Activity Report 2014

Team MAGNET

Machine Learning in Information Networks

RESEARCH CENTER
Lille - Nord Europe

THEME
Data and Knowledge Representation and Processing

Table of contents

Team MAGNET

Keywords: Machine Learning, Graphs, Networks, Natural Language

Creation of the Team: 2013 January 01.

1. Members

Research Scientist
Pascal Denis [Inria, Researcher]

Faculty Members
Marc Tommasi [Team leader, Univ. Lille III, Professor, HdR]
Rémi Gilleron [Univ. Lille III, Professor, HdR]
Mikaela Keller [Univ. Lille III, Associate Professor]
Fabien Torre [Univ. Lille III, Associate Professor]
Fabio Vitale [Univ. Lille III, Associate Professor]

PhD Students
Thomas Ricatte [Inria, until Oct 2014]
David Chatel [Inria]
Géraud Le Falher [Inria, since Oct 2014]
Pauline Wauquier [Cifre Clic and Walk]

Visiting Scientist
Paulin Melatagia Yonta [Univ. de Yaoundé I, Sep 2014]

Administrative Assistant
Julie Jonas [Inria]

Other
Ahmad Hosseini [Univ. Lille I, until May 2014]

2. Overall Objectives

2.1. Presentation

MAGNET is a new research group that aims to design new machine learning based methods geared towards mining information networks. Information networks are large collections of interconnected data and documents like citation networks and blog networks among others. We will then define new prediction methods for texts and networks of texts based on machine learning algorithms in graphs. Such algorithms include node and link classification, link prediction, clustering and probabilistic modeling of graphs. Envisioned applications include browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Application domains cover social networks for cultural data and e-commerce, and biomedical informatics.

3. Research Program

3.1. Introduction

The main objective of MAGNET is to develop original machine learning methods for networked data. We consider information networks in which the data are vectorial data and texts. We model such information networks as (multiple) (hyper)graphs wherein nodes correspond to entities (documents, spans of text, users, ...) and edges correspond to relations between entities (similarity, answer, co-authoring, friendship, ...). Our main research goal is to propose new learning algorithms to build applications like browsing, monitoring and recommender systems, and more broadly information extraction in information networks. Hence, we will investigate new learning algorithms for node clustering and node classification, link classification and link prediction. Also, we will search for the best hidden graph structure to be generated for solving a given learning task. We will base our research on generative models for graphs, on machine learning for graphs and on machine learning for texts. The challenges are the dimensionality of the input space, possibly the dimensionality of the output space, the high level of dependencies between the data, the inherent ambiguity of textual data and the limited amount of human labeling. An additional challenge will be to design scalable methods for large information networks. Hence, we will explore how sampling and randomization can be used in new machine learning algorithms. Also, active machine learning algorithms for graphs will be investigated.

On the first hand we want to design machine learning algorithms on graphs to solve problems in networks of texts and documents in natural language. The main originality of this research is to consider and take advantage of the setting of networked data exploiting the relationships between different data entities and, overall, the graph topology. On the second hand, in a concomitant way, we want to develop prediction models for graph-like data. This includes prediction, ranking and classification of links and nodes in an on-line or batch setting. The two objectives are intertwined, enrich each other and raise important scientific questions we want to focus on. Our research proposal is organized according to the following questions:

1. How to go beyond vectorial classification models in natural language oriented tasks?
2. How to adaptively build graphs with respect to the given tasks? How to create network from observations of information diffusion processes?
3. How to design methods able to achieve very good predictive accuracy without giving up on scalability?
4. How to go beyond strict node homophilic/similarity assumptions in graph-based learning methods?

3.2. Beyond vectorial models for NLP

One of our overall research objectives is to derive graph-based machine learning algorithms for natural language and text information extraction tasks. This section discusses the motivations behind the use of graph-based ML approaches for these tasks, the main challenges associated with it, as well as some concrete projects. Some of the challenges go beyond NLP problems and will be further developed in the next sections. An interesting aspect of the project is that we anticipate some important cross-fertilizations between NLP and ML graph-based techniques, with NLP not only benefiting from but also pushing ML graph-based approaches into new directions.

Motivations for resorting to graph-based algorithms for texts are at least threefold. First, online texts are organized in networks. With the advent of the web, and the development of forums, blogs, and micro-blogging, and other forms of social media, text productions have become strongly connected. Thus, documents on the web are linked through hyperlinks, forum posts and emails are organized in threads, tweets can be retweeted, etc. Additional connections can be made through users connections (co-authorship, friendship, follower, etc.). Interestingly, NLP research has been rather slow in coming to terms with this situation, and most work still focus on document-based or sentence-based predictions (wherein inter-document or inter-sentence structure is not exploited). Furthermore, several multi-document tasks exist in NLP (such as multi-document summarization and cross-document coreference resolution), but most existing work typically ignore document boundaries and simply apply a document-based approach, therefore failing to take advantage of the multi-document dimension [26], [28].

A second motivation comes from the fact that most (if not all) NLP problems can be naturally conceived as graph problems. Thus, NL tasks often involve discovering a relational structure over a set of text spans (words, phrases, clauses, sentences, etc.). Furthermore, the *input* of numerous NLP tasks is also a graph; indeed, most end-to-end NLP systems are conceived as pipelines wherein the output of one processor is in the input of the next. For instance, several tasks take POS tagged sequences or dependency trees as input. But this structured input is often converted to a vectorial form, which inevitably involves a loss of information.

Finally, graph-based representations and learning methods in principle appear to address some core problems faced by NLP, such as the fact that textual data are typically not independent and identically distributed, they often live on a manifold, they involve very high dimensionality, and their annotations is costly and scarce. As such, graph-based methods represent an interesting alternative, or at least complement, to structured prediction methods (such as CRFs or structured SVMs) commonly used within NLP. While structured output approaches are able to model local dependencies (e.g., between neighboring words or sentences), they cannot efficiently capture long distance dependencies, like forcing a particular n-gram to receive the same labeling in different sentences or documents for instance. On the other hand, graph-based models provide a natural way to capture global properties of the data through the exploitation of walks and neighborhood in graphs. Graph-based methods, like label propagation, have also been shown to be very effective in semi-supervised settings, and have already given some positive results on a few NLP tasks [9], [30].

