Expert Systems With Applications 68 (2017) 106–122

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

HeteClass: A Meta-path based framework for transductive classification of objects in heterogeneous information networks

Mukul Gupta*, Pradeep Kumar, Bharat Bhasker

Indian Institute of Management Lucknow, India

ARTICLE INFO

Article history:
Received 20 April 2016
Revised 9 October 2016
Accepted 10 October 2016
Available online 11 October 2016

Keywords:
Heterogeneous information network
Meta-path
Transductive classification
Weight learning

ABSTRACT

Transductive classification using labeled and unlabeled objects in a heterogeneous information network for knowledge extraction is an interesting and challenging problem. Most of the real-world networks are heterogeneous in their natural setting and traditional methods of classification for homogeneous networks are not suitable for heterogeneous networks. In a heterogeneous network, various meta-paths connecting objects of the target type, on which classification is to be performed, make the classification task more challenging. The semantic of each meta-path would lead to the different accuracy of classification. Therefore, weight learning of meta-paths is required to leverage their semantics simultaneously by a weighted combination. In this work, we propose a novel meta-path based framework, HeteClass, for transductive classification of target type objects. HeteClass explores the network schema of the given network and can also incorporate the knowledge of the domain expert to generate a set of meta-paths. The regularization based weight learning method proposed in HeteClass is effective to compute the weights of symmetric as well as asymmetric meta-paths in the network, and the weights generated are consistent with the real-world understanding. Using the learned weights, a homogeneous information network is formed on target type objects by the weighted combination, and transductive classification is performed. The proposed framework HeteClass is flexible to utilize any suitable classification algorithm for transductive classification and can be applied on heterogeneous information networks with arbitrary network schema. Experimental results show the effectiveness of the HeteClass for classification of unlabeled objects in heterogeneous information networks using real-world data sets.

1. Introduction

The information network is a most natural way of representing real-world entities/objects and their relationships. In these networks, real-world objects and their relationships are represented as nodes and links/edges respectively. Conventional networks are homogeneous in nature as these networks are composed of single type of objects and one type of relationship (Sun, Han, Yan, Yu, & Wu, 2011). However, most of the real-world information networks are heterogeneous in their natural setting (Gupta, Kumar, & Bhasker, 2015; Shi, Kong, Huang, Yu, & Wu, 2014). Heterogeneous information networks consist of more than one type of objects and/or relationships. Fig. 1(a) and (b) show examples of homogeneous and heterogeneous information networks respectively. Fig. 1(a) shows the snippet of a homogeneous information network of authors, related to each other by the co-author relationship. Fig. 1(b) shows the snippet of a bibliography dataset as a heterogeneous information network having different types of objects like papers, authors, and conferences with different relationships between them.

Conventional representation of linked information using homogeneous information network is very popular and convenient for various mining tasks. However, heterogeneous entities and their complex relationships cannot be represented using homogeneous information networks. Of late, researchers and practitioners are utilizing heterogeneous information networks in various mining tasks to get information nuggets (Deng, Lai, Wang, & Fang, 2012; Zhang, Hu, He, & Wang, 2015). The information rich heterogeneous information network gives better mining results as compared to its homogeneous transformation (Gupta et al., 2015; Shi et al., 2014; Sun et al., 2011). For example, Fig. 1(b) shows a heterogeneous information network which consists of authors, papers, and conferences. This network is information rich as compared to the homogeneous transformation that is shown in Fig. 1(a) and shows the co-author relationship between authors. For different mining tasks on heterogeneous information networks like clustering, classification, we need to measure the relatedness between objects.

* Corresponding author: Information Technology & Systems, Indian Institute of Management Lucknow, India.
E-mail addresses: fpm13008@iiml.ac.in (M. Gupta), pradeepkumar@iiml.ac.in (P. Kumar), bhasker@iiml.ac.in (B. Bhasker).

http://dx.doi.org/10.1016/j.eswa.2016.10.013
0957-4174/

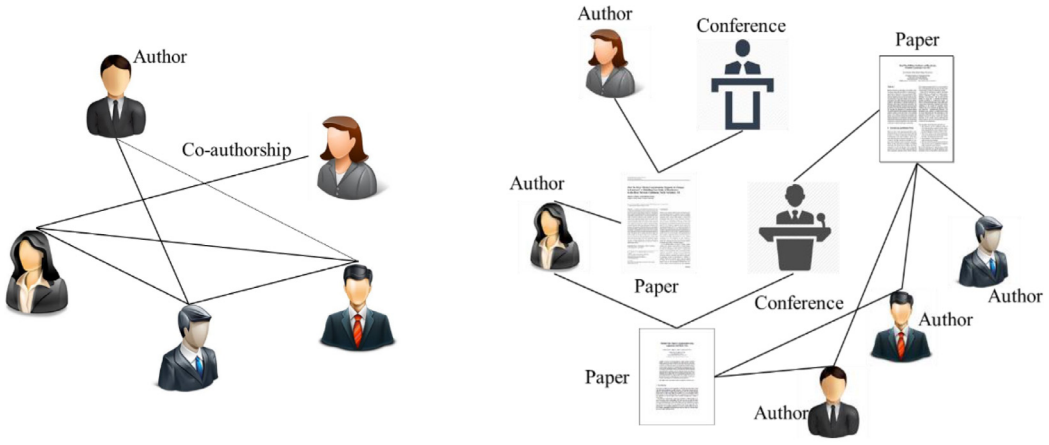


Fig. 1. Examples of homogeneous and heterogeneous information networks.

Conventional link-based measures like Personalized PageRank (Jeh & Widom, 2003; Liben-Nowell & Kleinberg, 2007), SimRank (Jeh & Widom, 2002) are not applicable to heterogeneous information networks due to the heterogeneity of objects and relationships in the networks (Gupta et al., 2015; Shi et al., 2014; Sun et al., 2011). Meta-path based relevance measures like PathSim, HeteSim, DPRel have been proposed recently to measures the relatedness between objects (Gupta et al., 2015; Shi et al., 2014; Sun et al., 2011). Using a meta-path based relevance measure, we can incorporate meta-path semantics while measuring the relatedness between objects.

In this work, we perform classification of objects in a heterogeneous information network. Classification is an important mining task as it is required for various applications like link prediction, community detection, and object recommendation (Kong, Yu, Ding, & Wild, 2012; Sun & Han, 2013). The problem of classification of objects using the link information present in a conventional homogeneous information network has been well studied and explored by researchers (Lu & Geetor, 2003; Macskassy & Provost, 2007; Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004; Zhu et al., 2003). Various relational and transductive classification methods for homogeneous information networks like weighted vote Relational Neighbor (wvRN) classifier (Macskassy & Provost, 2003), and Learning with Local and Global Consistency (LLGC) classifier (Zhou et al., 2004) have been proposed and utilized extensively. However, classification of objects in a heterogeneous information network is comparatively new. For classification of objects in a heterogeneous information network, we utilize transductive classification which is different from conventional classification. In conventional supervised classification, data objects are assumed to be independent and identically distributed; however, transductive classification utilizes the label correlation among a group of linked objects to decide the label of other objects in the network.

Transductive classification is highly challenging and complex in the case of heterogeneous information networks. Due to different object types and/or relationships between objects, the label information of objects of one type should not be utilized to determine the label of objects of another type (Angelova, Kasneci, & Weikum, 2012; Kong et al., 2012). In heterogeneous information networks, the label set of one object type is conceptually different from the label sets of other object types due to the different characteristics of objects of different types (Angelova et al., 2012; Kong et al., 2012). For example, in the case of a heterogeneous Flickr network, the label set concept for classification of photos will be different from the label set concept of Users (Angelova et al., 2012). Apart from that, the characteristics of heterogeneous data such as the complexity of the network structure, lack of features, and scarcity of labeled objects, also add to the difficulty in classifying objects in a heterogeneous information network (Ji, Sun, Danilevsky, Han, & Gao, 2010). Mining of such heterogeneous information network using conventional techniques is not feasible due to various semantics and subtleties present in the network that would be lost if not taken care of (Shi et al., 2014; Sun et al., 2011).

In traditional classification, supervised learning is performed using local features or attributes of objects. However, in most of the real-world heterogeneous information networks, there are no natural local features or attributes for objects (Ji et al., 2010). In heterogeneous information networks, if the link information is considered as attributes of objects, then it is likely that the dimensionality of the objects would be very high and with the increased number of objects, the data becomes sparse in nature (Ji et al., 2010; Luo, Guan, Wang, & Lin, 2014). However when some of the object types have features or attributes associated with them, the traditional classification techniques would not be applicable as their attributes would be incomparable due to the distinct feature/attribute space (Ji et al., 2010). That is why traditional classification techniques like Support Vector Machines, Logistic Regression, and Naïve Bayes are difficult to apply on heterogeneous information networks. Therefore, transductive classification of objects can be performed utilizing the labeled and unlabeled objects in the network to get the label information of the unlabeled objects.

In this work, we propose a framework called HeteClass for classification of objects of one type called target type in a heterogeneous information network. In this work, our focus is on the classification of one type of objects instead of all types of objects collectively. This problem setting exists in various real-world situations. The reason is due to the different label set concepts for different types of objects in the network (Kong et al., 2012). For example, in the case of a heterogeneous information network for bibliography, there would be several object types like author, conference, paper and keyword. The classification can be performed on any one type of objects like authors which would be termed as the target type. In the proposed framework HeteClass, various meta-paths are explored to incorporate semantics associated with those meta-paths while performing classification on target type objects. For classification, we perform weight learning of meta-paths to assign higher weights to meta-paths that would lead to good classification accuracy using prior label information. Weight learning of meta-paths is an important step in HeteClass framework as some of the meta-paths would result in good classification accuracy and therefore should be assigned higher weights as compared to other

meta-paths. If equal weights are assigned to all meta-paths, the classification accuracy would be reduced due to the impact of the semantics of meta-paths while performing classification.

To show the effectiveness of HeteClass, we use real-world bibliography dataset DBLP, and Flick Fashion 10,000 datasets. The proposed framework HeteClass is orthogonal to the algorithm utilized for classification in the framework i.e. its performance can be improved using any advanced classification algorithm. To demonstrate the effectiveness of HeteClass, we have utilized two different classification algorithms from different categories namely, LLGC (Zhou et al., 2004) and wvRN (Macskassy & Provost, 2003) classifiers. We also compare the performance of HeteClass with a recently proposed algorithm for heterogeneous information networks called HetPathMine (Luo et al., 2014) which can also utilize various meta-paths in the network for classification of target type objects. Experimental results show the effectiveness of HeteClass as compared to the aforementioned algorithms for classification.

The expected contribution from the study is a meta-path based novel framework HeteClass for classification of objects of target type in a heterogeneous information network where the label sets of object types are conceptually different. HeteClass explores the schema of the network to generate a set of meta-paths for consideration while performing classification of the target type objects. HeteClass allows incorporation of domain expert knowledge to select only those meta-paths that are significant and good enough to produce high classification accuracy. HeteClass performs the weight learning of meta-paths more effectively. Further, the proposed framework HeteClass is flexible and allows utilization of any suitable classification algorithm for classification of objects after generating a single homogeneous information network of target type objects.

In this work, we have utilized Personalized PageRank algorithm (Haveliwala, 2002) for classification of objects in the homogeneous network on the target type. However, in the proposed framework, instead of Personalized PageRank algorithm, we can utilize any advanced algorithm for classification which would increase the accuracy of classification. Experimental results show that the accuracy of the proposed framework is better as compared to LLGC, wvRN and HetPathMine algorithms.

The rest of the paper is organized as follows. In Section 2, we present the related work. Preliminaries to this study and formal problem definition are given in Section 3. The proposed framework HeteClass is explained in Section 4. In Section 5, the experimental setup is explained. In Section 6, results and discussion are presented. Finally, the conclusion and future research directions are presented in Section 7.

