Uncertain (Multi)graphs for Personalization Services in Digital Libraries

Claudio Taranto, Nicola Di Mauro, and Floriana Esposito

Department of Computer Science, University of Bari "Aldo Moro"
via E. Orabona, 4 - 70125, Bari, Italy
{claudio.taranto,ndm,esposito}@di.uniba.it

Abstract. Digital Libraries organized collections of multimedia objects in a computer processable form. They also comprise services and infrastructures to manage, store, retrieve and share objects. Among these services, personalization services represent an active and broad area of digital library research. A popular way to realize personalization is by using information filtering techniques aiming to remove redundant or unwanted information from data. In this paper we propose to use a probabilistic framework based on uncertain graphs in order to deal with information filtering problems. Users, items and their relationships are encoded in a probabilistic graph that can be used to infer the probability of existence of a link between entities involved in the graph. The goal of the paper is to extend uncertain graphs definition to multigraphs and to study whether uncertain graphs could be used as a valuable tool for information filtering problems. The performance of the proposed probabilistic framework is reported when applied to a real-world domain.

1 Introduction

Over the past years the information content have undergone a profound change in terms of information representation and services for the use of the contents. In particular, the information content has become heterogeneous, representing different information sources such as texts, images, audio and videos. The large number of these multimedia objects and their inherent complexity has led to the need of specific services for their management and interrogation. Digital Libraries organized digital collections of multimedia objects available online in computer processable form [4]. These libraries also comprise services and infrastructures to manage, store, retrieve and share objects. In [12] the authors identify many core topics focusing on Digital Libraries research area. These topics refers to the creation of digital libraries, applications (e-learning, health care, mobile learning), preservation of data and information organization and research.

In this paper we have decided to address the problem of information organizing and finding, focusing on the personalization services, representing an active and broad area of Digital Library research [19, 12, 10, 6]. Information filtering is a popular way to realize personalization, which can be classified into content-based filtering and collaborative filtering. The goal of this paper is to show how the use

of *uncertain graphs*, an increasingly important research topic [14, 22, 9], is useful to manage and solve some information filtering problems. In particular, we will see how relationships among users and among multimedia objects with their corresponding likelihood could be easily encoded adopting an uncertain graph. This probabilistic knowledge can then be used to infer the probability of existence of links between an user and an object involved in the graph. Predicting possible relationships between an user and a multimedia object can help to find useful information and to suggest multimedia objects that user could be interested in. The proposed probabilistic framework will be evaluated along its ability to represent multimedia objects, users and their relationships and to predict new relationships among the involved entities. The basic definition of uncertain graph will be extended to that of multigraphs in order to deal with multiple connection types between nodes. In particular, we will study the behavior of the system by varying the considered neighborhood of the nodes, by studying the inference accuracy, and by considering contextual information. Experimental results on real world data show that the proposed approach is promising.

2 Related works

Digital Libraries organizing digital collections of multimedia objects, are one of many examples of information overload problem. Information filtering systems, more broadly, aim at removing redundant or unwanted information. They aim at presenting relevant information and reducing the information overload, while improving the signal-to-noise ratio at the semantic level. Personalization services for Digital Library are a key component for the fruition of the contents. Their implementation is very close to the problem of recommender systems [1] and link prediction [7], since they share the same objective to filter relevant contents for the user. A recommender system performs information filtering to bring information items such as movies, music, books, news, images, web pages, tools to a user. This information is filtered so that it is likely to interest the user. It is possible to categorize a recommender system into five groups depending on the required knowledge as follows.

Content-based systems These systems analyse user preferences in order to create a profile. Using the user profile and a description of the multimedia objects, the system can identify one or more objects that are relevant to the user profile and therefore interesting for the user. The limitation of these systems is that they assume to have a significant number of preferences for each user in order to create the profile, a problem known as the cold-start problem [3].

