WSDM图方面论文总结

Inferring networks from opinions主题报告

作者是Jure Leskovec，个人主页：http://cs.stanford.edu/people/jure/
报告主要介绍的是他一年之前创办的一个公司：KOSEI。这个类似于一个广告平台，可以从walmart等app里面获取用户的点击购买信息，然后在另一些app的广告位显示商品广告，用户点击了商品即可跳转至walmart的商品页面。
建模部分用到了product graph。商品之间的关系，有两种，一是商品竞争(买了一个，再买另一个的概率降低)，一是商品搭配(买了一个，可能会搭配着买第二个)。根据共同浏览和共同购买，构建了商品关系图。作者从构建商品之间的关系强度，和对推荐给出解释两方面进行介绍。
构建商品之间的关系：
用到了链路预测，主题模型，以及商品分层的思想。作者先尝试用文本特征，评论的相似度，发现维度太高，容易过拟合；之后用LDA，并且结合最大似然来考虑商品之间的相关性。由于商品较多，两两计算相似度太耗时，所以作者对商品进行类目分层，同一类目内遍历算出最相关的商品，并用matching engine来提速。
推荐的解释：
找到那些最能解释这个商品关系的词，如：
主题模型用到的词，
逻辑回归量时用做预测的词，
最大似然时用到的短语。(推荐解释这部分讲的很粗略)

The power of random neighbors in social networks

作者首先提出了一个社交悖论：大多数人都认为，自己的朋友们的平均朋友数，要比自己的朋友数多。之前的人都是仅凭直觉进行分析，本文作者比较理论的分析了这个社交悖论的正确性，并在真实社交数据上验证了理论分析。另外作者也讨论了可能在影响力最大化方面的应用。
在小世界理论的前提下，每个点的边数服从power law分布，使得很多度比较小的点，会和度很大的点相连，度很大的点会拉高这些点的邻居的平均度。作者提出在power law因子在[1, 3]时，点的度和点的邻居的平均度存在一个渐进的比例差。
最后作者提出，这个理论在影响力最大方方面也能有所应用。作者随机选出一些点，与从这些点里面随机选出一些他们的邻居点做影响力的对比，然后证明，随机选它们邻居点的效果，要比随机选它们本身的效果好。

Sarcasm Detection on Twitter A Behavioral Modeling Approach

这篇文章主要讲twitter里面的反语检测。 传统的方法是用语义分析，作者利用用户表达反语时的行为特性以及历史信息，从行为学和心理学上构建了反语检测模型SCUBA，作者从人工标注标签的tweet文中提取特征，然后利用有监督的机器学习方法来进行学习分类。论文本质上是特征工程。模型用到的特征如下：

1. 获取推文中的词后，计算每个词的affect score和sentiment score。affect score通过查找词典获得，词典中分数越高的词表明这个词越容易让人愉悦；sentiment score通过sentiStrength方法计算；
2. 根据带hashtag的tweet文构建二元词典和三元词典(1中的为一元词典)，详细方法见Kouloumpis；
3. 根据用户历史tweet文的情绪波动；
4. 根据tweet文的可读性(词个数，音节个数，多音节词个数等)，词长度，词长度分布和历史tweet文长度对比；
5. 根据历史tweet文的情感分，历史推文的affect score和sentiment score；
6. tweet文的发推时间，与上衣tweet文的间隔时间，是否伴随着发誓等词；
7. 根据用户的语言熟练度，词汇能力和语法能力，历史中发反语的能力；
8. 根据用户对tweeter的使用熟练度；
9. 根据用户tweet文的书写风格。

Modeling and Predicting Retweeting Dynamics on Microblogging Platforms

本文提出了用增强泊松过程模型来预测微博的转发量。增强泊松过程(RPP)模型同时考虑三部分：
1. 固有属性，自身的受欢迎程度，是个常数；
2. 随时间变化t的关系，power-law分布；
3. 随已转发人数k的关系(richer-get-richer)，和k指数级相关。
以上共三个参数。另外，考虑到微博用户在不同的时间段，转发微博的次数很不一样，所以作者对微博的转发时间做了一个拉伸变换，便于学习泊松过程时的时间参数。模型利用每条微博的前一部分转发记录学习参数，然后再预测接下来的转发量。文章和spikeM模型有点像，先是指数级上升，后是幂律级下降。spikeM里面有考虑周期性，这个模型没有。

Can cascade be predicted

这篇文章是Jon Kleinberg，Jure Leskovec和Facebook合作写的一篇论文，发表于2014 WWW，分析如何在社交网络中预测图片等的转量。
作者先对比了个人用户和公共主页发的图片的转发量，结论显然是公共主页的图片的转发量较大，且转发量都服从幂律分布；然后分析转发子图(理论上是一个有向树，root是图片发布者)的wiener index(树中每两个点的平均距离)，个人用户的wiener index指数更高，说明个人用户的图片，分享路径一般更深，公共主页的图片的转发网络，则会相对类似于hub结构。
然后作者对预测问题做了简单的分析，直接预测转发量，不是很靠谱，所以作者提出了这样一个问题：如果知道某个图片已经被转发了k次，那么它最终被转发大于2k次的概率。由于转发量服从幂律分布，大于k次的图片的转发量的中位数是2k，所以随机猜对的概率是0.5.然后作者提取特征，用机器学习的方法进行训练，最终得到的准确率约为0.8，好于随机。使用的特征为：
原始内容，包括文字内容和图片内容
内容发布者的个人特征
转发者的特征，好友数，年龄等
网络结构的特征，k个转发者的结构特征，包括度，转发深度之类的信息
时间特征，转发的时间信息
最终发现，这些特征中，时间特征最为有效。(猜测是对于热门tweet来说，发布了之后马上会带来大量的转发，这类微博的最终转发量肯定大于2k，一般的冷门微博，断断续续的被转发，最终转发量大概率

小于2k) 最后作者做了一些其他的尝试，比如分析准确率随k的变化，各个特征随k的变化

On Integrating Network and Community Discovery

本文提出的是如何在不知道整个网络信息的情况下将网络发现和社区发现结合起来。访问网络中的点，会带来一定的cost，文章分析如何在只已知网络中部分信息时，迭代的进行网络发现和社区发现。
网络发现部分：
这部分涉及到选择新的需要发掘的点，指标有：
High purity：这个点的邻居最好都属于一个社群
Large observed degree：这个点的已知的边较多
Balanced communities：这个点能让社区大小均衡
社区检测部分：
新发掘一个点后，便更新原有的社区信息。作者提出利用EM算法来更新节点的社区信息。当点的数量变多时，EM算法效率降低，所以作者提出了一个局部更新的方法来进行提速。每加入一个新点时，只更新它和它的邻居点的社区信息，并周期性的用EM做全局更新来避免错误积累。

Inverting a Steady-State

WSDM2015的best paper。
平时我们考虑的是已知图和其转移矩阵，如何求得其恒稳分布，这篇文章反过来思考，如何在已知恒稳分布的情况下，推导其转移矩阵。
作者先将原始图转变为一个特殊的二部图，然后证明已知恒稳分布时，此二部图可以求得转移矩阵，最终可以求出对应的原始图的转移矩阵。论文的证明以及公式较多，不太容易看懂。