2. Related work

The problem of transductive classification of objects in a linked structure or network has received significant attention. The idea has been to use prior label information of a small number of objects and perform transductive classification on the objects in the network by exploiting the local and global structure of the network. Earlier studies by Lu and Geetor (2003), Macskassy and Provost (2003), Zhou et al. (2004), and Macskassy and Provost (2007) have performed the classification of objects using the local and/or global structure of the network. However, these approaches were designed for homogeneous information networks and are not directly applicable to heterogeneous information networks.

In heterogeneous information networks, the existence of different types of objects and/or relations makes the networks unsuitable for aforementioned link-based classification. When these approaches for linked based classification are applied directly on a heterogeneous information network, the classification accuracy would be low (Kong et al., 2012; Sun & Han, 2013). If we transform the heterogeneous information network into a corresponding homogeneous information network following a meta-path and then apply the aforementioned link based classification approaches, the accuracy of classification may be low if the selected meta-path semantically does not convey the meaningful relationship between target type objects. Since in a heterogeneous information network, there may be several meta-paths that have different semantic meaning, following these meta-paths we can convert the heterogeneous information network into a corresponding homogeneous information network consisting only of target type objects. Since we do not know in advance which path is good in terms of classification accuracy, we may end up following an ineffective path and that would lead to low classification accuracy. Also, by following only one meta-path, we may miss other semantically significant meta-paths that may contribute to good classification accuracy. Hence, consideration of all meta-paths starting and ending at the target type nodes is necessary for good classification accuracy.

For classification of objects in heterogeneous information networks, recently, different techniques have been proposed. In the work done by Rossi, de Andrade Lopes, and Rezende (2016), authors performed transductive classification of objects in the heterogeneous network. However, their approach is for bipartite networks. Angelova et al. (2012)) proposed a technique called Graffiti for classification of objects collectively in the network. They utilized the random walk based model in their approach. In their work, they considered the problem of classification of different types of nodes with the conceptually different label sets i.e. concept/semantic of even same label set would not be similar across different types of nodes as the labels are type specific (Angelova et al., 2012). In another work done by Ji et al. (2010), authors proposed GNetMine for transductive classification of objects in the networks. In their approach, they utilized the whole network for transductive classification. They assumed the same label set concept for all object types across the network. They discarded the semantics of label set between object types and transferred the label information across the network. For classification, they utilized the graph-based regularization framework. However, they did not consider the distinction between the label set concept of different object types. For example, the characteristics of the label set for authors would be different from the characteristics of the label set for conferences (Angelova et al., 2012).

Of late, Kong et al. (2012) proposed Heterogeneous Collective Classification (HCC) algorithm for classification of objects in heterogeneous information networks. In their work, authors performed collective classification of objects of one type called target type. However, they took one or more object types as feature objects for the target type. For example, in a bibliography dataset like DBLP, they took keywords of the paper as features for the author to be classified. Then they performed feature augmentation of objects by considering various meta-paths. Using the augmented features, they performed the classification in two steps: (1) Bootstrap by considering the node features, (2) Iterative inference by considering relational features. They, in their work, assumed the characteristics of the label set of one object type different from other object types, and they performed classification of one object type termed as target object type. Also, they considered various meta-paths in the network for feature set augmentation. However, there are some limitations with their approach. First, since they used feature based classification, a large number of labeled objects are required for model generation. In real-world scenarios acquiring label information for a large number of objects is costly. This makes HCC algorithm unsuitable for real-world situations where the task is to infer the label information of the unlabeled objects using a small number of labeled objects in a network. Second, they used one or more object types as features for the target type objects. However, it is difficult to determine the appropriate features for objects when we

Table 1
Description of symbols.

Symbol	Description
$G = (V, E)$, T_G	Heterogeneous information network (HIN) and its schema
A, R	Object types and relations in HIN
$\emptyset$, ψ	Object type and link type mapping function
R_C, $\mathcal{P}$	Composite relation and meta-path
$W_{A_iA_j}$, M	Adjacency matrix and weighted path matrix
$Sim_{\mathcal{P}}(a_i, a_j)$	Relatedness between objects a_i and a_j following meta-path $\mathcal{P}$
θ_k, w_k	Importance of meta-path $\mathcal{P}_k$ and weight of meta-path $\mathcal{P}_k$

have only link information. Next, in their work they considered meta-paths but they gave equal importance/weight to all meta-paths. However, we know that some of the meta-paths could be comparatively better than other meta-paths and therefore, should be given higher weights.

Recently, Luo et al. (2014) proposed HetPathMine algorithm for classification of target type objects in a heterogeneous information network. The authors in their work proposed weight learning for meta-paths to integrate various path semantics. They used transductive classification for labeling of objects in the network. They addressed the limitations of earlier works for classification in heterogeneous information networks when the label set of one object type is conceptually different from others. However, there are some limitations with HetPathMine. In HetPathMine, there is no method for generating meta-paths from network schema. Also, the weight learning in HetPathMine is unconstrained supervised learning using the prior label information which may lead to negative path weight and that would not be meaningful. Next, HetPathMine utilized PathSim for relatedness measurement which is not effective for relatedness measurement in heterogeneous information networks and can utilize only symmetric meta-paths of the network (Gupta et al., 2015; Shi et al., 2014).

To address the limitations of HetPathMine and earlier methods for classification in heterogeneous information networks when label sets are conceptually different for different object types, we propose a novel framework called HeteClass for classification of objects in heterogeneous information networks. The proposed framework HeteClass explores the network schema of the heterogeneous information network to generate a set of meta-paths that can be utilized in weight learning. HeteClass can also integrate the knowledge of domain expert, if available, to reduce the number of meta-paths for weight learning. The advantage of reducing the number of paths is that the computation time would decrease for weight learning. Since there may be some paths which would not lead to good classification accuracy, removing these paths from the set of paths generated would also increase the weight of effective paths as the total number of paths in the set would be reduced and the sum of the weights for all paths would be one. Also, the HeteClass allows using a classification algorithm of choice on the final homogenous network consisting of target type objects. This makes the HeteClass framework highly effective, which may lead to good accuracy. For this study, we compared the performance of HeteClass with LLGC and wvRN. We also compared the performance of HeteClass with HetPathMine to show the effectiveness of HeteClass. For experiments, we utilized the real-world bibliography dataset DBLP and Flickr Fashion 10,000 dataset.

3. Preliminaries and problem definition

In this section, the background and preliminaries for this work are presented. The formal definition is also given of transductive classification on target type nodes in the heterogeneous information network. Some important and frequently used notations are given in Table 1.

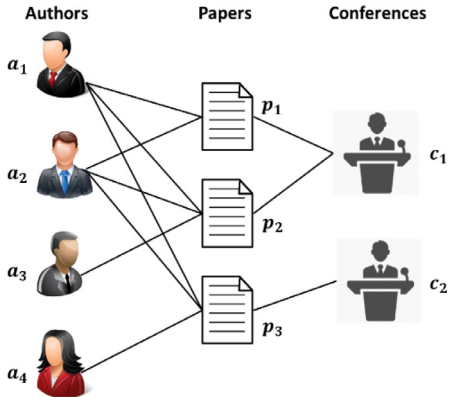


Fig. 2. Snippet of bibliography dataset represented as information network.

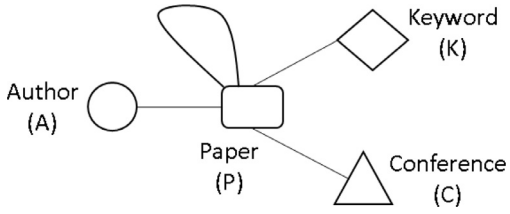


Fig. 3. Network schema of DBLP network.

This work utilizes information network as defined in Definition 1, which is similar to the definition by Sun et al. (2011).

Definition 1 (Information network)**.** 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 ψ: $E \rightarrow R$, 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 $\psi(e) \in R$.

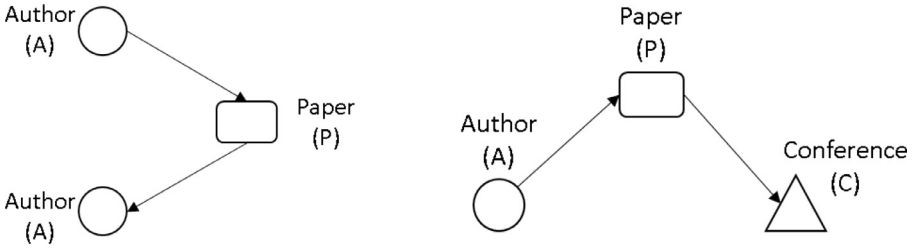
When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is a heterogeneous information network; otherwise, it is a homogeneous information network.

In the network snippet of a bibliography dataset, as shown in Fig. 2, we can see that there are three different types of nodes and two different relationships. In this network since more than one type of nodes and relationships exist, it is a heterogeneous information network. For a heterogeneous information network, we consider it's meta-level (i.e., schema level) representation for better understanding (Lao & Cohen, 2010) as defined in Definition 2.

Definition 2 (Network schema)**.** The network schema denoted as $T_G = (A, R)$, is a meta-level representation for a heterogeneous information network $G = (V, E)$ with object type mapping $\emptyset$: $V \rightarrow A$ and link type mapping ψ: $E \rightarrow R$, which is a directed graph over object types A and edges as relations from R.

A bibliography information network like DBLP is an example of a heterogeneous information network. This network consists of four types of entities/objects: papers (P), authors (A), conferences (C) and keywords (K). Fig. 3 shows the meta-level representation (network schema) of the DBLP network. Since the relationships are bidirectional, undirected links have been used in this network.

Definition 3 (Meta-path)**.** A meta-path $\mathcal{P}$ is defined on network schema $T_G = (A, R)$ and denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R_c = R_1 \circ R_2 \circ \cdots \circ R_l$ of length l between source object type A_1 and target object type A_{l+1} using composition operator $\circ$ on relations.



(a) Meta-path: APA (b) Meta-path: APC

Fig. 4. Meta-paths for DBLP network schema.

Two meta-paths *APA* (*Author – Paper – Author*) and *APC* (*Author – Paper – Conference*) of the DBLP network schema are shown in Fig. 4(a) and(b) respectively.

In a heterogeneous information network, two objects can be connected via different paths and these paths will have different semantic meanings. For example, meta-paths *APA* and *APCPA* in DBLP network schema are two different meta-paths connecting authors to authors i.e., source and target objects are same-typed in these two meta-paths. But these two meta-paths have different semantic interpretations. Meta-path *APA* means authors who are co-authors for papers; meta-path *APCPA* indicates authors publishing papers in the same conference. The different semantic meanings of different meta-paths will lead to different classification accuracy. The relatedness between authors following the meta-path *APA* emphasizes on the papers, whereas in the case of *APCPA*, conferences are emphasized. Therefore, the relatedness between objects in a heterogeneous information network depends on the meta-path followed.

If there are no multiple relations between objects, we can represent the meta-path using only object types such as $\mathcal{P} = (A_1A_2\cdots A_{l+1})$. A *path instance* $p = (a_1a_2\cdots a_{l+1})$ between a_1 and a_{l+1} in a network G is an instance of the meta-path $\mathcal{P}$, i.e., $p \in \mathcal{P}$, if for each a_i, $\emptyset(a_i) = A_i$ and for each link $e_i = \langle a_i, a_{i+1}\rangle$, $\psi(e_i) = R_i$. The *reverse path* of the meta-path $\mathcal{P}$, denoted as $\mathcal{P}^{-1}$, defines an inverse relation between object types. Likewise, the *reverse path instance* $p^{-1} \in \mathcal{P}^{-1}$ is the reverse path of p in G.

