# Learning Structure and Schemas from Heterogeneous Domains in Networked Systems Surveyed

Marenglen Biba and Fatos Xhafa

**Abstract.** With the continuous growing amount of digital documents in many different formats and with the increasing possibility to access these through internet-based technologies in distributed environments, there is strong motivation to develop robust methods to organize documents in large and repositories. In particular, the extremely large volume of document collections makes it unfeasible to manually handle such documents. In addition, most of the documents exist in an unstructured form and do not follow any schemas. Therefore, research efforts in this direction are being dedicated to automatically infer structure and schemas. This is essential in order to properly organize huge collections as well as to effectively and efficiently retrieve documents in in . This chapter presents a survey of the state-of-the-art methods for inferring structure from documents and schemas in networked environments. The survey is organized around important application domains such as bio-informatics, sensor networks, social networks, P2P systems, automation and control, transportation and privacy-preserving for which we analyze the recent developments on dealing with unstructured data in such domains.

## 1 Introduction

For a long time, , and research areas have focused mainly on structured domains. Structured information is usually easier to process by machines, however, there are still many research issues [14]. Some research works are concerned with the visual representation of documents, while improvements are being made in the area of

Marenglen Biba
Department of Computer Science, University of New York Tirana, Albania
e-mail: marenglenbiba@unyt.edu.al

Fatos Xhafa
Technical University of Catalonia,
Barcelona, Spain
e-mail: fatos@lsi.upc.edu

M. Biba and F. Xhafa (Eds.): Learning Structure and Schemas from Documents, SCI 375, pp. 1–16.
springerlink.com 

pattern recognition and to classify documents according to structure found in their layout [31]. On the other side, research works in the field of machine learning try to exploit attributes of documents and relationships among different documents to infer structures in large collections of documents taking into consideration also the uncertainty that may be present in the application domain [18].

Most of the information sources on the web contain information in the form of free text or heterogeneous documents distributed among different domains in networked environments. Recently, a growing body of research work is addressing the problem of recognizing structure and schemas in documents of various formats from heterogeneous domains. As in the case of structured information, machine learning, pattern recognition and data mining methods have the potential to uncover relevant hidden structures in heterogeneous data. However, most of the methods and techniques developed so far have focused only on the classical problem of discovering structures in documents and data, and should now re-consider their setting (and maybe their theoretical framework) in order to deal with unstructured information in heterogeneous domains. In particular new features, typical of heterogeneous distributed data, such as noise, incomplete data, missing attributes, stream data, private data, social data, etc., pose new challenges to the community of machine learning and data mining. In fact, dealing with heterogeneous data is a challenge *per se* even in non networked systems.

In this chapter we present some recent approaches in the literature that deal with inferring structure in heterogeneous collections of documents and data which are distributed in networked systems. The goal is to present the state-of-the-art in the area in order to emphasize some lessons learned, identify new research issues and challenges as well as opportunities for further developments. The growing body of research work dedicated to the problem of dealing with different types of data in heterogeneous domains is found across several application domains such as bioinformatics, sensor networks, social networks, P2P systems, automation and control, transportation and privacy-preserving. We dedicate a section to each of these areas considering recent developments and describing each of the approaches in order to outline opportunities for new research.

## 2 Learning Patterns in Sensor Networks

present large amounts of data spread over many physically distributed nodes. Machine learning and data mining techniques have the potential to deal with these kind of data. Due to the complexity of heterogeneous networked data, important challenges have arisen such as the need for run-time data aggregation, parallel computing, and distributed hypothesis formation [8]. One of the existing approaches in sensor networks is presented in [53] where the authors present an algorithm for finding distributed icebergs-elements that may have low frequency at individual nodes but high aggregate frequency (this is a problem that arises commonly in practice). The work in [7] addresses a major challenge in data mining applications where the full information about the underlying processes, such as sensor networks or large

online databases, cannot be practically obtained due to physical limitations such as low bandwidth or memory, storage, or computing power. They propose a framework for detecting anomalies from these large-scale data mining applications where the full information is not practically possible to obtain.

In [43] it is presented another approach for network management in large-scale randomly-deployed sensor networks, called Energy Map, which explores the inherent relationships between the energy consumption and the sensor operation. Through nonlinear manifold learning algorithms the approach visualizes the residual energy level of each sensor in a large scale network, infers the sensor locations and the current network topology through mining the collected residual energy data in a randomly-deployed sensor network, and explores the inherent relation between sensor operation and energy consumption to find the dynamic patterns from large volumes of sensor network data for network design.

In [39] and [40] the author proposes a declarative query language and data mining techniques to discover frequent event patterns and their spatial and temporal properties. In these works, raw streams of sensor readings are collected for later offline processing and analysis and in-network data mining techniques are explored to discover frequent event patterns and their spatial and temporal properties.

