WDS

DataScience@Web

机器学习讨论班

2018年暑期

图中节点相似度计算

丁金如

介绍内容

- 早期模型(Jaccard相似度，cosine相似度，min相似度)
- SimRank
- SimRank++
- P-Rank
- 基于SimRank模型的可扩展相似性搜索

早期模型

早期模型只考虑了俩对象的“公共邻节点数”与“度数”来评估其相似度。

$$S_{Jaccard}(u,v) = \frac{|I(u) \cap I(v)|}{|I(u) \cup I(v)|}$$

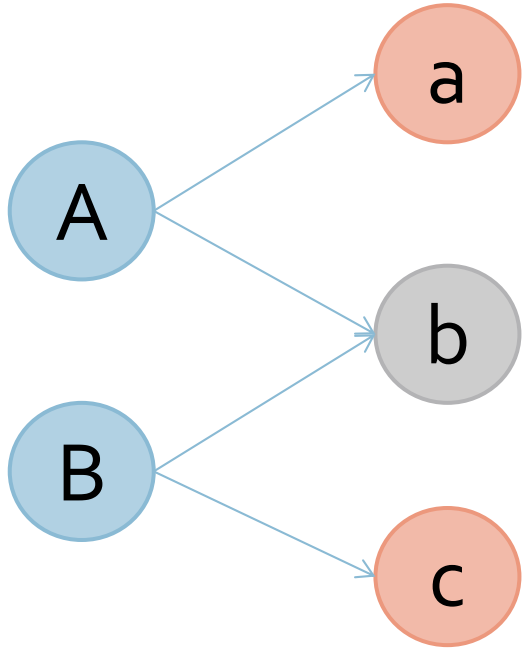
$I(u)$是u的入邻点集合

$$S_{cosine}(u,v) = \frac{|I(u) \cap I(v)|}{\sqrt{|I(u)| * |I(v)|}}$$

$$S_{min}(u,v) = \frac{|I(u) \cap I(v)|}{\min(|I(u)|, |I(v)|)}$$

SimRank

SimRank核心思想：如果两个实体被其相似的实体所引用(即有相似的入邻边结构)，那么这两个实体也相似。



SimRank

SimRank数字定义式：

$$s(a,b)=\begin{cases}1, & a=b\\ 0, & L(a)=\emptyset \ \ or \ \ L(b)=\emptyset\\ \dfrac{c}{|L(a)||L(b)|}\displaystyle\sum_{i=1}^{|L(a)|}\sum_{j=1}^{|L(b)|}s\big(L_i(a),L_j(b)\big), & otherwise\end{cases}$$

指向节点b的节点集合

C:阻尼系数
0<c<1
c=0.8 or c=0.6

SimRank computation

- SimRank矩阵计算
  - 将SimRank数学公式用矩阵乘法的形式表示
  - 利用MapReduce分布式并行编码实现
  - 优点：计算精度较高
  - 缺点：复杂度较高
- SimRank随机游走计算
  - 优点：复杂度较低
  - 缺点：有一定的随机性，精度较低

SimRank 迭代算法(矩阵算法)

$$S_0(a,b)=\begin{cases}0 & a\neq b\\ 1 & a=b\end{cases}$$

关于迭代次数k的单调不减函数

$$S_{k+1}(a,b)=\begin{cases}\frac{c}{|L(a)||L(b)|}\sum_{i=1}^{|L(a)|}\sum_{j=1}^{|L(b)|}S_k(L_i(a),L_j(b)) & a\neq b\\ 1 & a=b\end{cases}$$

