论文阅读报告

一、论文基本情况

论文名	DeepWalk: Online Learning of Social Representations				
作者	Bryan Perozzi Rami Al-Rfou Steven Skiena				
作者机构	Stony Brook University Department of Computer Science				
发表年份	2014				
来源（刊物名、会议名）	KDD				
论文可信力	期刊	EI/SCI		影响因子	
		Jcr 分区		CCF 分类	
	会议	CCF 分类	A	领域影响力	顶级数据挖掘学术会议
	引用次数	469			
阅读人	房慧文	阅读日期		2019.01.10	

二、论文主要工作

1. 论文概述

针对问题：社交网络中节点之间的连接比较稀疏，是比较典型的信息比较少的网络。使用机器学习的算法解决问题需要有大量的信息，为了将机器学习广泛应用到网络中，必须要先对信息比较少的网络进行处理。

解决办法：提出了新的、无监督的、独立于标签分布的（捕获结构信息时不考虑标签）、捕获图结构信息的算法 Deepwalk 学习网络中顶点的潜在表示。这些潜在表示将社会关系编码到连续的向量空间中，编码到向量空间后的社会关系，很容易应用到统计模型中。

实验结果：DeepWalk 能够对网络进行全局的观察，特别是在存在缺失信息的情况下。当已标记数据很少时，DeepWalk 的表示得到的 F1 分数比对比方法高出 10%。在一些实验中，当训练数据少于 60%时，DeepWalk 的表现能够胜过所有对比算法。

2. 论文提出的方法

首先通过随机游走产生一串节点组成的序列，得到网络局部信息。当图中节点遵循幂律分布时，短随机游走中顶点出现的频率也将遵循幂律分布。自然语言中单词出现的频率也遵循类似的分布，因此可以将一串节点类比为句子，每个节点类比为单词，使用对自然语言进行建模的 word2vec 对随机游走得到的序列进行建模。为了方便计算，对语言建模进行了以下条件放宽：使用单词来预测上下文；上下文由给定的单词左右两边的单词组成；不考虑句子中上下文出现的顺序，最大化出现在上下文中的所有单词的概率。

该算法由两个主要组件组成：一个随机游走生成器和一个更新程序。

Algorithm 1 DeepWalk(G, w, d, γ, t)

Input: graph $G(V, E)$
window size w
embedding size d
walks per vertex γ
walk length t
Output: matrix of vertex representations $\Phi \in \mathbb{R}^{|V| \times d}$

1: Initialization: Sample Φ from $\mathcal{U}^{|V| \times d}$
2: Build a binary Tree T from V
3: **for** $i = 0$ to γ **do**
4: $\quad \mathcal{O} = \text{Shuffle}(V)$
5: $\quad$ **for each** $v_i \in \mathcal{O}$ **do**
6: $\quad\quad \mathcal{W}_{v_i} = RandomWalk(G, v_i, t)$
7: $\quad\quad$ SkipGram(Φ, $\mathcal{W}_{v_i}$, w)
8: $\quad$ **end for**
9: **end for**

Algorithm 2 SkipGram(Φ, $\mathcal{W}_{v_i}$, w)

1: **for each** $v_j \in \mathcal{W}_{v_i}$ **do**
2: $\quad$ **for each** $u_k \in \mathcal{W}_{v_i}[j - w : j + w]$ **do**
3: $\quad\quad J(\Phi) = -\log \Pr(u_k \mid \Phi(v_j))$
4: $\quad\quad \Phi = \Phi - \alpha * \frac{\partial J}{\partial \Phi}$
5: $\quad$ **end for**
6: **end for**

算法参数：一个图 G(V,E)，游走窗口大小 w，产生的向量的长度（顶点表示的维度）d，对于每个点走的次数γ，走的长度 t。

输出：一个顶点表示矩阵φ，大小为|V|*d

DeepWalk 逐行介绍：

1：初始化φ

2：从 V 构建一个二叉树 T（应该是用来做分层 softmax 的）

3-9 行：将每次迭代看作是对数据进行“传递”，迭代次数由输入数据 γ 指定。每次循环中，先生成一个随机排序来遍历顶点；再得到从每个顶点开始的长度为 t 的随机游走 Wvi；最后，根据 Wvi，利用 SkipGram 算法实现表示更新。