Collaborative filtering systems Collaborative filtering systems are based on collecting and analysing a large amount of information about users behaviour and preferences, and predicting what users will like based on their similarity to other users. These systems ignore the representation of multimedia objects. The suggestion of objects can be done in three ways: *user-based* where

user preferences are compared with those of other most similar users; *item-based* using objects similar to those that the user has seen, and *hybrid* combination of the two approaches. These approaches are called *memory-based*. Collaborative filtering methods centred on computing the relationships between multimedia objects or between users. This approach may be viewed as computing a measure of proximity or a similarity between user and objects. A similar problem is the *link prediction* problem that wants to infer missing links from an observed network: in a number of domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist [8, 13, 17].

Demographic systems These systems create a user profile based on demographic information. The suggested new multimedia objects is retrieved by considering the user demographic information and ignoring information about the description of the objects.

Knowledge-based systems These systems use a user profile that the user has previously filled. In this profile the user explicitly indicates his preferences in order to guide the suggestions of the system.

Hybrid systems These hybrid recommendation systems combine the results of multiple recommendation systems in order to obtain a more accurate recommendation. They could be divided into *homogeneous recommendation systems* which combine the output from different versions of the same recommender system and *heterogeneous recommendation systems* which combines the output from different recommender systems.

Over the last few years uncertain graphs have become an important research topic [14, 20, 21]. In these graphs each edge is associated with an existence probability that quantifies the likelihood that the edge exists in the graphs. Using this representation it is possible to adopt the *possible world* semantics to model it. One of the main issues in uncertain graphs is how to compute the connectivity of the network. The network reliability problem [5] is a generalization of the pairwise reachability, in which the goal is to determine the probability that all pairs of nodes are reachable from one another. Unlike a deterministic graph in which the reachability function is a binary function indicating whether or not there is a path connecting two nodes, in the case of uncertain graphs the function assumes probabilistic values. In [14], the authors provide a list of alternative shortest path distance measures for uncertain graphs in order to discover the k closest vertices to a given one. Another work [11] try to deal with the concept of $x - y$ distance constraint reachability problem. In particular, given two vertices x and y, they try to solve the problem of computing the probability that the distance from x to y is less than or equal to a user-defined threshold. In order to solve this problem, they proposed an exact algorithm and two reachability estimators based on probability sampling.

In this paper the idea is to use the expressive power of uncertain graphs formalism to address the information filtering problem and to allow the user to find objects of interest. The approach proposed in this paper takes advantage of the encouraging results obtained in [16], where the uncertain graph formalism

has been applied to solve the problem of collaborative filtering. In this paper we extended that framework to *uncertain multigraph*, allowing us to represent many heterogeneous connections among the involved entities. In order to test the multigraph extension we applied the system on an extension of the Movielens dataset used in [16] and its performances have been tested adopting different metrics. Furthermore, the behavior of the system has been studied by varying the neighborhood of the nodes during the creation of the uncertain graph, and by varying the inference accuracy. Finally, we will introduce contextual information in the form of probabilistic edges to study whether it contributes to the improvement in the inference step.

3 Uncertain Multi-Graphs

Let $G = (V, E)$, be a graph where V is a collection of nodes and $E \in V \times V$ is the set of edges, or relationships, between the nodes.

Definition 1 (Uncertain multi-graph). *A* uncertain multi-graph *is a system* $G = (V, E, \Sigma, l_V, l_E, s, t, p_e)$*, where* (V, E) *is an directed graph,* V *is the set of nodes,* E *is the set of ordered pairs of nodes where e=(s,t),* Σ *is a set of labels,* $l_V : V \to \Sigma$ *is a function assigning labels to nodes,* $l_E : E \to \Sigma$ *is a function assigning labels to the edges,* $s : E \to V$ *is a function indicating the source node of an edge,* $t : E \to V$ *is a function indicating the target node of an edge, and* $p_e : E \to [0, 1]$ *is a function assigning* existence probability *values to the edges.*