The authors of [29] propose and evaluate distributed algorithms for data clustering in self-organizing ad-hoc sensor networks with computational, connectivity, and power constraints. One of the benefits of in-network data clustering algorithms is the capability of the network to transmit only relevant, high level information, namely models, instead of large amounts of raw data, also reducing drastically energy consumption. Finally, the work in [38] presents an exploration of different characteristics of sensor networks which define new requirements for knowledge discovery, with the common goal of extracting some kind of comprehension about sensor data and sensor networks, focusing on clustering techniques which provide useful information about sensor networks as it represents the interactions between sensors.

In [34] the authors propose a combination of a neural network based offline learning approach and online repuation update schemes to identify nodes reporting inconsistent data. The authors experimentally evaluate their scheme for two different network sizes and two different data patterns over the sensor field and the results show that their approach is successful in identifying multiple colluding malicious nodes without any false positives and false negatives.

## 3 Learning Structures in Biological Domains

Recent advances in computing, digital storage technologies and high throughput data acquisition technologies have led to a growing capability to gather and store huge volumes of data. For example, advances in high throughput sequencing and other data acquisition have resulted in gigabytes of DNA, protein sequence data, and gene expression data being gathered continuously. On the other side, the necessity for methods to automatically analyze large volumes of data has led to a growing

effort in machine learning and data mining communities to develop robust methods that work effectively on heterogeneous domains. For example, in [20], it was presented a mixture model associative artificial neural network that integrates two heterogeneous domain knowledge (Gene Ontology (GO) annotation and gene expression profiling) for discovery of genome-wide functional patterns. The presented experiments showed that association of these domains reduces analytical noises and produces a more meaningful functional grouping. In the same direction goes also another approach presented in [49]. Since in many biomedical modeling tasks a number of different types of data may influence predictions made by the model, an established approach to pursuing supervised learning with multiple types of data is to encode these different types of data into separate kernels and use multiple kernel learning. In [49] the authors present a simple iterative approach to multiple kernel learning (MKL), focusing on multi-class classification and show that the proposed method outperforms state-of-the-art results on an important protein fold prediction dataset and gives competitive performance on a protein subcellular localization task.

Another interesting approach is presented in [47], where the authors introduced a multi-label large-margin classifier that automatically learns the underlying intercode structure and allows the controlled incorporation of prior knowledge about medical code relationships. In addition to refining and learning the code relationships, their classifier can also use this shared information to improve its performance. Experiments on a publicly available dataset containing clinical free text and their associated medical codes showed that the proposed multi-label classifier outperforms related multi-label models in this problem.

Another line of research has been that of monitoring the occurrence of topics in a stream of events, such as a stream of news articles. This has led to several models of bursts in these streams, i.e., periods of elevated occurrence of events and there are several burst definitions and detection algorithms. The authors in [21] present a topic dynamics model for the large PubMed/MEDLINE database of biomedical publications, using the MeSH (Medical Subject Heading) topic hierarchy. They show that their model is able to detect bursts for MeSH terms accurately as well as efficiently.

Another challenge in bioinformatics is that of selecting genes that are differentially expressed and critical to a particular biological process. Important developments in data gathering technologies have made various data sources available such as mRNA and miRNA expression profiles, biological pathway and gene annotation, etc. Recent works have also shown that integration of multiple data sources helps enrich knowledge for selecting genes bearing significant biological relevance. One approach is the one proposed in [52], where multiple data sources are extracted into an intrinsic global geometric pattern which is used in covariance analysis for gene selection. Another approach is presented in [36] where the authors propose a machine learning technique to identify essential genes using the experimental data of genome-wide knock-out screens from one bacterial organism to infer essential genes of another related bacterial organism. They use a broad variety of topological features, sequence characteristics and co-expression properties potentially associated with essentiality, such as flux deviations, centrality, codon frequencies of the sequences, coregulation and phyletic retention.

The analysis of trends and topics in the biomedical literature is yet another hot research direction nowadays. Often the goal in bio-informatics is to identify potential diagnostic and therapeutic bio-markers for specific diseases. In [33] the presented approach integrates several data sources to provide the user with up-to-date information on current research in the field. The BioJournalMonitor is a decision support system that deploys state-of-the-art text mining technologies to provide added value on top of the original content, including named entity detection, relation extraction, classification, clustering, ranking, summarization, and visualization. The presented results suggest that early prediction of emerging trends is possible through a probabilistic topic models that can be used to annotate recent articles with the most likely MeSH terms.

## 4 Learning in Distributed Automation and Control Systems

Machine learning and data mining provide excellent methods and techniques for dealing with automation and control in a distributed setting. Here we explore some approaches that have proven successful in important areas such as transportation, fleets and automation control.