Given the above motivations, our first line of research will be to investigate how one can leverage an underlying network structure (e.g., hyperlinks, user links) between documents, or text spans in general, to enhance prediction performances for several NL tasks. We think that a "network effect", similar to the one that took place in Information Retrieval (with the Page Rank algorithm), could also positively impact NLP research. A few recent papers have already opened the way, for instance in attempting to exploit Twitter follower graph to improve sentiment classification [29].

Part of the challenge in this work will be to investigate how adequately and efficiently one can model these problems as instances of more general graph-based problems, such as node clustering/classification or link prediction discussed in the next sections. In a few cases, like text classification or sentiment analysis, graph modeling appears to be straightforward: nodes correspond to texts (and potentially users), and edges are given by relationships like hyperlinks, co-authorship, frienship, or thread membership. Unfortunately, modeling NL problems as networks is not always that obvious. From the one hand, the right level of representation will probably vary depending on the task at hand: the nodes will be sentences, phrases, words, etc. From the other hand, the underlying graph will typically not be given a priori, which in turn raises the question of how we construct it. Of course, there are various well-known ways to obtain similarity measures between text contents (and its associated vectorial data), and graphs can be easily constructed from those combined with some sparsification method. But we would like our similarity to be tailored to the task objective. An additional problem with many NLP problems is that features typically live in different types of spaces (e.g., binary, discrete, continuous). A preliminary discussion of the issue of optimal graph construction for semi-supervised learning in NLP is given in [9], [33]. We identify the issue of adaptative graph construction as an important scientific challenge for machine learning on graphs in general, and we will discuss it further in Section 3.3.

As noted above, many NLP tasks have been recast as structure prediction problems, allowing to capture (some of the) output dependencies. Structure prediction can be viewed as (set of) link prediction with global loss or dependencies, which means that graph-based learning methods can handle (at least, approximately) output prediction dependencies, and they can in principle capture additional more global dependencies given the right graph structure. How to best combine structured output and graph-based ML approaches is another challenge that we intend to address. We will initially investigate this question within a semi-supervised context, concentrating on graph based regularization and graph propagation methods. Within such approaches, labels are typically binary or they correspond to small finite set. Our objective is to explore how one propagates an exponential number of *structured labels* (like a sequence of tags or a dependency tree) through graphs. Recent attempts at blending structured output models with graph-based models are investigated in [30], [17]. Another related question that we will address in this context is how does one learn with *partial labels* (like partially specificied tag sequence or tree) and use the graph structure to complete the output structure. This

last question is very relevant to NL problems where human annotations are costly; being able to learn from partial annotations could therefore allow for more targeted annotations and in turn reduced costs [18].

The NL tasks we will mostly focus on are coreference resolution and entity linking, temporal structure prediction, and discourse parsing. These tasks will be envisonned in both document and cross-document settings, although we expect to exploit inter-document links either way. Choices for these particular tasks is guided by the fact that are still open problems for the NLP community, they potentially have a high impact for industrial applications (like information retrieval, question answering, etc.), and we already have some expertise on these tasks in the team. As a midterm goal, we also plan to work on tasks more directly relating to micro-blogging, such sentiment analysis and the automatic thread structuring of technical forums; the latter task is in fact an instance of rhetorical structure prediction [32].

We have already initiated some work on the coreference resolution problem in the context of ML graph-based approaches. We cast this problem as a spectral clustering problem. Given than features can be numerical or nominal, the definition of a good similarity measure between entities is not straightforward. As a first solution, we consider only numerical attributes to build a k-nn graph of mentions so that graph clustering methods can be applied. Nominal attributes and relations are introduced by means of soft constraints on this clustering. Constraints can have various forms and have the ability of going beyond homophily assumptions, taking into account for instance dissimilarity relationships. From this setting we derive new graph-based learning methods. We propose to study the modification of graph clustering and spectral embeddings to satisfy certain constraints induced by several types of supervision: (i) nodes belong to the same group or to different groups, and (ii) some groups are fully known while others have to be discovered. This semi-supervised graph clustering problem is studied in a batch and transductive setting. But interesting extensions can be investigated in an online and active setting.

3.3. Adaptive Graph Construction

In most applications, edge weights are computed through a complex data-modeling process and convey crucially important information for classifying nodes, which makes it possible to infer information related to each data sample even exploiting the graph topology solely. In fact, a widespread approach to the solution of several classification problems is representing the data through an undirected weighted graph in which edge weights quantify the similarity between data points. This technique for coding input data has been applied to several domains, including classification of genomic data ([27]), face recognition ([16]), and text categorization ([21]).

In some cases, the full adjacency matrix is generated by employing suitable similarity functions chosen through a deep understanding of the problem structure. For example TF-IDF representation of documents, the affinity between pairs of samples is often estimated through the cosine measure or the χ^2 distance. After the generation of the full adjacency matrix, the second phase for obtaining the final graph consists in an edge sparsification/reweighting operation. Some of the edges of the clique obtained in the first step are pruned and the remaining ones can be reweighted to meet the specific requirements of the given classification problem. Constructing a graph with these methods obviously entails various kinds of loss of information. However, in problems like node classification, the use of graphs generated from several datasets can lead to an improvement in accuracy performance ([34], [10], [11]). Hence, the transformation of a dataset into a graph may, at least in some cases, partially remove various kinds of irregularities present in the original datasets, while keeping some of the most useful information for classifying the data samples. Moreover, it is often possible to accomplish classification tasks on the obtained graph using a running time remarkably lower than is needed by algorithms exploiting the initial datasets, and a suitable sparse graph representation can be seen as a compressed version of the original data. This holds even when input data are provided in a online/stream fashion, so that the resulting graph evolves over time.

In this project we will address the problem of adaptive graph construction towards several directions. One is the question of choosing the best similarity measure given the objective learning task. This question is related to the question of similarity learning ([12]) which has not been considered in the context of graph

based learning. In the context of structured prediction, we will develop approaches where output structures are organized in graphs whose similarity is given by top-k outcomes of greedy algorithms.

