generated a number of random recommendations for each user for dissertations that were not yet downloaded by that user.

The results obtained are shown below:

System Description	LSA	Recall
Random (Baseline)		0.051
Content Based	No LSA	0.200
Collaborative (using similar users)	No LSA	0.291
Collaborative (finding dissertations similar to downloaded dissertations)	No LSA	0.399
Content Based	K=20	0.139
Collaborative (using similar users)	K=20	0.012
Collaborative (finding dissertations similar to downloaded dissertations)	K=20	0.058
Content Based	K=50	0.217
Collaborative (using similar users)	K=50	0.007
Collaborative (finding dissertations similar to downloaded dissertations)	K=50	0.050
Content Based	K=100	0.217
Collaborative (using similar users)	K=100	0.028
Collaborative (finding dissertations similar to downloaded dissertations)	K=100	0.036
Content Based	K=200	0.218
Collaborative (using similar users)	K=200	0.020
Collaborative (finding dissertations similar to downloaded dissertations)	K=200	0.067

The best performing system was the collaborative system that finds dissertations based on similarity to the downloaded dissertations (System 5 in the previous list). This produced a recall of 40%. The collaborative system that works by finding other similar

users produced a recall of 29%. Latent Semantic Analysis on the collaborative systems did not produce good results at all – the results hover around the random recommendation baseline results.

Content-based recommendation, which represents the 'standard' way of how recommendations are usually calculated produces a recall of 20%. Latent Semantic Analysis in this case is able to enhance recall by some 2%. Here, one should note as well that the system using content-based recommendation without LSA takes the longest time to produce the recommendations – making it the least feasible approach in an online operational system.

Future Work (Research):

The results obtained provide an indication of which approach is the most suitable to be used to generate personalised recommendations to user. However, one short-coming of this research is that the evaluation is not exhaustive enough. It is providing no indication on the accuracy of the relevant recommendations. As mentioned before, a user may not be aware at all of relevant dissertations. Therefore, the fact that a user has not downloaded a dissertation does not mean that he/she is not interested in it.

Ideally, a better evaluation can be performed by having the system generate recommendations for a user, and have the user provide feedback as to whether he/she are interested in the recommended dissertations. Otherwise, this may be performed more implicitly by logging whether the downloaded dissertations are from the list of recommendations.

One other possible way of evaluating the accuracy of recommendations via a reference dataset is to utilise a search-results log in addition to the download log. If the system logs the search results for all user queries, we would be able to identify those dissertations that feature on the search results, but are not downloaded by the user. Such dissertations can then be considered as not relevant for the user, and used to provide an indication of the accuracy of results.

Other future research, apart from having an enhanced evaluation, would be to develop an ensemble of different techniques. In this case, rather than issuing recommendations based on the output of only 1 system, the output of multiple systems would be combined in some way or another.

Another avenue for future research can be of appending the rows from the term-by-document matrix to the user-by-document matrix, and using this combined matrix to

calculate recommendations, and/or to profile the different users – i.e. identify each user's area of interest.

Future Collaboration:

Our plan is to use this research as the foundation for future collaboration between ourselves. The plan for the immediate future is to implement a personalised recommendation functionality to the PHD UNS digital library. Apart from providing a better service to its users, the recommendation functionality can be used to obtain better evaluation data.

It is our intention as well to embark on research on personalised search systems. This will be done in collaboration with a researcher from the University of Cyprus with whom Dragan had already collaborated with.

Our ultimate aim is to have a funded research project that focuses on digital libraries, and providing an enhanced personalised service to users.

Foreseen Publications:

Our plan to submit the described research and the obtained results to the second keystone conference.

Eventually, our intention is to pursue this research further and submit publications as we proceed.

19/4/2016

Dr. Joel Azzopardi
University of Malta

Date