Evidence recommendation in forensics based on cyclic meta-paths in heterogeneous information networks

Eugenia Koblents, Federico Álvarez

Grupo de Aplicación de Telecomunicaciones Visuales, Universidad Politécnica de Madrid,
Ramiro de Maeztu 7, Madrid, Spain, ekl@gatv.ssr.upm.es

Abstract

In this article we propose a novel approach for evidence recommendation in digital forensic applications. The so-called Evidence Graph (EG) is a Heterogeneous Information Network (HIN) constructed from entities and relationships extracted from large collections of low quality video footage, collected by surveillance cameras in relation to a crime. The EG has been developed in the context of the EU-funded project LASIE. It presents multiple types of nodes and edges and a generic network schema in which all types of nodes can be connected. In this work, relevant cyclic meta-paths are identified and an iterative scheme for evidence recommendation is proposed and evaluated on real data.

1 Introduction

In this work we address the problem of evidence recommendation in digital forensic applications. In particular, we focus on the efficient exploration of large amounts of video footage collected in relation to the investigation of a crime. When exploring large amounts of video footage, investigators often invest big amounts of time and effort watching videos searching for relevant content. However, usually most of the collected video content is actually irrelevant to the investigation.

The aim of this work is to develop an evidence recommendation scheme for forensic investigations in order to speed-up the data exploration process. The system recommends detections extracted from videos collected at different locations and times, that can be related to a suspect or person involved in a crime. The goal is to find high quality detections of the people of interest in order to help their identification and tracking.

The proposed approach is based on the analysis of Heterogeneous Information Networks (HINs), i.e., graphs containing nodes and edges of multiple types. The HIN under study is a so-called Evidence Graph (EG), a multilayer network created based on detections and relationships extracted from low quality videos, by means of advanced video analysis techniques. In particular, nodes represent detections of people and vehicles, and carry additional attributes including dominant color and spatial-temporal information. On the other hand, links represent closeness and similarity relationships among detections. Links can be present among all types of nodes, and thus the network schema of the EG is a complete graph.

The problem of HIN analysis has recently attracted the attention of data mining researchers due to the ubiquity of this kind of networks in practical applications [1]. Typical examples of HINs often used in the literature for evaluation purposes include the bibliographic DBLP and the movie IMDB networks [1]. However, HINs naturally appear in a great variety of other data analysis problems, such as the one addressed here.

Problems such as clustering, classification, ranking, link prediction and recommendation in HINs have been extensively studied during the past decade. In this work we focus on the closely related ranking and non-personalized recommendation problems. Several algorithms have been recently proposed for ranking [2, 1] and recommendation [3, 1] in HINs. However, the EG presents some particular features that hinder the application of many existing HIN mining techniques. The network schema of the EG is a complete graph, while most of the techniques proposed in the literature limit their scope to specific network structures, such as the star network schema. On the other hand, both the nodes and links in the EG present a high level of uncertainty, as they are generated based on the results of the analysis of low quality video. Finally, most of the existing recommendation techniques focus on personalized approaches, which are not applicable in this case as no personal scores are available for different users.

In this paper we identify a set of meta-paths which are relevant in this scenario, and propose an iterative procedure for non-personalized evidence recommendation based on the selected meta-paths. In particular, we focus on a particular type of restricted meta-paths, which describe cyclic closeness and similarity relationships, in order to identify reinforced relationships among detections and mitigate the effect of uncertainty in the original graph.

This paper is organized as follows. Section 2 introduces the theoretical concepts this work is based on and summarizes previous work relevant to this topic. Section 3 describes the structure and particular characteristics of the EG, and introduces the algorithm proposed for suspect re-identification and recommendation for related evidence. Section 4 presents the results obtained on the real dataset under study. Finally, section 5 summarizes the conclusions of the paper.

2 Background on HINs and meta-paths

The study of HINs has attracted increased attention of the data mining research community during the last decade and multiple HIN mining techniques have been applied on a variety of problems that require complex graph representations [1].

A heterogeneous information network is defined as a multi-typed directed or undirected graph, in which nodes and edges can be of different types. Most of the existing information network analysis techniques assume that the type of objects or links is unique. These homogeneous information networks are usually created by simply ignoring the heterogeneity of objects and links in real interacting systems. However, taking into account the rich semantics contained in complex multi-typed heterogeneous networks allows to extract meaningful information.

