

期刊和论文聚类新方法：网络 and 自然语言处理技术

吴金闪

北京师范大学系统科学学院

2019年5月6日



$$\frac{e^{\pm\beta H}}{Z}$$

等着你一起来提出新问题发展新方法

- ① 报告的目的
- ② 背景：论文聚类、网络、Node2Vec (word2vec)
- ③ 方法：基于Node2Vec的期刊聚类方法
- ④ 结果：类比测试、学科图谱、聚类、多样性
- ⑤ 讨论和致谢：推广到论文、结合论文内容
- ⑥ 带回家的消息：网络 and 自然语言处理能够帮助解决科学学领域的问题

报告的目的

- 介绍一下最近的用网络 and 自然语言处理来做期刊和论文聚类的工作
- 寻找合作者：你们提出问题，我们发展方法，或者一起来发展方法
- 顺便敬请期待我们在论文聚类上的结果

期刊和论文聚类

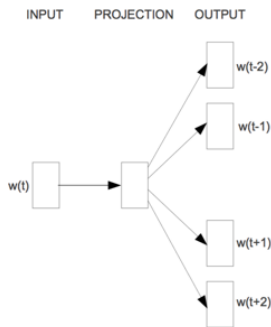
- 论文聚类就是把内容上接近的论文放在一起，形成层次性的类，主题、领域、学科
- 论文聚类可以用于论文评价，例如领域归一化
- 论文聚类可以用于把握科学发展的脉络
- 论文聚类可以用于把握科学领域之间的相互依赖关系
- 期刊聚类可以用于期刊评价，例如领域排序

期刊和论文聚类技术

- 基于引用关系（直接引用、共被引、同施引）得到任意两篇文章的相似性值
- 得到相似性值之后运用层次聚类等方法
- 假设：引用稠密的集团在内容上也接近，构成学科意义上的类
- 也有人考虑把文字上的相似性和引用上的相似性结合
- 参见Boyack等Scientometrics 111, 1223-1224(2017)，聚类问题专辑

word2vec

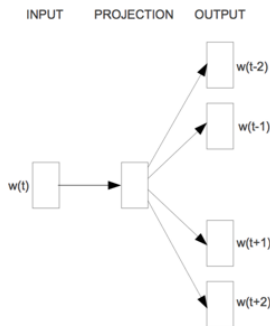
- 目标：给每个词一个矢量，要求意义相差比较大的其表示也相差比较大（内积很小）



Skip-gram

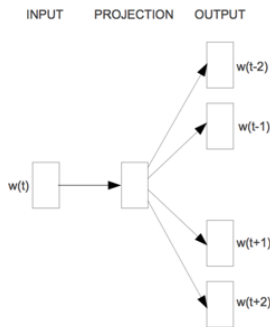
word2vec

- 目标：给每个词一个矢量，要求意义相差比较大的其表示也相差比较大（内积很小）
- 方法：每次通过读取一个词（已知词）周围的其他词（预测词），调整一个词的矢量表示，要求和这个已知词表示上（概率意义上）最接近的就是预测词



word2vec

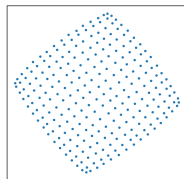
- 目标：给每个词一个矢量，要求意义相差比较大的其表示也相差比较大（内积很小）
- 结果举例：King - Man + Woman = Queen



Skip-gram

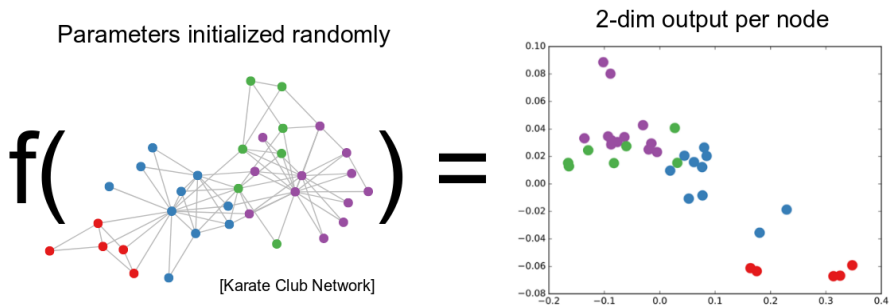
网络、Node2Vec

- 引用关系构成网络：论文（顶点）、施引（边、关系）
- 聚类实际上是整体现象：集团内部的连边比外部更紧密，需要考虑邻居的邻居...
- 网络上的顶点的聚类方法有经过或者跳过相似性的计算的
- Node2Vec算法的步骤和结果：
 - ① 通过随机行走把网络变成语料
 - ② 在语料上运行word2vec得到词语——顶点——的表示矢量
- 二维正方晶格的Node2Vec表示结果（t-SNE二维布局算法）