$$\lim_{k\to\infty}S_k(a,b)=S(a,b)\qquad(\forall a,b\in V)$$

$$|S(a,b)-S_k(a,b)|\leq C^k\qquad(\forall a,b\in V)$$

矩阵形式：

$$S=C\cdot Q\cdot S\cdot Q^T+(1-C)\cdot I_n$$

Q:概率转移矩阵，每一列和为1，如果从节点i可以转移到节点j,并且这样的节点i有n个，则$Q_{i,j}=\frac{1}{n}$

基于随机游走的SimRank

- 模型说明
  - 在图G中，顶点a和b之间的相似度取决于a和b在图中随机游走直至相遇所经过的路径的长度以及能相遇的次数。

$$s(a,b) = \sum_{t(a,b)\to(x,x)} P[t]C^{l(t)}$$

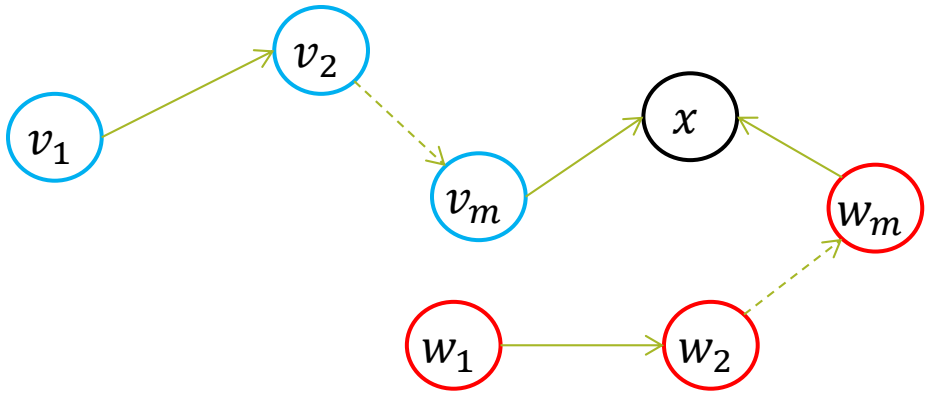
以a, b为起点的两条随机游走路径

$$s^{k+1}(a,b) = \sum_{t(a,b)\to(x,x),\, l(t)\le k+1} P[t]C^{l(t)}$$

a和b通过路径t_1和t_2在顶点x相遇的概率

首次在x点相遇，两条路径的长度

基于随机游走的SimRank



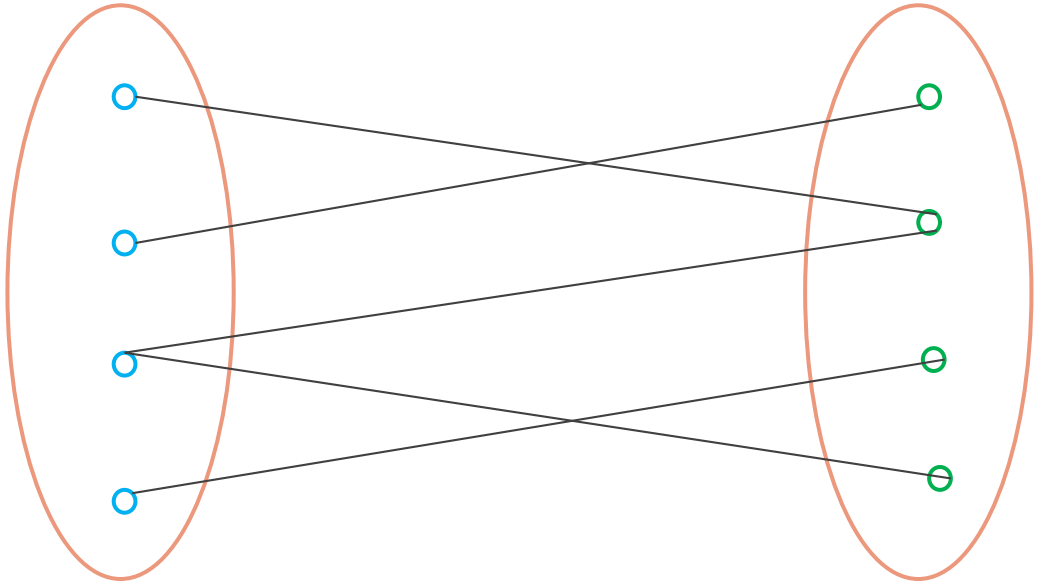
假设两条路径分别为$t_1 = (v_1, v_2, \cdots v_m, x)$, $t_2 = (w_1, w_2, \cdots w_m, x)$
其中$v_1 = a$,$w_1 = b$
游走t_1的概率$P[t_1] = \prod_{i=1}^{m} \frac{1}{|O(v_i)|}$
游走t_2的概率$P[t_2] = \prod_{i=1}^{m} \frac{1}{|O(w_i)|}$
a和b通过路径t_1和t_2在顶点x相遇的概率

