A New Relevance Measure for Heterogeneous Networks

Mukul Gupta(✉), Pradeep Kumar, and Bharat Bhasker

Indian Institute of Management, Lucknow, India
{fpm13008,pradeepkumar,bhasker}@iiml.ac.in

Abstract. Measuring relatedness between objects (nodes) in a heterogeneous network is a challenging and an interesting problem. Many people transform a heterogeneous network into a homogeneous network before applying a similarity measure. However, such transformation results in information loss as path semantics are lost. In this paper, we study the problem of measuring relatedness between objects in a heterogeneous network using only link information and propose a meta-path based novel measure for relevance measurement in a general heterogeneous network with a specified network schema. The proposed measure is semi-metric and incorporates the path semantics by following the specified meta-path. For relevance measurement, using the specified meta-path, the given heterogeneous network is converted into a bipartite network consisting only of source and target type objects between which relatedness is to be measured. In order to validate the effectiveness of the proposed measure, we compared its performance with existing relevance measures which are semi-metric and applicable to heterogeneous networks. To show the viability and the effectiveness of the proposed measure, experiments were performed on real world bibliographic dataset DBLP. Experimental results show that the proposed measure effectively measures the relatedness between objects in a heterogeneous network and it outperforms earlier measures in clustering and query task.

Keywords: Heterogeneous network · Meta-path · Clustering · Query task · Relevance measure

1 Introduction

Measuring relatedness between objects (represented as nodes) in a heterogeneous network for different mining tasks has gained the attention of researchers and practitioners due to the information rich results. To mine a heterogeneous network, many researchers transform the heterogeneous network into the corresponding homogenous network before applying different mining techniques [1–3]. However, mining of heterogeneous-transformed homogeneous network involves information loss as path semantics are lost [1]. Due to this, there is a surge of studies for measuring relatedness between objects. Relevance measurement for objects in a heterogeneous network has become important for various mining tasks like clustering, classification and query [1, 2].

S. Madria and T. Hara (Eds.): DaWaK 2015, LNCS 9263, pp. 165–177, 2015.
DOI: 10.1007/978-3-319-22729-0_13

Mining of heterogeneous information networks has become important as they are ubiquitous and play a critical role in modern information infrastructure [1, 3, 4]. For example, a bibliographic network can be modelled as a heterogeneous network and typically has nodes representing papers, authors, conferences and keywords. The relationship of nodes are represented by links in the network. Since in a heterogeneous network different typed objects and different typed relationships co-exist, so measuring relatedness between different typed objects, following different relationship paths, can give interesting results and may reflect the real nature of data. Figure 1 shows the network schema of DBLP database [2]. In Fig. 1, a bidirectional link exists between Paper and Author indicating that every paper is associated with author(s) and vice versa. Similarly bidirectional links exist between Paper and Keyword as well as Paper and Conference. A paper may be cited by another paper hence a self-loop exists at Paper node.

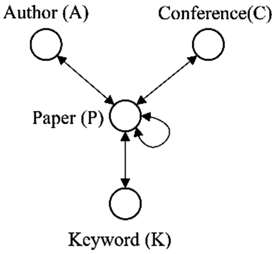


Fig. 1. Network schema of DBLP heterogeneous network data

Different mining and query tasks can be performed on a heterogeneous network to answer different questions. For example, in case of DBLP network, we can answer questions like "*Who are the peer researchers of a specified author?*", "*Who are the leading researchers in Information Retrieval area?*", "*How are Computer Science research areas structured?*" and many others. To answer these questions, we need to perform different mining tasks like clustering, classification, ranking etc. over different meta-paths (meta-level description of paths) [2–4]. For these mining tasks, measuring relatedness between objects is an important step. For example, in case of DBLP dataset, in order to answer the question like "*Who are the leading researchers in Information Retrieval area?*" we need to measure the relevance of Authors (A) with the Conferences (C) related to information retrieval area. Also, in case of clustering, we need to measure the relatedness between same typed objects following a relationship path involving different typed objects.

In this paper, we propose a meta-path based novel relevance measure which can measure the relatedness between same as well as different typed objects in a heterogeneous network. The proposed measure is semi-metric meaning it has important properties of reflexivity, symmetry and limited range [5]. In addition, the proposed measure does not require decomposition of an atomic relation when source and target objects are of different type and, thus, reduces the computational requirement [4]. Further, the proposed measure can be used to measure the relatedness between objects following any arbitrary meta-path. To validate the effectiveness of the proposed measure, we use real world bibliographic dataset from DBLP and perform clustering and query task [4, 6]. For comparison, we use other relevance measures which are semi-metric and applicable to heterogeneous networks, namely, HeteSim and PathSim [3, 4]. For clustering task, we use both PathSim and HeteSim and compare the performance with the proposed measure. But for query task, we compare the performance of the proposed measure only with HeteSim as PathSim cannot measure similarity between different typed objects [3].