The network schema of a HIN is a simple graph that helps to describe its structure, showing the possible relationships among the different types of nodes. Some common examples of HIN schemes are the bipartite, star and multi-relation network schemes [1].

Differently from homogeneous networks, nodes in a HIN can be connected via different so-called meta-paths, which carry different physical meanings. A meta-path of length l is denoted as $\mathcal{P} = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$, where $A_1, A_2, \ldots, A_{l+1}$ and $R_1, R_2, \ldots, R_l$ denote possibly different node and relation types. The meta-path $\mathcal{P}$ defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ among objects. A meta-path is said to be symmetric if the relation defined by it is symmetric. For simplicity of notation, in the following we denote meta-paths as $\mathcal{P} = A_1 A_2 \ldots A_{l+1}(R_1 R_2 \ldots R_l)$ or simply $R_1 R_2 \ldots R_l$ if the focus is on the link types only. We say that a path $p = a_1 a_2 \ldots a_{l+1}(r_1 r_2 \ldots r_l)$ between a_1 and a_{l+1} is an instance of $\mathcal{P}$ if it follows the meta-path $\mathcal{P}$.

The rich semantics of meta-paths are an important and distinctive characteristic of HINs. Meta-paths describe diverse semantic relationships among objects. For example, in a bibliographic network, authors can be similar in terms of common conferences or common co-authors, and both relationships are represented via different semantic meta-paths.

In this work we focus on meta-path dependent similarity search, recommendation and ranking tasks. Similarity search allows to identify semantically related items in a HIN, optionally considering a query node [4]. In the following, two different similarity measures used in this work are described, PathSim and HeteSim. Relevance recommendation aims at identifying multi-typed items related to a query object in terms of meaningful semantic relationships [3]. Finally ranking in HINs allows to identify the most relevant multi-typed nodes according to the selected meta-path [2].

2.1 Similarity measures in HINs

Similarity search aims at identifying similar objects in a network in terms of a similarity measure of interest. Similarity measures are the basis of many other data mining tasks, such as clustering, ranking and recommendation [1].

PageRank [5] is a well known link based approach that measures the similarity of objects based on their link structure in a homogeneous graph. Similarity measures among objects in HINs take into account the meta-paths connecting them in the network. Different meta-paths contain different semantic meanings, which may lead to different similarities. Thus, similarity measures in HINs are meta-path constrained.

PathSim is a meta-path based similarity measure first proposed in [4] to evaluate the similarity of same-typed objects based on symmetric paths. The intuition behind it is that two similar peer objects should not only be strongly connected, but also share comparable visibility. As the relation of peer should be symmetric, PathSim is limited to symmetric meta-paths [4].

Given a symmetric meta-path $\mathcal{P}$, PathSim between two objects of the same type x and y is defined as

$$PathSim(x, y | \mathcal{P}) =$$
$$\frac{2 \times |\{p_{x \to y} : p_{x \to y} \in \mathcal{P}\}|}{|\{p_{x \to x} : p_{x \to x} \in \mathcal{P}\}| + |\{p_{y \to y} : p_{y \to y} \in \mathcal{P}\}|},$$

where $p_{x \to y}$ is a path instance between x and y, $p_{x \to x}$ is that between x and x, and $p_{y \to y}$ is that between y and y [4]. Thus, given a meta-path $\mathcal{P}$, PathSim between x and y is defined in terms of the number of paths between them following $\mathcal{P}$, and the number of path instances between themselves [4].

The HeteSim similarity measure was proposed in [6] in order to measure the relevance of any object pair under arbitrary meta-paths. Given a relevance path P the HeteSim score between two objects x and y is defined as

$$HeteSim(x, y | \mathcal{P}) = \frac{1}{|O(x|R_1)||I(y|R_l)|} \sum_{i=1}^{|O(x|R_1)|} \sum_{j=1}^{|I(y|R_l)|}$$
$$HeteSim(O_i(x|R_1), I_j(y|R_l) | R_2 \circ \ldots \circ R_{l-1})$$

where $O(x|R_1)$ denotes the out-neighbors of x based on relation R_1, and $I(y|R_l)$ denotes the in-neighbors of y based on relation R_l. The computation of HeteSim needs to iterate over all pairs of $(O_i(x|R_1), I_j(y|R_l))$ of (x, y) along the path and sum up the relatedness of these pairs. It is normalized by the total number of out-neighbors of x and in-neighbors of y. HeteSim measures how likely x and y will meet at the same node when x goes along the path and y goes against the path [6].