Each edge $a = (u, v) \in E$ has a probability called *existence probability* $p_e(a)$ which expresses the probability that the edge a, between u and v, can exist in the graph. A particular case of uncertain graph is the *discrete graph*[1], where binary edges between nodes represent the presence or absence of a relationship between them, i.e., the existence probability value on all observed edges is 1.0. The semantic of an uncertain graph is the *possible world semantics* where we can imagine an uncertain graph G as a sampler of worlds, where each world is an instance of G. An instance of G is a discrete graph G' obtained by sampling from an uncertain graph G according to the probability distribution P_e, denoted as $G' \sqsubseteq G$, when each edge $a \in E$ is selected to be an edge of G' with probability $p_e(a)$. We can consider edges labeled with probabilities as mutually independent random variables indicating whether or not the corresponding edge belongs to a discrete graph.

Assuming independence among edges, the probability distribution over discrete graphs $G' = (V, E') \sqsubseteq G = (V, E)$ is given by

$$P(G'|G) = \prod_{a \in E'} p_e(a) \prod_{a \in E \setminus E'} (1 - p_e(a)). \tag{1}$$

Definition 2 (Simple path). *Given an uncertain graph* G*, a* simple path *of a length* k *from* u *to* v *in* G *is an acyclic path denoted as a sequence of edges*

[1] Sometimes called *certain graph*.

$p_{u,v} = \langle e_1, e_2, \dots e_k \rangle$, *such that* $e_1 = (u, v_1)$, $e_k = (v_{k_1}, v)$, *and* $e_i = (v_{i-1}, v_i)$ *for* $1 < i < k$.

Given an uncertain graph G, and $p_{u,v}$ a path in G from node u to node v, $\ell(p_{u,v}) = l(e_1)l(e_2)\cdots l(e_k)$ denotes the concatenation of labels of all the edges in $p_{u,v}$.

We adopt a *regular expression* $\mathtt{R}$ to denote what is the exact sequence of labels that the path must contain. In this way we are not interested in all the paths in the uncertain graph of length k but only in those who have exactly the labels expressed by the regular expression. Now we can define a language-constrained simple path.

Definition 3 (Language-constrained simple path). *Given an uncertain graph G and a* regular expression $\mathtt{R}$*, a* language constrained simple path *is a simple path p such that $\ell(p) \in L(\mathtt{R})$.*

3.1 Querying Uncertain Graphs

The concept of existence probability of an edge in an uncertain graph can be extended to paths. We want to calculate the probability that there exists a simple path between two nodes u and v, that is, querying for the probability that a randomly sampled discrete graph contains a simple path between u and v. More formally, the *existence probability* $P_e(q|G)$ of a simple path q in a probabilistic graph G corresponds to the marginal $P((q, G')|G)$ with respect to q:

$$P_e(q|G) = \sum_{G' \sqsubseteq G} P(q|G') \cdot P(G'|G) \tag{2}$$

where $P(q|G') = 1$ if there exits the simple path q in G', and $P(q|G') = 0$ otherwise. Hence, the existence probability of the simple path q is the probability that the simple path q exists in a randomly sampled discrete graph.

Definition 4 (Language-constrained simple path probability). *Given an uncertain graph G and a* regular expression $\mathtt{R}$*, the* language-constrained simple path probability *of $L(\mathtt{R})$ is*

$$P_e(q|L(\mathtt{R}), G) = \sum_{G' \sqsubseteq G} P(q|G', L(\mathtt{R})) \cdot P(G'|G) \tag{3}$$

where $P(q|G', L(\mathtt{R}) = 1$ if there exists a simple path q in G' such that $\ell(q) \in L(\mathtt{R})$, and $P(q|G', L(\mathtt{R})) = 0$ otherwise.