In [22] it is presented a distributed vehicle performance data mining system designed for commercial fleets. The MineFleet system analyzes high throughput data streams onboard the vehicle, generates the analytics, sends those to the remote server over the wide-area wireless networks and offers them to the fleet managers using stand-alone and web-based user-interface. MineFleet is probably one of the first commercially successful distributed data stream mining systems. Another approach was proposed in [27] called mobility-based clustering that deals with practical research on hot spots in smart city taking into consideration unique features, such as highly mobile environments, supremely limited size of sample objects, and the non-uniform, biased samples. The authors report performance of mobility-based clustering based on real traffic situations.

The authors in [54] deal with the critical problem in a crisis situation of how to efficiently discover, collect, organize, search and disseminate real-time disaster information. The proposed system exploits the latest advances in data mining technologies to analyze the integrated input data from different sources. Another interesting approach is that in [17] where a massive quantity of complex, dynamic, and distributed location traces is handled and mined to provide effective mobile sequential recommendation.

In another recent work, a novel approach was presented based on the theory of multiple kernel learning to detect potential safety anomalies in very large data bases of discrete and continuous data from world-wide operations of commercial fleets [11]. Their results show that the proposed algorithm uncovers operationally significant events in high dimensional data streams in the aviation industry which are not detectable using state of the art methods. Another interesting approach is that of [23] where it is presented a system based on Ubiquitous Data Mining (UDM)

concepts. It merges and analyses different types of information from crash data and physiological sensors to diagnose driving risks in real time.

An important feature of networked data is their uncertainty since sensors are typically expected to have considerable noise in their readings because of inaccuracies in data retrieval, transmission, and power failures. In [2] the authors propose a method for clustering uncertain data streams.

Another interesting approach is presented in [6] where the authors show that shortcomings of automating datacenters using closed-loop control can be addressed by replacing simple techniques of modeling and model management with more sophisticated techniques imported from statistical machine learning. An interesing approach is also presented in [45] where the authors present a novel framework that combines queueing networks and graphical models, allowing Markov chain Monte Carlo to be applied. The authors demonstrate the effectiveness on real-world data from a benchmark Web application.

Another challenging task has always been collective decision making. In particular, collective recognition is the problem of jointly applying multiple classifiers. The decisions are made about the class of an entity, situation, image, etc, and the joint decision is used to improve quality of the final decision by aggregation and coordination of different classifier decisions using a metalevel algorithm [19].

## 5 Learning Structures in Social Networks

Social networks have inspired a lot research in the machine learning and data mining community. This is due to the growing amount of available data that have to be analyzed. Moreover, most social networks have an outstanding marketing value and developing methods for viral marketing is a hot topic in the research community [13].

Mobile devices and wireless technologies have led to mobile social network systems which are increasingly becoming popular. In a mobile social network the spread of information and influence is in the form of word-of-mouth, therefore it is important to find a subset of influential individuals in a mobile social network such that targeting them initially (e.g. for marketing campaigns) will maximize the spread of the influence. Unfortunately, it has been shown that the problem of finding the most influential nodes is NP-hard. It has also been shown that a Greedy algorithm with provable approximation guarantees can give good approximation. However, it is computationally expensive, if not prohibitive, to run the greedy algorithm on a large mobile network. In [56] the authors propose a new algorithm called Community-based Greedy algorithm for mining top-K influential nodes. The proposed algorithm encompasses two components: 1) an algorithm for detecting communities in a social network; and 2) a dynamic programming algorithm for selecting communities to find influential nodes. Empirical experiments on a large real-world mobile social network show that their algorithm is more than an order of magnitudes faster than the state-of-the-art Greedy algorithm for finding top-K influential nodes and the error of their approximate algorithm is small.

Another line of research is the analysis of blog data. In [1] the authors present an approach that uses innovative ways to employ contextual information and collective wisdom to aggregate similar bloggers.

Social interactions that occur regularly typically correspond to significant yet often infrequent and hard to detect interaction patterns. To identify such regular behavior, the authors in [24] propose a new mining problem of finding periodic or near periodic subgraphs in dynamic social networks. They propose a practical, efficient and scalable algorithm to find such subgraphs that takes imperfect periodicity into account and demonstrate the applicability of their approach on several real-world networks and extract meaningful and interesting periodic interaction patterns.

Some recent developments regard the application of the concept of organizational structure to social network analysis which may well represent the power of members and the scope of their power in a social network. In [37], the authors propose a data structure, called Community Tree, to represent the organizational structure in the social network. They combine the PageRank algorithm and random walks on graph to derive the community tree from the social network. Experiments conducted on real data show that the methods are effective at discovering the organizational structure and representing the evolution of organizational structure in a dynamic social network.