The rest of the paper is organized as follows. Section 2 introduces the related work. In Sect. 3, we present our proposed relevance measure. An illustration is given in Sect. 4. Section 5 presents experimental setup and results. Finally, in Sect. 6 conclusion and future research directions are given.

2 Related Work

Information networks can be broadly classified as homogeneous and heterogeneous networks [1]. Mining of homogeneous networks has been attempted by many researchers. Research done by Jeh and Widom [7] and Page et al. [8] are two such examples. However, mining of heterogeneous networks is relatively an emerging research area and has been attempted by few researchers [1, 2]. Heterogeneous networks are richer in information as compared to their heterogeneous-transformed homogeneous counterpart [1]. Earlier, researchers have used various similarity/distance measures to compute the relatedness between objects [5, 9]. Conventional similarity/distance measures like Jaccard coefficient, Cosine similarity and Euclidean distance are features based measures and not suitable to use directly on heterogeneous networks [4, 5, 10]. Link based similarity measures such as PageRank [8] and SimRank [7] are widely accepted but these are limited to use in homogeneous networks.

Measuring relatedness between objects in a heterogeneous information network has gained momentum recently and there are only few measures which are directly applicable to heterogeneous networks and can incorporate the path semantics like PCRW (Path Constrained Random Walk), PathSim, HeteSim and AvgSim. PCRW, HeteSim and AvgSim use random walk based methods while PathSim uses path count for relevance measurement [3, 4, 11, 12]. For all the four measures, meta-path based approach is applied for measuring relatedness between objects [3, 4, 11, 12].

PCRW, proposed by Lao and Cohen [11], is not a symmetric measure, which means relatedness between two objects will not be equal in forward and reverse directions and it is a major limitation with PCRW [4]. Sun et al. [3] proposed PathSim to measure the relatedness between objects in a heterogeneous network. But the limitation with PathSim is that it can measure relatedness only between same typed objects following an even length symmetric meta-path. Since, in many practical applications it may be required to measure the relatedness between different typed objects like mining of movies database, Flickr, Social network data, in that situation, PathSim would not be applicable.

In order to measure the relatedness between different as well as same typed objects, Shi et al. [4] proposed HeteSim. HeteSim has shown improved performance as compared to both PCRW and PathSim [4]. Although better than PCRW and PathSim, HeteSim has its own limitation. The limitation with HeteSim is that in order to measure the relatedness between objects, if the length of meta-path is odd, we need to do the decomposition of atomic relations to make the length of meta-path even which is computationally intensive. Both PathSim and HeteSim are semi-metric. Another work done by Meng, X. et al. [12] proposed a random walk based measure AvgSim which, although, has symmetric property but is not semi-metric which makes it suitable only for limited mining techniques/algorithms.

The prior overview shows that relevance measurement for objects in a heterogeneous network is an area which has lot of potential for research. This motivated us to design a new relevance measure which is more effective as compared to earlier measures and also addresses their limitations.

3 A Novel Relevance Measure

Mining of a heterogeneous network can give information rich results and by following different paths, we can capture different semantics and subtleties which is not possible in case of heterogeneous-transformed homogeneous network. In this section, we propose a novel relevance measure to compute the relatedness between same as well as different typed objects for any arbitrary meta-path. Building upon the framework described by Sun et al. [3], first, we define information network and network schema, middle object type, middle relation type and weighted path matrix which will be used in formulating the new measure.

Definition 1 (Information Network and Network Schema). *An information network is defined as a directed graph* $G = (V, E)$ *with an object type mapping function* $\emptyset : V \rightarrow A$ *and a link type mapping function* $\psi : E \rightarrow R$*, where each object* $v \in V$ *belongs to one particular object type* $\emptyset(v) \in A$*, and each link* $e \in E$ *belongs to a particular relation type* $\psi(e) \in R$*. However, the network schema is the meta-level representation of* $G = (V, E)$ *which is a directed graph over object types A and relation types R, denoted as* $T_G = (A, R)$*.*

When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is heterogeneous information network; otherwise, it is homogeneous information network.