For classification of objects using HeteClass, we need to measure the relatedness between objects. For that, relevance measure DPRel (Gupta et al., 2015) is utilized. For measuring the relatedness between objects using DPRel, we need to transform a heterogeneous information network into a bipartite network consisting of only source and target type objects. For that, we need to compute the weighted path matrix as explained in Definition 4.

Definition 4 (Weighted path matrix)**.** For a heterogeneous information network and its schema level representation, a weighted path matrix M for meta-path $\mathcal{P} = (A_1A_2\cdots A_{l+1})$ 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 objects of type 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 $\mathcal{P}$, and $M[x_i, y_j] = M[y_j, x_i]$.

For the heterogeneous information network shown in Fig. 2, the calculation of weighted path matrix is shown in Fig. 5 following meta-path *APC* where source object type is A(Author) and target object type is C(Conference).

DPRel is defined in Definition 5. Then we show how to utilize DPRel for measuring the relatedness between objects.

Definition 5 (DPRel)**.** Given a meta-path $\mathcal{P} = (A_1A_2\cdots A_{l+1})$ such that A_1 and A_{l+1} are different object types, then for bipartite representation of heterogeneous information network that has only objects of type A_1 and A_{l+1}, the relatedness between source object

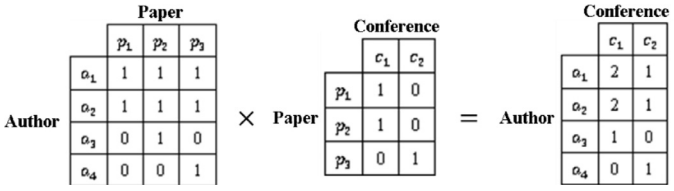


Fig. 5. Calculation of weighted path matrix.

$a_{1i} \in A_1$ and target object $b_{(l+1)j} \in A_{l+1}$ is:

$$DPRel\left(a_{1i}, b_{(l+1)j}|\mathcal{P}\right) = \frac{w\left(a_{1i}, b_{(l+1)j}\right)\left(\frac{1}{\deg(a_{1i})} + \frac{1}{\deg\left(b_{(l+1)j}\right)}\right)}{\frac{1}{\deg(a_{1i})}\sum_j w\left(a_{1i}, b_{(l+1)j}\right) + \frac{1}{\deg\left(b_{(l+1)j}\right)}\sum_i w\left(a_{1i}, b_{(l+1)j}\right)} \quad (1)$$

where $w(a_{1i}, b_{(l+1)j})$ is the value $M[a_{1i}, b_{(l+1)j}]$ from weighted path matrix i.e. the number of paths connecting objects $a_{1i} \in A_1$ and $b_{(l+1)j} \in A_{l+1}$ following the specified meta-path. $\deg(a_{1i})$ and $\deg(b_{(l+1)j})$ are node degrees of objects a_{1i} and $b_{(l+1)j}$ respectively in the bipartite representation.

Since in this work, we have to perform classification of objects, we need to measure the relatedness between same-typed objects i.e., source and target object type of a meta-path would be same. In this case, we measure the relatedness between same-typed objects as explained in Gupta et al. (2015). An example is given below.

For the heterogeneous information network shown in Fig. 2 and the weighted path matrix for this network shown in Fig. 5, following meta-path *APCPA*, we compute the relatedness between authors. The middle object type of the meta-path is C (Conference). Using this object type we divide the meta-path *APCPA* into two equal length sub-paths i.e., $\mathcal{P}_L = APC$ and $\mathcal{P}_R = CPA$. Using these meta-paths we compute the relatedness between authors as shown below:

$$Sim_{\mathcal{P}_L}(a_1, c_1) = DPRel(a_1, c_1|\mathcal{P}_L) = \frac{2\left(\frac{1}{2} + \frac{1}{3}\right)}{\left(\frac{1}{2} \times 3\right) + \left(\frac{1}{3} \times 5\right)} = 0.53$$

The full relatedness matrix between authors and conferences, following meta-path $\mathcal{P}_L = APC$ is shown in Fig. 6(a). The relatedness matrix between authors and conferences following meta-path $\mathcal{P}_R^{-1} = APC$ would be same as the matrix shown in Fig. 6(a) as this meta-path is symmetric. Using these two matrices, we compute the relatedness between authors as shown below:

$$\vec{X} = DPRel(a_1, \{c_1, c_2\}|\mathcal{P}_L = APC) = \{0.53, 0.33\}$$

$$\vec{Y} = DPRel\left(a_3, \{c_1, c_2\}|\mathcal{P}_R^{-1} = APC\right) = \{0.5, 0\}$$

$$Sim_{APCPA}(a_1, a_3) = \frac{\vec{X}.\vec{Y}}{\vec{X}^2 + \vec{Y}^2 - \vec{X}.\vec{Y}} = 0.71$$

Conference

Author	c_1	c_2
a_1	0.53	0.33
a_2	0.53	0.33
a_3	0.5	0
a_4	0	0.67

(a) Relatedness between authors and conferences

Author

Author	a_1	a_2	a_3	a_4
a_1	1	1	0.71	0.36
a_2	1	1	0.71	0.36
a_3	0.71	0.71	1	0
a_4	0.36	0.36	0	1

(b) Relatedness between authors

Fig. 6. Computing relatedness between authors using DPRel and following meta-path APCPA.

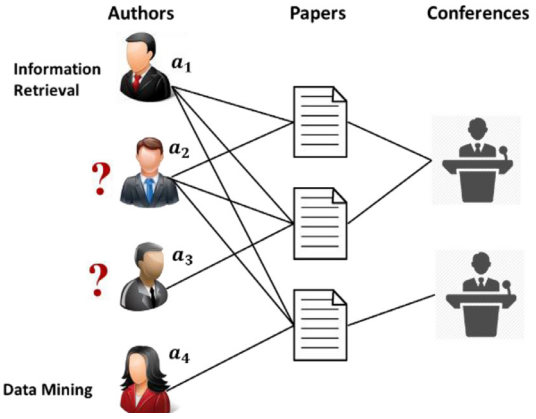


Fig. 7. Classification of authors.

The results of rest of the computations are shown in Fig. 6(b) in matrix form.

After computing the relatedness between objects on which classification is to be performed following different meta-paths, we create homogeneous information networks corresponding to those meta-paths. After that, we perform learning of weights for meta-paths and then perform a weighted combination of those individual networks. Then transductive classification is performed to label the unlabeled objects. The problem of transductive classification is formally defined in Definition 6.

Definition 6 (Transductive classification)**.** Given a heterogeneous information network $G = (V, E), V = U_{(i=1)}^{m} V_i$ where $V_i \in A_i, i = 1, \ldots, m$ and a subset of target data objects $V_T' \subset V_T \in A_T$ which are labelled with values $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ denoting class label to which each object belongs to, predict the labels for the unlabelled objects of target type $(V_T - V_T') \in A_T$.

For example, in Fig. 7, the target type objects are authors on which we have to perform the classification. Authors a_1 and a_4 are pre-labelled with class labels "Information Retrieval" and "Data Mining" respectively. The task is to classify authors a_2 and $a_3$using the pre-labelled information. Since following various paths we can connect authors to form homogeneous information networks, weight learning is required to determine the importance of various paths. The proposed framework HeteClass explores various meta-paths from the network schema and performs weight learning using the pre-labelled information to leverage the semantics of various meta-paths for classification of objects.

4. The framework of heteclass

Transductive classification utilizes dependency among data objects for relational learning to classify unlabeled objects. In this section, we present the proposed framework HeteClass for transductive classification on target type objects in a heterogeneous information network. The framework of HeteClass performs the classification task in two phases. In the first phase, a set of meta-paths is generated and knowledge of domain expert(s) is utilized (if available) to filter the meta-paths produced by exploring the input network schema. This would reduce the number of meta-paths to be explored for classification. In the second phase, weight learning is performed for the final set of met-paths produced by the first phase using prior label information in the network. After the weight learning, a single homogeneous information network is produced by a weighted combination of various homogeneous information networks corresponding to each meta-path. Then, using the transductive classification algorithm, the unlabelled set of objects in the network is classified. Fig. 8 shows the detailed framework of HeteClass. In this framework, there are two phases:

- Generation of meta-paths, and
- Weight learning and transductive classification.

4.1. Phase 1: generation of meta-paths

In this phase of the HeteClass framework, network schema of the heterogeneous information network is explored to generate all meta-paths that start from and end with the target object type. For example, Fig. 9(a) shows the network schema of a bibliography dataset. In this network schema, if the target object type is Author (A), the exploration of the network schema to generate meta-paths that start from A and end with A would be as shown in Fig. 9(b). Meta-paths generated in this example would be in the format of $author - * - author$ i.e., the start and end type objects would be same.

The pseudo code of the algorithm for meta-path generation is given in Algorithm 1. This algorithm takes the network schema of the heterogeneous information network and target object type on which classification is to be performed as input. It also takes as input the maximum length of meta-paths to be generated. The start node of the algorithm is the target object type. Algorithm 1 generates meta-paths by exploring the neighbors of the current node and if any node is the target object type, it generates the meta-path by storing the meta-path starting from the target object type and ending with the same. Algorithm 1 has two functions. The first function is used to generate the neighbors of the current node and the second function generates meta-paths by merging the two sub-paths that have the same end and start object type. This process is repeated for a number of iterations as shown in Fig. 9(b) which would be equal to the maximum length for meta-paths.

After generation of meta-paths, we can utilize domain expert knowledge for reducing the number of meta-paths by discarding the paths that would not lead to good classification accuracy. For a network, it might be possible to have a large number of meta-paths for weight learning. This would increase the computation

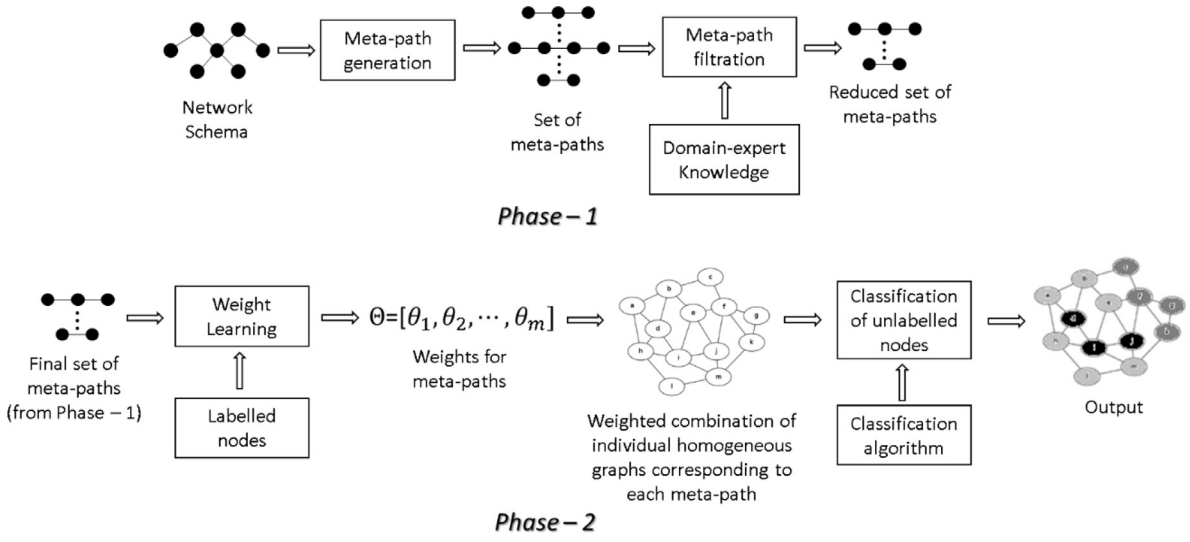
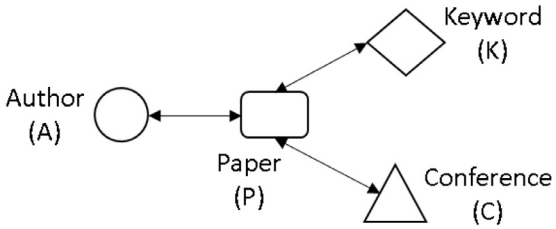
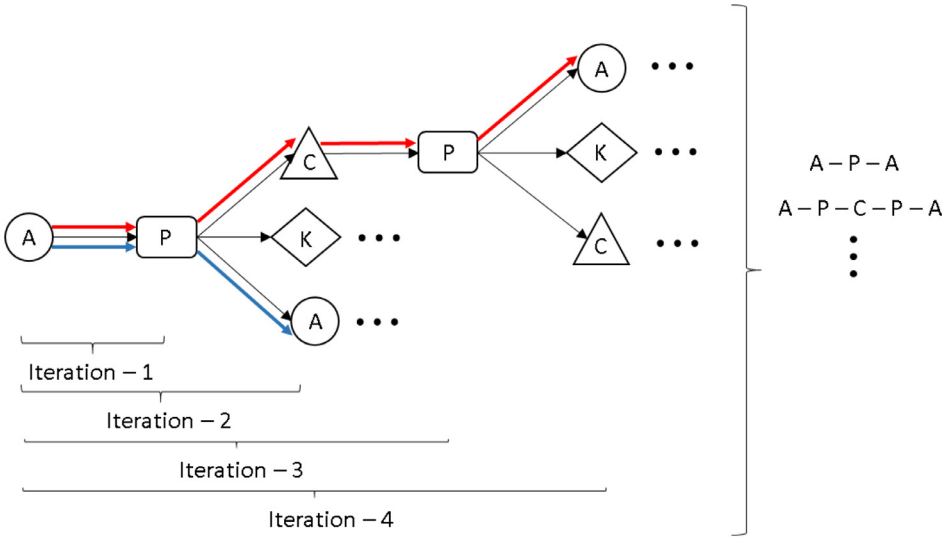