The existence probability computation adopting (2) or (3) is intensive and intractable for large graphs since the number of discrete graphs to be checked is exponential in the number of probabilistic edges. In order to overcome this

problem the solution is to approximate it using a Monte Carlo sampling approach [11] in which we do not generate all the possible certain graphs but only a random subset providing the following basic sampling estimator for $P_e(q|G)$:

$$P_e(q|G) \approx \widehat{P_e(q|G)} = \frac{\sum_{i=1}^{n} P(q|G')}{n} \tag{4}$$

We proposed, as reported in [16], an iterative depth first search procedure to check the path existence. When a node is just visited, we will sample all its adjacent edges and pushing them into the stack used by the iterative procedure. We will stop the procedure either when the target node is reached or when the stack is empty which means that there isn't a path between the two nodes. In this way we can avoid to sample all edges to check whether the graph contains the path.

4 Uncertain graphs for digital library

The task of Information Filtering in DL aims to suggest a new item for a user. In this way an user can find interesting content even if the size of the DL are prohibitive or there are no effective methods to search a particular item. A classical approach is to exploit the information deriving from the adoption of a neighbourhood model. As we have shown in the related works section the two widely used methods are the user-oriented and the item-based approaches. The former estimates unknown ratings exploiting past ratings of similar users, while the latter estimates a rating using known ratings made by the same user on similar items. Let U be a set of n users and I a set of m items. A rating r_{ui} indicates the preference by user u of item i, where high values mean stronger preference. Let S_u be the set of items rated from user u. A user-based approach predicts an unobserved rating $\widehat{r_{ui}}$ as follows:

$$\widehat{r_{ui}} = \overline{r_u} + \frac{\sum_{v \in U | i \in S_u} \sigma_u(u,v) \cdot (r_{vi} - \overline{r_v})}{\sum_{v \in U | i \in S_u} |\sigma_u(u,v)|} \tag{5}$$

where $\overline{r_u}$ represents the mean rating of user u, and $\sigma_u(u,v)$ stands for the similarity between users u and v, computed, for instance, using the Pearson correlation:

$$\sigma_u(u,v) = \frac{\sum_{a \in S_u \cap S_v} (r_{ua} - \overline{r_u}) \cdot (r_{va} - \overline{r_v})}{\sqrt{\sum_{a \in S_u \cap S_v} (r_{ua} - \overline{r_u})^2 \sum_{a \in S_u \cap S_v} (r_{va} - \overline{r_v})^2}}. \tag{6}$$

On the other side, item-based approaches predict the rating of a given item using the following formula:

$$\widehat{r_{ui}} = \frac{\sum_{j \in S_u | j \neq i} \sigma_i(i,j) \cdot r_{uj}}{\sum_{j \in S_u | j \neq i} |\sigma_i(i,j)|}, \tag{7}$$

where $\sigma_i(i,j)$ is the similarity between the item i and j.

The idea behind the neighbourhood model is to consider each object as a point of a network structure and to adopt a similarity function in order to connect the objects similar to each other. The limit of this approach is to consider only the direct connections among the entities involved in the domain, and ignoring all the information available from indirect connections and from the contextual information [18, 15]. As already presented in [16], the proposed approach is used to represent a dataset consisting of user ratings, $\mathcal{K} = \{(u, i, r_{ui}) | r_{ui} \text{ is known}\}$ with an uncertain graph and then performing inference on this graph to solve classical collaborative filtering tasks. In particular, in this paper we extended the uncertain graphs definition to that of multigraphs in order to be able to manage multiple connections among nodes.