Definition 2 (Middle Object Type). *For an even length meta-path* $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_{m-1}} A_m \xrightarrow{R_m} \dots \xrightarrow{R_{l-1}} A_l \xrightarrow{R_l} A_{l+1}$*, where l is even, the middle object type* A_m *is the object type which is equidistant from source object type* A_1 *as well as target object type* A_{l+1}*.*

For DBLP network schema shown in Fig. 1, consider the meta-paths *APA* ("Author-Paper-Author") and *APCPA* ("Author-Paper-Conference-Paper-Author") which have middle object types *P* ("Paper") and *C* ("Conference") respectively. All symmetric meta-paths of length more than one are essentially even length path and, therefore, have middle object type. For meta-path *APKPC*, which is not a symmetric path, the middle object type is *K* and for meta-path *APCP* there is no middle object type as the length of meta-path is odd. For a meta-path, using middle object type, we can divide the path into two equal length paths i.e.one from source to middle object type and other from middle to target object type. For example, path *APCPA* can be divided into *APC* and *CPA*.

Definition 3 (Middle Relation Type). *For an odd length meta-path* $A_1 \xrightarrow{R_1} A_2 \dots \xrightarrow{R_m} \dots A_l \xrightarrow{R_l} A_{l+1}$*, where l is odd, the middle relation type* R_m *is the relation type which has equal number of preceding and succeeding relation types from source to target object type.*

For meta-path *APCP* in DBLP network schema which has odd length, the middle relation type is *PC*.

For measuring relatedness between source and target object, we can transform a heterogeneous network into a bipartite network consisting of only source and target type objects. For that, we need to compute the weighted path matrix as given in Definition 4.

Definition 4 (Weighted Path Matrix). *For a heterogeneous network and its schema level representation, a weighted path matrix M for meta-path* $P = (A_1A_2 \ldots A_{l+1})$ *is defined as* $M = W_{A_1A_2} \times W_{A_2A_3} \times \ldots \times W_{A_lA_{l+1}}$*, where* $W_{A_iA_j}$ *is the adjacency matrix between object types* A_i *and* A_j*.* $M(x_i, y_j)$ *represents the number of path instances between objects* $x_i \in A_1$ *and* $y_j \in A_{l+1}$ *following meta- path P.*

Now, we present the proposed relevance measure for measuring relevance for objects in a heterogeneous network.

Definition 5. *Given a meta-path* $P = (A_1A_2 \ldots A_{l+1})$*, for bipartite representation of a heterogeneous network that has only source and target type objects, relatedness between source object* $a_i \in A_1$ *and target object* $b_j \in A_{l+1}$ *is:*

$$Rel\left(a_i, b_j|P\right) = \frac{w_{a_ib_j}\left(\frac{1}{\deg(a_i)} + \frac{1}{\deg(b_j)}\right)}{\frac{1}{\deg(a_i)}\sum_j w_{a_ib_j} + \frac{1}{\deg(b_j)}\sum_i w_{a_ib_j}} \tag{1}$$

where $w_{a_ib_j}$ *is the value* $M(a_i, b_j)$ *from the weighted path matrix i.e. the number of paths connecting objects* a_i *and* b_j*.* $deg(a_i)$ *and* $deg(b_j)$ *are the degree of objects* a_i *and* b_j *respectively in bipartite representation.*

There can be three different cases for computing relatedness using proposed measure based on type of objects and length of meta-path. Below, we present all the three cases and present the formula for calculating the relevance using Definition 5.

Case I: Relevance Measurement for Different Typed Objects. When source and target objects are of different type, then regardless of length of the meta-path, the relatedness between source object $a_i \in A_1$ and target object $b_j \in A_{l+1}$ following meta-path $P = (A_1A_2 \ldots A_{l+1})$ is calculated by first computing the bipartite representation using Definition 4. Then we can calculate the relatedness $DPRel\left(a_i, b_j|P\right)$ between source and target objects using Definition 5.

$$DPRel\left(a_i, b_j|P\right) = Rel\left(a_i, b_j|P\right) \tag{2}$$

Case II: Relevance Measurement for Same Typed Objects When Path Length is Even. When source and target objects are of same type and length of meta-path is even then the relatedness between source object $a_i \in A_1$ and target object $a_j \in A_{l+1}$ is calculated as follows. First, we find the middle object type of meta-path *P* using Definition 2. In this case, the length of meta-path is even so $P = P_LP_R = (A_1A_2 \ldots A_m)(A_mA_{m+1} \ldots A_{l+1})$. For $P_L = A_1A_2 \ldots A_m$, we calculate

