International Journal of Engineering Research and Application *www.ijera.com ISSN : 2248-9622 Vol. 8, Issue 9 (Part -V) Sep 2018, pp 111-123*

RESEARCH ARTICLE OPEN ACCESS

Mining Link: A Survey

Dr.Saaroj Kumar Rout[1],Pragyan Paramita Mahala[2], Dr.Sujit Kumar Panda[3]
[1,3]*Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar*
[2]*Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar*

ABSTRACT
Many datasets of interest today are best described as a linked collection of interrelated objects. These may repre- sent homogeneous networks, in which there is a single-object type and link type, or richer, heterogeneous networks, in which there may be multiple object and link types (and possibly other semantic information) . Examples of homo- geneous networks include single mode social networks, such as people connected by friendship links, or the WWW, a collection of linked web pages. Examples of heterogeneous networks include those in medical domains describing pa- tients, diseases, treatments and contacts, or in bibliographic domains describing publications, authors, and venues. Link rnininq refers to data mining techniques that explicitly con- sidertheselinkswhenbuildingpredictiveordescriptivemod- els of the linked data. Commonly addressed link mining tasks include object ranking, group detection, collective clas- sification, link prediction and subgraph discovery. While network analysis has been studied in depth in particular ar- eas such as social network analysis, hypertext mining, and webanalysis, onlyrecentlyhastherebeenacross-fertilization of ideas among these different communities. This is an ex- citing, rapidly expanding area. In this article, we review some of the common emergingthemes.

I. INTRODUCTION

"Links," or more genetically relationships, among data in- stances are ubiquitous. These links often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object. In some cases, not all links will be observed; therefore, we may be inter- ested in predicting the existence of links between instances. In other domains, where the links are evolving over time, our goal may be to predict whether a link will exist in the future, given the previously observed links. By taking links into ac- count, more complex patterns arise as well. This leads to other challenges focused on discovering substructures, such as communities, groups, or commonsubgraphs.

Traditional data mining algorithms such as association rule mining, market basket analysis, and cluster analysis com- monly attempt to find patterns in a dataset characterized by a collection of independent instances of a single rela- tion. This is consistent with the classical statistical infer- ence problem of trying to identify a model given a indepen- dent, identically distributed (IID) sample. One can thinkofthis process as learning a model for the node attributes of a homogeneous graph while ignoring the links between the nodes.

A key emerging challenge for data mining is tackling the problemofminingrichlystructured,heterogeneous datasets. These kinds of datasets are best described as networks or graphs. The domains often consist of a variety of object types; the objects can be linked in a variety of ways. Thus, the graph may have different node and edge (or hyperedge) types. Naively applying traditional statistical inference pro- cedures, which assume that instances are independent, can lead to inappropriate conclusions about the data /57 . Care must be taken that potential correlations due to links are handled appropriately. In fact, object linkage is knowledge that should be exploited. This information can be used to improve the predictive accuracy of the learned models: at- tributes of linked objects are often correlated, and links are more likely to exist between objects that have some com- monality. In addition, the graph structure itself may be an important element to include in the model. Structural properties such as degree and connectivity can be important indicators.

Link rnininq is a newly emerging research area that is at the intersection of the work in link analysis [58, 40] , hypertext and web mining /16] , relational learning and inductive logic programming [38a , and graph mining [23] . We use the term link mining to put a special emphasis on the links—moving them up to first-class citizens in the data analysis endeavor. In recent years, there have been several workshop series de- voted to topics related to link mining. One of the earliest workshops was the 1998 AAAI Fall Symposium on AI and Link Analysis [58] . Other workshop series include the work- shops on Statistical Relational Learning [48, 49, 28] , Multi-Relational Data Mining [65, 39, 36, 37] , LinkKDD [35, 1,3] , Link Analysis, Counter-terrorism and Security [104,26,103] , and Mining Graphs, Trees and Sequences [94,66, 85]. The objective of this survey is to provide a perspective on re- search within the relevant communities that are addressing current link mining challenges. Link mining encompasses a wide range of tasks; therefore, our review will cover the core challenges addressed by a majority of ongoing research in the field. We begin by describing some of the challenges in data representation for link mining. Then we progress through eight link mining tasks that can be broadly categorized as tasks that focus on objects, links, or graphs (Table 1). Fi- nally, we close with a discussion of areas that we believe have not yet received sufhcientattention.

Table 1: A taxonomy of common link mining tasks.

1. Object-RelatedTasks
(a) Link-Based ObjectRanking
(b) Link-Based ObjectClassification
(c) Object Clustering (GroupDetection)
(d) Object Identification (EntityResolution)
2. Link-RelatedTasks
(a) LinkPrediction
3. Graph-RelatedTasks
(a) SubgraphDiscovery
(b) GraphClassification
(c) Generative Models forGraphs

II. DATAREPRESENTATION

While data representation and feature selection are signifi- cant issues for traditional machine learning algorithms, data representation for linked data is even more complex. Con- sider a simple example from Singh et al. [101] of a social network describing actors and their participation in events. Such social networks are commonly called a ffili ation net- works [112] , and are easily represented by three tables rep- resenting the actors, the events, and the participation re- lationships. Even this simple structure can be represented as several distinct graphs. The most natural representa- tion is a bipartite graph, with a set of actor nodes, a set of event nodes, and edges that represent an actor's participa- tion in an event. Other representations may enable different insights and analysis. For example, we may construct a net- work in which the actors are nodes and edges correspond to actors who have participaterl in an event together. This rep- resentation allows us to perform a more actor-centric anal- ysis. Alternatively, we may represent these relations as a graph in which the events are nodes, and events are linked if they have an actor in common. This representation may allow us to more easily see connections betweenevents.