4.1 Uncertain graph construction

In order to costruct an unceratin graph from raw data, we start by analyzing the set of ratings $\mathcal{K} = \{(u, i, r_{ui}) | r_{ui} \text{ is known}\}$. For each user in $\mathcal{K}$ we add a node with label *user* and for each item in $\mathcal{K}$ a node with label *item*. As in the approach based on the neighbourhood model we add the connections among nodes. We add two kind of connections: *simU* and *simI*. For the *simU* connections, for each user u we added an edge between u and the k most similar users to u. The probability of the edge *simU* connecting two users u and v is computed as:

$$P(\texttt{simU}(u, v)) = \sigma_u(u, v) \cdot w_u(u, v) \tag{8}$$

where $\sigma_u(u, v)$ is the Pearson correlation between the vectors of ratings corresponding to the set of items rated by both user u and user v, and $w_u(u, v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}$, where S_u is the set of items rated from user u. For the *simI* connections, for each item i we added an edge between i and the most k similar items to i. The probability of the edge `simI` connecting the item i to the item j has been computed as:

$$P(\texttt{simI}(i, j)) = \sigma_i(i, j) \cdot w_i(i, j), \tag{9}$$

where s_{ij} is the Pearson correlation between the vectors corresponding to the histogram of the set of ratings for the item i and the item j, and $w_i(i, j) = \frac{|\overline{S}_i \cap \overline{S}_j|}{|\overline{S}_i \cup \overline{S}_j|}$, where $\overline{S}_i$ and $\overline{S}_j$ are the set of users rating the item i and j.

In this paper we adopt a multigraph, hence we can describe multiple connections between two nodes. Supposing that users and items are described using a set of features, for each item i, we can add an edge with label `simI`f with respect to the feature f, between i and the most k similar items to i. In particular, the probability of the edge `sim`f connecting the item i to the item j could be computed as:

$$P(\texttt{sim}f(i, j)) = \frac{|i_f \cap j_f|}{|i_f| + |j_f| + 1}, \tag{10}$$

where i_f is the value of the feature f for the item i. For instance, if a film is described using its genres and actors, the previous formula may be used to compute a similarity between films based on actors and genres. With a similar argument we can add an edge between the user u and the most k similar user to u with respect to a given feature.

The edges labelled $\mathtt{r}_k$ have probability equal to 1.0 denoting a specific vote of a user relative to an object. Now that we have an uncertain graph we can predict an unknown rating $\widehat{r_{ui}}$ solving the following maximization problem:

$$\widehat{r_{ui}} = \arg\max_j P(\mathtt{r}_j(u,i)|G), \tag{11}$$

where $\mathtt{r}_j(u,i)$ is the unknown link with label $\mathtt{r}_j$ between the user u and the item i. Adopting this approach we can simulate user-based collaborative filtering by querying the probability of the paths, starting from a user node and ending to an item node, belonging to the regular expression $L_i = \{\mathtt{simU}^1\mathtt{r}_i^1\}$. In particular, predicting the probability of the rating j as $P(\mathtt{r}_j(u,i)$ in (11) corresponds to compute the probability $P(q|G)$ for a query path in L_i, i.e., computing $P(L_i|G)$ as in (3):

$$\widehat{r_{ui}} = \arg\max_j P(\mathtt{r}_j(u,i)|G) \approx \arg\max_j P(L_j|G). \tag{12}$$

We can simulate item-based collaborative filtering in the same way by computing the probability of the paths belonging to the regular expression $L_i = \{\mathtt{r}_i^1\mathtt{simI}^1\}$. Adopting a regular expression based approach we can construct any type of query: simple, complex, hybrid (combining a user-based and an item-based approach) and exploiting contextual information.

5 Experiments

In order to validate the proposed approach the HetRec2011[2] dataset has been used. This dataset is an extension of the Movielens dataset and contains user ratings expressing preferences for different movies. The dataset contains 2113 users, 10197 films and 855598 ratings. The ratings are one of 10 distinct values ranging from 0.5 to 5.0 with increments of 0.5. The meta-data available include user-movie tag information, movie genres, movie directors, country assignments, and aggregate statistics of audience and critics ratings. The dataset has been divided in training and testing data. The testing part includes the last four ratings for each user, while the training part includes all the previous ones. Then, the validation procedure has been conducted following the steps: a) creating the uncertain graph from the training data as reported in Section 4; b) defining a regular expression corresponding to a specific information filtering task; and c) testing the ratings reported in the testing dataset $\mathcal{T}$ by computing, for each pair $(u,i) \in \mathcal{T}$ the predicted rating as in Equation (12) and comparing the prediction with the true rating as reported in $\mathcal{T}$. In this particular dataset we have a uncertain graph with nodes labeled as `user` or `film`. There are edges