Fig. 8. HeteClass framework.



(a) Network schema of bibliography dataset



(b) Generation of meta-paths for target object type

Fig. 9. Meta-paths generation from the network schema.

time. However, if we remove the paths that would lead to low classification accuracy, the computation time would be reduced. Also, by removing ineffective paths, weight learning would become more effective as more weight would be assigned to effective paths.

4.2. Phase 2: weight learning and transductive classification

For the set of meta-paths produced from phase-1 using Algorithm 1 of HeteClass, weight learning is performed using prior label information. The weight learning for meta-paths is important as each meta-path has a semantic meaning, which is different from other meta-paths. For example, meta-path APA semantically signifies the co-author relationship. However, meta-path APCPA semantically means authors publishing papers in the same conference. Each meta-path would yield a different accuracy of classification and higher weights should be assigned to meta-paths that result in good classification accuracy. We have posed the problem of weight learning for meta-paths as an optimization problem as defined in

Algorithm 1 Generation of meta-paths.

Input:
T_G: network schema of heterogeneous information network in adjacency matrix form
A_T: target object type
l_{max}: maximum length for meta-paths to be generated
Output:
$Path\_set = \{\mathcal{P}_1, \mathcal{P}_2, \ldots\}$: set of meta-paths from T_G in form of $A_T - * - A_T$
Begin
1. $Path\_set \leftarrow create\_list()$ //create a dynamic array of empty lists
2. $first\_list \leftarrow list(A_T)$ //create a list having target object type
3. ***for*** $i \leftarrow 1$ to l_{max} ***do***
4. $\quad second\_list \leftarrow next\_list(first\_list)$ //function call
5. $\quad com\_list \leftarrow merge\_list(first\_list, second\_list)$ //function call
6. $\quad first\_list \leftarrow com\_list$
7. $\quad$ ***for*** $j \leftarrow 1$ to $length(first\_list)$ ***do***
8. $\quad\quad$ ***if*** last element of $first\_list[j] == A_T$ then
9. $\quad\quad\quad Path\_set \leftarrow Path\_set \cup first\_list[j]$
10. $\quad\quad$ ***end if***
11. $\quad$ ***end for***
12. ***end for***

$next\_list(list\_one)$ //function
{
1. $list\_two \leftarrow create\_list()$
2. $nodes \leftarrow NULL$
3. ***for*** $i \leftarrow 1$ to $length(list\_one)$ ***do***
4. $\quad nodes \leftarrow nodes \cup last\ element\ of\ list\_one[i]$
5. ***end for***
6. $nodes \leftarrow unique(nodes)$ //take only unique nodes
7. ***for*** $i \leftarrow 1$ to $length(nodes)$ ***do***
8. $\quad neb \leftarrow neighbors\ of\ nodes[i]$ //find neighbours
9. $\quad$ ***for*** $j \leftarrow 1$ to $length(neb)$ ***do***
10. $\quad\quad$ //creation of meta-paths by concatenation of two elements and
11. $\quad\quad$ insertion into array of lists
12. $\quad\quad list\_two \leftarrow list\_two \cup concatenate(nodes[i], neb[j])$
13. $\quad$ ***end for***
14. ***end for***
15. *return(list_two)*

}

$merge\_list(list\_one, list\_two)$ //function
{
1. $mlist \leftarrow create\_list()$
2. ***for*** $i \leftarrow 1$ to $length(list\_one)$ ***do***
3. $\quad$ ***for*** $j \leftarrow 1$ to $length(list\_two)$ ***do***
4. $\quad\quad$ ***if*** last element of $list\_one[i]$ == first element of $list\_two[j]$ then
5. $\quad\quad\quad r \leftarrow$ merge $list\_one[i]$ and $list\_two[j]$
6. $\quad\quad\quad mlist \leftarrow mlist \cup r$
7. $\quad\quad$ ***end if***
8. $\quad$ ***end for***
9. ***end for***
10. *return(mlist)*

}
End

Eq. (2).

$$\Theta^* = \underset{\Theta=\{\theta_1,\theta_2,\ldots,\theta_K\}}{\operatorname{argmin}} \; L(\Theta) \tag{2}$$

Here, $\theta_k, k = 1, \ldots, K$ is the importance of the meta-path $\mathcal{P}_k, k = 1, \ldots, K$ and $L(\Theta)$ is the loss function, defined below in Eq. (3), that is to be minimized.

$$L(\Theta) = \frac{1}{2} \sum_{v_i, v_j \in V_T', i \neq j} \left\| 1 - Sign\left(v_i, v_j\right) \sum_{k=1}^{K} \theta_k Sim_{\mathcal{P}_k}\left(v_i, v_j\right) \right\|_2^2 + \frac{\lambda}{2} \|\Theta\|_2^2$$
$$s.t.\ \theta_k \geq 0, \forall k = 1, \ldots, K \tag{3}$$

Here, v_i and v_j are the labelled objects and are of target type objects on which classification is to be performed. λ is the regularization parameter and $\| \cdot \|$ is the ℓ^2- norm. The function $Sign()$ is defined in Eq. (4).

$$Sign\left(v_i, v_j\right) = \begin{cases} 1, & v_i, v_j \in C_T \text{ and } i \neq j \\ -1, & otherwise \end{cases} \tag{4}$$

It returns $+1$ if both objects have the same label, otherwise it returns -1. The function $Sim_{\mathcal{P}_k}()$ computes the relatedness between target type objects following meta-path $\mathcal{P}_k$. For relatedness computation, we utilize the relevance measure DPRel (Gupta et al., 2015).

Algorithm 2 Weight learning for meta-paths.

Input:
$G = (V, E)$: heterogeneous item information network
$\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K\}$: set of meta-paths from G
λ: regularization parameter
Sim_{P_k}: relatedness matrix for target type objects using *DPRel* following meta-path $P_k, k = 1, \ldots, K$
ϵ:convergence tolerance
Output:
$W = \{w_{\mathcal{P}_1}, w_{\mathcal{P}_2}, \ldots, w_{\mathcal{P}_K}\}$: weights of meta-paths
Begin
1. initialize $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\} > 0$
2. ***repeat***
3. $\quad \Theta_{old} = \Theta$;
4. $\quad$ Calculate $\theta_k, \forall k = 1, \ldots, K$ using Eq. (6)
5. $\quad \theta_k' = \max(0, \theta_k), \forall k = 1, \ldots, K$;
6. $\quad \Theta = [\theta_1', \theta_2', \ldots, \theta_K']$ using $\theta_k', \forall k = 1, \ldots, K$
7. ***until*** $|\Theta - \Theta_{old}| \leq \epsilon$
8. ***for*** $q \leftarrow 1$ to K ***do***
9. $\quad w_{\mathcal{P}_k} = \frac{\theta_k}{\sum_k \theta_k}$
10. ***end for***
End

The value of loss function defined in Eq. (2), is high for a meta-path if, following that meta-path, the relatedness between differently labeled objects are high. However, if a meta-path connects objects with the same label with high relatedness, then the importance assigned to that meta-path would be high. So, the idea is to maximize the correlations between objects with the same label and to minimize the correlation between differently labeled objects. To solve the optimization problem defined in Eq. (2), we take the partial derivative of loss function with respect to $\theta_k, k = 1, \ldots, K$ and equate it to zero to get the value of $\theta_k, k = 1, \ldots, K$ as defined in Eq. (5).

$$\frac{\partial L(\Theta)}{\partial \theta_k} = 0 \tag{5}$$

Now, we solve Eq. (5) and get the value of θ_k as given in Eq. (6). Using this $\theta_k, k = 1, \ldots, K$ equation, we determine the optimal value of the weights iteratively as given in Algorithm 2.

$$\theta_k = \frac{\sum_{v_i, v_j \in V_T', i \neq j} Sign\left(v_i, v_j\right) Sim_{\mathcal{P}_k}\left(v_i, v_j\right) f\left(v_i, v_j\right)}{\lambda + \sum_{v_i, v_j \in V_T',\ i \neq j} Sign^2\left(v_i, v_j\right) Sim^2_{\mathcal{P}_k}\left(v_i, v_j\right)} \tag{6}$$

where $f\left(v_i, v_j\right) = \left(1 - Sign\left(v_i, v_j\right) \sum_{r \neq k} \theta_r Sim_{\mathcal{P}_k}\left(v_i, v_j\right)\right)$.

In Algorithm 2, all importance values are first initialized with positive values greater than zero. Then, using Eq. (6), we compute the importance of each path using the importance value of other meta-paths. We take the maximum between zero and the new value of importance of meta-path to ensure that the importance does not acquire negative value. This process is iterated till the importance values are converged. To check the convergence condition, we use convergence tolerance with the appropriate value. After that, we compute the weight of meta-paths by normalizing the importance values.

After getting the weights of meta-paths, a single homogeneous information network is created on target type objects using meta-paths. For each meta-path, the corresponding homogeneous information network, represented using the relatedness matrix, is multiplied with the weight of that meta-path and summation is performed to get a single homogenous network as given below in

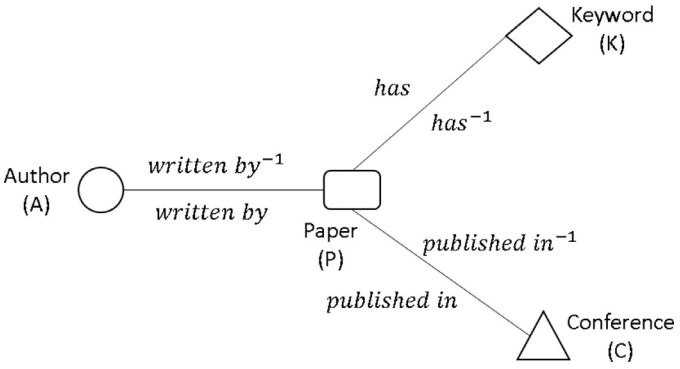


Fig. 10. Network Schema for DBLP database.