This flexibility in the representation of a graph arises from a basic graph representation duality. This duality is illus- trated by the following simple example: Consider a data set represented as a simple C = (0, L), where 0 is the set of objects (i.e., the nodes or vertices) and L is the set of links (i.e., the edges or hyperedges) . The graph G(0, L) can be transformed into a new graph C'(0', L'), in which the links 1; , I j in G are objects in G' and there exists an link between o;, oj C 0' if and only if I; and /j share an ob- ject in G. This basic graph duality illustrates one kind of simple data representation transformation. For graphs with multiple node and edge types, the number of possible trans- formations becomes immense. Typically, these reformula- tions are not considered as part of the link mining process. However, the representation chosen can have a significant impact on the quality of the statistical inferences that can be made. Therefore, the choice of an appropriate represen- tation is actually an important issue in effective link mining, and is often more complex than in the case where we have IID data instances. In the following sections, we willassume thatadatarepresentationhasbeenselected,thatthe desig- nation of the objects or nodes in the graph has been made, and that the links or edges in the graph have been definerl. However, when applying link mining to real world domains, one should not underestimate the effort required in choosing an appropriaterepresentation.

III. LINK-BASED OBJECTRANKING

Perhaps the most well known link mining task is that of link-based object ranking (LBR) , which is a primary focus of the link analysis community. The objective of LBR is to exploit the link structure of a graph to order or prioritize the set of objects within the graph. Much of this research focuses on graphs with a single object type and a single link type.

In the context of web information retrieval, the PageR ank [91a] and HITS [64] algorithms are the most notable ap- proaches to LBR. PageRank models web surfing as a random walk where the surfer ranrlomly selects anrl follows links and occasionally jumps to a new web page to start another traversal of the link structure. The rank of a given web page in this context is the fraction of time that the random web surfer would spend at the page if the random process were iterated ad infinitum. This can be determined by computing the steady-state distribution of the randomprocess.

HITS assumes a slightly more complex process, modeling the web as being composed of two types of web pages: Notts and authorities. Hubs are web pages that link to many au- thoritative pages. Authorities are web pages that are linked to by many hubs. Each page in the web is assigned hub and authority scores. These scores are computed by an it- erative algorithm that updates the scores of a page based on the scores of pages in its immediate neighborhood. This approach bears a relation to PageRank with two separate rantlom walks—one with hub transitions and one with au- thority transitions—on a corresponding bipartite graph of hubs and authorities [73, 95, 84a] . The hub and authority scores are the steady-state distributions of the respective randomprocesses.

Since the introduction of PageRank and HITS, a number of algorithms have been proposed that are variations on these basic themes. Bharat and Henzinger /8] and Chakrabarti ct al. [17 propose modifications to HITS that exploit web page content to weight pages and links based on relevance. Ng et al. [83,84] analyze the stability of PageRank and HITS to small perturbations in the link structure and present mod- ifications to HITS that yield more stable rankings. Haveli- wala PSI] and Jeh and Widom [56] propose topic-sensitive PageRank algorithms that identify topically authoritative web pages efhciently at query time. Ding et a1. [29] pro- posesaunifierlframeworkencompassingbothPageRank anr1 HITS and presents several new ranking algorithms within this algorithm class with closed-form solutions. Cohn and Chang [20 introduce a probabilistic analogue to HITS based on probabilistic latent semantic indexing, where the model attempts to explain the link structure in terms of a small set of latent factors. Cohn and Hofmann [21] and Richard- son and Domingos [98] present probabilistic models inspired by HITS and PageRank, respectively, that incorporate both content and linkstructure.

In the domain of social network analysis (SNA) , LBR is a core analysis task. The objective is to rank orderindividu-als in a given social network in terms of a measure of their importance, referred to as centrnfitp. Measures of centrality have been the subject of research in the SNA community for decades [112] . These measures characterize some aspect of the local or global network structure as seen from a given in- dividual’s position in the network. They range in complexity from local measures such as degree centrality [43] , which is simply the vertex degree, to global measures such as eigen- vector/power centrality [12], which use spectral methods to characterize the importance of individuals based on their connectedness to other importantindividuals.

In the above work, the common goal is a global ranking of objects in a static graph produced using a specified mea- sure. Notable variations from this theme inclucte approaches that rank objects relative to one or more relevant objects in the graph [55,114,105] and methods that rank objects over time in dynamic graphs [89,S8] . Jeh and Widom [55] propose a metric for assessing the similarity of two objects basect on the degree to which they link to similar objects. The similarity between two objects in a directed or bipartite graph is computed using a random walk formulation. Sun et al. [105] in this issue propose a related object ranking approach for relevance search and anomaly detection that combines random walks and graph partitioning to improve scalability. White and Smyth [114a] define and evaluate a host of metrics to compute the similarity between a given object and one or more reference objects in agraph.

Ranking objects in dynamic graphs that capture event data such as email, telephone calls, or publications introduces new challenges. In contrast to ranking methods for static settings that produce a single rank, the goal is to track the changes in object rank over time as new events unfold. Static ranking methods can be applied to aggregated event data over various time intervals, but this aggregation removes the time ordering of events, and the sparse link structure over a given time interval limits the utility of the resulting ranks. O’Madadhain and Smyth [89] and O’Madadhain et al. [88] in this issue propose a series of desired algorithmic properties for dynamic object ranking, discuss the limita- tions of notable static ranking algorithms, and introduce a ranking algorithm based on potential flow that satisfies the specifiedrequirements.

IV. LINK-BASED OBJECTCLASSIFICATION

Traditional machine learning has focused on the classifica- tion of data consisting of identically structured objects that are typically assumed to be IID. Many real-world datasets, however, lack this homogeneity of structure. In the link- based object classification (LBC) problem, a data graph G = (0, L) is composed of a set objects 0 connected to each other via a set of links L. The task is to label the members of 0 from a finite set of categorical values. The discerning feature of LBC that makes it different from traditional clas- sification is that in many cases, the labels of related objects tend to be correlated. The challenge is to design algorithms for collective classification that exploit such correlations and jointly infer the categorical values associated with the ob- jects in thegraph.