3 Iterative evidence recommendation

In this section the HIN under study, namely the EG, is described, the concept of cyclic meta-paths is introduced and an algorithm proposed for evidence recommendation based on cyclic meta-paths is presented.

3.1 Evidence Graph

The EG is a HIN representing detections that appear in a large collection of videos and the interrelationships among them. The EG has been created based on a real video dataset provided by the Metropolitan Police of London for the design and implementation of a forensic data analysis system in the framework of the LASIE research project. The dataset consists of a
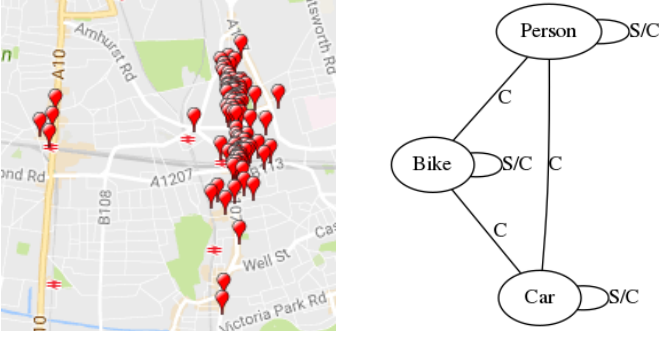
Figure 1. Location of the videos used for the EG creation and EG network schema.

collection of 140 videos recorded by a set of 44 cameras during the riots in London in 2010. The locations of the cameras are depicted in Figure 1 (*left*).

This dataset contains sensible information, including images of people and vehicles that could allow for their identification. For this reason, a data sharing agreement has been signed among the involved parties and competent Data Protection Authorities have been informed of the processing of personal data. Additionally, the UPM ethical committee has validated the processing of this dataset.

The EG is built from data produced by advanced video analysis tools, including person and vehicle recognition and tracking, logo and plate detection, etc. These tools produce a very large number of detections of different people and vehicles. Each node in the EG represents a detected track of a person or vehicle, appearing in contiguous frames in a video. Thus, many nodes in the graph can represent different detections of the same person or vehicle. Additionally, the nodes can be annotated with attributes such as dominating color, spatial-temporal information, the presence of a logo, etc. Nodes in the EG are thus potential pieces of evidence that can be relevant for a criminal investigation.

The EG interconnects these objects with various types of edges. In particular, in this work we focus on the closeness and visual similarity links. Closeness is defined as the coexistence relationship between two tracks in the same frame of a video. Nodes of different types can be liked via the closeness relationship, for example, a car and a person can be linked as *close* in the EG if they appear together in some frame. On the other hand, similarity links connect nodes of the same type that present visual similarity, which is computed based on color histograms of the representative detections of both tracks. Both closeness and similarity links can carry weights denoting the strength of the relationship.

The EG is stored in an OrientDB [7] graph database and contains a total of 50K person tracks, 20K vehicle tracks, 700K similarity and 500K closeness edges. The EG can be accessed via an OrientDB interface in python, java, etc.

We denote nodes of *person* type as $P_1, P_2, \ldots$, and nodes of *car* or *bike* types as $V_1, V_2, \ldots$. On the other hand, we denote similarity links as S and closeness links as C. Figure 1

(*right*) depicts the network schema of the EG, which is a complete graph in which all types of nodes can be connected by different types of links. This network schema does not fit any of the standard network structures described in the literature (bipartite, multi-relational or star schema). The EG presents some other particular features that make it different from many networks described in the literature and hinder the application of existing techniques:

- Many networks widely studied in the literature (e.g., the bibliographic network [1]), contain only certain entities and relationships. On the contrary, nodes and links in the EG present a high degree of uncertainty, due to errors in the detection and tracking process, and the limitations of the similarity computations based on color histograms.

- While most of the existing HIN recommendation techniques focus on personalized approaches [1], in this problem we focus on non-personalized recommendations, since no ratings or any other source of explicit or implicit information on personal preferences is available.

- Due to the vast amount of nodes and links in the EG, in this scenario ranking and recommendation techniques that only require local computations are preferred.