Social networks often involve multiple relations simultaneously. People usually construct an explicit social network by adding each other as friends, but they can also build implicit social networks through daily actions like commenting on posts, or tagging photos. The authors in [15] address this problem: given a real social networking system which changes over time, do daily interactions follow any pattern? They model the formation and co-evolution of multi-modal networks proposing an approach that discovers temporal patterns in peoples social interactions. They show the effectiveness of the approach on two real datasets (Nokia FriendView and Flickr) with 100,000 and 50,000,000 records respectively, each of which corresponds to a different social service, and spans up to two years of activity.

The formation of implicit groups is an interesting related problem here. Although users of online communication tools do not usually categorize their contacts into groups such as ”family”, ”co-workers”, or ”jogging buddies”, they implicitly cluster contacts through their interactions with them, forming implicit groups. The authors in [41] describe the implicit social graph which is formed by users’ interactions with contacts and groups of contacts, and which is distinct from explicit social graphs in which users explicitly add other individuals as their ”friends”. They introduce an interaction-based metric for estimating a user’s affinity to his contacts and groups and propose a novel friend suggestion algorithm that exploits a user’s implicit social graph to generate a friend group, given a small seed set of contacts which the user has already labeled as friends. Their experimental results prove the importance of both implicit group relationships and interaction-based affinity ranking in suggesting friends.

Also, the study of critical nodes appears to be an interesting problem related to inhibiting diffusion of complex contagions such as rumors, undesirable fads and mob behavior in social networks by removing a small number of nodes (called critical

nodes). The authors in [42] develop efficient heuristics for these problems and perform empirical studies of their performance on three well known social networks, namely epinions, wikipedia and slashdot.

Another important task in an online community is to observe and track the popular events, or topics that evolve over time in the community. Existing approaches have usually focused on either the burstiness of topics or the evolution of networks, but have usually ignored the interplay between textual topics and network structures. In [25] the authors formally define the problem of popular event tracking in online communities (PET), focusing on the interplay between texts and networks. They propose a novel statistical method that models the popularity of events over time, taking into consideration the burstiness of user interest, information diffusion on the network structure, and the evolution of textual topics. The approach models the influence of historic status and the dependency relationships in the graph through a Gibbs Random Field. Empirical experiments with two different communities and datasets (i.e., Twitter and DBLP) show that their approach is effective.

Current social networks are continuously growing. This makes them hard to analyze and in some cases intractable. One solution to this is compressing social networks so that we can substantially facilitate mining and advanced analysis of large social networks. The optimal solution would be to compress social networks in a way that they still can be queried efficiently without decompression. For example, we should still be able to perform neighbor queries efficiently (which search for all neighbors of a query vertex), as these are the most essential operations on social networks. The problems has been addressed in [32] where the authors propose an social network compression approach based on a novel Eulerian data structure using multi-position linearizations of directed graphs. Their approach seems to be the first that can answer both out-neighbor and in-neighbor queries in sublinear time and they verfiy their design with an extensive empirical study on more than a dozen benchmark real data sets.

Finally, an important task often essential in some social networks is the discovery of communities. Usually, the given scenario is the one where communities need to be discovered with only reference to the input graph. However, for many interesting applications one is interested in finding the community formed by a given set of nodes. In [44] the authors study a query-dependent variant of the community-detection problem, which they call the community-search problem: given a graph G, and a set of query nodes in the graph, the goal is to find a subgraph of G that contains the query nodes and it is densely connected. A measure of density is proposed based on minimum degree and distance constraints, and an optimum greedy algorithm is developed for this measure. The authors characterize a class of monotone constraints and they generalize the algorithm to compute optimum solutions satisfying any set of monotone constraints. Finally they modify the greedy algorithm and present two heuristic algorithms that find communities of size no greater than a specified upper bound. The experimental evaluation on real datasets demonstrates the efficiency of the proposed algorithms and the quality of the solutions they obtain.

## 6 Learning Structures in Peer-to-Peer Networks

There research in machine learning and data mining on analyzing data in Peer-to-Peer (P2P) networks is attracting a lot of attention of researchers. We will briefly describe here some recent developments in this exciting area.

In [3] the authors proposed a novel P2P learning framework for concept drift classification, which includes both reactive and proactive approaches to classify the drifting concepts in a distributed manner. Their empirical study shows that the proposed technique is able to effectively detect the drifting concepts and improve the classification performance.

The authors of [5] presented an algorithm for learning parameters of Gaussian mixture models (GMM) in large P2P environments that can be used for a variety of well-known data mining tasks in distributed environments such as clustering, anomaly detection, target tracking, and density estimation, which are necessary for many emerging P2P applications in bio-informatics, web-mining and sensor networks.