$Rel(a_i, b_k|P_L)$ for source object a_i and all middle objects $b_k \in A_m, \forall k$. Similarly, for target object a_j, we calculate $Rel(a_j, b_k|P_R^{-1})$ following $P_R^{-1} = A_{l+1}A_l \ldots A_m$. Then, we calculate relatedness $DPRel\left(a_i, a_j|P\right)$ between a_i and a_j as follows using Tanimoto coefficient [5]. Here middle objects are taken as attribute objects.

$$X = Rel(a_i, b_k|P_L) \tag{3}$$

$$Y = Rel(a_j, b_k|P_R^{-1}) \tag{4}$$

$$DPRel(a_i, a_j|P) = \frac{X.Y}{|X|^2 + |Y|^2 - X.Y} \tag{5}$$

where X and Y are the vectors of relevance of a_i and a_j respectively with all middle objects $b_k \in A_m, \forall k$ in the meta-path.

Case III: Relevance Measurement for Same Typed Objects When Path Length is Odd. This situation occurs rarely when length of the meta-path is odd for same type objects [3, 4]. In this case, we first find the middle relation type using Definition 3 and decompose instances of that middle relation type to create middle objects as described by Shi et al. [4]. This will result in a meta-path of even length. After that we will follow the same process as in case 2.

Thus, only in third case i.e. relevance measurement for same typed objects when path length is odd, we are required to do decomposition of atomic relations.

3.1 Properties of the Proposed Measure

Our proposed measure is semi-metric which makes it useful for many applications involving heterogeneous networks. Before applying the proposed measure, we need to convert the heterogeneous network into a bipartite network consisting only of source and target object types using Definition 4. Since in case of same source and target object types, we convert the odd length meta-path case into even length case by decomposing atomic middle relations, so there is no need to give a separate proof for this case. The proof of semi-metric properties [5] are given below for the two cases i.e. when the source and target objects are of same type and of different types.

Case I: Relevance Measurement for Different Typed Objects
Property 1 (Limited Range). $0 \le DPRel(a_i, b_j|P) \le 1$

Proof: Relatedness between two objects $a_i \in A_1$, and $b_j \in A_{l+1}$ following meta-path $P = A_1A_2 \ldots A_{l+1}$

$$DPRel\left(a_i, b_j|P\right) = \frac{w_{a_ib_j}\left(\frac{1}{\deg(a_i)} + \frac{1}{\deg(b_j)}\right)}{\frac{1}{\deg(a_i)}\sum_j w_{a_ib_j} + \frac{1}{\deg(b_j)}\sum_i w_{a_ib_j}}$$

If there is no path between a_i and b_j, then,$w_{a_ib_j} = 0$ and $Rel\left(a_i, b_j|P\right) = 0$

If a_i and b_j are connected to only each other but with no other node then, $\deg\left(a_i\right) = \deg\left(b_j\right) = 1$ and $\sum_j w_{a_ib_j} = \sum_i w_{a_ib_j} = w_{a_ib_j}$, therefore,

$$DPRel\left(a_i, b_j|P\right) = 2 \times \frac{w_{a_ib_j}}{w_{a_ib_j} + w_{a_ib_j}} = 1$$

The relevance value can never be negative since degree of a node and weight $w_{a_ib_j}$ can never be negative in case of meta-path based framework. Also, the degree of a node can never be zero because no isolated node will be present in the network.

Thus,

$$0 \le DPRel(a_i, b_j|P) \le 1$$

Property 2 (Reflexivity). $DPRel\left(a_i, b_j|P\right) = 1 iff a_i = b_j$

Proof: Since a_i and b_j are of different type therefore $a_i = b_j$ means that their connection patterns are same. This can happen only when they are connected to only each other.