LBC has received considerable attention recently. Chakra- barti et a1. [18] consider the problem of classifying related news items in the Reuters dataset. They were among the firsttonoticethatexploitingclasslabelsofrelatedobjects aids classification, whereas exploiting features of related ob- jects can actually form classification accuracy. Oh et a1. [87] report similar results on a collection of encyclopedia arti- cles: simply incorporating words from neighboring docu- ments was not helpful, while making use of the predicted class of neighboring documents was helpful. Lafferty ct al. [71] introduce conditional random fields (CRF), which extend traditional maximum entropy models for LBC in the restricted case where the data graphs are chains. Taskar et al. [107] extend Lafferty et a1.'s approach [71a] to the case where the data graphs are arbitrary graphs. Neville and Jensen 80] propose simple LBC algorithms to classify cor- porate datasets with rich schemas that produce graphs with heterogeneous objects, each with its own distinct set of fea- tures. Lu and Getoor [76 extend simple machine learn- ing classifiers to perform LBC by introducing new features that measure the rlistribution of class labels in the Markov blanket of the object to be classified. In addition to the machine learning community, the computer vision and nat- ural language communities have also studied the LBC prob- lem. Rosenfeld et al. [99] proposed relaxation labeling, an inference algorithm later used by Chakrabarti et al. [18] to perform link-based classification. Hummel and Zucker [53a] present one of many approaches for exploring relaxation la- beling theoretically. Lafferty ct a1. [71] proposed CRFs for use in part-of-speech tagging, a task in natural language processing.

V. GROUPDETECTION

A third object-centric task is group detection. The goal of group detection is to cluster the nodes in the graph into groups that share common characteristics. A range of tech- niques have been presented in various communities to ad- dress this general problem. In recent years, a central chal- lenge has been to rlevelop scalable methods that can exploit increasingly complex graphs to aid the knowledge discovery process.

Consider first the case where the graph contains objects and links of a single type, without attributes. Many of the tech- niques for identifying groups in this scenario can be classified as either agglomerative or divisive clustering methods. The task of blockmodeling of social network analysis (SNA) in- volves partitioning social networks into sets of individuals, called positions, that exhibit similar sets of links to others in the network [112a] . A similarity measure is defined between link sets and agglomerative clustering is used to identify the positions. Spectral graph partitioning methods address the group detection problem by identifying an approximately minimal set of links to remove from the graph to achieve a given number of groups /82; 30]. In a related vein, Gib- son ct al. [50] have shown that the rlominant eigenvectors of the HITS authority matrix provide a natural decomposi- tion of web community structure. Other recent approaches for group detection use a measure of edge betweenness, de- rived from Freeman's notion of betweenness centrality [43] , to identify links connecting groups [109] . Links with high edge betweenness are incrementally removed to partition the graph.

In contrast to the above methods, where group assignments are deterministic, a number of approaches for group detec- tion have been introduced that are based on the concept of stochastic blockmodeling from SNA. In stochastic blockmod- eling, the observed social network is assumed to be a realiza- tion from a pair-dependent stochastic blockmodel [112, 86] . Positions for the individuals in the network are treaterl as IID random variables, and relational links of a given type be- tween two individuals are random variables dependent solely on the positions of the individuals they link. Nowicki and Snijders [86] propose a general stochastic blockmodelling ap- proach admitting directed, valued relations and an arbitrary number of positions. Gibbs sampling is used to infer the pos- terior distribution for positions. Kemp et a1. fi1 remove the need to specify the number of positions a priori; instead, the number of positions is inferred directly from the data. Wolfe and Sensen [115] extend the general stochastic blockmod- elling approach by allowing an individual to have multiple position types; this provides the flexibility to model multiple roles that an individual may have in differentcontexts.

To address group detection challenges in the intelligence and law enforcement domains, methods are needed that can exploit volumes of multi-relational data to detect indica- tors of collaboration. Several recent efforts have proposed methods to address such challenges. Adibi et al. [2] pro- pose a hybrid approach that initially posits potential groups using knowledge-based reasoning techniques and then aug- ments these hypotheses with additional candidates based on observed interactions that indicate likely association. Ku- bica et al. [69a] presents a generative model for multi-type link generation given group membership and individual at- tribute information. Maximum likelihood estimation is used to identify the most likely chart mapping individuals to their respective group memberships. In later work Kubica ct a1. [68] introduce a scalable version of this approach that uses a method similar to k-means clustering to significantly accelerate group discovery, while retaining the unrlerlying generative model. Most recently, Wang et a1. [110] propose a generalization of the general stochastic blockmodelling ap- proach that allows joint inference of groups and topics based on observed relationships and their text ua1 attributes. Such a model provides a mechanism to connect an observed rela- tionship with its underlyingcontext.

VI. ENTITYRESOLUTION

The final object-centric task is entity resolution, which in- volves identi fyinp the set of objects in a domain. The goal of entity resolution is to determine which references in the data refer to the same real-world entity. Examples of this problem arise in databases (deduplication, data integration), natural language processing (co- reference resolution, object consoli- dation) , personal information management, and other fields. The problem has been defined with many variations; in the most general form, neither the domain entities nor the num- ber of such entities is assumed to be known. Traditionally, entity resolution has been viewed as a pair-wise resolution problem, where each pair of references is independently re- solvent as being co-referent or otherwise, depending on the similarity of their attributes. Recently, there has been sig- nificant interest in the use of links for improved entity res- olution. The central idea is to consider, in addition to the attributes of the references to be resolved, the other refer- ences to which these are linked. These links may be, for example, co-author links between author references in bibli- ographic data, hierarchical links between spatial references in geo-spatial data, or co-occurrence links between name ref- erences in natural languagedocuments.

Theuseoflinksforresolutionwasfirstexploredindatabases. Ananthakrishna ct a1. [6] introduce a method for deduplica- tion using links in rlata warehouse applications where there is a dimensional hierarchy over the link relations. More re- cently, Kalashnikov et a1. [59] enhance feature-based sim- ilarity between an ambiguous reference and the many en- tity choices for it with linkage analysis between the entities, such as afhliation and co-authorship. However, while these approaches consider links for entity resolution, only the at- tributesoflinkedreferencesareconsideredanddifferentres- olution decisions are still takenindependently.

In contrast, collective entity resolution approaches have also been proposed in databases [9,34] , where one resolution decision affects another if they are linked. Bhattacharya and Getoor [9, 10a] propose different measures for linkage similarity in graphs and show how these can be combined with attribute similarity for collective entity resolution in collaboration graphs. Dong et a1. [34] collectively resolve entities of multiple types by propagating evidence over links in a dependencygraph.