基于Node2Vec的期刊聚类方法

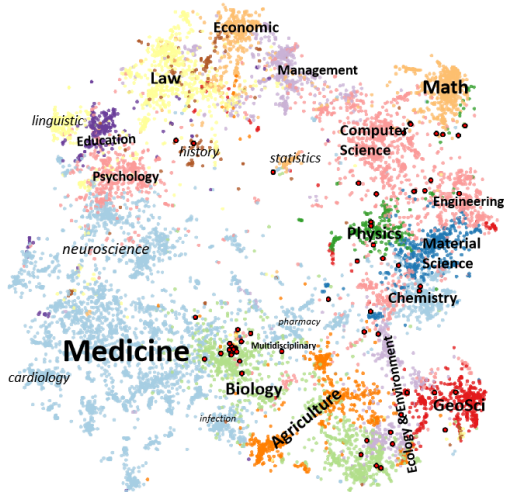
- ① 收集数据构建期刊引用网络，有权重，JCR2017
- ② 通过随机行走把期刊引用网络变成语料，运行word2vec得到期刊的表示矢量
- ③ 从矢量表示得到任意两个期刊的相似性
- ④ 运行层次聚类算法从相似性矩阵得到层次聚类



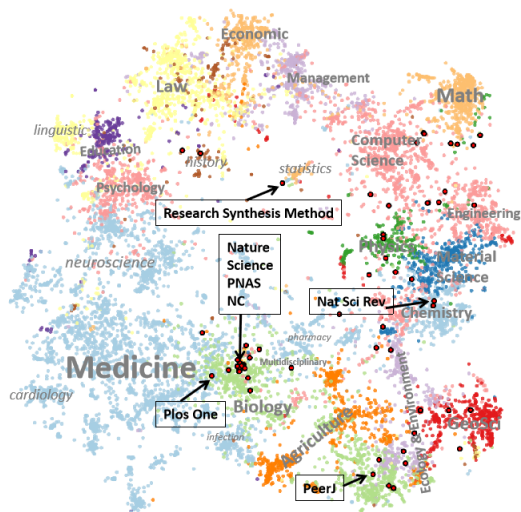
基于Node2Vec的期刊聚类方法，续

- ① 把矢量表示呈现为二维的形式，Science Map
- ② 把聚类以后的结果和主要结果对比，AMI相似性
- ③ 更详细的聚类结果的比较，没有做
- ④ 偶然发现：拿矢量表示来算算多样性
- ⑤ 在论文层次方法相同，并且可以用word2vec直接处理文字相似性