Table 2
Class label distribution of authors in DBLP "four-area" dataset.

Class	# Authors
Database	1,197
Data Mining	745
Artificial Intelligence	1,109
Information Retrieval	1,006

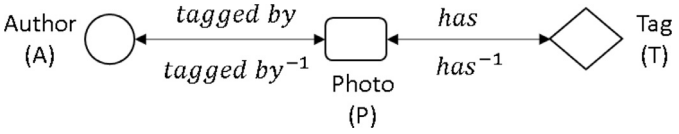


Fig. 11. Network Schema for Flickr Fashion 10,000 dataset.

Eq. (7).

$$Sim_{\mathcal{P}_1,\ldots,\mathcal{P}_K} = \sum_{k=1}^{K} w_k \times Sim_{\mathcal{P}_k} \tag{7}$$

Now, on this single homogeneous information network, which is the weighted combination of different homogeneous information networks, we perform transductive classification using prior label information. For that, we utilize Personalized PageRank (Haveliwala, 2002). However, our framework can utilize any transductive classification algorithm; by taking more advanced and accurate algorithm we can improve the classification accuracy.

5. Experimental setup

To validate the effectiveness of the proposed framework HeteClass for classification of objects in a heterogeneous information network, we utilized DBLP "four-area" data set which is a bibliography database, and a subset of Flickr Fashion 10,000 data set. We compared the performance of HeteClass with the algorithms discussed in Section 5.2. All experiments were performed on a system with Intel Core i5 processor and 4 GB RAM using R version 3.0.3.

5.1. Description of datasets

In this section, we present the description of the DBLP "four-area" dataset and Flickr Fashion 10,000 dataset used for classification.

5.1.1. DBLP "four-area" dataset

For performance comparison, we utilized DBLP database,[1] which is a computer science bibliography database. DBLP database can be modeled as a heterogeneous information network (Ji et al., 2010). This network dataset consists of four types of objects and three different relationships. The network schema of DBLP dataset is shown in Fig. 10. The network schema of DBLP database has four different types of objects represented as nodes in the network i.e., *Author* (A), *Conference* (C), *Keyword* (K) and *Paper* (P). Three different relationships, represented as links between objects, exist in the network. A bidirectional link between *Paper* and *Author* nodes indicates that every paper in the database has been written by some author(s) and vice-versa. Similarly, bidirectional links exist between *Paper* and *Keyword* as well as *Paper* and *Conference* nodes. The link from *Author* node to *Paper* node represents the relationship "*written by*" which means the paper has been written by the author and inverse relationship "*writen by*$^{-1}$" means that the author has written the paper. Likewise, other relationships in the network can be explained in forward and backward directions. When a relationship between two nodes in a network is bidirectional, we can use an undirected link to represent that relationship. For example, the relationship between author and paper nodes is in forward and in the backward direction; therefore, we can use an undirected link to show the relationship.

In our experiments, we utilized DBLP "four-area" dataset,[2] which is a subset of the DBLP database and has conferences in four research area classes: Artificial Intelligence, Information Retrieval, Database and Data Mining. This dataset has been frequently utilized in various studies (Gupta et al., 2015; Ji et al., 2010; Shi et al., 2014). The dataset contains 20 conferences, 14,475 authors, 14,376 papers and 8,920 keywords with 170,794 links in total. In this dataset, 4,057 authors, 100 papers, and all 20 conferences are labeled with one of the four research area classes. For accuracy evaluation, we need to have the ground truth (label information) for objects to be classified. The label sets are conceptually different for different object types (Angelova et al., 2012; Kong et al., 2012). In this work, we perform transductive classification on authors as we have label information for authors and number of authors is large enough to draw a meaningful conclusion from experiments. Table 2 shows the four classes and a corresponding number of authors in each class in the data set.

For experiments, we created two networks consisting of 2,500, and 4,057 authors having label information. These two datasets are named DBLP – 1, and DBLP – 2 respectively. The description of these two datasets is given in Table 3. The authors in these datasets are selected randomly from the set of labeled authors. DBLP – 2 has all the labeled authors.

5.1.2. Flickr fashion 10,000 dataset

To show the effectiveness of HeteClass, we utilized another real-world dataset named Flickr Fashion 10,000.[3] This dataset consists of 32,398 photos (URLs of images on Flickr[4]) related to various fashion categories. These photos have been categorized into 262 distinct fashion classes containing at least 10 photos in a class and at most 200 photos in a class. Each photo in the dataset has been tagged by authors. The total number of distinct tags is 56,275. Various meta-information related to photos like the number of favorites, the number of comments, and geo-location are available in the dataset. However, we utilized only author and tag information related to photos as these are the relevant information suitable for classification task in our experiments. The resulting heterogeneous information network has the network schema as shown in Fig. 11.

In the network schema, three different types of nodes are there i.e. *Author* (A), *Photo* (P), and *Tag* (T). Also, two bidirectional re-

[1] http://dblp.uni-trier.de/.

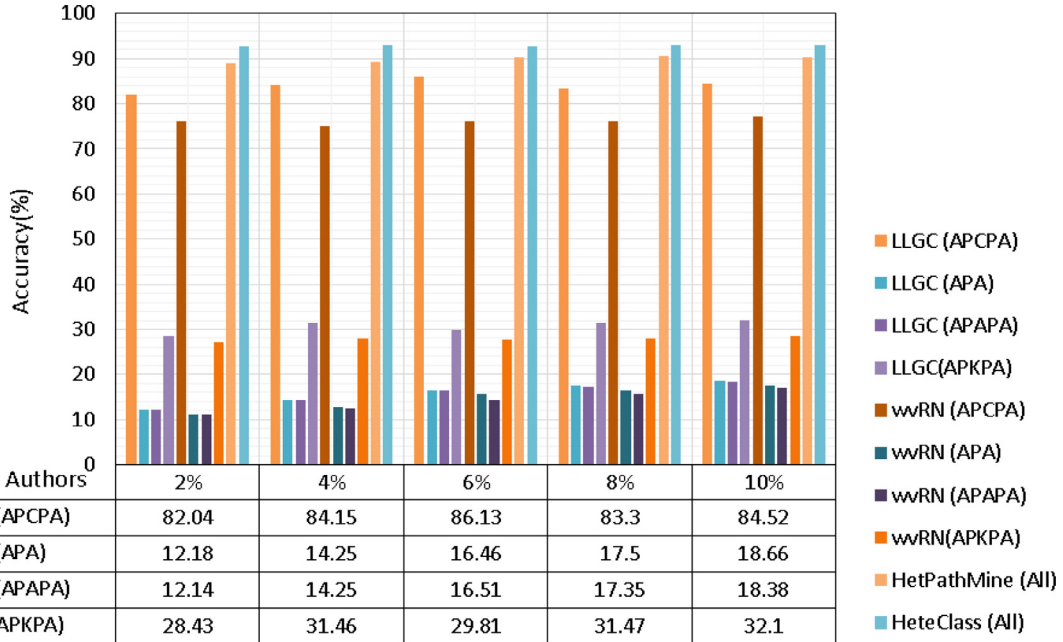
[2] http://web.engr.illinois.edu/~mingji1/.

[3] http://www.st.ewi.tudelft.nl/~bozzon/fashion10000dataset/.

[4] https://www.flickr.com/.

Table 3
Description of the two DBLP datasets utilized for experiments.

	Authors	Papers	Conferences	Keywords	Total Objects	Total Links
DBLP – 1	2,500	9,899	20	7,316	19,735	100,963
DBLP – 2	4,057	14,328	20	8,898	27,303	148,246



Labeled Authors	2%	4%	6%	8%	10%
LLGC (APCPA)	82.04	84.15	86.13	83.3	84.52
LLGC (APA)	12.18	14.25	16.46	17.5	18.66
LLGC (APAPA)	12.14	14.25	16.51	17.35	18.38
LLGC(APKPA)	28.43	31.46	29.81	31.47	32.1
wvRN (APCPA)	76.06	74.97	76.17	76.16	77.33
wvRN (APA)	11.24	12.85	15.78	16.55	17.69
wvRN (APAPA)	11.01	12.54	14.28	15.66	16.9
wvRN(APKPA)	27.1	28.07	27.79	28.09	28.57
HetPathMine (All)	89.1	89.32	90.34	90.53	90.2
HeteClass (All)	92.8	92.93	92.86	92.93	92.88

Fig. 12. Accuracy results for DBLP – 1.

Table 4
Class label distribution of Photos in the subset of Flickr Fashion 10,000 dataset utilized for experiments.

Class	# Photos	Class	# Photos
Haute Couture	206	Jeans	199
Band Collar	200	Mitre	199
Boat Neck	200	Robe	199
Cope	200	Style Line	199
Jodhpurs	200	Alb	198
Rubber Glove	200	Toile	197
Umbrella	200	Tuxedo	197
Wetsuit	200	Sari	196
Dirndl	199	Sash	196
Hijab	199	Toga	196

lationships are present in the network schema. For example, between Author and Photo nodes, the relationship "*tagged by*" indicates that the photo has been tagged by author(s). For our experiments, we utilized a subset of this Flickr Fashion 10,000 dataset. The extracted subset contains total 3,980 photos categorized into 20 classes. The class label distribution of these photos has been shown in Table 4.

For experiments, we created two networks of the extracted subset consisting of 2,500, and 3,980 photos having label information and performed classification of photos. These two datasets are named Fashion – 1, and Fashion – 2 respectively. The photos in these two datasets are selected randomly from the set of labeled photos. The description of these two datasets is given in Table 5. Fashion – 2 has all the labeled photos.

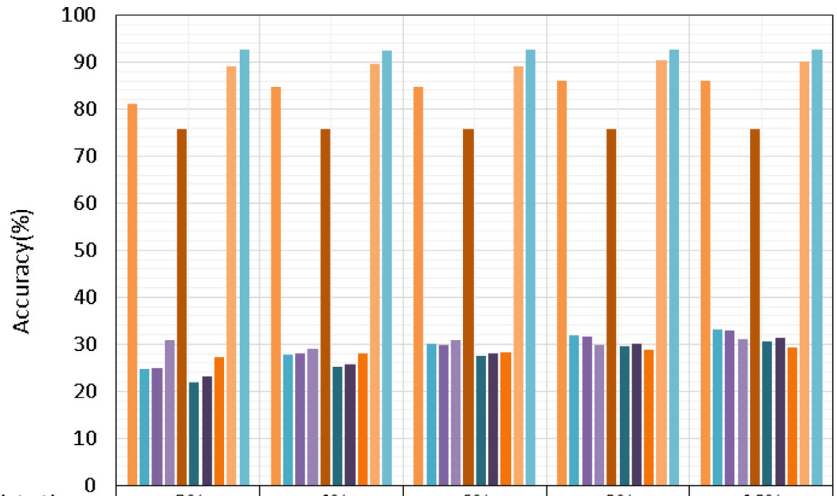
Table 5
Description of the two Flickr Fashion 10,000 datasets utilized for experiments.

	Photos	Authors	Tags	Total Objects	Total Links
Fashion – 1	2,500	539	7,061	10,100	47,327
Fashion – 2	3,980	734	9,296	14,010	75,015

For performance evaluation of algorithms on each dataset, we randomly select *x*%, (where $x = 2, 4, 6, 8$ and 10) of the labelled objects of the target type from that dataset, and use their label information as prior knowledge. Using the prior knowledge, we perform classification of the rest of the objects of target type and evaluate the accuracy of the algorithms. This process is repeated 10 times for each value of *x* and average value of accuracy is reported in the results.