In machine learning, probabilistic models that take into ac- count interaction between different entity resolution deci- sions have been proposed for named entity recognition in natural language processing and for citation matching. Li et al. [74] address the problem of disambiguating “entity mentions,” potentially of multiple types, in the context of unstructured textual documents. Parag ct a1. [102a] use the idea of merging evidence to allow the flow of reasoning be- tween linked pair-wise decisions over multiple entity types. In addition, models have been proposed that explicitly con- sider links among references for collective resolution [92, 11, 25] . Pasula ct al. [92] propose a generic probabilistic rela- tional model framework for the citation matching problem. Culotta and McCallum [25] construct a conditional random field model of deduplication that captures linked dependen- cies between references of multiple types. Bhattacharya ct al. all] adapt the Latent Dirichlet model for documents and topics and extend it to propose a generative group model for unsupervised collective entityresolution.

VII. LINKPREDICTION

We next turn to edge-related tasks. Link prediction is the problemofpredictingtheexistenceofalinkbetweent woen- tities, based on attributes of the objects and other observed links. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events [88a] , such as email, tele- phone calls and co-authorship; and predicting semantic rela- tionships such as "advisor-of" based on web page links and content [24, 108a] . Most often, some links are observed, and one is attempting to predict unobserverl links, or there is a temporal aspect: a snapshot of the set of links at time t is given and the goal is to predict the links at time t +1.

This problem is often viewed as a simple binary classifica- tion problem: for any two potentially linked objects o; and o j , predict whether I ; jis l or 0. One approach is to make this prediction entirely based on structural properties of the network. Liben-Nowell and Kleinberg [75a] present a survey of predictors based on different graph proximity measures. Other approaches make use of attribute information for link prediction. Popescul ct al. [93] introduce a structured logis- tic regression model that can make use of relationalfeaturesto predict the existence of links. The relational features are definedviadatabasequeries;theauthorsshowhowto search overthespaceofrelationalfeatures.O'Madarlhainct a1.[88,90] construct local conditional probability models, based on attribute and structuralfeatures.

Link prediction is hard because most interesting linked data sets are sparse. As pointed out by many researchers [46, 88,97] , one of the difficulties in building statistical models for edge prediction is that the prior probability of a link is typically quite small. This causes difficulty both in model evaluation and, more importantly, in quantifying the level of confidence in the predictions. Rattigan and Jensen [97] in this issue discuss some of thesechallenges.

One way to improve the quality of the predictions is to make the predictions collectively. A number of approaches define a single probabilistic model over the entire link graph, la- bels, and edges. These joint models of network structure are often based on models such as Markov random fields [19]. In the simplest case, where there is a set of objects O, with attributes A, and edges E among the objects, the MRF models a joint distribution over the set of edges E, P(A), or a distribution conditioned on the attributes of the nodes, P E

Richer models, based on relational rep- resentations, are possible, such as Relational Markov Net- works[108]and,morerecently,MarkovLogicNetworks [33]. Models based on directed graphical models are also possible. Getoor et al. [47] describe several approaches for handling link uncertainty in probabilistic relationalmodels.

A discerning feature of these latter approaches is that they perform probabilistic inference to make inferences about the links. This allows them to capture the correlations among the links. They can also be used for other tasks, such as link- based classification. Ideally this makes for more accurate predictions. However, model-based probabilistic approaches have a computational price: exact inference is generally in- tractable,soapproximateinferencetechniquesarenecessary.

VIII. SUBGRAPHDISCOVERY

An area of data mining that is related to link mining is the work on subgraph discovery. This work attempts to find interesting or commonly occurring subgraphs in a set of graphs. Discovery of these patterns may be the sole purpose of the systems, or the discovered patterns may be used for graph classification (Section9).

One line of work attempts to find frequent subgraphs [54, 70,116a] . Many of these approaches exploit the Apriori prop- erty [4] from frequent item set mining. Typically, there is a candidate generation phase followed by a matching phase. Naive matching requires a subgraph isomorphism test, so efhcient algorithms are needed here as well. Inokuchi ct al. [54] describe AGM, an Apriori-based algorithm that finds all induced subgraphs in a graph database satisfying a min- imum support. Kuromachi ct al. [70] improve on AGM by using an adjacency representation of the graph data and de- scribing new optimizations to candidate substructure gener- ation. Yan et al. [116] describe gspan, which avoids the cost of candidate generation by first mapping each graph to a depth-first search code and lexicographically ordering these codes, then performing DFS on the search tree defined by this lexicographicordering.

Other approaches come from the inductive logic program- ming (ILP) community [79,72a] . One early success wasthe work of Dehaspe et al. [27] , who applied techniques from inductive logic programming to finding frequent patterns in a toxicologyrlomain.

Another line of work focuses on efficient subgraph genera- tion and compression-based heuristic search [22, 70] . Sub- due [22] , the earliest work in this area, uses an MDL-based heuristic to guide the search for subgraphs. Subdue has been used for both subgraph discovery and graph classifi- cation [23]. As another example, Graph-Based Induction (GBI) compresses the input graph by chunking the vertex pairs that appear frequently [117] . Both of these approaches use a greedy local approach in their search for frequent sub- structures. Ketkar et al. [62] compare these approaches to ILPapproaches.

IX. GRAPHCLASSIFICATION

Unlike link-based object classification, which attempts to label nodes in a graph, graph classification is a supervised learning problem in which the goal is to categorize an entire graph as a positive or negative instance of a concept. This is one of the earliest tasks addressed within the context of applying machine learning and data mining techniques to graph data. Graph classification does not typically require collective inference, as is needed for classifying objects and edges, because the graphe are generally assumed to be incte- pendentlygenerated.