Tagging information is often an important feature to exploit in analyzing test documents. In many application areas involving classification of text documents, web users participate in the tagging process and the collaborative tagging results in the formation of large scale P2P systems which can function, scale and self-organize in the presence of highly transient population of nodes and do not need a central server for co-ordination. In [16] it is presented a P2P classifier learning system for extracting patterns from text data where the end users can participate both in the task of labeling the data and building a distributed classifier on it. The approach is based on a novel distributed linear programming based classification algorithm which is asynchronous in nature. The authors provide extensive empirical results on text data obtained from the online repository of NSF Abstracts Data.

Another important challenge in data mining over P2P networks is the right data representation. In [58] the authors describe an approach to collaborative feature extraction, selection and aggregation in distributed, loosely coupled domains. The authors focus on scenarios in which a large number of loosely coupled nodes apply data mining to different, usually very small and overlapping, subsets of the entire data space. The goal is to learn a set of local concepts and not to find a global concept. The paper proposes two models for collaborative feature extraction, selection and aggregation for supervised data mining. One is based on a centralized P2P architecture, and the other on a fully distributed P2P architecture. The comparison of both models is performed on a real word data set.

An important direction for research is the self-reorganization in P2P networks. In [4] the authors employ machine learning feature selection in a novel manner: to reduce communication cost thereby providing the basis of an efficient neighbor selection scheme for P2P overlays. In addition, their method enables nodes to locate and attach to peers that are likely to answer future queries with no a priori knowledge of the queries.

Finally, an important challenge is that fully distributed data mining algorithms build global models over large amounts of data distributed over a large number

of peers in a network, without moving the data itself. The difficulty of the problem stands in implementing good quality models with an affordable communication complexity, while assuming as little as possible about the communication model. In [35] the authors describe a conceptually simple, yet powerful generic approach for designing efficient, fully distributed, asynchronous, local algorithms for learning models of fully distributed data. The key idea proposed in the paper is that many models perform a random walk over the network while being gradually adjusted to fit the data they encounter, using stochastic gradient descent search. The authors demonstrate their approach by implementing the support vector machine method and by experimentally evaluating its performance in various failure scenarios over different benchmark datasets.

## 7 Learning and Privacy-Preserving in Distributed Environments

Privacy is a major issue in data mining applications and very often privacy considerations often constrain data mining projects. For this reason, a growing amount of research is being dedicated to the problem of analyzing data from distributed environments under privacy requirements. Most distributed data mining applications, such as those dealing with health care, finance, counter-terrorism and homeland security, use sensitive data from distributed databases held by different parties.

One of the recent approaches proposes mining of association rules where transactions are distributed across sources [46]. In this scenario each site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules. The sites must not reveal individual transaction data. The authors present a two-party algorithm for efficiently discovering frequent patterns, without either site revealing individual transaction values.

Some research work has considered the possibility of using multiplicative random projection matrices for privacy preserving distributed data mining. The work in [26] considers the problem of computing statistical aggregates like the inner product matrix, correlation coefficient matrix, and Euclidean distance matrix from distributed privacy sensitive data possibly owned by multiple parties. The authors propose an approximate random projection-based technique to improve the level of privacy protection while still preserving certain statistical characteristics of the data.

An interesting scenario arises when different parties like businesses, governments, and others, may wish to benefit from cooperative use of their data, but privacy regulations and other privacy concerns may prevent them from sharing the data. Privacy-preserving data mining provides solutions through distributed data mining algorithms in which the underlying data is not revealed. An interesting approach is presented in [57] where a Bayesian network structure is learned for distributed heterogeneous data. In the proposed setting, two parties owning confidential databases wish to learn the structure of Bayesian network on the combination of their databases without revealing anything about their data to each other.

The authors give an efficient and privacy-preserving version of the K2 algorithm to construct the structure of a Bayesian network for the parties' joint data.

One of the approaches towards privacy-preserving data mining is to adapt existing successful knowledge discovery algorithms so that they can deal with privacy issues. One of the most widely used classification methodologies in data mining and machine learning is support vector machine classification. The work in [50] proposes privacy-preserving solution for support vector machine classification. The solution constructs the global classification model from the data distributed at multiple parties, without disclosing the data of each party to others. It is assumed that data is horizontally partitioned: each party collects the same features of information for different data objects.

Traditional research on preserving privacy in data mining focuses on time-invariant privacy issues. However with time series data mining, snapshot-based privacy solutions should take into consideration the addition of the time dimension. The work in [55] shows that current techniques to preserve privacy in data mining are not effective in preserving time-domain privacy. They show with real data, that the data flow separation attack on privacy in time series data mining, which is based on blind source separation techniques from statistical signal processing, is effective. The authors propose possible countermeasures to the data flow separation attack in the paper.