5.2. Algorithms for comparison and evaluation

To show the effectiveness of the proposed framework HeteClass, we compared its performance with two algorithms: Learning with Local and Global Consistency (LLGC) (Zhou et al., 2004) and



Labeled Authors	2%	4%	6%	8%	10%
LLGC (APCPA)	81.09	84.67	84.86	86.14	86.12
LLGC (APA)	24.63	27.74	30.02	31.98	33.27
LLGC (APAPA)	24.89	27.93	29.83	31.63	32.91
LLGC(APKPA)	30.79	28.98	30.92	29.69	31.17
wvRN (APCPA)	75.68	75.75	75.76	75.79	75.82
wvRN (APA)	21.73	25.18	27.38	29.52	30.55
wvRN (APAPA)	23.14	25.7	27.99	30.17	31.3
wvRN(APKPA)	27.2	27.89	28.15	28.83	29.32
HetPathMine (ALL)	89.13	89.66	89.17	90.47	90.07
HeteClass (ALL)	92.6	92.56	92.58	92.59	92.58

Fig. 13. Accuracy results for DBLP – 2.

weighted-vote Relational Neighbour classifier (wvRN) (Macskassy & Provost, 2003).

LLGC is a graph-based transductive algorithm for classification of nodes. It utilizes the link structure of the network to propagate label information to the rest of the nodes for classification. LLGC utilizes the local as well as the global structure of the network for classification. However, wvRN is a simple relational learning algorithm, which utilizes only the local structure of the network for classification of nodes. These two algorithms are very popular and widely utilized for classification task on networks. Neither LLGC nor wvRN is directly applicable to a heterogeneous information network. We can apply these algorithms only after transforming the heterogeneous information network into a homogeneous information network by following a meta-path. In our experiments, by following different meta-paths, we transform a heterogeneous information network into the corresponding homogeneous information network, which consists of only target type objects. Then, we apply LLGC and wvRN algorithms on each homogeneous information network to perform classification on target type objects. We also compare the performance of HeteClass with HetPathMine algorithm by Luo et al. (2014). HetPathMine determines the weights of different meta-paths and combines the corresponding network into a single homogeneous information network. Then classification is performed on the target type objects.

To evaluate the performance of a classification algorithm, we utilize accuracy measure (Han, Kamber, & Pei, 2011). Accuracy measures compute the accuracy of the algorithm for classification of unlabeled objects. It can be computed as the ratio of correctly classified objects to the total number of unlabelled objects (Han et al., 2011). Accuracy measure is defined as follows in Eq. (8):

$$Accuracy = \frac{\#Correctly\ classified\ objects}{N} = \frac{\sum_{i=1}^{K} a_i}{N} \quad (8)$$

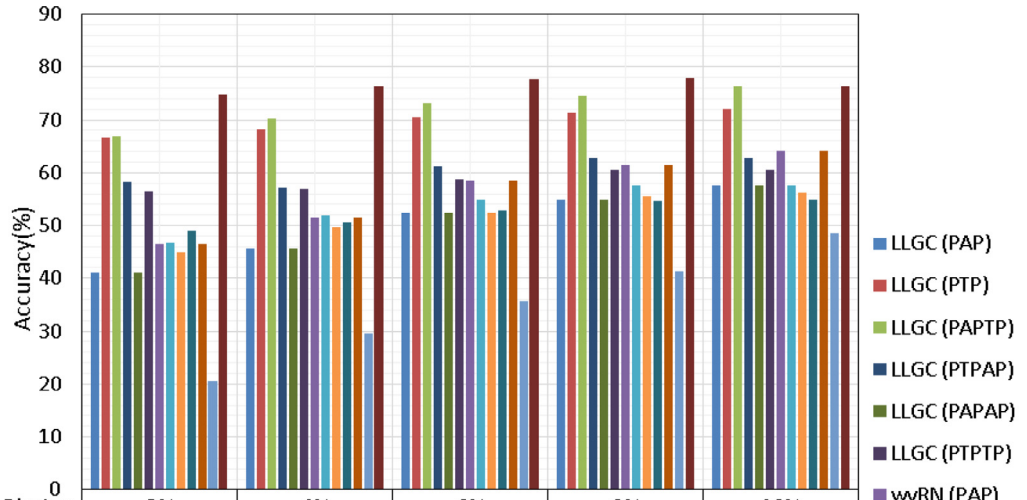
where N is the number of unlabelled objects need to be classified, K is the number of classes in the dataset and a_i is the number of objects correctly classified to its actual class.

Table 6
Final set of meta-paths generated by Phase – 1 of HeteClass for DBLP – 1 and DBLP – 2 datasets.

Meta-path	Length	Symmetric
Author – Paper – Author (APA)	2	Yes
Author – Paper – Conference – Paper – Author (APCPA)	4	Yes
Author – Paper – Author – Paper – Author (APAPA)	4	Yes
Author – Paper – Keyword – Paper – Author (APKPA)	4	Yes

5.3. Meta-path generation and selection

To utilize the semantics of various meta-paths, HeteClass generates a set of meta-paths (in Phase – 1) that can be combined to form the homogeneous information network consisting of objects that need to be classified. For our experiments, we consider the network schema of DBLP as shown in Fig. 10 for HeteClass to generate the set of meta-paths. In this work, we consider meta-paths of length up to four since higher length meta-paths would be highly noisy for the classification process (Kong et al., 2012; Shi et al., 2014) and affect the accuracy of results. The final set of meta-paths generated and utilized in our experiments for dataset DBLP – 1 and DBLP – 2 is shown in Table 6. The four meta-paths utilized for experiments are symmetric (Gupta et al., 2015).



Labeled Photos	2%	4%	6%	8%	10%
LLGC (PAP)	41.13	45.66	52.38	54.93	57.47
LLGC (PTP)	66.7	68.13	70.5	71.4	72.14
LLGC (PAPTP)	66.75	70.18	73.18	74.52	76.33
LLGC (PTPAP)	58.23	57.1	61.2	62.86	62.67
LLGC (PAPAP)	41.15	45.56	52.32	54.98	57.47
LLGC (PTPTP)	56.38	56.93	58.82	60.6	60.54
wvRN (PAP)	46.55	51.54	58.54	61.37	64.23
wvRN (PTP)	46.7	51.96	54.76	57.61	57.67
wvRN (PAPTP)	45.03	49.59	52.42	55.58	56.28
wvRN (PTPAP)	49.1	50.59	52.94	54.64	54.76
wvRN (PAPAP)	46.55	51.54	58.54	61.37	64.23
wvRN (PTPTP)	20.54	29.48	35.69	41.39	48.55
HeteClass (All)	74.65	76.42	77.68	77.95	76.27

Fig. 14. Accuracy results for Fashion – 1.

Table 7
Final set of meta-paths generated by Phase – 1 of HeteClass for Fashion – 1 and Fashion – 2 datasets.

Meta-path	Length	Symmetric
Photo – Author – Photo (PAP)	2	Yes
Photo – Tag – Photo (PTP)	2	Yes
Photo – Author – Photo – Tag – Photo (PAPTP)	4	No
Photo – Tag – Photo – Author – Photo (PTPAP)	4	No
Photo – Author – Photo – Author – Photo (PAPAP)	4	Yes
Photo – Tag – Photo – Tag – Photo (PTPTP)	4	Yes

For Fashion – 1 and Fashion – 2 datasets, the set of meta-paths generated are listed in Table 7. For these datasets, total six meta-paths are generated in Phase – 1 of HeteClass considering network schema shown in Fig. 11. Out of six meta-paths, four meta-paths are symmetric and the rest two meta-paths are asymmetric.

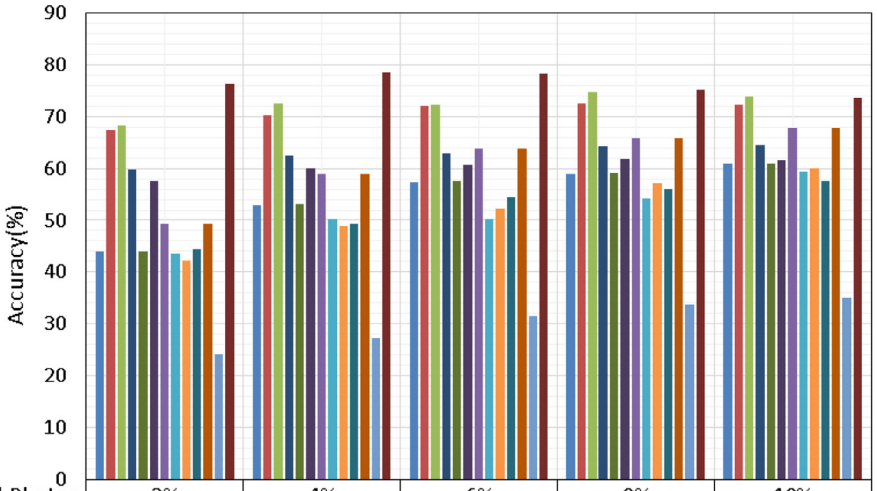
6. Results and discussion

In this section, we present the results of the comparison of HeteClass with other algorithms. Since LLGC and wvRN algorithms cannot combine the semantics of meta-paths, we apply these algorithms to the homogeneous information networks created by following each meta-path individually. However, HetPathMine can perform weighted combination of the networks corresponding to the symmetric meta-paths (Luo et al., 2014; Sun et al., 2011). Therefore, HetPathMine utilizes only symmetric meta-paths simultaneously.

6.1. Results for DBLP – 1 and DBLP – 2 datasets

The accuracy results for DBLP – 1 and DBLP – 2 datasets are shown in Figs. 12 and 13 respectively. From the figures, it is clear that HeteClass outperforms the considered algorithms for classification of authors for these two datasets. HeteClass has consistently outperformed other algorithms and its performance is stable. HeteClass gives an improvement in accuracy of more than 3% over HetPathMine. From the results, we can see that for LLGC and wvRN algorithms, meta-path APCPA performs better as compared to the rest of the meta-paths. The accuracy of LLGC and wvRN is not good for meta-paths APA and APAPA. The reason might be that these paths are not capturing the semantics required for good classification accuracy. However, meta-path APKPA has given better performance for LLGC and wvRN as compared to meta-paths APA and APAPA.

This shows that the classification of objects in a heterogeneous information network by leveraging semantics of various meta-paths is more effective than transforming the heterogeneous information network into a homogeneous information network following a single meta-path. HeteClass combines the semantics of various meta-paths and performs the classification of objects more accurately.



Labeled Photos	2%	4%	6%	8%	10%
LLGC (PAP)	43.91	52.94	57.41	58.98	60.96
LLGC (PTP)	67.46	70.36	72	72.5	72.18
LLGC (PAPTP)	68.37	72.49	72.34	74.67	73.75
LLGC (PTPAP)	59.84	62.44	62.87	64.3	64.48
LLGC (PAPAP)	44.04	53.06	57.47	59.07	60.96
LLGC (PTPTP)	57.55	60.09	60.64	61.9	61.53
wvRN (PAP)	49.22	58.92	63.73	65.78	67.77
wvRN (PTP)	43.56	50.19	50.28	54.29	59.25
wvRN (PAPTP)	42.12	48.89	52.25	57.19	60.05
wvRN (PTPAP)	44.46	49.23	54.51	55.92	57.52
wvRN (PAPAP)	49.22	58.92	63.73	65.78	67.77
wvRN (PTPTP)	24.11	27.19	31.57	33.65	34.99
HeteClass (All)	76.32	78.58	78.24	75.14	73.58

Fig. 15. Accuracy results for Fashion – 2.