In this situation, $\sum_j w_{a_ib_j} = \sum_i w_{a_ib_j} = w_{a_ib_j}$ and $\deg\left(a_i\right) = \deg\left(b_j\right) = 1$. Therefore,

$$DPRel\left(a_i, b_j|P\right) = 2 \times \frac{w_{a_ib_j}}{w_{a_ib_j} + w_{a_ib_j}} = 1$$

Property 3 (Symmetry). $DPRel\left(a_i, b_j|P\right) = DPRel\left(b_j, a_i|P^{-1}\right)$

Proof:

$$DPRel\left(a_i, b_j|P\right) = \frac{w_{a_ib_j}\left(\frac{1}{\deg\left(a_i\right)} + \frac{1}{\deg\left(b_j\right)}\right)}{\left(\frac{1}{\deg\left(a_i\right)}\sum_j w_{a_ib_j} + \frac{1}{\deg\left(b_j\right)}\sum_i w_{a_ib_j}\right)}$$

$$= \frac{w_{b_ja_i}\left(\frac{1}{\deg\left(b_j\right)} + \frac{1}{\deg\left(a_i\right)}\right)}{\left(\frac{1}{\deg\left(b_j\right)}\sum_i w_{b_ja_i} + \frac{1}{\deg\left(a_i\right)}\sum_j w_{b_ja_i}\right)} = DPRel\left(b_j, a_i|P^{-1}\right)$$

Case II: Relevance Measurement for Same Typed Objects. Since, in case of same typed objects, we are converting the odd path length case into the even path length case by decomposing atomic middle relations, therefore, there is no need to give separate proof for the odd length case. For same typed objects, we calculate relatedness by calculating first the relevance of the source object a_i with middle objects $b_k, \forall k$ i.e. $Rel(a_i, b_k|P_L)$ and relevance of the target object a_j with middle objects $b_k, \forall k$ i.e. $Rel(a_j, b_k|P_R^{-1})$ where meta-path $P = P_L P_R = (A_1 A_2 \ldots A_m)(A_m A_{m+1} \ldots A_{l+1})$. Then we use Tanimoto coefficient [5] to measure the relatedness between a_i and a_j following meta-path P.

$$X = Rel(a_i, b_k|P_L)$$

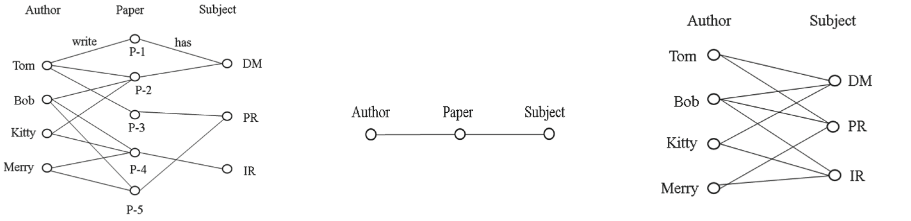
$$Y = Rel\left(a_j, b_k|P_R^{-1}\right)$$

$$DPRel(a_i, a_j|P) = \frac{X.Y}{|X|^2 + |Y|^2 - X.Y}$$

Since the Tanimoto coefficient has all the three properties [5] i.e. reflexivity, symmetry and limited range, therefore, in case of same type objects all three properties are automatically proven.

4 Illustration

In order to explain the working of the proposed measure, we take the following heterogeneous network as an example shown in Fig. 2(a). In this network, there are three types of nodes i.e. Author (A), Paper (P), and Subject (S). The semantic relationship between author and paper is different from relationship between paper and subject. The semantic relationships are bidirectional. For example, an author "*writes*" paper or a paper is "*written by*" authors. Therefore, we have taken undirected links in our example. The network schema and the bipartite representation following meta-path *APS* for this example are shown in Fig. 2(b) and (c) respectively. The calculation of weighted path matrix is shown in Fig. 3 (a).



(a) Example network

(b) Network schema

(c) Bipartite representation for path *APS*

Fig. 2. Example network, its schema, and bipartite representation

Now, we calculate the relatedness between different typed objects *Tom* and *DM* following meta-path *APS* i.e. "Author-Paper-Subject". In this case, regardless of the length of meta-path, decomposition of atomic relations is not required as we are calculating relatedness between different type objects.

$$DPRel\,(Tom, DM|APS) = \frac{w_{Tom,DM}\left(\frac{1}{\deg(Tom)} + \frac{1}{\deg(DM)}\right)}{\left(\frac{w_{Tom,DM}+w_{Tom,PR}+w_{Tom,IR}}{\deg(Tom)} + \frac{w_{Tom,DM}+w_{Bob,DM}+w_{Kitty,DM}+w_{Merry,DM}}{\deg(DM)}\right)}$$