In the next section we propose a set of meta-paths relevant in this scenario and an iterative algorithm for evidence recommendation that addresses the previous challenges.

3.2 Cyclic meta-paths in the EG

In this paper we focus on two different evidence recommendation scenarios. On one hand we address the person re-identification problem, which aims at finding unseen detections of a person of interest in other frames or videos, in possibly different locations and times. The goal in this case is to find video captures where the person can be more easily identified, due to a closer capture, a better angle, better light conditions, etc. On the other hand, we are interested in identifying other entities (people or vehicles) that repeatedly appear close to the person of interest and may thus be related to it.

In this application, the goal is to identify and locate the people involved in acts of violence and vandalism during the riots in London in 2010. However, the same approach can be used in the investigation of missing person cases, where the goal is to identify and locate the suspect, the victim, and other people or vehicles possibly linked to the crime.

In this application we focus on local query-based rather than global network analysis, in order to allow for efficient computations and the generation of fast recommendations. The user selects a set of initial query EG nodes, representing relevant detections of a person of interest. The recommender system then retrieves other potentially relevant detections.

As already mentioned in the previous section, in the analysis of the EG we deal with a high level of uncertainty, both in the nodes and links. As an intuitive way of managing the uncertainty in the graph, we propose to focus on cyclic meta-paths. The idea is to identify cyclic similarity and closeness

Path name	Involved nodes and links types	Meta-path interpretation
PP(S), VV(S)	$P_1 \xrightarrow{S} P_2, V_1 \xrightarrow{S} V_2$	Direct similarity relationship among people (or vehicles).
PPP(SSS)	$P_1 \xrightarrow{S} P_2 \xrightarrow{S} P_3 \xrightarrow{S} P_1$	Cyclic similarity of length 3 among people.
PPPP(SSSS)	$P_1 \xrightarrow{S} P_2 \xrightarrow{S} P_3 \xrightarrow{S} P_4 \xrightarrow{S} P_1$	Cyclic similarity of length 4 among people.
PPPP(CSCS)	$P_1 \xrightarrow{C} P_2 \xrightarrow{S} P_3 \xrightarrow{C} P_4 \xrightarrow{S} P_1$	Cyclic closeness to find similar pairs of people.
PVVP(CSCS)	$P_1 \xrightarrow{C} V_1 \xrightarrow{S} V_2 \xrightarrow{C} P_2 \xrightarrow{S} P_1$	Cyclic closeness to find similar pairs person-vehicle.

Table 1. Relevant meta-paths for evidence recommendation and their interpretation.

relationships and rank highly those nodes in the EG that belong to strong cycles, as they are more likely to have trustful relationships with a query node.

A cyclic meta-path $\mathcal{C}$ satisfies the constraint that the first and last nodes in the path coincide, i.e., $\mathcal{C} = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_1$. For instance, the meta-path PPP(SSS) represents the linear path $P_1 \xrightarrow{S} P_2 \xrightarrow{S} P_3 \xrightarrow{S} P_4$ with the cycle constraint $P_1 = P_4$. In cyclic meta-paths the number of nodes and links is the same. Cyclic meta-paths can be viewed as a particular case of constrained meta-paths, defined in [8]. Additional constraints can be added to the definition of the meta-path, for instance, on the color or spatial-temporal attributes of nodes. However, in this work we only consider cycle constraints.

The semantic meta-paths under study in this work are described in Table 1. The first meta-path of interest corresponds to the direct similarity relationship among nodes of the same type (people or vehicles). This basic relationship has been used in the simulations for comparison purposes. The cyclic similarity paths PPP(SSS) and PPPP(SSSS), of length 3 and 4 respectively, allow to identify reliable similarity relationships among detections. Finally, the cyclic meta-paths PPPP(CSCS) and PVVP(CSCS) allow to detect similarity among pairs of people or people-vehicles, that repeatedly appear together in the videos. Similarly, cyclic meta-paths relating any other types of nodes can be defined.

3.3 Evidence recommendation algorithm

The goal of the proposed evidence recommendation procedure is twofold. On one hand, it allows to recommend evidence (EG nodes) visually *similar* to one or more query nodes. On the other hand, it recommends evidence *related* to the nodes in terms of similarity and closeness relationships.