Three main approaches to graph classification have been ex- plored. Thèse are based on feature mining on graphs, induc- tive logic programming (ILP), anrl defining graph kernels. Feature mining on graphs uses methods related to those de- scribed in the previous section on subgraph discovery, Sec- tion 8. Feature mining on graphs is usually performed by finding all frequent or informative substructures in the graph instances. Thèse substructures are used for transforming the graph data into data represented as a single table, and then traditional classifiers are used for classifying the instances. As an example of an ILP approach, King et al. [63s first map the graph data describing mutagenesis into a relational rep- resentation. Their logical representation uses relations such as rerte+(9raphId, Verteæld, Verte+Label, Verte:sAttri #utes) and edge (grapliId,verte:sId1, rerte+Id2, BondLabel), and then uses an ILP system to find a hypothesis in thisspace.

Finding all frequent substructures is usually computation- ally prohibitive. An alternative approach makes use of ker- nel methods. Both Gärtner and Kashima describe graph kernels based on a measure of the walks on the graphs [44,60] . Gärtner [44] countfi walkfi with equal initial and termi- nal labels, whereas Kashima [60] looks at the probability of random walks with equal label sequences. A Gärtner [45]s surveys kernel methods for structureddata.

X. GENERATIVE MODELS FORGRAPHS

Generative morlels for a range of graph and dependency types have been studied extensively in the social network analysis community. For directed graphe with a single ob- ject and link type, there are several major classes of random graph distributions discussed in the literature: Bernoulli graph distributions, conditional uniform graph distributions, dyadic dependence distributions, and p+ models. Bernoulli graphs [41] (also known as Erdös-Rényi models or random graphs) are by far the simplest generative models. They assume that the random variables (/,) that indicate the existence of directed edges among the objects o; and oj are IID. When the probability of link existence equals 0.5, the distribution is often referred to as the uniform random graph distribution. Conditional uniform graph distributions [112] define uniform distributions over sets of graphs with spec- ified structural characteristics, such as a fixed number of links, out-degrees, or in-degrees. Dyadic dependence distri- butions [111a assume that only the dyads (l i j , 1 ji) are de- pendent and define multinomial distributions over the dyad states. P+ models assume that links sharing at least one object in common are dependent. Generative models ad- mitting dependency structures that are more general than Markov graphs have been introduced as well, along with models for multiple object and link types and dynamic net- works with a varying link structure and number of objects [14,52].

In recent years, significant attention has focused on studying the structural properties of networks such as the World Wide Web, online social networks, communication networks, cita- tion networks, and biological networks. Across these various networks, general patterns such as power law degree distri- butions, small graph diameters, and community structure are observed. These observations have motivated the search for general principles governing such networks US. Airoldi et al. [5] in this issue review sampling algorithms for a num- ber of the common network types such as scale free networks 7 , small-world networks [113] , core-periphery [13] , and cel- lular networks [42] that exhibit such attributes. In contrast to the random process models from the social network anal- ysis literature, many of these generative models are specified in procedural form, which is viewed as beneficial when the goal is to understand how power law degree distributions, for example, can naturally emerge in dynamic graphs over time. Chakrabarti [15] presents a taxonomy of recently proposed graphgenerators.

Finally, we note several generative models of link structure presented in the machine learning community that address a variety of application contexts. Kubica et a1. introduces a generative model for observed links among individ- uals given their underlying group memberships. Kubica et al. [67] present a link generation model for link analysis and collaboration queries that admits different link types and temporal information. Getoor et a1. [47] introduce proba- bilistic relational models, which that provide a unified gen- erative model for objects and link structure. Neville and Jensen [81] define a probabilistic relational model that rep- resents a joint distribution over objects, links and latent groups.

XI. OPEN ISSUES AND PROMISING AREAS FOR FUTURERESEARCH

In this survey, we have often described each link mining task in isolation. More generally, component link mining algo- rithms may be part of a larger knowledge discovery process. As we move from one domain to another, the processing reQuirements will change, but the need to compose the algo- rithms in a unified process will remain. Ideally, as we move from data conditioning to more complex inference tasks, we would like to propagate uncertainty throughout the process. One approach that solves this problem, in theory, is to de- fine a full probabilistic model; this the approach taken by Getoor et a1. [47] and Taskar et a1. [108]. However, this approach is not always desirable or feasible. As argued by Senator [100] in this issue, in addition to addressing spe- cific link mining tasks, it is equally important to consider how to effectively compose link mining algorithms to ad- dress a spectrum of knowledge discovery tasks. Ultimately, system performance is determined by the interplay among the components; therefore, it is critical to investigate how these component dependencies will shape the overall perfor- mance.

When considering the overall knowledge discovery process, it is important to keep in mind that many aspects of the pro- cess are dynamic. The dynamism, which can extend from the data to the user’s needs, interests, and beliefs, implies that a number of link mining algorithms will be applied re- peatedly and incrementally. We often envision applying link mining algorithms to the entire graph. While this is desir- able in some applications, it does not make sense when a user is interested in only a small subgraph. Therefore, it is important to develop methods supporting focused, incre- mental application of linkmining.

One interestingresearchdirectioninthisareaisquery-based classification using links. Most collective classification ap- proaches consider the dataset in its entirety as one linked instance of objects, performing prediction/classification for all of these objects jointly. When a user is interested in classifying only a small subset of these objects, it is worth- while to classify other objects only if they are helpful in correctly classifying the objects of interest via the link struc- ture. Given this goal, a query-based collective inference technique needs to first extract the links and objects that aremostrelevantforansweringthequeryapproximatelyand then perform collective classification only on the extracted subgraph. Identification of relevant subgraphs can also be helpful for incremental classification when new objects and links are added to an existing graph with classified objects. Link mining often needs to be performed on data from mul- tiple sources; therefore, information integration and recon- ciliation are important components of the link mining pro- cess. Furthermore, it is important to integrate the data (re)formulation more directly into the link process process. While there has been some work that integrates the statisti- cal approaches to link mining with the meta-data discovery and mapping [31] , there is much more to bedone.