A different way we envision adaptative graph construction is in the context of semi-supervised learning. Partial supervision can take various forms and an interesting and original setting is governed by two currently studied applications: detection of brain anomaly from connectome data and polls recommendation in marketing. Indeed, for these two applications, a partial knowledge of the information diffusion process can be observed while the network is unknown or only partially known. An objective is to construct (or complete) the network structure from some local diffusion information. The problem can be formalized as a graph construction problem from partially observed diffusion processes. It has been studied very recently in [23]. In our case, the originality comes either from the existence of different sources of observations or from the large impact of node contents in the network.

We will study how to combine graphs defined by networked data and graphs built from flat data to solve a given task. This is of major importance for information networks because, as said above, we will have to deal with multiple relations between entities (texts, spans of texts, ...) and also use textual data and vectorial data. We have started to work on combining graphs in a semi supervised setting for node classification problems along the PhD thesis of T. Ricatte. Future work include combination geared by semi-supervision on link prediction tasks. This can be studied in an active learning setting. But one important issue is to design scalable approaches, thus to exploit locality given by the network. Doing this we address another objective to build non uniformly parameterized combinations.

3.4. Prediction on Graphs and Scalability

As stated in the previous sections, graphs as complex objects provides a rich representation of data. Often enough the data is only partially available and the graph representation is very helpful in predicting the unobserved elements. We are interested in problems where the complete structure of the graph needs to be recover and only a fraction of the links is observed. The link prediction problem falls into this category. We are also interested in the recommendation and link classification problems which can be seen as graphs where the structure is complete but some labels on the links (weights or signs) are missing. Finally we are also interested in labelling the nodes of the graph, with class or cluster memberships or with a real value, provided that we have (some information about) the labels for some of the nodes.

The semi-supervised framework will be also considered. A midterm research plan is to study how graph-based regularization models help for structured prediction problems. This question will be studied in the context of NLP tasks, as noted in Section 3.2, but we also plan to develop original machine learning algorithms that have a more general applicability. Inputs are networks whose nodes (texts) have to be labeled by structures. We assume that structures lie in some manifold and we want to study how labels can propagate in the network. One approach is to find smooth labeling function corresponding to an harmonic function on both manifolds in input and output. We also plan to extend our results on spectral clustering with must-link and cannot-link constraints in two directions. We have proposed a batch method with an optimization problem based on an adaptive spectral embedding with respects to constraints. We want to extend this approach to an on-line and active setting where a flow of graphs (each one is a document) is given as input. In the case of large graphs, we also consider the case where partial supervision consists in the knowledge of few clusters.

Scalability is one of the main issue in the design of new prediction algorithms working on networked data. It has gained more and more importance in recent years, because of the growing size of the most popular networked data that are now used by millions of people. In such contexts, learning algorithms whose computation time scales quadratically, or slower, in the number of considered data objects (usually nodes or vertices, depending on the given task) should be considered impractical.

These observations lead to the idea of using graph sparsification techniques in order to work on a part of the original network for getting results that can be easily extended and used for the whole original input. A sparsified version of the original graph can often be seen as a subset of the initial input, i.e. a suitably selected input subgraph which forms the training set (or, more in general, it is included in the training set). This holds even for the active setting.

A simple example could be to find a spanning tree of the input graph, possibly using randomization techniques, with properties such that we are allowed to obtain interesting results for the initial graph dataset. We have started to explore this research direction for instance in [31]. This approach leaves us with the problem of choosing a good spanning tree, taking into account that the setting could be adversarial (e.g, in the online case the presentation and the assignment of the labels are both arbitrary). A suitable use of the randomization power becomes therefore remarkably significant. Moreover, it is interesting to observe that running a prediction algorithm on a sparsified version of the input dataset allows the parallelization of prediction tasks. In fact, given a prediction task for a networked dataset, in a preliminary phase one could run a randomized graph sparsification method in parallel on different machines. For example, in the case of the spanning tree use, one could then draw several spanning trees at the same time, each on a different computer. This way it is possible to simultaneously run different prediction experiments on the same task and aggregating the obtained results at the end, with several methods (e.g. simply by majority vote) in order to increase the robustness and accuracy predictions.

At the level of the mathematical foundations, the key issue to be addressed in the study of (large-scale) random networks also concerns the segmentation of network data into sets of independent and identically distributed observations. If we identify the data sample with the whole network, as it has been done in previous approaches [22], we typically end up with a set of observations (such as nodes or edges) which are highly interdependent and hence overly violate the classic i.i.d. assumption. In this case, the data scale can be so large and the range of correlations can be so wide, that the cost of taking into account the whole data and their dependencies is typically prohibitive. On the contrary, if we focus instead on a set of subgraphs independently drawn from a (virtually infinite) target network, we come up with a set of independent and identically distributed observations—namely the subgraphs themselves, where subgraph sampling is the underlying ergodic process [13]. Such an approach is one principled direction for giving novel statistical foundations to random network modeling. At the same time, because one shifts the focus from the whole network to a set of subgraphs, complexity issues can be restricted to the number of subgraphs and their size. The latter quantities can be controlled much more easily than the overall network size and dependence relationships, thus allowing to tackle scalability challenges through a radically redesigned approach.

We intend to develop new learning models for link prediction problems. We have already proposed a conditional model in [20] with statistics based on Fiedler values computed on small subgraphs. We will investigate the use of such a conditional model for link prediction. We will also extend the conditional probabilistic models to the case of graphs with textual and vectorial data by defining joint conditional models. Indeed, an important challenge for information networks is to introduce node contents in link ranking and link prediction methods that usually rely solely on the graph structure. A first step in this direction was already proposed in [19] where we learn a mapping of node content to a new representation constrained by the existing link structure and applied it for link recommendation. This approach opens a different view on recommendation by means of link ranking problems for which we think that non parametric approaches should be fruitful.

Regarding link classification problems, we plan to devise a whole family of active learning strategies, which could be based on spanning trees or sparse input subgraphs, that exploit randomization and the structure of the graph in order to offset the adversarial label assignment. We expect these active strategies to exhibit good accuracies with a remarkably small number of queried edges, where passive learning methods typically break down. The theoretical findings can be supported by experiments run on both synthetic and real-world (Slashdot, Epinions, Wikipedia, and others) datasets.