Sharing data among multiple parties, without disclosing the data, is an important issue. In [51] it is presented an approach for sharing private or confidential data where multiple parties, each with a private data set, want to collaboratively conduct association rule mining without disclosing their private data. The approach is based on homomorphic encryption techniques to exchange the data while keeping it private. The proposed solution is distributed, i.e., there is no central, trusted party accessing all the data. Another similar problem regarding distributed association rule mining is presented in [48] where the authors come up with a protocol based on a new semi-trusted mixer model. Their protocol can protect the privacy of each distributed database against the coalition up to n-2 other data sites or even the mixer if the mixer does not collude with any data site.

It should be noted that many of the existing techniques have strict assumptions on the involved parties which need to be relaxed in order to reflect the real-world requirements. In [12] the authors present a distributed scenario where the data is partitioned vertically over multiple sites and the involved sites would like to perform clustering without revealing their local databases. They propose a new protocol for privacy preserving k-means clustering based on additive secret sharing and show that the new protocol is more secure than the state of the art.

Interesting research has been dedicated to distributed environments where the participants in the system may also be mutually mistrustful. The work in [30] discusses the design and security requirements for large-scale privacy-preserving data mining (PPDM) systems in a fully distributed setting, where each client possesses its own records of private data. The authors argue in favor of using some well-known cryptographic primitives, borrowed from the literature on Internet elections. They

also show how their approach can be used as a building block to obtain Random Forests classification with enhanced prediction performance.

A powerful approach is extending the privacy preservation notion to original learning algorithms. An interesting effort that addresses this problem is presented in [9] where the authors focus on preserving the privacy in an important learning model, multilayer neural networks. They present a privacy-preserving two-party distributed algorithm of backpropagation which allows a neural network to be trained without requiring either party to reveal her data to the other.

Finally, a large body of research has been devoted to address corporate-scale privacy concerns related to social networks. The main concern has been on how to share social networks without revealing the identities or sensitive relationships of users. An interesting work is proposed in [28] that addresses privacy concerns arising in online social networks from the individual users viewpoint. The authors propose a framework to compute the privacy score of a user, which indicates the potential privacy risk caused by his participation in the network. The definition of privacy score satisfies the following: the more sensitive the information revealed by a user, the higher his privacy risk. Also, the more visible the disclosed information becomes in the network, the higher the privacy risk. The authors develop mathematical models to estimate both sensitivity and visibility of the information.

## 8 Conclusion

In this chapter we presented a survey of the current state-of-the-art methods for inferring structure and schemas from documents in heterogeneous networked environments. The rapidly growing volume of available digital documents of various formats and the possibility to access these through internet-based technologies in distributed environments, have led to the necessity to develop solid methods to properly organize and structure documents in large digital libraries and repositories. Specifically, since the extremely large volumes make it impossible to manually organize such documents and since most of the documents exist in an unstructured form and do not follow any schemas, most of the efforts in this direction are dedicated to automatically infer structure and schemas that can help to better organize huge collections. This is essential in order for these documents to be effectively and efficiently retrieved in heterogeneous domains in networked system.

