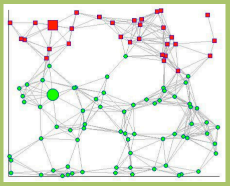
Adaptation of Graph-Based Semi-Supervised Methods to Large-Scale Text Data

Frank Lin and William W. Cohen

Language Technologies Institute & Machine Learning Department, School of Computer Science, Carnegie Mellon University

We like graph-based SSL…



They are efficient, effective, and fun…

The Problem with text (and also other kinds of) data:

The importance of a Web page is an inherently subjective matter, which depends on the readers…

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use…

You're not cool just because you have a lot of followers on twitter, get over yourself…

cool	web	search	make	over	you
0	4	8	2	5	3
0	8	7	4	3	2
1	0	0	0	1	2

Mostly non-zero - any two documents are likely to have a word in common – dense!

🟧	🟥	🟩	🟪	
123	27	125	-	🟪
56	23	-	125	🟩
77	-	23	27	🟥
-	77	56	123	🟧

O(n^2) time to construct

O(n^2) space to store

> O(n^2) time to operate on

Harmonic Functions (HF)
and a family of related methods:

- ❖ Gaussian fields and harmonic functions classifier (Zhu et al. 2003)
- ❖ Weighted-voted relational network classifier (Macskassy & Provost 2007)
- ❖ Weakly-supervised classification via random walks (Talukdar et al. 2008)
- ❖ Adsorption (Baluja et al. 2008)
- ❖ Learning on diffusion maps (Lafon & Lee 2006)
- ❖ and others …

But with the right SSL method and similarity function, GSSL can be done efficiently and exactly using **Implicit Manifolds!**

Just pick your SSL method:

Harmonic Functions (HF)

MultiRankWalk (MRW)

… and a similarity function:

Inner Product

Cosine Similarity

Bipartite Graph Walk

Implicit Manifolds can be applied whenever the algorithm + the similarity function can be decomposed into ***sparse matrix-vector multiplications***.

Simple examples:

HF: $V^{t+1} \leftarrow D^{-1}FF^TV^t$

MRW: $V^{t+1} \leftarrow (1-\alpha)FF^TD^{-1}V^t + \alpha R$

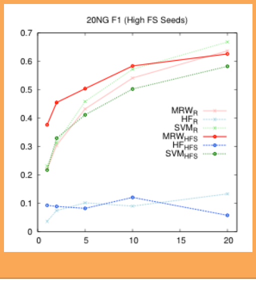
MultiRankWalk (MRW)
and a family of related methods:

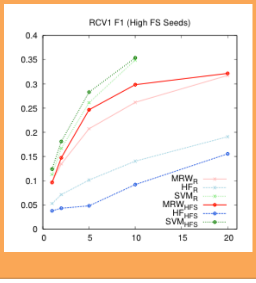
- ❖ Partially labeled classification using Markov random walks (Szummer & Jaakkola 2001)
- ❖ Learning with local and global consistency (Zhou et al. 2004)
- ❖ Graph-based SSL as a generative model (He et al. 2007)
- ❖ Ghost edges for classification (Gallagher et al. 2008)
- ❖ and others …

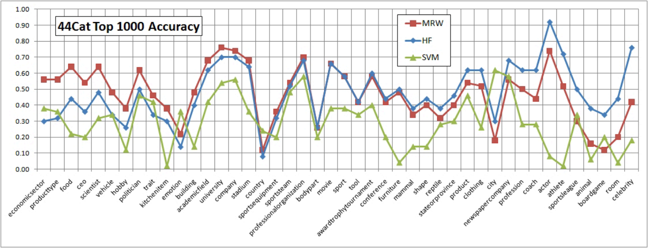
Okay, so what?

- ❖ A principled framework under which we can apply GSSL efficiently text (and other kinds of non-network) data
- ❖ A set of tools (2 general propagation GSSL methods + 3 similarity functions)
- ❖ The "ad-hoc" propagation method that worked well for you can now be connected to a greater body of work
- ❖ We know when we don't need to sparsify the matrix and still get the same results!

Hmm… they seem efficient, but how well do they work?







Much thanks to NSF, NIH, and Google!