$$
\begin{array}{c|ccccc}
 & \text{P-1} & \text{P-2} & \text{P-3} & \text{P-4} & \text{P-5} \\ \hline
\text{Tom} & 1 & 1 & 1 & 0 & 0 \\
\text{Bob} & 0 & 1 & 0 & 1 & 1 \\
\text{Kitty} & 0 & 1 & 0 & 1 & 0 \\
\text{Merry} & 0 & 0 & 0 & 1 & 1
\end{array}
\times
\begin{array}{c|ccc}
 & \text{DM} & \text{PR} & \text{IR} \\ \hline
\text{P-1} & 1 & 0 & 0 \\
\text{P-2} & 1 & 0 & 0 \\
\text{P-3} & 0 & 1 & 0 \\
\text{P-4} & 0 & 0 & 1 \\
\text{P-5} & 0 & 1 & 0
\end{array}
=
\begin{array}{c|ccc}
 & \text{DM} & \text{PR} & \text{IR} \\ \hline
\text{Tom} & 2 & 1 & 0 \\
\text{Bob} & 1 & 1 & 1 \\
\text{Kitty} & 1 & 0 & 1 \\
\text{Merry} & 0 & 1 & 1
\end{array}
$$

(a) Calculation of weighted path matrix for meta-path APS

	DM	PR	IR
Tom	0.588	0.333	0
Bob	0.286	0.333	0.333
Kitty	0.357	0	0.417
Merry	0	0.417	0.417

(b) Relatedness between Authors and Subjects following meta-path APS

	Tom	Bob	Kitty	Merry
Tom	1	0.579	0.383	0.209
Bob	0.579	1	0.662	0.744
Kitty	0.383	0.662	1	0.366
Merry	0.209	0.744	0.366	1

(c) Relatedness between Authors following meta-path $APSPA$

Fig. 3. Calculation of relevance

$$= \frac{2(\frac{1}{2}+\frac{1}{3})}{(\frac{2+1+0}{2}+\frac{2+1+1+0}{3})} = 0.588$$

The results of rest of these computations in matrix form are shown in Fig. 3 (b).

Next, we show how to calculate relatedness between same typed objects. Now, we calculate relatedness between authors by following meta-path $P = APSPA$. Since, we are measuring relatedness between same typed objects and the length of meta-path is even, therefore, there is no need to decompose atomic relations in this case. We first calculate the relatedness of authors with middle object type of this meta-path i.e. Subject (S). Then we use Tanimoto coefficient to calculate the relatedness between authors. Here $P = APSPA = P_L P_R = (APS)(SPA)$. Therefore, $P_L = APS$ and $P_R^{-1} = APS$.

$$X = R(Tom, \{DM, PR, IR\}|APS) = \{0.588, 0.333, 0\}$$

$$Y = R(Bob, \{DM, PR, IR\}|APS) = \{0.286, 0.333, 0.333\}$$

$$DPRel(Tom, Bob|APSPA) = \frac{X.Y}{|X|^2 + |Y|^2 - X.Y} = 0.579$$

The results of rest of these computations in matrix form are shown in Fig. 3(c).

5 Experimental Setup and Results

To show the viability and the effectiveness of the proposed measure DPRel, we take four area DBLP dataset collected from the website http://web.engr.illinois.edu/~mingji1/ [6]. For comparison, we take only PathSim and HeteSim as these are the only two meta-path based measures which are semi-metric and directly applicable to heterogeneous networks. All experiments were performed on a system with Intel Core i5 processor and 8 GB RAM using R version 3.0.3.

5.1 Dataset

The DBLP dataset used in our experiment is a subset of data available on DBLP website [6]. The dataset used in our experiment involves major conferences in four research areas: Database, Data Mining, Information Retrieval and Artificial Intelligence which naturally forms four classes. The dataset used in our experiment contains 14376 papers, 20 conferences, 14475 authors and 8920 keywords (terms) with 170794 links in total and the dataset is stored in plain text file. In this dataset, 4057 authors, all 20 conferences and 100 papers are labelled with one of the four research area classes. The network schema for DBLP is shown in Fig. 4. Since citation information is not present in this dataset, therefore, paper node in the schema has no self-loop.

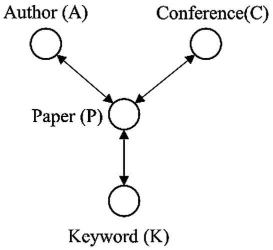


Fig. 4. Network schema of DBLP heterogeneous network data

5.2 Performance Comparison for Clustering

For clustering task, we use three meta-paths *APCPA*,*CPAPC* and *PAPCPAP* for clustering of authors, conferences and papers respectively [4]. We use Partition Around Medoid (PAM) [9, 10] and Affinity Propagation (AP) [13] for clustering and take value of $k = 4$ for PAM as there are four natural classes. The results of comparison are given in Tables 1 and 2. For performance evaluation, we use F-Measure, Normalized Mutual Information (NMI), Cluster Purity and Adjusted Rand Index (ARI) [9]. In clustering, we do not perform clustering of keywords since the four areas are very much overlapping in terminology used. Therefore, the clustering accuracy would be too low in case of clustering of keywords and would not be able to capture the essence of comparison of performance of three measures.