We are interested in studying generative models for graph labeling, exploiting the results obtained in p-stochastic model for link classification (investigated in [15]) and statistical model for node label assignment which can be related to tree-structured Markov random fields [24].

In developing our algorithms, we focus on providing theoretical guarantees on prediction accuracy and, at the same time, on computational efficiency. The development of methods that simultaneously guarantee optimal accuracy and computational efficiency is a very challenging goal. In fact, the accuracy of most methods in the literature is not rigorously analyzed from a theoretical point of view. Likewise, tight time and space complexity

bounds are not generally provided. This contrasts with the need to manage extremely large relational datasets like, e.g., snapshots of the World Wide Web.

3.5. Beyond Homophilic Relationships

In many cases, the algorithms devised for solving node classification problems are driven by the following assumption: linked entities tend to be assigned to the same class. This assumption, in the context of social networks, is known as homophily ([14], [25]) and involves ties of every type, including friendship, work, marriage, age, gender, and so on. In social networks, homophily naturally implies that a set of individuals can be parted into subpopulations that are more cohesive. In fact, the presence of homogeneous groups sharing interests is one of the most significant reasons for affinity among interconnected individuals, which suggests that, in spite of its simplicity, this principle turns out to be very powerful for node classification problems in general networks.

Recently, however, researchers have started to consider networked data where connections may also carry a negative meaning. For instance, disapproval or distrust in social networks, negative endorsements on the Web. Concrete examples are provided by certain types of online social networks. Users of Slashdot can tag other users as friends or foes. Similarly, users of Epinions can give positive or negative ratings not only to products but also to other users. Even in the social network of Wikipedia administrators, votes cast by an admin in favor or against the promotion of another admin can be viewed as positive or negative links. More examples of signed links are found in other domains, such as the excitatory or inhibitory interactions between genes or gene products in biological networks.

Although the introduction of signs on graph edges appears like a small change from standard weighted graphs, the resulting mathematical object, called signed graph, has an unexpectedly rich additional complexity. For example, the spectral properties of signed graphs, which essentially all sophisticated node classification algorithms rely on, are different and less known than those of their unsigned counterparts. Signed graphs naturally lead to a specific inference problem that we have discussed in previous sections: link classification. This is the problem of predicting the sign of links in a given graph. In online social networks, this may be viewed as a form of sentiment analysis, since we would like to semantically categorize the relationship between individuals.

Another way to go beyond homophily between entities will be studied using our recent model of hypergraphs with bipartite hyperedges [4]. A bipartite hyperedge connects two ends which are disjoint subsets of nodes. Bipartite hyperedges is a way to relate two collections of (possibly heterogeneous) entities represented by nodes. In the NLP setting, while hyperedges can be used to model bags of words, bipartite hyperedges are associated with relationships between bags of words. But each end of bipartite hyperedges is also a way to represent complex entities, gathering several attribute values (nodes) into hyperedges viewed as records. Our hypergraph notion naturally extends directed and undirected weighted graph. We have defined a spectral theory for this new class of hypergraphs and opened a way to smooth labeling on sets of nodes. The weighting scheme permits to weight the participation of each node to the relationship modeled by bipartite hyperedges accordingly to an equilibrium condition. This is exactly that equilibrium condition that provides a competition between nodes in hyperedges and allows interesting modeling properties that go beyond homophily and similarity over nodes. (Theoretical analysis of our hypergraphs exhibits tight relationships with signed graphs). Following this competition idea, bipartite hyperedges are like matches between two teams and examples of applications are team creation. The basic tasks in which we are interested in are hyperedge classification, hyperedge prediction, node weight prediction. Finally, hypergraphs also represent a way to summarize or compress large graphs in which there exists highly connected couples of (large) subsets of nodes.

To conclude, we plan to go beyond the homophilic bias from the algorithmic as well as from the modeling point of view. We will consider new kind of modeling and learning biases provided by graphs with negative weights (signed graphs) and hypergraphs. We will study their spectral properties, smoothness measures of (node or edge) labeling. Sampling and walking also need to be reconsidered. From the machine learning perspective, we will study edge and node labeling in batch and online settings. In connection with our main targeted applications, we will mainly consider unsupervised and semi-supervised situations. We think that

allowing negative weights and advanced relationships on nodes will also lead to space efficient representations of graphs.

4. Application Domains

4.1. Overview

Our main targeted applications are browsing, monitoring and mining in information networks. Such discovered structures would also be beneficial to predicting links between users and texts which is at the core of recommender systems. All the learning tasks considered in the project such as node clustering, node and link classification and link prediction are likely to yield important improvements in these applications. Application domains cover social networks for cultural data and e-commerce, and biomedical informatics.

5. New Software and Platforms

5.1. CoRTex

Participants: Pascal Denis [correspondent], David Chatel.

CoRTex is a LGPL-licensed Python library for Noun Phrase coreference resolution in natural language texts. This library contains implementations of various state-of-the-art coreference resolution algorithms, including those developed in our research. In addition, it provides a set of APIs and utilities for text pre-processing, reading the main annotation formats (ACE, CoNLL and MUC), and performing evaluation based on the main evaluation metrics (MUC, B-CUBED, and CEAF). As such, CoRTex provides benchmarks for researchers working on coreference resolution, but it is also of interest for developers who want to integrate a coreference resolution within a larger platform. This project is hosted on Inria gforge: https://gforge.inria.fr/projects/cortex/.

6. New Results

6.1. Highlights of the Year

We developed a new framework for high order learning [4].

We have illustrated the usefulness of automatically annotated examples in complex learning supervised by few training examples [2], [1].

We propose a new algorithm for semi-supervised spectral clustering and apply it to the NLP task of noun phrase coreference resolution [6].