The input to the recommendation algorithm is a set of query nodes Q_i, $i = 1, \ldots, N_Q$, provided by the user, representing one or multiple detections of a person of interest. A local query subnetwork is constructed by traversing the EG from the query nodes with a traversal depth of two links.

The two proposed recommendation features focus on two different kinds of semantic meta-paths. The first one corresponds to ranking according to the cyclic similarity meta-paths PPP(SSS) and PPPP(SSSS), while the second one corresponds to the cyclic meta-paths PPPP(CSCS) and PVVP(CSCS). In both cases, the output of the algorithm is a ranked list of K

recommended EG nodes, which maximize the selected similarity measure with the query nodes.

For the recommendation of *similar* evidence, the PathSim similarity measure along the cyclic similarity meta-paths of length 3 and 4 is computed. A weighted homogeneous network consisting of the people nodes involved in the cyclic meta-paths is then constructed. The weighted version of the PageRank algorithm, namely TextRank [9], is then applied to rank the nodes in the local similarity subnetwork of the query nodes and the nodes with the highest scores are recommended. The accepted recommendations are included in the set of query nodes for the next iteration. This is an iterative procedure in which the local similarity network is extended along the iterations, based on the accepted recommendations.

For the recommendation of *related* evidence, the HeteSim similarity measure has been used to evaluate the strength of the existing cycles in the local network of the query nodes, as HeteSim is suitable for asymmetric meta-paths. A weighted heterogeneous subnetwork with N_T track nodes N_j, for $j = 1, \ldots, N_T$, is constructed containing only the obtained cycles. For the PPPP(CSCS) meta-path we are interested in detections P_3, which correspond to person nodes that satisfy the CS and SC meta-paths with the query node P_1. On the other hand, for the PVVP(CSCS) we focus on detections V_2, which represent vehicles linked via the CS and SC meta-paths with the query node P_1. Finally, we rank the nodes in the generated subnetwork containing only cycles, and recommend the nodes that maximize the score w_j computed as the total HeteSim measure with the set of query nodes along the selected meta-path $\mathcal{P}$, i.e.,

$$w_j = \sum_{i=1}^{N_Q} HeteSim(Q_i, N_j | \mathcal{P}), \quad j = 1, \ldots, N_T \quad (1)$$

The user validates the recommendations by selecting the relevant ones. For *similar* evidence recommendation, the accepted nodes are added to the set of query nodes to extend the local similarity subnetwork. Rejected nodes are not recommended again in future iterations. The proposed algorithm is summarized in Table 2.

The proposed algorithm is related to the relevance recommendation algorithm HeteRecom [3], which automatically identifies relevant semantic meta-paths in HINs, starting from a query node. However, most of the possible semantic meta-paths are not meaningful in this scenario, due to the high uncertainty in the EG links. In particular, linear meta-paths may

attain high relevance weights based on the graph topology but actually lead to severe error propagation, as many of the similarity links are not reliable. For this reason, cyclic meta-paths have been considered in this work instead.

Table 2. Evidence recommendation algorithm.

The input to the algorithm is a set of N_Q query nodes representing detections of a person of interest, as well as the semantic meta-path of interest. The output at each iteration is a ranked list of K recommended detections.

At each iteration:

1. Obtain the local subnetwork of the N_Q query nodes by traversing the EG with a traversal depth of two links.

2. Identify the selected cyclic meta-paths in the query subnetwork, construct a graph containing only nodes and links involved in cyclic relationships and recommend the nodes with the highest rank:

   - For the recommendation of *similar* evidence, cyclic similarity meta-paths of length 3 and 4 are considered. The resulting network containing only similarity cycles is homogeneous and ranking is performed via the weighted PageRank algorithm [9].

   - For the recommendation of *related* evidence, cyclic meta-paths of the form CSCS are considered and ranking is performed on the set of nodes P_3 or V_2 (when ranking related people or vehicles, respectively) according to equation (1).

3. The user selects relevant recommendations. Accepted *similar* EG nodes are added to the set of query nodes.

4 Simulation results

The proposed algorithm has been implemented in python and run on the EG created from the available video dataset. The person track highlighted in Figure 2 with a black bounding box has been used as an initial query node.

The proposed recommendation algorithm has been run for 5 iterations, recommending $K = 5$ new EG nodes at each iteration, yielding a total of 25 recommendations. The obtained results are shown in Table 3, together with the acceptance ratios obtained for the different recommendation approaches.