[2] `http://ir.ii.uam.es/hetrec2011/datasets.html`

between two `film` nodes labeled as `simF` or `simFA`, and edges with label `simU` or `simUG` between two `user` nodes. These edges are added using the procedure presented in the previous section. In particular, `simF` denotes the probability that two films could be similar and it has been computed using (9), while `simU` indicates the probability that two users are similar computed with (8). `simFA` edges connecting two films whose probability has been computed using (10), in particular `simFA` has been computed using the actors of the films. `simUG` connects two users with a probability corresponding to the similarity computed using the histogram of the rated films' genres. For each rating $(u, i, r_{ui} = k)$ belonging to the training set there is an edge between the user u and the film i whose label is $\mathtt{r}_k$. The goal is to predict the correct rating for each instance belonging to the testing set $\mathcal{T}$. The predicted rating has been computed using a Monte Carlo approach by sampling certain graphs and adopting the function in (12).

The accuracy of the proposed framework has been evaluated according to the *mean absolute error* (MAE) and to the *root mean squared error* (RMSE), that are the two most commonly applied evaluation metrics for rating predictions. Given N computed rating predictions the functions are computed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\widehat{r_{ui}} - r_{ui}| \tag{13}$$

and

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{r_{ui}} - r_{ui})^2} \tag{14}$$

In order to evaluate the framework we proposed to query the paths belonging to the regular expressions reported in Table 1. The first language constrained simple paths L_1 corresponds to solve a user-based information filtering problem, while the third language L_3 gives us the possibility to simulate an item-based information filtering approach. As we can see from Table 2 results improve when we go from a user-based approach to a item-based in terms of MAE. We can see also that adopting languages L_2 and L_4, that consider contextual edges amongs users or items, we have improving results. In the second experiment, we proposed to extend the basic languages L_3 and L_4 in order to consider a neighbourhood with many nested levels. In particular, instead of considering the direct neighbours only, we inspect the uncertain graph following a path with a maximum length of two edges (L_5,L_6) and three edges (L_7). As we can see in Table 3 languages L_5,L_6 and L_7, where we extend the neighborhood of the explored graph, when compared with languages L_3 and L_4 achieved better results. Furthermore, languages L_8, L_9 and L_{10} corresponds to a hybrid system combining both user-based and item-based approach, whose corresponding results are shown in Table 4. In each table, reporting the MAE results, the first column reports the neighbourhood of the most similar nodes introduced in the graph for each similarity function, and the second column reports the number of sampling adopting for each languages.

$$
\begin{array}{l}
L_1 = \{\texttt{simU}^1 \texttt{r}_k^1\} \\
L_2 = \{\texttt{simU}^1 \texttt{r}_k^1\} \cup \{\texttt{simUG}^1 \texttt{r}_k^1\} \\
L_3 = \{\texttt{r}_k^1 \texttt{simF}^1\} \\
L_4 = \{\texttt{r}_k^1 \texttt{simF}^1\} \cup \{\texttt{r}_k^1 \texttt{simFA}^1\} \\
L_5 = \{\texttt{r}_k^1 \texttt{simF}^n : 1 \leq n \leq 2\} \\
L_6 = \{\texttt{r}_k^1 \texttt{simF}^n : 1 \leq n \leq 2\} \cup \{\texttt{r}_k^1 \texttt{simFA}^n : 1 \leq n \leq 2\} \\
L_7 = \{\texttt{r}_k^1 \texttt{simF}^n : 1 \leq n \leq 3\} \cup \{\texttt{r}_k^1 \texttt{simFA}^n : 1 \leq n \leq 3\} \\
L_8 = \{\texttt{simU}^1 \texttt{r}_k^1\} \cup \{\texttt{r}_k^1 \texttt{simF}^1\} \\
L_9 = \{\texttt{simU}^n \texttt{r}_k^1 : 1 \leq n \leq 2\} \cup \{\texttt{r}_k^1 \texttt{simF}^n : 1 \leq n \leq 2\} \\
L_{10} = \{\texttt{simU}^n \texttt{r}_k^1 : 1 \leq n \leq 3\} \cup \{\texttt{r}_k^1 \texttt{simF}^n : 1 \leq n \leq 3\}
\end{array}
$$