SkipGram:

这个算法是语言模型中，最大化窗口 w 中出现的词的概率的方法（梯度下降）,外层循环是对这个序列中的每个词进行操作，内层循环是对每个词的窗口大小为 w 的词序列进行操作。具体操作是用一个似然函数 J(φ）表示φ，然后求导，用梯度下降的方法更新，（这个α应该是学习率）。

3. 论文仿真实验及结果

数据集：BlogCatalog 是博客作者的社交关系网络。标签代表作者提供的主题类别。Flickr 是照片分享网站用户之间的联系网络。标签代表用户的兴趣组，如“黑白照片”。YouTube 是流行的视频分享网站用户之间的社交网络。 这里的标签代表喜欢不同类型视频（例如动漫和摔跤）的观众群体。

对比算法：SpectralClustering、Modularity、EdgeCluster、wvRN、Majority

实验设计：实验中通过多标签分类任务来评估算法的性能。从数据集中随机抽样标记节点的一部分(TR)，并将其用作训练数据。 其余的节点被用作测试。 重复这个过程 10 次，并报告 Macro-F1 和 Micro-F1 的平均性能。

实验结果：

1.BlogCatalog

	% Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
Micro-F1(%)	DeepWalk	**36.00**	**38.20**	**39.60**	**40.30**	**41.00**	**41.30**	41.50	41.50	42.00
	SpectralClustering	31.06	34.95	37.27	38.93	39.97	40.99	**41.66**	**42.42**	**42.62**
	EdgeCluster	27.94	30.76	31.85	32.99	34.12	35.00	34.63	35.99	36.29
	Modularity	27.35	30.74	31.77	32.97	34.09	36.13	36.08	37.23	38.18
	wvRN	19.51	24.34	25.62	28.82	30.37	31.81	32.19	33.33	34.28
	Majority	16.51	16.66	16.61	16.70	16.91	16.99	16.92	16.49	17.26
Macro-F1(%)	DeepWalk	**21.30**	**23.80**	25.30	26.30	27.30	27.60	27.90	28.20	28.90
	SpectralClustering	19.14	23.57	**25.97**	**27.46**	**28.31**	**29.46**	**30.13**	**31.38**	**31.78**
	EdgeCluster	16.16	19.16	20.48	22.00	23.00	23.64	23.82	24.61	24.92
	Modularity	17.36	20.00	20.80	21.85	22.65	23.41	23.89	24.20	24.97
	wvRN	6.25	10.13	11.64	14.24	15.86	17.18	17.98	18.86	19.57
	Majority	2.52	2.55	2.52	2.58	2.58	2.63	2.61	2.48	2.62

Table 2: Multi-label classification results in BlogCatalog

在实验中，将 BlogCatalog 网络上的训练比率(TR)从 10%提高到 90%，粗体数字表示每列中最高的性能。

结果分析：

DeepWalk 的性能始终优于 EdgeCluster，Modularity 和 wvRN。DeepWalk 在只有 20%的节点被标记时的性能，比这些方法在 90%的数据时被标记的情况下执行得更好。SpectralClustering 的性能更具竞争力，但是当 Macro-F1(TR≤20%)和 Micro-F1(TR≤60%)上的标记数据稀疏时，DeepWalk 仍然表现优异。