$$P[t] = P[t_1]P[t_2] = \prod_{i=1}^{m} \frac{1}{|O(v_i)||O(w_i)|}$$

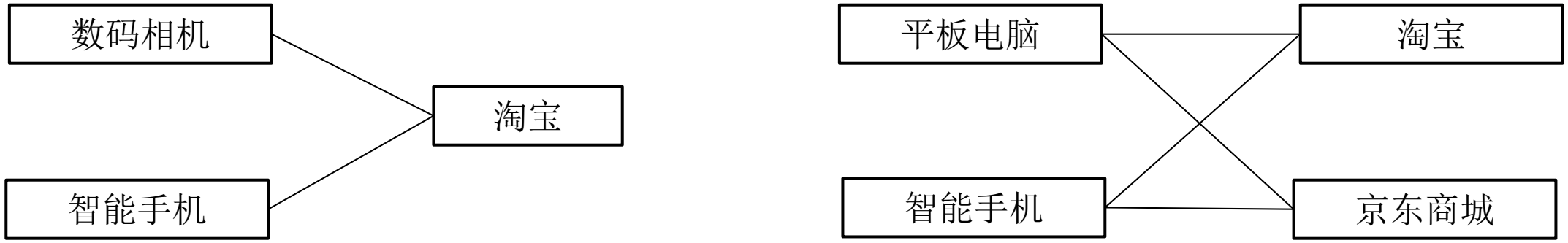
SimRank缺点

- SimRank在完全二部图中，算法计算出来的分数和人的直观是不一致的

二部图：设$G = (V, E)$是一个无向图，如果顶点V可分割为两个互不相交的子集(A, B)，并且图中的每条边(i, j)所关联的两个顶点i和j分别属于这两个不同的顶点集$(i\ in\ A, j\ in\ B)$，则称图G为一个二部图。



有更多的证据
（共同连接对象）



迭代	数码相机-智能手机	平板电脑-智能手机
1	0.8	0.4
2	0.8	0.56
3	0.8	0.624
4	0.8	0.6496
5	0.8	0.65984
6	0.8	0.663936
7	0.8	0.6655744

SimRank++

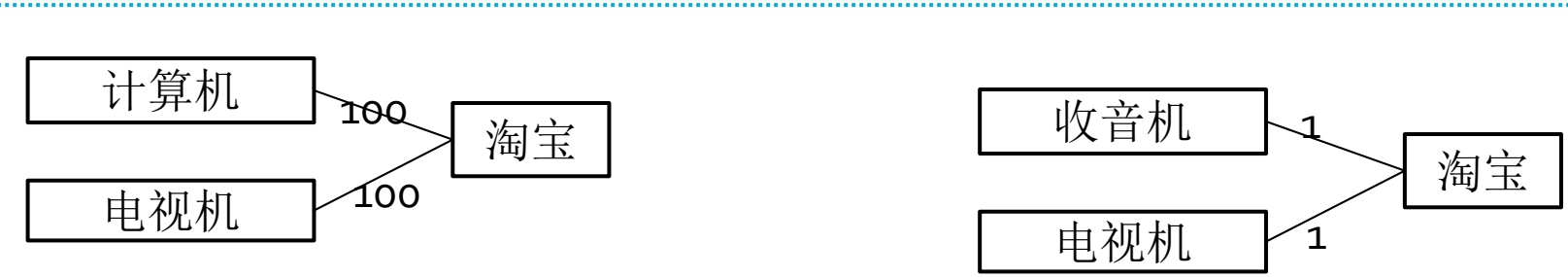
- 主要改进原SimRank的两个方面
  - 考虑了边的权值
  - 考虑了子集节点相似度的证据(共同连接对象)

1.原 SimRank 算法，对于边的归一化权重，是用的比较笼统的关联的边数分之一来度量，并没有考虑不同的边可能有不同的权重度量

2.原 SimRank 算法，只要认为有边相连，则为相似。却没有考虑到如果共同相连的边越多，则意味着两个节点的相似度会越高。

SimRank++

- 考虑边的权值
  - 如果两个节点对应的权值的方差相等，那么权值较大的节点对之间的相似性较高
  - 如果两个节点对应的权值的方差不相等，那么方差较小并且权值较大的节点对之间的相似性较高



