Graffiti: Node Labeling in Heterogeneous Networks

Ralitsa Angelova Gjergji Kasneci Fabian M. Suchanek Gerhard Weikum
Max-Planck Institute for Informatics, Saarbruecken, Germany
angelova, kasneci, suchanek, weikum@mpi-inf.mpg.de

Abstract

We introduce a multi-label classification model and algorithm for labeling heterogeneous networks, where nodes belong to different types and different types have different sets of classification labels. We present a graph-based approach which models the mutual influence between nodes in the network as a random walk. When viewing class labels as "colors", the random surfer is "spraying" different node types with different color palettes; hence the name Graffiti. We demonstrate the performance gains of our method by comparing it to three state-of-the-art techniques for graph-based classification.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Classifier design and evaluation; G.3 [**Probability and statistics**]: Markov processes

General Terms

Graph Classification, Link Analysis

1. INTRODUCTION

Heterogeneous graphs are developing faster than ever. In the context of the rapidly growing social networks (e.g. *Flickr, Deli.cio.us, LibraryThing, LinkedIn*), heterogeneous graphs are formed by encoding users, their postings like photos, bookmarks, book descriptions, ratings, etc., and other contextual information as graph nodes, belonging to different node types but co-existing and mutually influencing each other in the graph [3, 4, 5, 7, 11, 12]. Heterogeneous graphs are formed in many other areas like medical domains (containing information about patients, treatments, diseases, contacts) or e-commerce platforms (representing the complex interactions between different types of nodes like users, products, customers, rating, reviews, etc.).

In this paper we present a novel graph-based algorithm that aims at classifying each node as belonging to one or more labels from a finite set of *type-specific* labels. Let us use a toy example to illustrate the goal of the algorithm. If you consider the social network *flickr*, a user is more likely to belong to the class *Outdoor activity fans* if she has frequently used tags belonging to the class *Mountain* and/or she has viewed, commented and posted photos classified as *Nature Photography*. In this example, the graph $G = (V, E)$ contains nodes from set $V = \{V_1 \cup V_k \ldots \cup V_t\}$ that belong to one of three types V_k (here $t = 3$): user, photo or tag. The relations between them are represented as edges E which either connect two nodes of the same type or connect two nodes belonging to different types. To ease referring to the edges, we will use the abbreviation *S-edges* for the same-type edges and *X-edges* for the cross-type edges. An example of an S-link formation would be a friendship connection between two users. Note that heterogeneous graphs contain an overwhelming number of X-links and few S-links. Therefore, graph-based classification methods, designed to derive class labels based on direct links between the graph nodes will degrade in their performance, because most direct links connect different types of nodes with type-specific classes. The model we propose is able to leverage the prominent influence of X-links, as a powerful tool to facilitate class inference across different type nodes.

Let us think of class labels as colors. We model the mutual influence of nodes as a random walk, in which an intelligent surfer carries different color spray cans, and while walking in the graph is coloring all its nodes by spraying different node types with different color palettes. The color she would choose and the types of links she would follow to reach other nodes is described in a mathematically sound framework outlined below.

2. FRAMEWORK AND ALGORITHM

Graffiti models the influence among nodes of the same type by looking at their common neighbors, which typically belong to other types. We base the strength of the mutual influence between nodes $v, v' \in V_k$ on two criteria:

- The bigger the overlap of shared neighbors, the higher the influence of v and v' on each other.
- The bigger the mutual influence among shared neighbors, the higher the influence of v and v' on each other.

These criteria are incorporated in the random walk the Graffiti surfer takes by giving him the possibility at each node to either perform a two-hop walk, spreading colors to same-type nodes via heterogeneous common neighbors (thus following two X-edges), perform a one-hop walk to any node of v's heterogeneous neighborhood (following a X- or S- edge), or choose at random any node from G (perform a random jump).

The model associates each node with two vectors: a *prior class-label probability distribution* of v, denoted by $\lambda(v) = (\lambda_1(v), \ldots, \lambda_n(v))^T$, smoothed over the domain of all possible classes $\mathcal{C}$; and a vector $\mu(v)$ that captures the *class-specific (graph-topological) importance* of a node. The latter

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

measure represents a *topic-biased* authority, where $\mu_i(v)$ describes the importance of v for class c_i.

When performing a two-hop step from v to a node $v' \in$ type V_k, the surfer reaches into her pocket and decides to use the same color c_i she used at v, in order to paint v', with a probability that is proportional to the authority of $\mu_i(v)$. When the surfer decides to use another color c_j for v', she chooses that color with probability proportional to $\mu_i(v)$, dampened by $\lambda_j(v)$ (thus changing the color from c_i to c_j). That is, the prior for color c_j of node v has a certain stickiness. The random surfer spreads equal portions of the selected color on all nodes of V_k that share V_l neighbors with v.

Similarly, when the surfer decides to perform a one-hop walk from v to a neighboring node $v' \in V_k \cup V_l$, she can use any color c_j with a probability proportional to $\mu(v)$, dampened by $\lambda_j(v)$.