Table 1. Language constrained simple paths used for the HetRec2011 dataset.

Neighborhood	**Sampling**	L_1	L_3	L_2	L_4
5	100	1.0070	0.9878	0.7493	0.7316
5	500	1.0040	0.9840	0.7314	0.7300
10	100	0.9740	0.9661	0.6850	0.6788
10	500	0.9687	0.9631	0.6745	0.6720
15	100	0.9446	0.9404	0.6545	0.6521
15	500	0.9395	0.9380	0.6526	0.6488
20	100	0.9383	0.9308	0.6415	0.6409
20	500	0.9297	0.9263	0.6390	0.6339

Table 2. MAE with the languages L_1, L_2 and L_3.

Neighborhood	**Sampling**	L_2	L_4	L_5	L_6	L_7
5	100	0.7493	0.7316	0.6940	0.6911	0.6761
5	500	0.7314	0.7300	0.6812	0.6809	0.6633
10	100	0.6850	0.6788	0.6503	0.6404	0.6311
10	500	0.6745	0.6720	0.6309	0.6282	0.6225
15	100	0.6545	0.6521	0.6305	0.6227	0.6207
15	500	0.6526	0.6488	0.6176	0.6168	0.6140
20	100	0.6415	0.6409	0.6217	0.6196	0.6173
20	500	0.6390	0.6339	0.6162	0.6150	0.6087

Table 3. MAE with the languages L_2,L_4,L_5,L_6 and L_7.

Neighborhood	**Sampling**	L_8	L_9	L_{10}
5	100	0.7187	0.6781	0.6629
5	500	0.7100	0.6706	0.6564
10	100	0.6662	0.6386	0.6211
10	500	0.6609	0.6255	0.6111
15	100	0.6361	0.6201	0.6196
15	500	0.6322	0.6102	0.6072
20	100	0.6255	0.6179	0.6160
20	500	0.6237	0.6050	0.5912

Table 4. MAE with the languages L_8,L_9 and L_{10}.

Table 5 shows the results on HetRec2011 dataset, using a 10-fold cross-validation, comparing the proposed framework with respect to neighborhood-based recommendation methods reported in [2]. The approach proposed in [2] exploit also the tags assigned by the users in order to extract latent semantics by using Latent Semantic Analysis. The first recommender was based on collaborative filtering using the cosine similarity to build user neighbourhoods, the second uses content analysis on latent topic analysis, while the third was based on a simple average rating. As we can see in Table 5, even without using tag information, the obtained results adopting our system are better than, or comparable to, those obtained with the approaches exploited in [2].

Method	**RMSE**
Average Recommender Rating [2]	1.0880
Content Analysis [2]	0.9436
Collaborative Filtering [2]	0.8876
L_7	0.9071
L_9	0.9005
L_{10}	0.8891

Table 5. RMSE error on HetRec2011 adopting 10-fold cross-validation

6 Conclusions

In this paper a framework based on uncertain (multi)graphs able to deal with information filtering problems in DL has been presented. The evaluation of the proposed approach has been reported by applying it to a real world dataset and proving its validity in solving simple and complex information filtering tasks when compared with respect to other competing systems. In particular, we studied the behavior of the system by varying the neighborhood considered for each node, by varying the inference accuracy, and by considering contextual information. We have noticed that the contextual information provides a very strong improvement, especially for those regular expressions that make use of short paths and that consider the similarity of users and objects as something detached from the context.