Detections directly connected to the query node via the 8 strongest similarity links are represented in the first row. It can be observed that most of them are visually similar, but only 3 out of a total of 10 detections actually correspond to the same person (shown with a check mark in the table), yielding an acceptance ratio of 30%.

The detections listed in the second row of Table 3 are obtained based on the PageRank scores computed for the local similarity subnetwork of the query node. Out of a total of 25 PageRank-based recommendations, 8 detections correspond

track #59:939

Figure 2. The track corresponding to the initial query node.

to the person of interest, yielding an acceptance ratio of 32%, which is only slightly higher than the result obtained with direct similarity search.

The detections shown in the third row are recommendations based on cyclic similarity links that really represent the person of interest. In this case, 12 relevant detections of the person of interest have been obtained out of a total of 25 recommendations. The proposed algorithm thus retrieves a 48% of relevant results in this case, yielding a 60% improvement with respect to the direct similarity search, and a 50% improvement with respect to the PageRank results. It can be observed that some of the obtained detections have better quality than others and allow to identify the person of interest more easily. However, all of the detections, even the poor ones, allow to reconstruct the trajectory of the suspect, based on the spatial-temporal attributes of the involved nodes.

The recommendations shown in the fourth row have been obtained based on the PPPP(CSCS) meta-path, that is, they correspond to people that repeatedly appear in the videos close to the person of interest. It can be observed that some of them appear in the query track video and also in other videos. In this case, from 25 recommendations, 9 really correspond to people that appear together with the person of interest at least twice, yielding an acceptance ratio of 36%.

Finally, in the fifth row, detections of vehicles appearing close to the suspect in different videos are shown. These recommendations have been obtained based on the PVVP(CSCS) cyclic meta-path. In this last case, from a set of 25 recommendations, at least 6 do correspond to vehicles that appear close to the person of interest in at least two different tracks. In the rest of cases the relevance of recommendations is hard to asses given the low quality of the detections. For this reason, in this case we consider an acceptance ratio of at least 24%.

Both CSCS meta-paths allow to identify other potential suspects or vehicles related to a criminal activity. The investigator can select some of the recommended detections and further apply the proposed techniques to find other detections, that are similar or related to the new query nodes.

Direct similarity	✓	✓	✓	✗	✗	✗	✗	✗	3/10 = 0.30
PageRank	✓	✓	✓	✓	✓	✓	✓	✓	8/25=0.32
Cyclic similarity	✓	✓	✓	✓	✓	✓	✓	✓	12/25=0.48
Related people	✓	✓	✓	✓	✓	✓	✓	✓	9/25=0.36
Related vehicles	✓	✓	✓	✓	✓	✓	?	?	$> 6/25 = 0.24$

Table 3. Recommended results based on direct similarity, PageRank, cyclic similarity meta-paths PPP(SSS) and PPPP (SSSS), and related people and vehicles detections obtained based on the PPPP(CSCS) and PVVP(CSCS) meta-paths, respectively.

5 Conclusions

In this paper, an evidence recommendation approach for a specific type of Heterogeneous Information Network known as the Evidence Graph (EG) has been proposed. The EG represents detections of people and vehicles extracted from videos and relationships among them. A set of cyclic semantic meta-paths has been identified and an iterative evidence recommendation technique has been proposed. The new method has been evaluated on an EG constructed based on a real video dataset.

Acknowledgements

References

[1] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2017.

[2] Y. Li, C. Shi, Yu P. S., and Q. Chen. Hrank: a path based ranking method in heterogeneous information network. In *International Conference on Web-Age Information Management*, pages 553–565. Springer, 2014.

[3] C. Shi, C. Zhou, X. Kong, P. S. Yu, G. Liu, and B. Wang. Heterecom: a semantic-based recommendation system in heterogeneous networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1552–1555. ACM, 2012.

[4] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[6] C. Shi, X. Kong, Y. Huang, P. S. Yu, and Bin Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2479–2492, 2014.

[7] http://orientdb.com.

[8] C. Shi, Y. Li, P. S. Yu, and B. Wu. Constrained-meta-path-based ranking in heterogeneous information network. *Knowledge and Information Systems*, 49(2):719–747, 2016.

[9] R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *EMNLP*, volume 4, pages 404–411, 2004.