Another promising arena in which to apply link mining is the Semantic Web. In this issue, Ramakrishnan ct al. /96] de- scribe methods for discovering interesting subgraphs based on semantic information associated with the edges. There has been some other work in this area, for example Mad- che and Staab [77] and Doan et al. [32a] , but there is much more to be done. As information extraction techniques con- tinue to improve, one area for future research is combining information extraction with techniques from link mining to help to construct the Semantic Web, and another area for future research is how semantic and ontological information can help in link miningendeavors. As the amount of data grows and the number of sources expands, techniques from link mining can help us discover patterns and build useful prediction systems. Link mining research holds promise for many different areas, including commercial and business enterprises, personal information management, web search and retrieval, medicine and bio- informatics, and law and security enforcement. However, as cautioned by Sweeney 106 , as we develop thistechnol- ogy, privacy and information-access control issues and policy must be considered, not just as an afterthought, but as an integral part of the solution.

XII. CONCLUSION

More and more domains of interest

today are best described as a linked collection or network of interrelated heteroge- neous objects. Data mining algorithms have typically ad- dressed the discovery of patterns in collections of IID in- stances. Link mining is an emerging area within data min- ing that is focused on finding patterns in data by exploiting and r+pliritly madeling the links among the tlata instances. We have surveyed several of the more well studied link min- ing tasks: link-based object ranking, link-based object clas- sification, group detection, entity resolution, link predic- tion, subgraph discovery, graph classification, andgenera-

REFERENCES

[1]. J. Adibi, H. Chalupsky, M. Grobelnik, N. Milic-Frayling, and D. Mladenic. KDD Workshop on LinkAnalysis and Group Detection. 2004.

[2]. J. Adibi, H. Chalupsky, E. Melz, and A. Valente. The KOJAK group _nder: Connecting the dots via integratedknowledge-based and statistical reasoning. In Innovative Applications of Arti_cial Intelligence Con-ference, 2004.

[3]. J. Adibi, M. Grobelnik, D. Mladenic, and P. Pantel.KDD Workshop on Link Discovery: Issues, Ap-proaches and Applications. 2005.

[4]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In International Conference on VeryLarge Data Bases, pages 487|499, Sept. 1994.

[5]. E. M. Airoldi and K. M. Carley. Sampling algorithms for pure network topologies. SIGKDD Explorations,7(2), 2005.

[6]. R. Ananthakrishna, S. Chaudhuri, and V. Ganti.Eliminating fuzzy duplicates in data warehouses. InInternational Conference on Very Large Databases (VLDB), Hong Kong, China, 2002.

[7]. A. L. Barab_asi and R. Albert. Emergence of scaling in random networks. Science, 286:509{512, 1999.

[8]. K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment.In ACM SIGIR International Conference on Research and Development in Information Retrieval, pages 104{111, 1998.

[9]. I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In SIGMOD 2004 Work-shop on Research Issues on Data Mining and Knowledge Discovery, June 2004.

[10]. I. Bhattacharya and L. Getoor. Entity resolution in graphs. Technical Report 4758, Computer Science Department,University of Maryland, 2005.

[11]. I. Bhattacharya and L. Getoor. A Latent dirichlet model for unsupervised entity resolution. In SIAM In-ternational Conference on Data Mining, 2006. To Appear.

[12]. P. Bonacich. Power and centrality: A family of measures.American Journal of Sociology, 92(5):1170{1182, 1987.

[13]. S. P. Borgatti and M. G. Everett. Models of core / periphery structures. Social Networks, 21:375{395, 1999.

[14]. P. J. Carrington, J. Scott, and S. Wasserman. Models and Methods in Social Network Analysis. CambridgeUniversity Press, Cambridge, 2005.

[15]. D. Chakrabarti. Tools for Large Graph Mining. PhD thesis, School of Computer Science, Carnegie MellonUniversity, 2005.

[16]. S. Chakrabarti. Mining the Web. Morgan Kaufman,2002.

[17]. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg,P. Raghavan, and S. Rajagopalan. Automatic resourcelist compilation by analyzing hyperlink structure and associated text. In International World Wide Web Conference (WWW), 1998.

[18]. S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In SIGMODInternational Conference on Management of Data,pages 307{318, 1998.

[19]. R. Chellappa and A. Jain. Markov random _elds: the-ory and applications. Academic Press, Boston, 1993.

[20]. D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In InternationalConference on Machine Learning (ICML), pages 167{ 174. Morgan Kaufmann, San Francisco, CA, 2000.

[21]. D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity.In Neural Information Processing Systems 13, 2001.

[22]. D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and backgroundknowledge. Journal of Arti_cial Intelligence Research,1:231|255, 1994.

[23]. D. J. Cook and L. B. Holder. Graph-based data mining. IEEE Intelligent Systems, 15(2):32{41, 2000.

[24]. M. Craven, D. DiPasquo, D. Freitag, A. McCallum,T. Mitchell, K. Nigam, and S. Slattery. Learning toconstruct knowledge bases from the world wide web.

[25]. A. Culotta and A. McCallum. Joint deduplication of multiple record types in relational data. In FourteenthConference on Information and Knowledge Manage-ment (CIKM), 2005.

[26]. G. V. Cybenko and J. Srivastava. SIAM Workshop on Link Analysis, Counterterrorism and Privacy. 2004.

[27]. L. Dehaspe, H. Toivonen, and R. King. Finding frequent substructures in chemical compounds. In In-ternational Conference on Knowledge Discovery and Data Mining, pages 30|36, 1998.

[28]. T. Dietterich, L. Getoor, and K. Murphy. ICML Work-shop on Statistical Relational Learning and its Con-nections to Other Fields. 2004.

[29]. C. Ding, X. He, P. Husbands, H. Zha, and H. D. Simon.PageRank, HITS and a uni_ed framework forlink analysis. In ACM SIGIR Conference on Research and Development in Information Retrieval, pages 353{354, 2002.

[30]. C. H. Q. Ding. Spectral clustering, 2004.http://crd.lbl.gov/ cding/Spectral/.

[31]. A. Doan, P. Domingos, and A. Y. Halevy. Learning to match the schemas of data sources: A multistrategyapproach. Machine Learning, 50(3), 2003.

[32]. A. Doan, J. Madhavan, P. Domingos, and A. Halevy.Learning to map between ontologies on the semanticweb. In International World Wide Web Conference,2002.