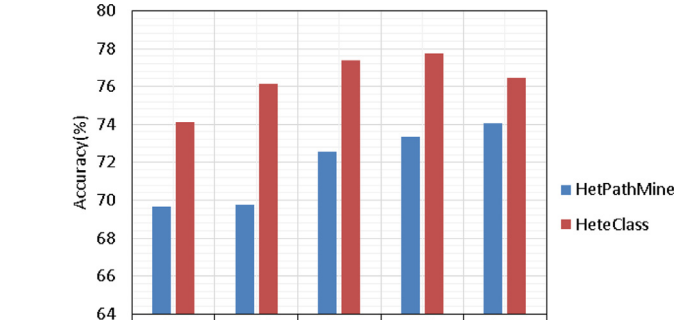
6.2. Results for fashion – 1 and fashion – 2 datasets

For Fashion – 1 and Fashion – 2 datasets, the six meta-paths listed in Table 7 are utilized. However, HetPathMine cannot utilize all six meta-paths as there are two meta-paths which are asymmetric (i.e. PAPTP and PTPAP). Since HetPathMine works only for symmetric meta-paths (Luo et al., 2014; Sun et al., 2011), therefore, we separately compare the performance of HetPathMine with HeteClass using only symmetric meta-paths.

6.2.1. Comparison of heteclass with wvRN and LLGC using all six meta-paths

Figs. 14 and 15 shows the results of the accuracy of classification for datasets Fashion – 1 and Fashion – 2 respectively. From the results, it is clear that the performance of HeteClass is better as compared to wvRN and LLGC following all six meta-paths. We can also see that the performance of HeteClass is stable and it consistently gives the almost same accuracy of classification for different sizes of training data.

From the results, we can see that the meta-path PAPTP gives the good accuracy for LLGC algorithm. However, for wvRN, meta-paths PAP and PAPAP gives the good accuracy for classification. From the results, we can see that leveraging simultaneously the semantics of all meta-paths gives the better results than following a single meta-path. We can also see that LLGC algorithm, for meta-path PAPTP, has performed slightly better than HeteClass for Fashion – 1 and Fashion – 2 datasets when labeled photos were 10% (i.e. for $x = 10$).

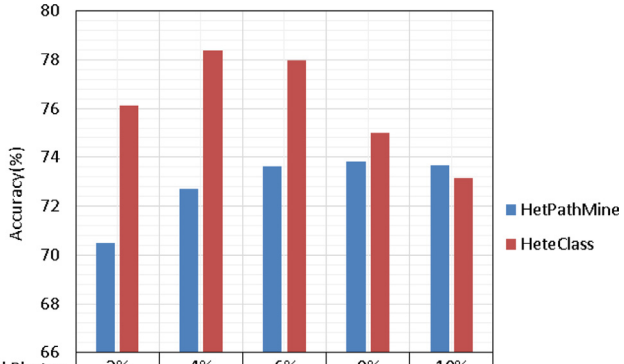


Labeled Photos	2%	4%	6%	8%	10%
HetPathMine	69.66	69.76	72.54	73.36	74.04
HeteClass	74.14	76.16	77.4	77.74	76.43

Fig. 16. Accuracy results for Fashion – 1.

6.2.2. Comparison of heteclass with hetpathmine using only symmetric meta-paths

Since meta-paths PAPTP and PTPAP are asymmetric, therefore, we cannot utilize these meta-paths for HetPathMine as it can use only symmetric meta-paths (Luo et al., 2014). Therefore, we have separately compared the performance of HeteClass with HetPathMine by considering only symmetric meta-paths. The results for datasets Fashion – 1 and Fashion – 2 are shown in Figs. 16 and 17 respectively. For this comparison, we utilized meta-paths PAP, PTP, PAPAP, and PTPTP for both HeteClass and HetPathMine. From



Labeled Photos	2%	4%	6%	8%	10%
HetPathMine	70.48	72.72	73.64	73.82	73.67
HeteClass	76.11	78.37	77.96	75	73.17

Fig. 17. Accuracy results for Fashion – 2.

the results, we can see that HeteClass outperforms HetPathMine for both datasets. The classification accuracy of HeteClass is about 4% better than the accuracy of HetPathMine.

Also, from the results shown in Figs. 14–17, we can see that the performance of HeteClass when utilizing all six meta-paths is slightly better than the performance when only symmetric meta-paths are utilized. This shows that the full utilization of semantics could be better than partially utilizing it and HeteClass is able to do that.

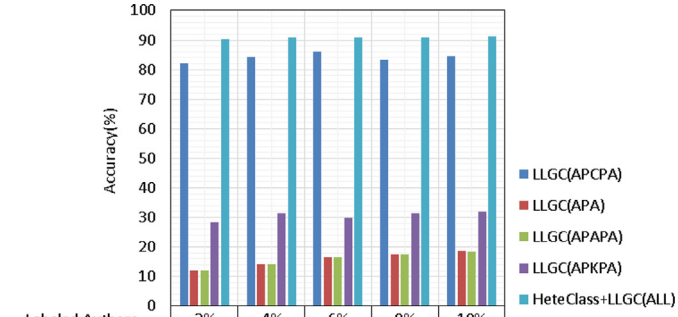
6.3. Significance of classification algorithm in heteclass framework

In the framework of HeteClass, after forming the homogeneous network in phase-2, we can apply the classification algorithm for classification of unlabeled target type objects in the network. The performance of HeteClass depends on and orthogonal to the classification algorithm utilized in the framework i.e. the performance of HeteClass can be improved by taking an advanced classification algorithm.

However, in our experimental results, the performance gain of HeteClass as compared to other algorithms is not only because of the chosen classification algorithm but also because of the ability of the framework to extract the various semantics in the network following different meta-paths and leveraging them simultaneously. To show this ability of the proposed framework HeteClass, we performed experiments on DBLP and Fashion datasets using LLGC as classification algorithm in the HeteClass framework instead of Personalized PageRank algorithm. Then, we compared the accuracy results of HeteClass utilizing LLGC (named as *HeteClass+LLGC*) with the accuracy results of LLGC. HeteClass+LLGC would be able to leverage all the meta-paths simultaneously, however, LLGC can utilize only one meta-path at a time.

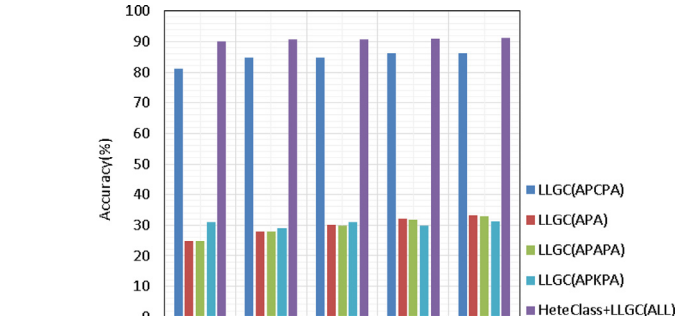
6.3.1. Comparing Heteclass + LLGC with LLGC for DBLP – 1 and DBLP – 2 datasets

The results of the comparison of HeteClass+LLGC with LLGC for datasets DBLP – 1 and DBLP – 2 are shown in Figs. 18 and 19 respectively. As we can see that the performance of HeteClass + LLGC is better as compared to the performance of LLGC following any of the meta-paths taken for experiments for both DBLP – 1 and DBLP – 2 datasets. It shows that the performance gain of HeteClass is not only because of the classification algorithm in the HeteClass framework but also due to the power of the framework to leverage the various semantics in the network.



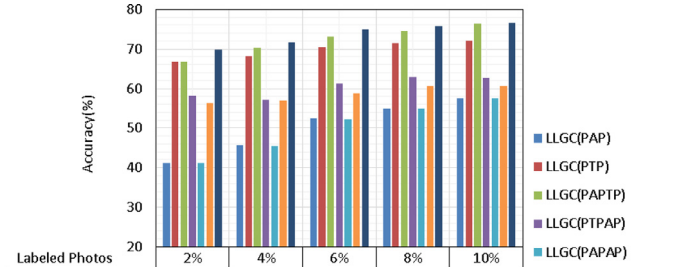
Labeled Authors	2%	4%	6%	8%	10%
LLGC(APCPA)	82.04	84.15	86.13	83.3	84.52
LLGC(APA)	12.18	14.25	16.46	17.5	18.66
LLGC(APAPA)	12.14	14.25	16.51	17.35	18.38
LLGC(APKPA)	28.43	31.46	29.81	31.47	32.1
HeteClass+LLGC(ALL)	90.19	91.03	91.05	90.93	91.13

Fig. 18. Accuracy results for DBLP – 1 using LLGC as classification algorithm in HeteClass framework.



Labeled Authors	2%	4%	6%	8%	10%
LLGC(APCPA)	81.09	84.67	84.86	86.14	86.12
LLGC(APA)	24.63	27.74	30.02	31.98	33.27
LLGC(APAPA)	24.89	27.93	29.83	31.63	32.91
LLGC(APKPA)	30.79	28.98	30.92	29.69	31.17
HeteClass+LLGC(ALL)	90.22	90.59	90.57	91.01	91.12

Fig. 19. Accuracy results for DBLP – 2 using LLGC as classification algorithm in HeteClass framework.

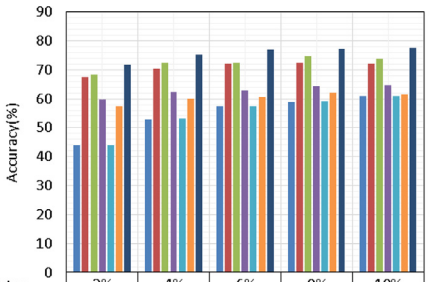


Labeled Photos	2%	4%	6%	8%	10%
LLGC(PAP)	41.13	45.66	52.38	54.93	57.47
LLGC(PTP)	66.7	68.13	70.5	71.4	72.14
LLGC(PAPTP)	66.75	70.18	73.18	74.52	76.33
LLGC(PTPAP)	58.23	57.1	61.2	62.86	62.67
LLGC(PAPAP)	41.15	45.56	52.32	54.98	57.47
LLGC(PTPTP)	56.38	56.93	58.82	60.6	60.54
HeteClass+LLGC(ALL)	69.9	71.7	74.93	75.78	76.61

Fig. 20. Accuracy results for Fashion – 1 using LLGC as classification algorithm in HeteClass framework.

6.3.2. Comparing Heteclass+LLGC with LLGC for fashion – 1 and fashion – 2 datasets

For Fashion – 1 and Fashion – 2 datasets, the results of the comparison are shown in Figs. 20 and 21 respectively. For these datasets also the performance of HeteClass+LLGC is better as compared to the performance of LLGC.



Labeled Photos	2%	4%	6%	8%	10%
LLGC(PAP)	43.91	52.94	57.41	58.98	60.96
LLGC(PTP)	67.46	70.36	72	72.5	72.18
LLGC(PAPTP)	68.37	72.49	72.34	74.67	73.75
LLGC(PTPAP)	59.84	62.44	62.87	64.3	64.48
LLGC(PAPAP)	44.04	53.06	57.47	59.07	60.96
LLGC(PTPTP)	57.55	60.09	60.64	61.9	61.53
HeteClass+LLGC(ALL)	71.9	75.24	77.09	77.37	77.57

Fig. 21. Accuracy results for Fashion – 2 using LLGC as classification algorithm in HeteClass framework.

Thus, these experiments prove that the performance gain of HeteClass is not only because of the classification algorithm in the framework but also because of the power of the framework to leverage simultaneously the various semantics present in the network. It also proves that it is important to consider and leverage the various semantics present in the network to improve the classification accuracy.

6.4. Weight learning for Meta-paths in heteclass

For leveraging the semantics of various meta-paths, it is required to determine the weight of each meta-path as some paths may be good for classification than others. Since higher weight should be assigned to paths which lead to good classification accuracy, weight learning is an important step. HeteClass performs weight learning from the labeled data for various meta-paths and assigns the highest weight to the meta-path which gives the highest accuracy among all the meta-paths.