6.2. Higher-order Learning with Graphs

Along the thesis of THOMAS RICATTE, in [4] and [8], we propose methods for learning from interactions between groups in networks. We propose a proper extension of graphs, called hypernode graphs as a formal tool able to model group interactions. A hypernode graph is a collection of weighted relations between two groups of distinct nodes. Weights quantify the individual participation of nodes to a given relation. We define Laplacians and kernels for hypernode graphs and and prove that they strictly generalize over graph kernels and hypergraph kernels. We prove that hypernode graphs correspond to signed graphs such that the matrix $D - W$ is positive semidefinite. As a consequence, homophilic relations between groups may lead to non homophilic relations between individuals. We define the notion of connected hypernode graphs and a resistance distance for connected hypernode graphs. We propose spectral learning algorithms on hypernode graphs allowing to infer node ratings or node labelings. As a proof of concept, we model multiple players games with hypernode graphs and we define skill rating algorithms competitive with specialized algorithms.

6.3. Natural Language Processing

In [6] (presented by DAVID CHATEL at the ECML–PKDD and CAp'2014 conferences) we propose a new algorithm for semi-supervised spectral clustering and apply it to the task of noun phrase coreference resolution. The main insight is in the inclusion of pairwise constraints into spectral clustering: our algorithm learns a new representation space for the data together with a distance in this new space. The representation space is obtained through a constraint-driven linear transformation of a spectral embedding of the data, and constraints are expressed with a Gaussian function that locally reweights the similarities in the projected space. A global, non-convex optimization objective is then derived and the model is learned via gradient descent techniques. Our algorithm is evaluated on the CoNLL-2012 coreference resolution shared task dataset, and shows some encouraging results.

In [2] and [1], we develop a new approach for the automatic identification of so-called implicit discourse relations. Specifically, our system combines hand labeled examples and automatically annotated examples based on explicit relations using several simple methods inspired by work in domain adaptation. Our system is evaluated empirically on the Annodis corpus, a French corpus annotated with discourse structures. Our system yields significant performance gains compared to only using hand-labeled data or using only automatically annotated data.

6.4. Ongoing work

6.4.1. Adaptive Graph Construction

We worked on developing a new algorithm in order to construct a graph in a adaptive way for a specific task. More precisely, we looked for a metric learning algorithm that could depend on the target task. Previous works on metric learning ([12]) aim at learning a relevant metric using a linear approach, which cannot capture the non-linearity of the data. Our approach, instead, aims at learning the most appropriate non-linear data projection for the target task. For this purpose, we train a neural network with relative constraints depending on the target task and a target classic metric (e.g. euclidean distance, cosine similarity, ...), in order to make the metric meaningful for the new data representation and our target task.

6.4.2. Correlation Clustering and Similarity/Dissimilarity Links

From a mathematical point of view, signed networks are graphs whose edges carry a sign representing the positive or negative nature of the relationship between the incident nodes. These structures are extremely useful for modeling, at the same time, similarity and dissimilarity object relationships. Given an undirected signed graphs, in the Correlation Clustering problem the goal is to find a node partition into clusters minimizing the number of negative (dissimilarity) edges linking two nodes within the same cluster and the number of positive (similarity) edges between different clusters.

We focused on devising an algorithm able to solve the Correlation Clustering problem for general input signed graphs (if the input is a complete signed graph the problem is proven to be much easier). One of the main objective of this work is the use of the proposed algorithm for creating a learner able to predict the unknown edges signs of a given signed graph. This prediction task is known as Link Classification in signed graphs. In fact, given an undirected signed graph whose edge set is split into training and test set, we could use the Correlation Clustering solution working for general input graphs for partitioning the training set and using the node partition generated for predicting the test edge signs. Moreover, one could exploit such an algorithm for developing new strategies for the Link Classification problem operating within the online and active Machine Learning setting.

Since the node set partitioning turns out to be strictly related to the Link Classification problem, we also focused on the very challenging goal of obtaining a deep understanding of the complex interplay between Link and Node Classification. More precisely, we investigated the relationships between the Vapnik Chervonenkis dimension of any given set of hypothesis space of node and edge similarity functions operating within this framework.

6.4.3. Ranking from Pairwise Sets of User Preferences

Given a set of objects (vertices of a graph) and a set of pairwise preference labels between objects (directed edges connecting vertices) which may be non-transitive due to irrationality or arbitrary noise, what is a correct way to sample preference labels for ordering the set of objects? This long standing open problem is, as far as we know, unsolved when each pairwise preference labels refers to two (disjoint) sets of objects (vertices). This framework can be easily motivated considering that quite often, in many real world contexts, users express their preferences between sets of items rather than single items, and turns out to be strictly connected with our recent model of hypergraphs with bipartite hyperedges [4]. We are working on devising a new algorithm able to rank a given set of items (graph node set) when only comparisons between sets containing at least 2 items are allowed. This challenging and interested problem is, as far as we know, quite novel and can be studied within different Machine Learning setting (online, batch, active, ...). The preliminaries results we are obtaining, when setting the cardinality of the item sets equal to 2, are encouraging and indicate that it could be possible to extend our strategies in order to deal with larger item sets.

6.5. Other results

In this section we provide the results we obtained that are not related with our main research directions.

In [3] we study the problem of learning sequential top-down tree-to-word transducers (STWs). First, we present a Myhill-Nerode characterization of the corresponding class of sequential tree-to-word transformations (STW). Next, we investigate what learning of stws means, identify fundamental obstacles, and propose a learning model with abstain. Finally, we present a polynomial learning algorithm.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Grants with Industry

7.1.1. Cifre SAP (2011-2014)

Participants: Thomas Ricatte, Marc Tommasi, Rémi Gilleron [correspondent].

In business intelligence information systems, one of the first tasks is to acquire and clean internal data and then enrich them with additional sources of informations. This preprocessing step is well planned and specialized for fixed analysis and fixed dashboards. The subject of our collaboration with SAP was included in general objective that (i) specializes this preprocessing task in order to deal with external data coming from networked data like social networks and open relational data, and (ii) simplifies the adaptation of the processing step evolving data analysis tasks. We have focused on the task of merging information acquired from many input data sources represented as graphs, with the final objective of providing a unique graph representation of all data or data models. This research has lead to new graph combination algorithms, but has also raised the need for representing and managing high order relations using graph-like techniques.

RÉMI GILLERON supervises the PhD thesis (Cifre) of Thomas Ricatte together with Yannick Cras from SAP.