From the results, it is clear that DPRel performs better as compared to HeteSim and PathSim for clustering of authors, conferences and papers in case of PAM. For AP, in case of paper and conference, DPRel performs better. We also see that for clustering of authors and conferences, performance of DPRel is high but for clustering of papers performance is low. The reason might be the selection of meta-path. The accuracy of clustering and other mining tasks depend upon the meta-path selected. The relatedness between conferences are measured using meta-path *CPAPC* which means conferences sharing same authors. Likewise, meta-path *APCPA* which means authors publishing papers in same conferences effectively capture the similarity of authors. Since, in both cases of authors and conferences, the similarity is captured appropriately by meta-paths, the performance of PAM and AP are high in both cases. However, in case of papers since similarity is captured by referenced authors (i.e., the *APCPA* path) which is not effectively measuring the similarity of papers, therefore, the performance of PAM and AP are low in this case. This shows that the performance highly depends upon the selection of meta-path apart from the accuracy of clustering algorithm.

Table 1. Comparison for clustering task using PAM

	Path	Precision	Recall	F-Measure	NMI	Cluster purity	ARI
DPRel	*APCPA*	0.8383	0.8392	**0.8387**	**0.737**	**0.912**	**0.7965**
HeteSim		0.6829	0.7615	0.72	0.6062	0.7693	0.5285
PathSim		0.8319	0.8384	0.8351	0.7322	0.9095	0.7907
DPRel	*CPAPC*	0.9	0.9	**0.9**	**0.9058**	**0.95**	**0.8588**
HeteSim		0.8167	0.8	0.8082	0.8073	0.9	0.7072
PathSim		0.7571	0.75	0.7536	0.7585	0.85	0.5983
DPRel	*PAPCPAP*	0.628	0.6199	**0.6239**	**0.5064**	**0.77**	**0.4632**
HeteSim		0.5358	0.5104	0.5228	0.3899	0.7	0.3326
PathSim		0.4872	0.5269	0.5063	0.3479	0.65	0.2746

Table 2. Comparison for clustering task using Affinity Propagation clustering

	Path	Precision	Recall	F-Measure	NMI	Cluster Purity	ARI
DPRel	*APCPA*	0.785	0.8014	0.7931	0.6748	0.8812	0.7349
HeteSim		0.8654	0.8693	**0.8674**	**0.7759**	**0.9283**	**0.8289**
PathSim		0.7834	0.7999	0.7916	0.674	0.8802	0.7342
DPRel	*CPAPC*	0.9	0.9	**0.9**	**0.9058**	**0.95**	**0.8588**
HeteSim		0.8167	0.8	0.8082	0.8073	0.9	0.7072
PathSim		0.8167	0.8	0.8082	0.8073	0.9	0.7072
DPRel	*PAPCPAP*	0.6334	0.6233	**0.6283**	**0.5161**	**0.77**	**0.4723**
HeteSim		0.5151	0.4899	0.5022	0.335	0.68	0.2559
PathSim		0.5791	0.556	0.5673	0.4054	0.73	0.4135

5.3 Performance Comparison for Query Task

Using query task, we can evaluate the effectiveness of DPRel for different typed objects in heterogeneous network. Since PathSim cannot measure the relatedness between different typed objects [3], therefore, we compare the performance of DPRel only with HeteSim. Using labelled subset of DBLP dataset, we measure the relatedness of conferences with authors following two meta-paths: *CPA* and *CPAPA*. For each conference, we rank authors according to its relevance value. We compute the AUC (Area Under ROC Curve) score based on the labels of authors and conferences to evaluate the performance of DPRel and HeteSim. For comparison, we take 7 representative conferences out of 20 and their AUC score values are listed in Table 3 for DPRel and HeteSim. We can see that for both meta-paths, performance of DPRel is better for all 7 conferences as compared to HeteSim (Table 3).