**Acknowledgment.** The authors would like to thank the Brain Gain program of the Albanian Government for the support.

## References


1. Agarwal, N., Liu, H., Subramanya, S., Salerno, J.J., Yu, P.S.: Connecting Sparsely Distributed Similar Bloggers. In: Proc. of Ninth IEEE International Conference on Data Mining, pp. 11–20 (2009)
2. Aggarwal, C., Yu, P.: A framework for clustering uncertain data streams. In: Proc. of 24th International Conference on Data Engineering, Cancún, México (2008)
3. Ang, H.H., Gopalkrishnan, V., Ng, W.K., Hoi, C. H.: On classifying drifting concepts in P2P networks. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS, vol. 6321, pp. 24–39. Springer, Heidelberg (2010)
4. Beverly, R., Afergan, M.: Proceedings of USENIX Tackling Computer Systems Problems with Machine Learning Techniques (SysML 2007) Workshop, Cambridge, MA (April 2007)
5. Bhaduri, K., Srivastava, A.N.: A Local Scalable Distributed Expectation Maximization Algorithm for Large Peer-to-Peer Networks. In: Proc. of Ninth IEEE International Conference on Data Mining, pp. 31–40 (2009)
6. Bodik, P., Griffith, R., Sutton, C., Fox, A., Jordan, M.I., Patterson, D. A.: Statistical Machine Learning Makes Automatic Control Practical for Internet Datacenters. In: Workshop on Hot Topics in Cloud Computing, HotCloud 2009 (2009)
7. Budhaditya, S., Pham, D., Lazarescu, M., Venkatesh, S.: Effective Anomaly Detection in Sensor Networks Data Streams. In: Proc. of Ninth IEEE International Conference on Data Mining, pp. 722–727 (2009)
8. Cantoni, V., Lombardi, L., Lombardi, P.: Challenges for Data Mining in Distributed Sensor Networks. In: Proc. of 18th International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 1000–1007 (2006)
9. Chen, T., Zhong, S.: Privacy-preserving backpropagation neural network learning. IEEE Transactions on Neural Networks 20(10), 1554–1564 (2009)
10. Das, S., Egecioglu, O., Abbadi, A.E.: Anonymizing weighted social network graphs. In: Proc. of IEEE 26th International Conference on Data Engineering (ICDE), pp. 904–907 (2010)
11. Das, S., Matthews, B.L., Srivastava, A.N., Oza, N.C.: Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 25-28. ACM, USA (2010)
12. Doganay, M.C., Pedersen, T.B., Saygin, Y., Savas, E., Levi, A.: Distributed privacy preserving k-means clustering with additive secret sharing. In: Proc. of the 2008 International Workshop on Privacy and Anonymity in Information Society, Nantes, France, March 29-29 (2008)
13. Domingos, P.: Mining Social Networks for Viral Marketing. IEEE Intelligent Systems 20(1), 80–82 (2005)
14. Domingos, P.: Structured Machine Learning: Ten Problems for the Next Ten Years. Machine Learning 73, 3–23 (2008)
15. Du, N., Wang, H., Faloutsos, C.: Analysis of large multi-modal social networks: Patterns and a generator. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS, vol. 6321, pp. 393–408. Springer, Heidelberg (2010)
16. Dutta, H., Zhu, X., Mahule, T., Kargupta, H., Borne, K., Lauth, C., Holz, F., Heyer, G.: TagLearner: A P2P Classifier Learning System from Collaboratively Tagged Text Documents. In: Proc. of Ninth IEEE International Conference on Data Mining Workshops, pp. 495–500 (2009)

17. Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., Pazzani, M.: An energy-efficient mobile recommender system. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
18. Getoor, L., Taskar, B.: Introduction to statistical relational learning. MIT Press, Cambridge (2007)
19. Gorodetskiy, V.I., Serebryakov, S.V.: Methods and algorithms of collective recognition. Automation and Remote Control 69(11), 1821–1851 (2008)
20. He, J., Dai, X., Zhao, P.X.: Mixture Model Adaptive Neural Network for Mining Gene Functional Patterns From Heterogenous Knowledge Domains. International Journal of Information Technology and Intelligent Computing (2007)
21. He, D., Parker, D.S.: Topic dynamics: an alternative model of bursts in streams of topics. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 25-28. ACM, Washington (2010)
22. Kargupta, H., Sarkar, K., Gilligan, M.: MineFleet: an overview of a widely adopted distributed vehicle performance data mining system. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
23. Krishnaswamy, S., Loke, S.W., Rakotonirainy, A., Horovitz, O., Gaber, M. M.: Towards Situation-awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application. In: Proc. of Conference on Intelligent Vehicles and Road Infrastructure (IVRI 2005), February 16-17, University of Melbourne (2005)
24. Lahiri, M., Berger-Wolf, T.Y.: Mining Periodic Behavior in Dynamic Social Networks. In: Proc. of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19. IEEE Computer Society, Pisa (2008)
25. Lin, C.X., Zhao, B., Mei, Q., Han, J.: PET: a statistical model for popular events tracking in social communities. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, ACM, New York (2010)
26. Liu, K., Kargupta, H., Ryan, J.: Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering 18(1), 92–106 (2006)
27. Liu, S., Liu, Y., Ni, L.M., Fan, J., Li, M.: Towards mobility-based clustering. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
28. Liu, K., Terzi, E.: A Framework for Computing the Privacy Scores of Users in Online Social Networks. In: Proc. of the Ninth IEEE International Conference on Data Mining, pp. 288–297, 932–937 (2009)
29. Lodi, S., Monti, G., Moro, G., Sartori, C.: Peer-to-Peer Data Clustering in Self-Organizing Sensor Networks. In: Intelligent Techniques for Warehousing and Mining Sensor Network Data, pp. 179–212. IGI Global (2010)
30. Magkos, E., Maragoudakis, M., Chrissikopoulos, V., Gritzalis, S.: Accurate and large-scale privacy-preserving data mining using the election paradigm. Data and Knowledge Engineering 68(11), 1224–1236 (2009)
31. Marinai, S., Fujisawa, H. (eds.): Machine Learning in Document Analysis and Recognition. SCI, vol. 90. Springer, Heidelberg (2008)
32. Maserrat, H., Pei, J.: Neighbor query friendly compression of social networks. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28. ACM, New York (2010)