7.1.2. Cifre Clic and Walk (2013-2016)

Participants: Pauline Wauquier, Marc Tommasi, Mikaela Keller [correspondent].

We start a one to one cooperation with the CLIC AND WALK company that makes marketing surveys by consumers (called clicwalkers). The goal of the company is to understand the community of clicwalkers (40 thousands in one year) and its evolution with two objectives: the first one is to optimize the attribution of surveys to clicwalkers, and the second is to expand company's market to foreign countries. Social data can be obtained from social networks (G+, Facebook, ...) but there is no explicit network to describe the clicwalkers community. But users activity in answering surveys as well as server logs can provide traces of information diffusion, geolocalisation data, temporal data, sponsorship, ...We will study the problem of adaptive graph construction from the clicwalkers network. Node (users) classification and clustering algorithms will be applied. For the problem of survey recommendations, the problem of teams constitution in a bipartite graphs of users and surveys will be studied. Random graph modeling and generative models of random graphs will be one step towards the prediction of the evolution of clicwalkers community.

MIKAELA KELLER and MARC TOMMASI supervise the PhD thesis (Cifre) of PAULINE WAUQUIER on graph-based recommendation together with Guillaume André from CLIC AND WALK.

8. Partnerships and Cooperations

8.1. Regional Initiatives

MARC TOMMASI and PASCAL DENIS supervise the PhD thesis of DAVID CHATEL on semi-supervised spectral clustering. The PhD is funded by Inria and the "Région Nord – Pas de Calais".

MARC TOMMASI belongs to the scientific committee involved in the process of building the IDEX proposal for Lille.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. ANR Lampada (2009-2014)

Participants: Marc Tommasi [correspondent], Rémi Gilleron, Fabien Torre.

The Lampada project on “Learning Algorithms, Models and sPArse representations for structured DAta” is coordinated by Tommasi from Mostrare. Our partners are the SEQUEL project of Inria Lille Nord Europe, the LIF (Marseille), the HUBERT CURIEN laboratory (Saint-Etienne), and LIP6 (Paris). More information on the project can be found on http://lampada.gforge.inria.fr/.

8.2.2. Competitivity Clusters

We are part of FUI HERMES (2012-2015), a joint project in collaboration with many companies (Auchan, KeyneSoft, Cylande, ...). The main objective is to develop a platform for contextual customer relation management. The project started in November 2012.

8.2.3. EFL

PASCAL DENIS is an associate member of the Laboratoire d’Excellence *Empirical Foundations of Linguistics* (EFL), http://www.labex-efl.org/.

8.2.4. Conseil national des universités

FABIEN TORRE is elected for "CNU section 27 (informatique)" since Oct. 2011

8.3. European Initiatives

8.3.1. Collaborations in European Programs, except FP7 & H2020

Program: ERC Advanced Grant
Project acronym: STAC
Project title: Strategic conversation
Duration: Sept. 2011 - Aug. 2016
Coordinator: Nicholas Asher, CNRS, Université Paul Sabatier, IRIT (France)
Other partners: School of Informatics, Edinburgh University; Heriot Watt University, Edinburgh
Abstract: STAC is a five year interdisciplinary project that aims to develop a new, formal and robust model of conversation, drawing from ideas in linguistics, philosophy, computer science and economics. The project brings a state of the art, linguistic theory of discourse interpretation together with a sophisticated view of agent interaction and strategic decision making, taking advantage of work on game theory.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

We invited Prof. Claudio Gentile (University of Insubria, Italy) in January (he gave a talk on "Online Clustering of Bandits in a Social Network") and in June.

Prof. Nicolò Cesa-Bianchi (University of Milan, Italy) visited us in July (he gave a talk on "The Online Approach to Machine Learning").

Finally, we invited Prof. Mark Herbster (University College London) in July (he gave a talk on "Online Approximate Prediction at the Limit of Zero Temperature in an Ising Model") and November.

8.4.2. Visits to International Teams

In May FABIO VITALE visited the Department of Computer Science of the University of Milan, collaborating with Prof. Nicolò Cesa-Bianchi.

9. Dissemination

9.1. Promoting Scientific Activities

PASCAL DENIS served as member of the program committee of ACL 2014, COLING 2014, EACL 2014, and EMNLP 2014.

RÉMI GILLERON served as member of the program committee of CAp'2014.

MARC TOMMASI served as member of the program committee of CAp'2014.

FABIEN TORRE served as member of the program committee of ICPRAM 2015, EGC 2015, CluCo 2015, ECAI 2014, ICPRAM 2014.

9.1.1. Reviewer

MIKAELA KELLER served as member of the program committee of NIPS 2014 and CAp'2014, of the scientific committee reviewing Digiteo post-doctoral grant proposal.

MIKAELA KELLER was also reviewer for EGC 2015.

FABIO VITALE was reviewer for NIPS 2014.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Master MOCAD: PASCAL DENIS, Extraction d'information, 18h, M2, Université Lille 1, France

Master MOCAD: DAVID CHATEL, Extraction d'information, 3h30, M2, Université Lille 1, France

Master MIASHS: RÉMI GILLERON, Bases de données avancées, 24h, M1, Université de Lille, France

Master MIASHS: RÉMI GILLERON, Programmation R (1), 24h, M1, Université de Lille, France

Master MIASHS: RÉMI GILLERON, Programmation R (2), 24h, M1, Université de Lille, France

Master MIASHS: RÉMI GILLERON, Algorithmes de fouille de données, 36h, M1, Université de Lille, France

Licence MIASHS: RÉMI GILLERON, Programmation Python, 36h, L1, Université de Lille, France

Licence : MIKAELA KELLER, Introduction to computer science and programming, 46h, L1 and L2, Université Lille 3, France

Licence : MIKAELA KELLER, Introduction to Statistical Learning, 54h, L3, Université Lille 3, France

Licence : MIKAELA KELLER, C2i, 55h, L2-L3, Université Lille 3, France

Licence MIASHS: MARC TOMMASI, Réseaux, 64h, L1 and L2, Université de Lille, France

Licence MIASHS: MARC TOMMASI, Programmation client, 24h, L2, Université de Lille, France