Table 3. Comparison for query task using AUC

Conference	*CPA*		*CPAPA*	
	DPRel	*HeteSim*	*DPRel*	*HeteSim*
KDD	**0.8117**	0.8111	**0.8296**	0.827
SIGIR	**0.9522**	0.9507	**0.9456**	0.9402
SIGMOD	**0.7674**	0.7662	**0.8016**	0.7934
VLDB	**0.8282**	0.8262	**0.8589**	0.8477
ICDE	**0.7296**	0.7282	**0.7709**	0.7648
AAAI	**0.812**	0.8109	**0.8215**	0.8113
IJCAI	**0.8771**	0.8754	**0.8911**	0.8785

5.4 Time Complexity Analysis

Let $\boldsymbol{n}$ be the average number of objects of one type in the network. Then, for DPRel, the space complexity would be $\boldsymbol{O}(\boldsymbol{n}^2)$ to store the relevance matrix. Let $\boldsymbol{d}$ be the average degree of a node in the network. Then, for a specified meta-path of length $\boldsymbol{l}$, the time complexity for HeteSim would be $\boldsymbol{O}(\boldsymbol{ld}^2\boldsymbol{n}^2)$, since for all node pairs (i.e.$\boldsymbol{n}^2$), the relevance is calculated along the relevance path before the matrix pair multiplication [4]. However, in case of DPRel, relevance is calculated after getting the bipartite representation (i.e. after doing the multiplication of matrices). Therefore, the time complexity of calculating DPRel would be $\boldsymbol{O}(\boldsymbol{ln}^2 + \boldsymbol{dn}^2)$ which is far less than the time complexity of HeteSim. Also, in case of DPRel we need to do the decomposition operation only in the case of same typed objects when path length is odd. This property further improves the efficiency of the DPRel as compared to HeteSim.

6 Conclusion and Future Research Directions

In this paper, we proposed a novel meta-path based measure, DPRel, to compute the relatedness between objects in a heterogeneous information network. The proposed measure addresses the limitations of earlier measures and is able to measure the relatedness between same as well as different typed objects. In this paper, we comparatively and systematically examined the performance of DPRel and compared the effectiveness with two well-known relevance measures, PathSim and HeteSim. From the experiments performed on real world bibliographic dataset DBLP, it is clear that DPRel outperforms PathSim and HeteSim in clustering and query tasks. Our work has following main contributions: First, in this work, we study different relevance measures in heterogeneous networks and address the problems of earlier meta-path based measures. Second, we propose a novel measure for relevance measurement in general heterogeneous networks. Our proposed measure is semi-metric, therefore, has applicability in real world applications. Third, the proposed measure can measure the relatedness between objects of different as well as same typed objects following any arbitrary meta-path. Fourth, the proposed measure has no need to decompose atomic relation while computing relatedness between different typed objects following any arbitrary path. Finally, our proposed measure performs better than other measures.

Future research directions include a dynamic programming based approach for fast computation of DPRel to compute the relevance in heterogeneous networks. Also, apart from DBLP dataset, the proposed measure can also be tested on other heterogeneous datasets emerging from social networking sites like, Facebook, Twitter etc.

References

1. Huang, Y., Gao, X.: Clustering on heterogeneous networks. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **4**(3), 213–233 (2014)
2. Sun, Y., Han, J.: Mining heterogeneous information networks: a structural analysis approach. ACM SIGKDD Explor. Newsl. **14**(2), 20–28 (2013)
3. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: VLDB (2011)
4. Shi, C., Kong, X., Huang, Y., Philip, S.Y., Wu, B.: HeteSim: a general framework for relevance measure in heterogeneous networks. IEEE Trans. Knowl. Data Eng. **26**(10), 2479–2492 (2014)
5. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, London (2009)
6. Ji, M., Sun, Y., Danilevsky, M., Han, J., Gao, J.: Graph regularized transductive classification on heterogeneous information networks. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part I. LNCS, vol. 6321, pp. 570–586. Springer, Heidelberg (2010)
7. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford University Database Group (1998)
9. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
10. Kumar, P., Raju, B.S., Radha Krishna, P.: A new similarity metric for sequential data. Int. J. Data Warehouse. Min. **6**(4), 16–32 (2010)
11. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained random walks. Mach. Learn. **81**(1), 53–67 (2010)
12. Meng, X., Shi, C., Li, Y., Zhang, L., Wu, B.: Relevance measure in large-scale heterogeneous networks. In: Chen, L., Jia, Y., Sellis, T., Liu, G. (eds.) APWeb 2014. LNCS, vol. 8709, pp. 636–643. Springer, Heidelberg (2014)
13. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)