Finally, when the surfer decides to randomly jump to any node u in the graph, she chooses a color c_i to paint u with probability proportional to $\lambda_i(u)/|V|$.

This random walk model includes the following components:

- the probability that the random surfer performs a random jump J,
- the probability that she follows an X-link path from node v to a node v' of the same type, spraying the same color as in the starting node v - $F^2_{same\ color}$,
- the probability that she follows an X-link path from node v to a node v' of the same type, spraying another color $F^2_{change\ color}$,
- and the probability of following a link to any node of the heterogeneous neighborhood F^1.

Finally, the importance of a node $v \in V_k$ with respect to class c_i is given by

$$\mu_i(v) = P[v, c_i, J] + \sum_{v' \in V_k} P[v', v, c_i, c_i, F^2_{same\ color}]$$

$$+ \sum_{v' \in V_k} \sum_j P[v', v, c_j, c_i, F^2_{change\ color}] + \sum_j P[u, v, c_j, c_i, F^1].$$

We prove that Graffiti's random walk corresponds to a stochastic process and has desired properties like convergence and a unique solution. Details are omitted due to shortage of space. We implemented and tested the framework using real-world data from the social network *flickr*.

3. EXPERIMENTAL VALIDATION

We crawled a subgraph from the *flickr* network where people share personal photos and attach tags to photos. Our crawl contains approximately 231,000 nodes, roughly 4,000 users, 123,000 photos and 104,000 tags. There are close to 3,000,000 links, out of which only 23,000 are same-type edges. We use the same set of class labels for all three node types. Note that even if two nodes from different types have the same label, they are conceptually different, since the labels are type-specific (i.e. User.c_i is different from Photo.c_i). We use the following labels: 1) Animals, 2) Birds, 3) Architecture, 4) Portrait, and 5) Nature. These are explicit category names in *flickr* itself.

We compare the performance of different classifiers based on the **precision/recall break-even point** (hence, also equal to the micro-averaged $F1$ **measure**) [10] , which is a typical measure for the effectiveness of a multi-label classifier. As we produce a vector of predictions, we would also like to test to what extent the order of class probabilities in the ground truth is reflected in the prediction vector. Therefore, we compared the rankings provided by the ground truth vector and the prediction vector by using the **normalized discounted cumulative gain** (NDCG).We also compared both vectors in terms of ranking the class probabilities using the well-known **Kendall's tau** measure.

We compare the performance of five classifiers: a text only classifier, in our case a Naive Bayesian (NB) classifier, which serves as initializer of the topical distribution per node in the graph; a hybrid classifier (HC), that uses the links in a node's neighborhood to enhance its node representation and later on applies a weighted voting scheme to decide on the most probable class of a node [9]; the proposed method Graffiti that takes as initial class distribution the predictions returned by the NB text only classifier (the baseline); a topical PageRank approach [8] (TPR); and an iterative Relaxation Labeling (RL) approach from the family of graphical models, that builds on, but extends and generalizes earlier work [1, 2].

Table 1 shows results for an experiment with a training set of 1,000 labeled nodes and a test set of 230,000 unlabeled nodes. Methods requiring initialization use the same baseline. All gains in Table 1 are statistically significant with a t-test level of 0.05.

System	microF1	NDCG	K.Tau
NB	0.5195	0.7797	0.3060
Graffiti	**0.5407**	**0.7930**	**0.3278**
TPR	0.5293	0.7863	0.3160
RL	0.2830	0.6767	0.2051
HC	0.4781	0.7579	0.2685

Table 1: Performance comparison

Overall, these and other experiments (omitted here for space restrictions) provide clear evidence for the superiority of the proposed method Graffiti.

4. REFERENCES

[1] Angelova,R., Weikum,G. Graph-based Text Classification: Learn from Your Neighbors. SIGIR, 2006.
[2] Chakrabarti,S., Dom,B.E., Indyk,P. Enhanced hypertext categorization using hyperlinks. SIGMOD, 1998.
[3] Getoor, L. Link mining: a new data mining challenge. ACM SIGKDD, 2003
[4] Getoor,L. and Taskar, B. Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning). The MIT Press, 2007.
[5] Getoor,L., Diehl,C.P. Link mining: a survey. KDD, 2005.
[6] Jarvelin, K., Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. ACM SIGIR, 2000.
[7] Macskassy, S. A., Provost, F. Classification in Networked Data: A toolkit and a univariate case study. Journal of Machine Learning, 8(May):935-983, 2007.
[8] Nie,L., Davison,B.D., Qi,X. Topical link analysis for web search. SIGIR, 2006.
[9] Oh,H.-J., Myaeng,S.H., Lee,M. A practical hypertext catergorization method using links and incrementally available class information. SIGIR, 2000.
[10] Thorsten, Joachims. Transductive Inference for Text Classification using Support Vector Machines. ICML, 1999.
[11] Wang,F., Zhang,C. Label Propagation Through Linear Neighborhoods. ICML'06.
[12] Wu T.-F., Lin, C.-J., Weng,R.C. Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of Machine Learning Research, 2004.