References

1. Adomavicius, G., Tuzhilin, A.: 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 17(6), 734–749 (2005)
2. Bothos, E., Christidis, K., Apostolou, D., Mentzas, G.: Information market based recommender systems fusion. In: Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems. pp. 1–8. ACM (2011)

3. Burke, R.: The adaptive web. chap. Hybrid web recommender systems, pp. 377–408. Springer-Verlag (2007)
4. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model. Foundations for Digital Libraries (2007)
5. Colbourn, C.J.: The Combinatorics of Network Reliability. Oxford University Press (1987)
6. Gao, F., Xing, C., Du, X., Wang, S.: Personalized service system based on hybrid filtering for digital library. Tsinghua Science & Technology 12(1), 1–8 (2007)
7. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explorations 7(2), 3–12 (2005)
8. Goldberg, D.S., Roth, F.P.: Assessing experimentally derived interactions in a small world. Proceedings of the National Academy of Sciences 100(8), 4372–4376 (2003)
9. Hintsanen, P., Toivonen, H.: Finding reliable subgraphs from large probabilistic graphs. Data Min. Knowl. Discov. 17(1), 3–23 (2008)
10. Itmazi, J.A., Megías, M.G.: Using recommendation systems in course management systems to recommend learning objects. Int. Arab J. Inf. Technol. 5(3), 234–240 (2008)
11. Jin, R., Liu, L., Ding, B., Wang, H.: Distance-constraint reachability computation in uncertain graphs. Proc. VLDB Endow. 4, 551–562 (2011)
12. Nguyen, S.H., Chowdhury, G.: Digital library research (1990-2010): A knowledge map of core topics and subtopics. In: ICADL. pp. 367–371 (2011)
13. Popescul, A., Ungar, L.H.: Statistical relational learning for link prediction. In: IJCAI03 Workshop on Learning Statistical Models from Relational Data (2003)
14. Potamias, M., Bonchi, F., Gionis, A., Kollios, G.: k-nearest neighbors in uncertain graphs. Proc. VLDB Endow. 3, 997–1008 (2010)
15. Taranto, C., Di Mauro, N., Esposito, F.: Probabilistic inference over image networks. Italian Research Conference on Digital Libraries 2011 CCIS 249, 1–13 (2011)
16. Taranto, C., Di Mauro, N., Esposito, F.: Uncertain graphs meet collaborative filtering. In: 3rd Italian Information Retrieval Workshop (2012)
17. Taskar, B., Wong, M.F., Abbeel, P., Koller, D.: Link Prediction in Relational Data. In: in Neural Information Processing Systems (2003)
18. Witsenburg, T., Blockeel, H.: Improving the accuracy of similarity measures by using link information. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Ras, Z.W. (eds.) ISMIS. Lecture Notes in Computer Science, vol. 6804, pp. 501–512. Springer (2011)
19. Zhen-ming, Y., Tianhao, Y., Jia, Z.: A social tagging based collaborative filtering recommendation algorithm for digital library. In: ICADL. pp. 192–201 (2011)
20. Zou, Z., Gao, H., Li, J.: Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 633–642. ACM (2010)
21. Zou, Z., Li, J., Gao, H., Zhang, S.: Finding top-k maximal cliques in an uncertain graph. International Conference on Data Engineering pp. 649–652 (2010)
22. Zou, Z., Li, J., Gao, H., Zhang, S.: Mining frequent subgraph patterns from uncertain graph data. IEEE Transactions on Knowledge and Data Engineering 22, 1203–1218 (2010)