Fig. 22 shows the accuracy of HeteClass for combined as well as individual networks corresponding to different meta-paths for DBLP – 1 and DBLP – 2 datasets. From these results, we can see that the accuracy of HeteClass corresponding to meta-path *Author – Paper – Conference – Paper – Author* (APCPA) is highest for these datasets. HeteClass is able to learn this subtlety and assigns the highest weight to this meta-path. Table 8 shows the weights assigned to different meta-paths by HeteClass and HetPathMine. HeteClass has assigned the highest weight (0.87 ∼ 0.96) to the meta-path APCPA. HetPathMine has also assigned highest weight to the same meta-path APCPA; however, HetPathMine has assigned only around 0.43 ∼ 0.58 wt to meta-path APCPA. Since meta-path APCPA gives significantly better classification accuracy among all the meta-paths considered in these experiments, the weight assigned to this meta-path should be close to 1. HeteClass has assigned the weight to meta-path APCPA close to 1.

For Fashion – 1 and Fashion – 2 datasets, the accuracy of HeteClass for the networks following individual meta-path and all meta-paths simultaneously is shown in Fig. 23(a) and (b) respectively. For comparison with weight learning process of HetPathMine, we have taken only the symmetric meta-paths while performing the classification of objects. From the figures, we can see that the paths *Photo – Tag – Photo* (PTP) and *Photo – Tag – Photo – Tag – Photo* (PTPTP) has given almost equal accuracy and better than other two paths. The weights assigned to different meta-paths by HeteClass and HetPathMine are listed in Table 9.

From Table 9, we can see that HetPathMine has assigned same weights to the paths PTPTP and PTP as they have given almost equal accuracy. Similarly, paths PAPAP and PAP have been assigned almost equal weights but less than the weights of paths PTP and PTPTP. From the results, we can understand that the reason for weight assignment is that the accuracy given by paths PAP and PAPAP is low as compared to paths PTP and PTPTP. However, HetPathMine has assigned the highest weight to the path PTPTP and path PTP has been assigned very low weight, even though following that path we can get high accuracy. This shows that the weight learning process of HeteClass is more effective than the weight learning process of HetPathMine.

6.5. Discussion

From the results on both datasets i.e. DBLP and Flickr Fashion, we can understand that meta-path based classification on target type objects is effective. A meta-path contains the semantic which could be highly significant for classification or other mining tasks. Therefore, we should consider the semantics of meta-paths and should leverage the semantics of various meta-paths simultaneously for effective results.

The proposed framework, HeteClass, demonstrates that leveraging the various semantics present in the network can improve the classification accuracy on target type objects. The method proposed in this work for assigning weights to various meta-paths considered for a weighted combination of various semantics is

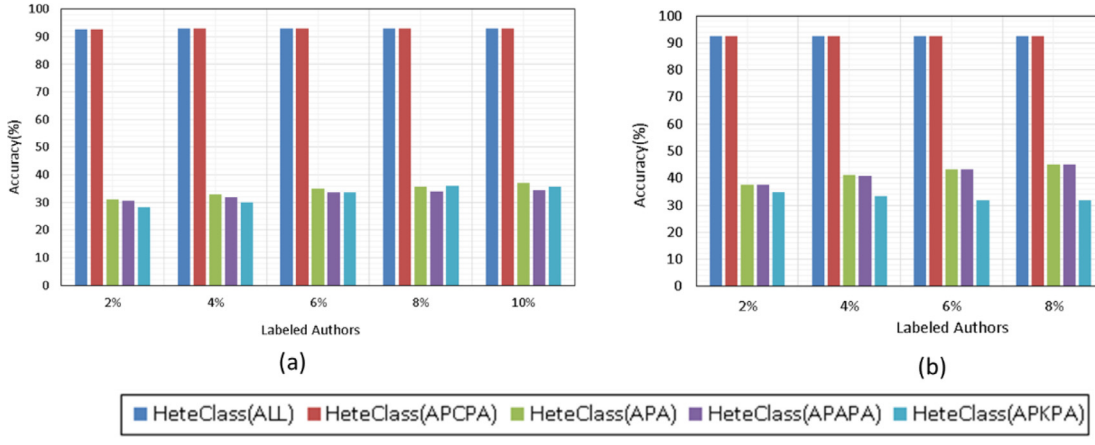


Fig. 22. Accuracy results of HeteClass for individual as well as for weighted combinations of networks corresponding to different meta-paths for DBLP – 1 and DBLP – 2 datasets.

Table 8
Assignment of weights to different meta-paths for DBLP datasets.

Meta-path	Weights assigned by HeteClass	Weights assigned by HetPathMine
Author – Paper – Author(APA)	0.001 ∼ 0.01	0.24 ∼ 0.39
Author – Paper – Conference – Paper – Author (APCPA)	0.87 ∼ 0.96	0.43 ∼ 0.58
Author – Paper – Author – Paper – Author (APAPA)	0.01 ∼ 0.08	0.08 ∼ 0.13
Author – Paper – Keyword – Paper – Author (APKPA)	0.001 ∼ 0.18	0.1 ∼ 0.26

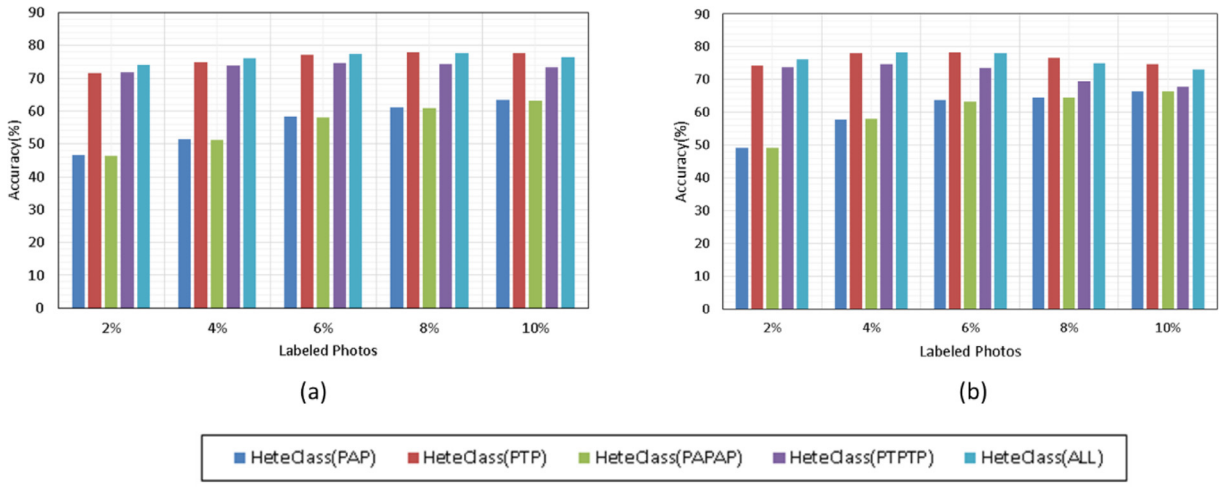


Fig. 23. Accuracy results of HeteClass for individual as well as for weighted combinations of networks corresponding to different meta-paths for Fashion dataset.

Table 9
Assignment of weights to different meta-paths for Fashion datasets.

Meta-path	Weights assigned by HeteClass	Weights assigned by HetPathMine
Photo – Author – Photo (PAP)	0.15 ∼ 0.2	0 ∼ 0.23
Photo – Tag – Photo (PTP)	0.2 ∼ 0.33	0 ∼ 0.1
Photo – Author – Photo – Author – Photo (PAPAP)	0.12 ∼ 0.19	0.1 ∼ 0.3
Photo – Tag – Photo – Tag – Photo (PTPTP)	0.23 ∼ 0.35	0.5 ∼ 0.9

effective and applicable for datasets from different domains. The weights assigned by HeteClass to different meta-paths are close to the real-world understanding. Also, the flexibility of HeteClass framework to utilize an algorithm of choice for classification of objects in Phase – 2 of the framework, makes HeteClass more useful for different domains as we can choose the appropriate algorithm for classification as per the context.

7. Conclusion and future research directions

In this paper, we studied the problem of transductive classification on target type objects in heterogeneous information networks. For transductive classification on objects, we propose a novel framework HeteClass, which is different from earlier approaches for classification in heterogeneous information networks. The proposed framework explores the network schema to generate a set of meta-paths for classification. By incorporating the knowledge of domain expert using HeteClass, we can reduce the set of meta-paths to select only those meta-paths that would be effective for classification. This would eventually reduce the computation time. HeteClass leverages the semantic subtleties of various meta-paths by weight learning and performing the weighted combination of various networks corresponding to each meta-path. Experimental studies performed in this paper demonstrate the effectiveness of the HeteClass as compared to the baseline algorithms. HeteClass also performs better as compared to the HetPathMine algorithm in terms of classification accuracy and weight learning. We can get meaningful insights and knowledge extraction by improving the classification accuracy using the HeteClass framework.

Interesting future research directions include the extension of the proposed framework for multi-label classification problem. In many real-world applications, a target type object in a heterogeneous network may acquire multiple labels. For example, in the case of classification of movies according to genres is a multi-label classification problem as a movie may have multiple genres. Also, the proposed framework can be extended to, first, find the most informative objects in the heterogeneous network and then acquiring label information for those objects so that the overall classification accuracy is improved for unlabeled objects. Also, it would be interesting to see the behavior of the proposed framework for those datasets which have objects having highly skewed class distribution.

References

Angelova, R., Kasneci, G., & Weikum, G. (2012). Graffiti: Graph-based classification in heterogeneous networks. *World Wide Web, 15*(2), 139–170.
Deng, Z. H., Lai, B. Y., Wang, Z. H., & Fang, G. D. (2012). PAV: A novel model for ranking heterogeneous objects in bibliographic information networks. *Expert Systems with Applications, 39*(10), 9788–9796.
Gupta, M., Kumar, P., & Bhasker, B. (2015). A new relevance measure for heterogeneous networks. In *Big data analytics and knowledge discovery* (pp. 165–177). Springer International Publishing.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.
Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web* (pp. 517–526). ACM.

Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web* (pp. 271–279). ACM.

Jeh, G., & Widom, J. (2002). 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).

Ji, M., Sun, Y., Danilevsky, M., Han, J., & Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Machine learning and knowledge discovery in databases* (pp. 570–586). BerlinHeidelberg: SpringerACM.

Kong, X., Yu, P. S., Ding, Y., & Wild, D. J. (2012). Meta path-based collective classification in heterogeneous information networks. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1567–1571). ACM.

Lao, N., & Cohen, W. W. (2010). Relational retrieval using a combination of path–constrained random walks. *Machine learning, 81*(1), 53–67.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology, 58*(7), 1019–1031.

Lu, Q., & Getoor, L. (2003). Link-based classification. In *ICML* (pp. 496–503). AAAI.

Luo, C., Guan, R., Wang, Z., & Lin, C. (2014). HetPathMine: A novel transductive classification algorithm on heterogeneous information networks. In *Advances in information retrieval* (pp. 210–221). Springer International Publishing.

Macskassy, S. A., & Provost, F. (2003). A simple relational classifier. In *Workshop on Multi-Relational Data Mining* (pp. 64–76).

Macskassy, S. A., & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *The Journal of Machine Learning Research, 8*, 935–983.

Rossi, R. G., de Andrade Lopes, A., & Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management, 52*(2), 217–257.

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

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). PathSim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB*.

Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: A structural analysis approach. *ACM SIGKDD Explorations Newsletter, 14*(2), 20–28.

Zhang, M., Hu, H., He, Z., & Wang, W. (2015). Top-k similarity search in heterogeneous information networks with x-star network schema. *Expert Systems with Applications, 42*(2), 699–712.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in neural information processing systems, 16*(16), 321–328.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML: Vol. 3* (pp. 912–919).