通过以上两点可以看出，算法的优势在于，只有小部分图表被标记时，具有强大的性能。

2.Flickr

	% Labeled Nodes	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Micro-F1(%)	DeepWalk	**32.4**	**34.6**	**35.9**	**36.7**	**37.2**	**37.7**	**38.1**	**38.3**	**38.5**	**38.7**
	SpectralClustering	27.43	30.11	31.63	32.69	33.31	33.95	34.46	34.81	35.14	35.41
	EdgeCluster	25.75	28.53	29.14	30.31	30.85	31.53	31.75	31.76	32.19	32.84
	Modularity	22.75	25.29	27.3	27.6	28.05	29.33	29.43	28.89	29.17	29.2
	wvRN	17.7	14.43	15.72	20.97	19.83	19.42	19.22	21.25	22.51	22.73
	Majority	16.34	16.31	16.34	16.46	16.65	16.44	16.38	16.62	16.67	16.71
Macro-F1(%)	DeepWalk	**14.0**	17.3	**19.6**	**21.1**	**22.1**	**22.9**	**23.6**	**24.1**	**24.6**	**25.0**
	SpectralClustering	13.84	**17.49**	19.44	20.75	21.60	22.36	23.01	23.36	23.82	24.05
	EdgeCluster	10.52	14.10	15.91	16.72	18.01	18.54	19.54	20.18	20.78	20.85
	Modularity	10.21	13.37	15.24	15.11	16.14	16.64	17.02	17.1	17.14	17.12
	wvRN	1.53	2.46	2.91	3.47	4.95	5.56	5.82	6.59	8.00	7.26
	Majority	0.45	0.44	0.45	0.46	0.47	0.44	0.45	0.47	0.47	0.47

Table 3: Multi-label classification results in Flickr

在实验中，将 Flickr 网络上的训练比率(TR)从 1%变为 10%。 这相当于在整个网络中有大约 800 到 8000 个节点标记用于分类。 表 3 给出了实验结果，粗体数字表示每列中最高的性能。

结果分析：

对于 Micro-F1，DeepWalk 的性能至少要比其他算法高出 3%。DeepWalk 可以比其他算法少 60%的培训数据。
在 Macro-F1 中表现也相当不错，最接近的是 SpectralClustering。
3.YouTube

	% Labeled Nodes	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Micro-F1(%)	DeepWalk	**37.95**	**39.28**	**40.08**	**40.78**	**41.32**	**41.72**	**42.12**	**42.48**	**42.78**	**43.05**
	SpectralClustering	—	—	—	—	—	—	—	—	—	—
	EdgeCluster	23.90	31.68	35.53	36.76	37.81	38.63	38.94	39.46	39.92	40.07
	Modularity	—	—	—	—	—	—	—	—	—	—
	wvRN	26.79	29.18	33.1	32.88	35.76	37.38	38.21	37.75	38.68	39.42
	Majority	24.90	24.84	25.25	25.23	25.22	25.33	25.31	25.34	25.38	25.38
Macro-F1(%)	DeepWalk	**29.22**	**31.83**	**33.06**	**33.90**	**34.35**	**34.66**	**34.96**	**35.22**	**35.42**	**35.67**
	SpectralClustering	—	—	—	—	—	—	—	—	—	—
	EdgeCluster	19.48	25.01	28.15	29.17	29.82	30.65	30.75	31.23	31.45	31.54
	Modularity	—	—	—	—	—	—	—	—	—	—
	wvRN	13.15	15.78	19.66	20.9	23.31	25.43	27.08	26.48	28.33	28.89
	Majority	6.12	5.86	6.21	6.1	6.07	6.19	6.17	6.16	6.18	6.19

Table 4: Multi-label classification results in YouTube

YouTube 网络规模大，更接近现实世界网络。SpectralClustering 和 Modularity 不能用于这种规模的网络。

在实验中，训练比率(TR)从 1%变化到 10%，粗体数字表示每列中最高的性能。
结果分析：

从实验中可以看出，DeepWalk 明显优于其他算法：DeepWalk 可以扩展到大图，并且在这样一个稀疏标记的环境中执行得非常好。

三、存在问题

1. 自己对该论文不太理解的地方

2. 该论文方法本身的问题

只适用于无权网络，DeepWalk 期望具有更高二阶邻近度的节点产生类似的低维表示。（LINE 适合任意类型的信息网络，且 LINE 保留了对一阶相似性和二阶相似性的敏感度）

随机游走难以达到深度和广度的平衡。（node2vec 类似于 deepwalk，主要的创新点在于改进了随机游走的策略，定义了两个参数 p 和 q，在 BFS 和 DFS 中达到一个平衡，同时考虑到局部和宏观的信息，并且具有很高的适应性）

四、论文评价

1. 论文亮点

把自然语言处理模型 word2vec 的方法应用到网络的节点表示中,通过 word2vec 的方法把网络学习为向量的潜层表示

2. 评价

领域内经典论文