33. Morchen, F., Dejori, M., Fradkin, D., Etienne, J., Wachmann, B., Bundschus, M.: Anticipating annotations and emerging trends in biomedical literature. In: Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27. ACM, New York (2008)
34. Mukherjee, P., Sen, S.: Using learned data patterns to detect malicious nodes in sensor networks. In: Rao, S., Chatterjee, M., Jayanti, P., Murthy, C.S.R., Saha, S.K. (eds.) ICDCN 2008. LNCS, vol. 4904, pp. 339–344. Springer, Heidelberg (2008)
35. Ormandi, R., Hegedu, I., Jelasity, M.: Asynchronous Peer-to-peer Data Mining with Stochastic Gradient Descent. In: Proceedings of 17th International European Conference on Parallel and Distributed Computing, EuroPar 2011, Bordeux, France (2011)
36. Plaimas, K., Eils, R., Konig, R.: Identifying essential genes in bacterial metabolic networks with machine learning methods. In: BMC Systems Biology 2010, vol. 4, p. 56 (2010)
37. Qiu, J., Lin, Z., Tang, C., Qiao, S.: Discovering Organizational Structure in Dynamic Social Network. In: Proc. of the Ninth IEEE International Conference on Data Mining, pp. 932–937 (2009)
38. Rodrigues, P.P., Gama, J., Lopes, L.: Knowledge Discovery for Sensor Network Comprehension. In: Intelligent Techniques for Warehousing and Mining Sensor Network Data, pp. 179–212. IGI Global (2010)
39. Römer, K.: Discovery of frequent distributed event patterns in sensor networks. In: Verdone, R. (ed.) EWSN 2008. LNCS, vol. 4913, pp. 106–124. Springer, Heidelberg (2008)
40. Romer, K.: Distributed Mining of Spatio-Temporal Event Patterns in Sensor Networks. In: EAWMS / DCOSS 2006, San Francisco, USA, pp. 103–116 (June 2006)
41. Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., Merom, R.: Suggesting friends using the implicit social graph. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28. ACM, New York (2010)
42. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Selecting information diffusion models over social networks for behavioral analysis. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS, vol. 6323, pp. 180–195. Springer, Heidelberg (2010)
43. Song, C.: Mining and visualising wireless sensor network data Source. International Journal of Sensor Networks archive 2(5/6), 350–357 (2007)
44. Sozio, M., Gionis, A.: The community-search problem and how to plan a successful cocktail party. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28. ACM, New York (2010)
45. Sutton, C., Jordan, M.I.: Learning and Inference in Queueing Networks. In: Conference on Artificial Intelligence and Statistics, AISTATS (2010)
46. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: Proc. of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002)
47. Yan, Y., Fung, G., Dy, J.G., Rosales, R.: Medical coding classification by leveraging inter-code relationships. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28. ACM, New York (2010)
48. Yi, X., Zhang, Y.: Privacy-preserving distributed association rule mining via semi-trusted mixer. Data and Knowledge Engineering 63(2), 550–567 (2007)

49. Ying, Y., Campbell, C., Damoulas, T., Girolami, M.: Class Prediction from Disparate Biological Data Sources Using an Iterative Multi-kernel Algorithm. In: 4th IAPR International Conference on Pattern Recognition in Bioinformatics, Sheffield (2009)
50. Yu, H., Jianga, X., Vaidya, J.: Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: Proc. of the 2006 ACM Symposium on Applied computing, April 23-27, Dijon, France (2006)
51. Zhan, J., Matwin, S., Chang, L.: Privacy-preserving collaborative association rule mining. Journal of Network and Computer Applications 30(3), 1216–1227 (2007)
52. Zhao, Z., Wang, J., Liu, H., Ye, J., Chang, Y.: Identifying biologically relevant genes via multiple heterogeneous data sources. In: Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27. ACM, New York (2008)
53. Zhao, H., Lall, A., Ogihara, M., Jun, X.: Global iceberg detection over distributed data streams. In: Proc. of IEEE 26th International Conference on Data Engineering, ICDE (2010)
54. Zheng, L., Shen, C., Tang, L., Li, T., Luis, S., Chen, S., Hristidis, V.: Using data mining techniques to address critical information exchange needs in disaster affected public-private networks. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
55. Zhu, Y., Fu, Y., Fu, H.: On privacy in time series data mining. In: Proc. of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Osaka, Japan, May 20-23 (2008)
56. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
57. Wright, R., Yang, Z.: Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004)
58. Wurst, M., Morik, K.: Distributed feature extraction in a p2p setting: a case study. Future Generation Computer Systems 23(1), 69–75 (2007)