[33]. P. Domingos and M. Richardson. Markov logic: A unifying framework for statistical relational learning. InICML-2004 Workshop on Statistical Relational Learn-ing and its Connections to Other Fields, pages 49|54,2004.

[34]. X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In ACMSIGMOD International Conference on Management of Data, pages 85{96, 2005.

[35]. S. Donoho, T. Dybala, M. Grobelnik, N. Milic-Frayling, and D. Mladenic. KDD Workshop on LinkAnalysis for Detecting Complex Behavior. 2003.

[36]. S. Dzeroski and H. Blockeel. KDD Workshop on Multi-Relational Data Mining. 2004.

[37]. S. Dzeroski and H. Blockeel. KDD Workshop on Multi-Relational Data Mining. 2005.

[38]. S. Dzeroski and N. Lavrac, editors. Relational Data Mining. Kluwer, Berlin, 2001.

[39]. S. Dzeroski, L. D. Raedt, and S. Wrobel. KDD Work-shop on Multi-Relational Data Mining. 2003.

[40]. R. Feldman. Link analysis: Current state of the art,2002.

[41]. O. Frank and K. Nowicki. Exploratory statistical analysis of networks. Annals of Discrete Mathematics,55:349{366, 1993.

[42]. T. Frantz and K. M. Carley. A formal characterization of cellular networks. Technical Report CMU-ISRI-05-109, Carnegie Mellon University, 2005.

[43]. L. Freeman. Centrality in social networks: Conceptual clari_cations. Social Networks, 1:215{239, 1979.

[44]. T. G • artner. Exponential and geometric kernels for graphs. In NIPS Workshop on Unreal Data: Princi-ples of Modeling Nonvectorial Data, 2002.

[45]. T. G • artner. A survey of kernels for structured data. SIGKDD Explorations, 5(1):49{58, 2003.

[46]. L. Getoor. Link mining: a new data mining challenge.SIGKDD Explorations, 5(1):84{89, 2003.

[47]. L. Getoor, N. Friedman, D. Koller, and B. Taskar.Learning probabilistic models of link structure. Jour-nal of Machine Learning Research, 3:679{707, 2003.

[48]. L. Getoor and D. Jensen. AAAI Workshop on Learn-ing Statistical Models from Relational Data. AAAIPress, 2000.

[49]. L. Getoor and D. Jensen. IJCAI Workshop on Learn-ing Statistical Models from Relational Data. 2003.

[50]. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In ACM Con-ference on Hypertext and Hypermedia, pages 225{234,1998.

[51]. T. H. Haveliwala. Topic-sensitive PageRank. In In-ternational Conference on the World Wide Web(WWW), pages 517{526, 2002.

[52]. M. Huisman and T. A. B. Snijders. Statistical analysis of longitudinal network data with changing composition.Sociological Methods and Research, 32:253{287,2003.

[53]. R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. IEEE Transactions onPattern Analysis and Machine Intelligence, pages 267{287, 1983.

[54]. A. Inokuchi, T. Washio, and H. Motoda. An Aprioribased algorithm for mining frequent substructuresfrom graph data. In European Conference on Prin-ciples and Practice of Knowledge Discovery and DataMining, pages 13|23, 2000.

[55]. G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In ACM SIGKDD In-ternational Conference on Knowledge Discovery and Data Mining, pages 538{543, 2002.
[56]. G. Jeh and J. Widom. Scaling personalized web search. In International Conference on the World Wide Web (WWW), pages 271{279, 2003.
[57]. D. Jensen. Statistical challenges to inductive inference in linked data. In Seventh International Workshop on Arti_cial Intelligence and Statistics, 1999.
[58]. D. Jensen and H. Goldberg. AAAI Fall Symposium on
[59]. D. V. Kalashnikov, S. Mehrotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In SIAM International Conference on Data Min-ing, April 21{23 2005.
[60]. H. Kashima and A. Inokuchi. Kernels for graph classi-_cation. In ICDM Workshop on Active Mining, 2002.
[61]. C. Kemp, T. L. Gri_ths, and J. B. Tenenbaum. Discovering latent classes in relational data. Technical ReportAI Memo 2004-019, Massachusetts Institute of Technology, September 2004.
[62]. N. Ketkar, L. Holder, and D. Cook. Comparison of graph-based and logic-based multi-relational data mining. SIGKDD Explorations, 7(2), December 2005.
[63]. R. D. King, S. H. Muggleton, A. Srinivasan, and M. J. E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. National Academy of Sciences, 93(1):438{442, January 1996.
[64]. J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604{632,1999.
[65]. A. Knobbe and D. van der Wallen. ECML/PKDD Workshop on Multi-Relational Data Mining. 2001.
[66]. J. N. Kok and T. Washio. ECML/PKDD Workshop on Mining Graphs, Trees and Sequences. 2004.
[67]. J. Kubica, A. Moore, D. Cohn, and J. Schneider. cGraph: A fast graph-based method for link analysis and queries. In IJCAI 2003 Text-Mining and Link- Analysis Workshop, August 2003.
[68]. J. Kubica, A. Moore, and J. Schneider. Tractable group detection on large link data sets. In The Third IEEE International Conference on Data Mining, pages 573{576, 2003.
[69]. J. Kubica, A. Moore, J. Schneider, and Y. Yang.Stochastic link and group detection. In Eighteenth Na-tional Conference on Arti_cial Intelligence, pages 798{804. American Association for Arti_cial Intelligence, 2002.
[70]. M. Kuramochi and G. Karypis. Frequent subgraph discovery. In IEEE International Conference on DataMining, pages 313{320, 2001.
[71]. J. La_erty, A. McCallum, and F. Pereira. Conditional random _elds: Probabilistic models for segmentingand labeling sequence data. In Proc. of ICML-01, 2001.
[72]. N. Lavra_c and S. D_zeroski. Inductive Logic Program- ming: Techniques and Applications. Ellis Horwood, 1994.
[73]. R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC e_ect.Computer Networks, 33(1{6):387{401, 2000.
[74]. X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identi_able entities.AI Magazine. Special Issue on Semantic Integration, 2005.
[75]. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In International Con-ference on Information and Knowledge Management (CIKM), pages 556{559, 2003.
[76]. Q. Lu and L. Getoor. Link-based classi_cation. In In-ternational Conference on Machine Learning, 2003.
[77]. A. Madche and S. Staab. Ontology learning for the semantic web. IEEE Intelligent Systems, 16(2):72|79, March/April 2001.
[78]. T. Matsuda, T. Horiuchi, H. Motoda, and T. Washio.Extension of graph-based induction for general graph structured data. In PAKDD, pages 420|431, 2000.
[79]. S. Muggleton, editor. Inductive Logic Programming.Academic Press, 1992.
[80]. . Neville and D. Jensen. Iterative classi_cation in relational data. In Proc. AAAI-2000 Workshop on Learn- ing Statistical Models from Relational Data. AAAI Press, 2000.
[81]. J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In IEEE Inter-national Conference on Data Mining (ICDM), 2005.
[82]. M. E. J. Newman. Detecting community structure in networks. European Physical Journal B, 38:321{330,2004.
[83]. A. Y. Ng, A. X. Zheng, and M. I. Jordan. Link analysis,eigenvectors and stability. In