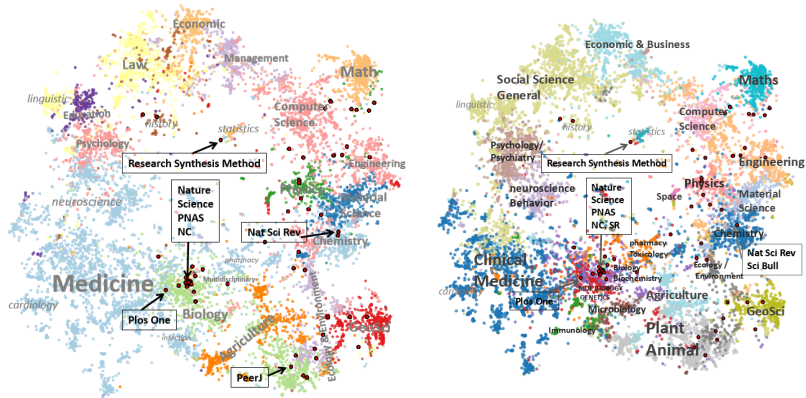
Science Mapping, t-SNE二维布局算法



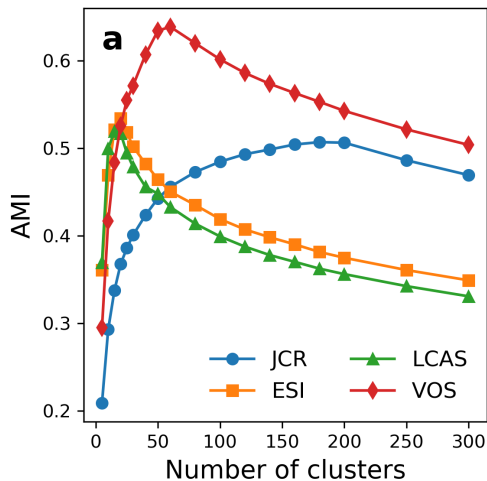
把所谓的综合性期刊放上去



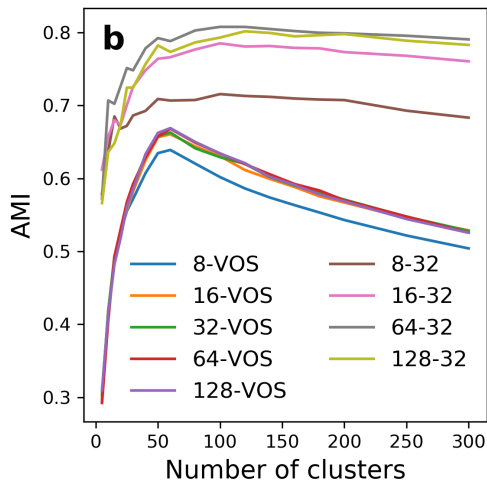
和ESI期刊分类比较一下



比较一下分类



如果你还想知道表示的维数 (32) 话

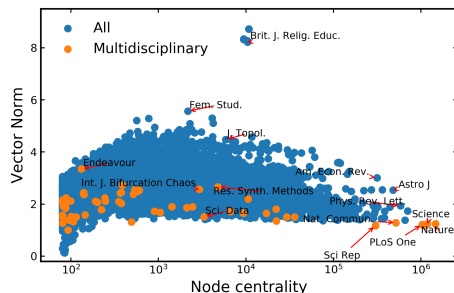
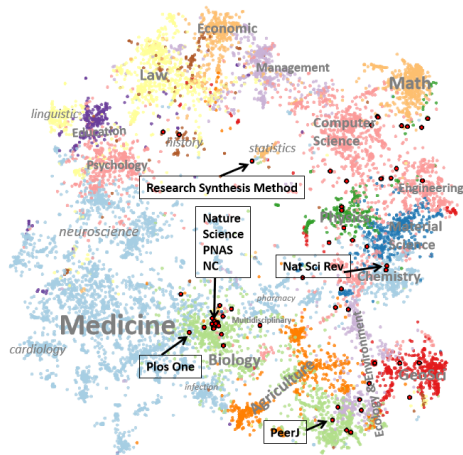


King - Man + Woman = Queen 类比测试

Example	Test 1	Test 2	Test 3
King	<i>PLoS Comput. Biol.</i>		
Man	<i>Nat. Cell Biol.</i>		
Woman	<i>Phys. Rev. Lett.</i>	<i>Genome Biol.</i>	<i>J. Neurosci.</i>
Queen	<i>J. Stat. Mech. Theory Exp</i>	<i>Bioinformatics</i>	<i>NeuroImage</i>
	<i>Phys. Rev. E</i>	<i>BMC Bioinformatics</i>	<i>Biol. Cybern.</i>
	<i>Fluctuation Noise Lett.</i>	<i>J. Comput. Biol.</i>	<i>Front. Comput. Neurosci.</i>
	<i>EPL</i>	<i>BioData Min.</i>	<i>Cereb. Cortex</i>
	<i>Eur. Phys. J. B</i>	<i>J. Bioinform. Comput. Biol.</i>	<i>J. Comput. Neurosci.</i>