Licence : MARC TOMMASI, C2i, 25h, L2, Université de Lille, France

Licence : MARC TOMMASI, Culture numérique, 30h, L2, Université de Lille, France

Master : FABIEN TORRE, langages statiques du web, 37h, M1, Université Lille 3, France

Master : FABIEN TORRE, algorithmique et programmation php pour le web, 75h, M1, Université Lille 3, France

Master : FABIEN TORRE, algorithmique et programmation pour l'extraction d'information, 55h, M2, Université Lille 3, France

Master : FABIEN TORRE, javascript langage dynamique du web, 38h, M2, Université Lille 3, France

Licence : FABIO VITALE, C2i, 25h, L1, Université Lille 3, France

Licence MIASHS (SHS) : FABIO VITALE, Intelligence Artificielle, 31h30, L2, Université Lille 3, France

Master MIASHS (SHS) : FABIO VITALE, Classification non supervisée, 30h, M1, Université Lille 3, France

Licence SHDS (SHS) : FABIO VITALE, TICE, 18h, L1, Université Lille 3, France

Licence SQ (SHS) : FABIO VITALE, Algorithmique de graphes, 19h, L3, Université Lille 3, France

FABIEN TORRE teaches a course at EPAT 2014 (École de Printemps sur l'Apprentissage arTificiel) on Machine Learning Collaborative Methodes.

RÉMI GILLERON is in charge of Master Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales (MIASHS) at Université de Lille.

MARC TOMMASI is a member of the "conseil de l'UFR MIME".

FABIEN TORRE is "secrétaire du CHSCT (Comité hygiène, sécurité et conditions de travail) de Lille 3" since Dec. 2012.

9.2.2. Supervision

PhD: THOMAS RICATTE, Hypernode graphs : a new data structure for higher order learning, Université Lille 3, since May 2011, defense scheduled for Jan. 2015, RÉMI GILLERON

PhD in progress: DAVID CHATEL, Supervised Spectral Clustering and Information Diffusion in Graphs of Texts, Université Lille 1, since Sept. 2012, PASCAL DENIS and MARC TOMMASI

PhD in progress: PAULINE WAUQUIER, Recommendation in Information Networks, Université Lille 1, since Dec. 2013 supervised by MARC TOMMASI and MIKAELA KELLER

PhD: Jean Decoster, Programmation logique inductive pour la classification et la transformation de documents semi-structurés, Université Lille 1, since Sept. 2009, defense on July 2014, FABIEN TORRE, RÉMI GILLERON and SLAWOMIR STAWORKO

PhD in progress: EMMANUEL LASSALLE, Improved Coreference Resolution with Feature Space Learning, Université Paris-Diderot, since Sept. 2010, PASCAL DENIS and LAURENCE DANLOS (Université Paris-Diderot)

PhD in progress: CHLOÉ BRAUD, Discourse Relation Identification from Labeled and Unlabeled Data, Université Paris-Diderot, since Sept. 2011, PASCAL DENIS and LAURENCE DANLOS (Université Paris-Diderot)

Master in progress: MATHIEU DEHOUCK, Propagating structured labels for Dependency Parsing, Université Lille 1, supervised by PASCAL DENIS

PhD in progress: GÉRAUD LE FALHER, Machine Learning in Signed Graphs, Inria Lille – Nord Europe, since Oct. 2014, MARC TOMMASI, FABIO VITALE and CLAUDIO GENTILE (University of Insubria, Italy).

MARC TOMMASI wrote a report for the Phd thesis of Daniel Bernades ("LIP6", France).

9.2.3. Juries

MIKAELA KELLER was a member of the following selection/hiring committees: PhD committee of J.P. Peyrache, Université de St Etienne (France), MdC hiring committees at Université Paris-Sud, and MdC hiring committees at Université Paris 6 (France).

MARC TOMMASI was member of the following selection/hiring committees: UPMC (Paris 6): Professor, St Etienne (France): Assistant professor and Lille (France): Inria young researchers (CR2).

MARC TOMMASI was also reviewer for Crédit Impot Recherche.

FABIEN TORRE participated in the selection committees "COS" for the 27 MCF 0133 and 27 MCF 0470 position (spring 2014, Lens, France).

9.3. Popularization

MARC TOMMASI was invited to give seminars at (i) Inria Rennes and (ii) LiX Colloquium on data science.

MIKAELA KELLER was involved in "Fête de la science" as a "chercheur itinérant".

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] C. BRAUD, P. DENIS. *Identifier les relations discursives implicites en combinant données naturelles et données artificielles*, in "Traitement Automatique des Langues", December 2014, vol. 55, n^{o} 1, 31 p. , https://hal.inria.fr/hal-01094346

International Conferences with Proceedings

[2] C. BRAUD, P. DENIS. *Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification*, in "coling", Dublin, Ireland, August 2014, https://hal.inria.fr/hal-01017151

[3] G. LAURENCE, A. LEMAY, J. NIEHREN, S. STAWORKO, M. TOMMASI. *Learning Sequential Tree-to-Word Transducers*, in "8th International Conference on Language and Automata Theory and Applications", Madrid, Spain, Springer, March 2014, https://hal.inria.fr/hal-00912969

[4] T. RICATTE, R. GILLERON, M. TOMMASI. *Hypernode Graphs for Spectral Learning on Binary Relations over Sets*, in "ECML/PKDD - 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Nancy, France, Machine Learning and Knowledge Discovery in Databases, September 2014, Paper accepted for publication at ECML/PKDD 2014, https://hal.inria.fr/hal-01017025

Conferences without Proceedings

[5] D. CHATEL, P. DENIS, M. TOMMASI. *Clustering Spectral avec Contraintes de Paires réglées par Noyaux Gaussiens*, in "CAP 2014", Saint-Etienne, France, July 2014, https://hal.inria.fr/hal-01105339

[6] D. CHATEL, P. DENIS, M. TOMMASI. *Fast Gaussian Pairwise Constrained Spectral Clustering*, in "ECML/PKDD - 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases", Nancy, France, September 2014, pp. 242 - 257 [*DOI :* 10.1007/978-3-662-44848-9_16], https://hal.inria.fr/hal-01017269