International JointConference on Arti_cial Intelligence (IJCAI), pages 903{910, 2001.
[84]. A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In ACM SIGIR Conferenceon Research and Development in Information Retrieval, 2001.
[85]. S. Nijssen, T. Meinl, and G. Karypis. ECML/PKDD Workshop on Mining Graphs, Trees and Sequences.2005.
[86]. K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. Journal of theAmerican Statistical Association, 96(455):1077{1087, 2001.
[87]. H.-J. Oh, S. H. Myaeng, and M.-H. Lee. A practical hypertext catergorization method using links andincrementally available class information. In Interna-tional ACM SIGIR Conference on Research and De-velopment in Information Retrieval, pages 264{271,2000.
[88]. J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for even-based network
[89]. J. O'Madadhain and P. Smyth. EventRank: A framework for ranking time-varying networks. In KDDWorkshop on Link Discovery (LinkKDD): Issues, Ap-proaches and Applications, 2005.
[90]. J. O'Madadhain, P. Smyth, and L. Adamic. Learning predictive models for link formation. Presented atthe International Sunbelt Social Network Conference,February, 2005.
[91]. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web.Technical report, Stanford University, 1998.
[92]. H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser.Identity uncertainty and citation matching. InAdvances in Neural Information Processing Systems 15. MIT Press, 2003.
[93]. A. Popescul and L. H. Ungar. Statistical relational learning for link prediction. In IJCAI Workshop onLearning Statistical Models from Relational Data,2003.
[94]. L. D. Raedt and T. Washio. ECML/PKDD Workshop on Mining Graphs, Trees and Sequences. 2003.
[95]. D. Ra_ei and A. O. Mendelzon. What is this page known for? Computing web page reputations. In In-ternational World Wide Web Conference (WWW), pages 823{835. North-Holland Publishing Co., 2000.
[96]. C. Ramakrishnan, W. Milnor, M. Perry, and A. Sheth. Discovering informative connection subgraphs inmulti-relational graphs. SIGKDD Explorations, 7(2),December 2005.
[97]. M. Rattigan and D. Jensen. The case for anomalous link discovery. SIGKDD Explorations, 7(2), December2005.
[98]. M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and ContentInformation in PageRank. In Advances in Neural Information Processing Systems 14. MIT Press, 2002.
[99]. A. Rosenfeld, R. Hummel, and S. Zucker. Scene labelling by relaxation operations. IEEE Transactions onSystems, Man and Cybernetics, 6(6), 1976.
[100]. T. Senator. Link mining applications: Progress and challenges. SIGKDD Explorations, 7(2), 2005.
[101]. L. Singh, L. Getoor, and L. Licamele. Pruning social networks using structural properties and descriptiveattributes. In International Conference on Data Min-ing, 2005.
[102]. P. Singla and P. Domingos. Multi-relational record linkage. In KDD Workshop on Multi-Relational DataMining, Seattle, WA, August 2004.
[103]. D. Skillicorn and K. Carley. SIAM Workshop on Link Analysis, Counterterrorism and Security. 2005.
[104]. A. Srivastava, D. Barbara, H. Kargupta, and V. Kumar.SIAM Workshop on Data Mining for Countert-errorism and Security. 2003.
[105]. J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos.Relevance search and anomaly detection in bipartitegraphs. SIGKDD Explorations, 7(2), December 2005.
[106]. L. Sweeney. Privacy-enhanced linking. SIGKDD Ex-plorations, 7(2), 2005.
[107]. B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In Proc. of UAI, pages 485{492, Edmonton, Canada, 2002.
[108]. B. Taskar, M.-F.Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In Neural Information Processing Systems Conference, Vancouver, Canada,December 2003.
[109]. J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as Spectroscopy: Automated Discovery of Com-munity Structure within Organizations. Kluwer, B.V.,Deventer, The Netherlands, The Netherlands, 2003.
[110]. X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations

and text. In KDD Work- shop on Link Discovery, August 2005.
[111]. S. Wasserman. Conformity of two sociometric relations. Psychometrika, 52:3{18, 1987.
[112]. S. Wasserman and K. Faust. Social Network Analy-sis: Methods and Applications. Cambridge UniversityPress, Cambridge, 1994.
[113]. D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. Nature, 393:440{442, 1998.
[114]. S. White and P. Smyth. Algorithms for estimating relative importance in networks. In ACM SIGKDD In-ternational Conference on Knowledge Discovery and Data Mining, pages 266{275, 2003.
[115]. A. P. Wolfe and D. Jensen. Playing multiple roles:Discovering overlapping roles in social networks. InICML-04 Workshop on Statistical Relational Learning and its Connections to Other Fields, 2004.
[116]. X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In International Conference on DataMining, 2002.
[117]. K. Yoshida, H. Motoda, and N. Indurkhya. Graphbased induction as a uni_ed learning framework. Jour-nal of Applied Intelligence, 4(3):297{316, July 1994.