偶然发现：矢量长度的含义

- 回到词，最经常出现的词长度短，没有特异性。期刊也类似吗？

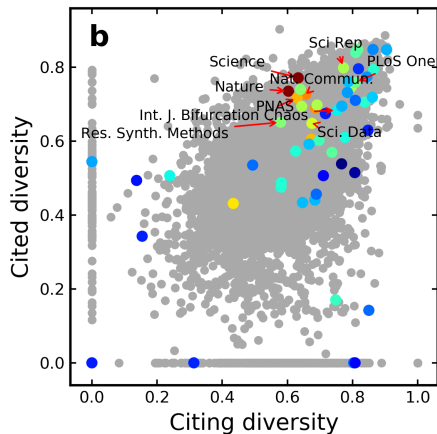
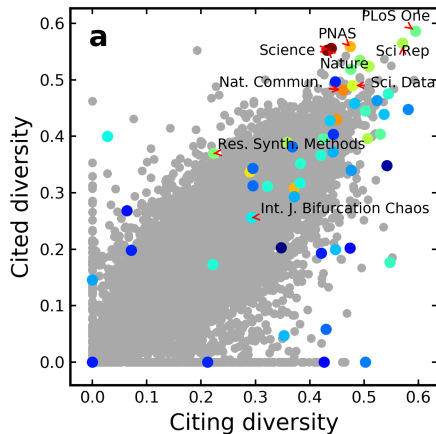


基于施引和被引相似性的多样性

- Rao-Stirling多样性公式 $D = \sum_{i \neq j} p_i p_j (1 - s_{ij})$
- 通常基于直接引用网络的相似性: $s_{ij}^{NB} = \cos(c_i^{dc}, c_j^{dc})$
- 换成基于全局信息的矢量表示的相似性: $s_{ij}^G = \cos(x_i^{n2v}, x_j^{n2v})$

多样性结果比较

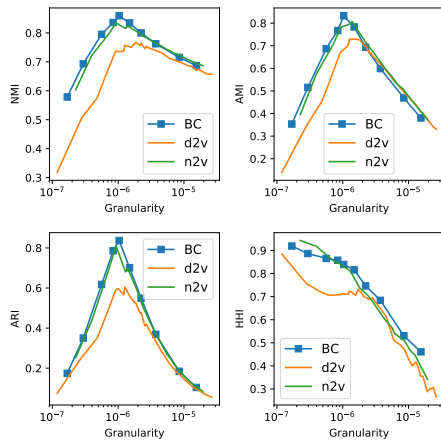
- 看起来我们的全局矢量更有道理，多样的跳出来了而不是埋藏其中



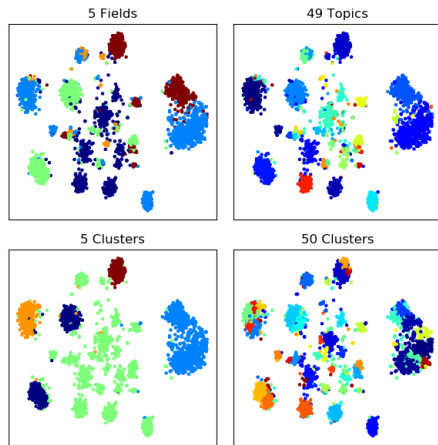
在期刊聚类上的初步结果说明

- 包含全局信息的矢量表示在聚类和多样性上有一定优势
- 多样性本身用什么指标本研究暂时没有考虑
- 其他聚类算法本研究暂时没有考虑
- 展望: Node2vec, word2vec可以自然地综合考虑引用和文本相似性, 并且包含了直接和间接联系
- 在论文层次的聚类, 以及其他问题上呢?

论文层次的初步结果



论文层次的初步结果，续



带回家的消息：网络与自然语言处理能够帮助解决科学学领域的问题

- 科学学的深入研究依赖于对内容的理解，形式数据远远不够
- 对内容的理解需要依靠自然语言处理技术、学科知识表示技术
- 形式数据的表示需要网络科学
- 学科知识表示需要网络科学
- 你来问问题，我们来发展方法
- 一起提出新问题、发展新方法
- 让我们一起推动科学学的发展

致谢，感谢您的时间与注意力

- Zhesi Shen, Fuyou Chen, Liying Yang, Jinshan Wu, Node2Vec representation for clustering journals and as a possible measure of diversity, JDIS, in press(2019).
- 感谢合作者：沈哲思、陈福佑、史双青、杨立英等
- 要了解我更多，请访问“系统科学人之吴金闪”，吴金闪在北师大系统科学学院的主页
- 研究工作的整理在这里“Big Physics（大数据大物理研究小组）”