[7] T. RICATTE, R. GILLERON, M. TOMMASI. *Hypernode Graphs for Spectral Learning on Binary Relations over Sets*, in "Conférence Francophone sur l'Apprentissage Automatique (Cap 2014)", Saint-Etienne, France, July 2014 [*DOI :* 10.1007/978-3-662-44851-9_42], https://hal.inria.fr/hal-01104618

Other Publications

[8] T. RICATTE, R. GILLERON, M. TOMMASI. *Hypernode Graphs for Learning from Binary Relations between Groups in Networks*, December 2014, Networks: From Graphs to Rich Data, NIPS Workshop, https://hal.inria.fr/hal-01088036

References in notes

[9] A. ALEXANDRESCU, K. KIRCHHOFF. *Graph-based learning for phonetic classification*, in "IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007", pp. 359-364

[10] M.-F. BALCAN, A. BLUM, P. P. CHOI, J. LAFFERTY, B. PANTANO, M. R. RWEBANGIRA, X. ZHU. *Person Identification in Webcam Images: An Application of Semi-Supervised Learning*, in "ICML2005 Workshop on Learning with Partially Classified Training Data", 2005

[11] M. BELKIN, P. NIYOGI. *Towards a Theoretical Foundation for Laplacian-Based Manifold Methods*, in "Journal of Computer and System Sciences", 2008, vol. 74, n^{o} 8, pp. 1289-1308

[12] A. BELLET, A. HABRARD, M. SEBBAN. *A Survey on Metric Learning for Feature Vectors and Structured Data*, in "CoRR", 2013, vol. abs/1306.6709

[13] P. J. BICKEL, A. CHEN. *A nonparametric view of network models and Newman–Girvan and other modularities*, in "Proceedings of the National Academy of Sciences", 2009, vol. 106, pp. 21068–21073

[14] P. BLAU. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*, MACMILLAN Company, 1977, http://books.google.fr/books?id=jvq2AAAAIAAJ

[15] N. CESA-BIANCHI, C. GENTILE, F. VITALE, G. ZAPPELLA. *A Linear Time Active Learning Algorithm for Link Classification*, in "Proc of NIPS", 2012, pp. 1619-1627

[16] H. CHANG, D.-Y. YEUNG. *Graph Laplacian Kernels for Object Classification from a Single Example*, in "Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2", Washington, DC, USA, CVPR '06, IEEE Computer Society, 2006, pp. 2011–2016, http://dx.doi.org/10.1109/CVPR.2006.128

[17] D. DAS, S. PETROV. *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*, in "ACL", 2011, pp. 600-609

[18] E. R. FERNANDES, U. BREFELD. *Learning from Partially Annotated Sequences*, in "ECML/PKDD", 2011, pp. 407-422

[19] A. FRENO, G. C. GARRIGA, M. KELLER. *Learning to Recommend Links Using Graph Structure and Node Content*, in "NIPS Workshop on Choice Models and Preference Learning", 2011

[20] A. FRENO, M. KELLER, G. C. GARRIGA, M. TOMMASI. *Spectral Estimation of Conditional Random Graph Models for Large-Scale Network data*, in "Proc. of UAI 2012", Avalon, États-Unis, 2012, http://hal.inria.fr/hal-00714446

[21] A. B. GOLDBERG, X. ZHU. *Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization*, in "Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing", Stroudsburg, PA, USA, TextGraphs-1, Association for Computational Linguistics, 2006, pp. 45–52, http://dl.acm.org/citation.cfm?id=1654758.1654769

[22] A. GOLDENBERG, A. X. ZHENG, S. E. FIENBERG. *A Survey of Statistical Network Models*, Foundations and trends in machine learning, Now Publishers, 2010, http://books.google.fr/books?id=gPGgcOf95moC

[23] M. GOMEZ-RODRIGUEZ, J. LESKOVEC, A. KRAUSE. *Inferring networks of diffusion and influence*, in "Proc. of KDD", 2010, pp. 1019-1028

[24] M. HERBSTER, S. PASTERIS, F. VITALE. *Online Sum-Product Computation Over Trees*, in "Proc. of NIPS", 2012, pp. 2879-2887

[25] M. MCPHERSON, L. S. LOVIN, J. M. COOK. *Birds of a Feather: Homophily in Social Networks*, in "Annual Review of Sociology", 2001, vol. 27, n^{o} 1, pp. 415–444, http://dx.doi.org/10.1146/annurev.soc.27.1.415

[26] A. NENKOVA, K. MCKEOWN. *A Survey of Text Summarization Techniques*, in "Mining Text Data", Springer, 2012, pp. 43-76

[27] H. SHIN, K. TSUDA, B. SCHÖLKOPF. *Protein functional class prediction with a combined graph*, in "Expert Syst. Appl.", March 2009, vol. 36, n^{o} 2, pp. 3284–3292, http://dx.doi.org/10.1016/j.eswa.2008.01.006

[28] S. SINGH, A. SUBRAMANYA, F. C. N. PEREIRA, A. MCCALLUM. *Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models*, in "ACL", 2011, pp. 793-803

[29] M. SPERIOSU, N. SUDAN, S. UPADHYAY, J. BALDRIDGE. *Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph*, in "Proceedings of the First Workshop on Unsupervised Methods in NLP", Edinburgh, Scotland, 2011

[30] A. SUBRAMANYA, S. PETROV, F. C. N. PEREIRA. *Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models*, in "EMNLP", 2010, pp. 167-176

[31] F. VITALE, N. CESA-BIANCHI, C. GENTILE, G. ZAPPELLA. *See the Tree Through the Lines: The Shazoo Algorithm*, in "Proc of NIPS", 2011, pp. 1584-1592

[32] L. WANG, S. N. KIM, T. BALDWIN. *The Utility of Discourse Structure in Identifying Resolved Threads in Technical User Forums*, in "COLING", 2012, pp. 2739-2756

[33] K. K. YUZONG LIU. *Graph-Based Semi-Supervised Learning for Phone and Segment Classification*, in "Proceedings of Interspeech", Lyon, France, 2013

[34] X. ZHU, Z. GHAHRAMANI, J. LAFFERTY. *Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions*, in "Proc. of ICML", 2003, pp. 912-919