SimRank++

- 引入一个新函数$p(*,*,)$表示图中节点间的转移概率

$$p(a,a) = 1 - \sum_{i \in E(a)} p(a,i)$$

$$p(a,i) = spread(i) \cdot normalized\_weight(a,i) \quad \forall i \in E(a)$$

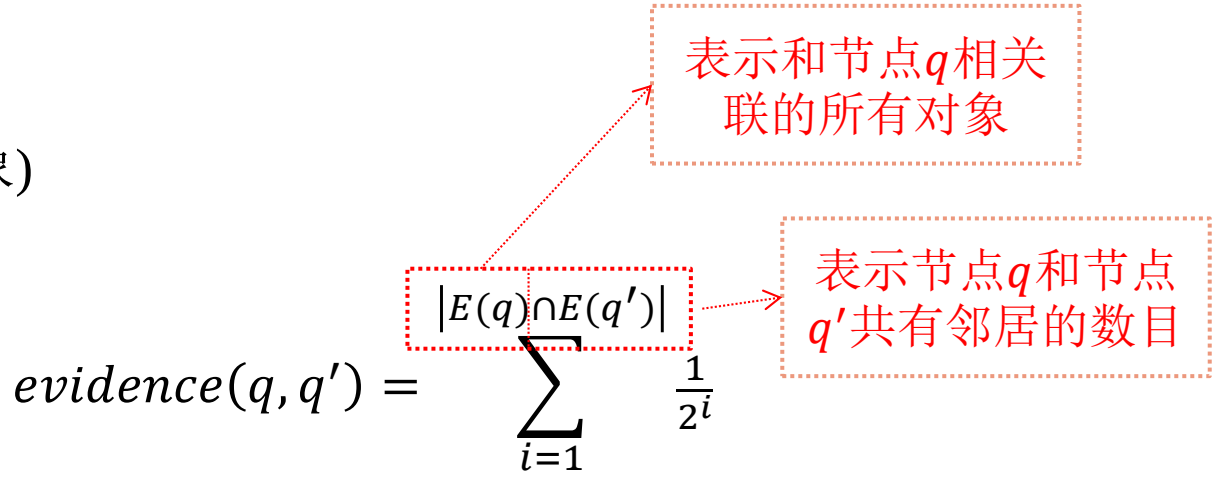
其中：

$$spread(i) = \frac{1}{variance(i)}$$

$$normalized\_weight(a,i) = \frac{w(a,i)}{\sum_{j \in E(a)} w(a,j)}$$

SimRank++

- 添加证据因子(考虑相关联的对象)



$|E(q) \cap E(q')|$和$evidence(q,q')$呈正比关系，$|E(q) \cap E(q')|$越大，$evidence(q,q')$越接近1。

SimRank++

- SimRank++算法的矩阵运算公式

$$S^k = \begin{cases} cP^T S^{k-1} P + I_n - Diag\left(diag\left(cP^T S^{k-1} P\right)\right), & k > 0 \\ I_n, & k = 0 \end{cases}$$

- 通过MapReduce并行计算模型

P-Rank

- SimRank在定义两个对象间的相似度时仅考虑了它们的入邻点结构，而忽略了其出邻节点结构。

- P-Rank的的中心思想
  - 两个不同的对象若引用了相似的对象，则它们相似(出邻点递归)
  - 两个不同的对象若被相似的对象所引用，则它们相似(入邻点递归)
  - 每个对象都与它自身最相似

P-Rank

C_{in}与$C_{out} \in (0,1)$分别为入边和出边方向的阻尼系数

$$S(u,v) = 1$$

$$S(u,v) = \frac{\lambda \cdot C_{in}}{|I(u)||I(v)|} \sum_{i=1}^{|I(u)|} \sum_{j=1}^{|I(v)|} S(I_i(u), I_j(v)) + \frac{(1-\lambda) \cdot C_{out}}{|O(u)||O(v)|} \sum_{i=1}^{|O(u)|} \sum_{j=1}^{|O(v)|} S(O_i(u), O_j(v))$$

[0,1]是一个权重参数

(1)若$I(u)$或$I(v) = \emptyset$,则“入邻点部分”为0

P-Rank

- P-Rank的缺点
  - 计算精度无法控制(证明了迭代的收敛性，但迭代的精度仍然未知)
  - 权重系数与阻尼因子对P-Rank稳定性会有影响

当$C_{in} > C_{out}$时，λ的值越大，P-Rank方程越不稳定
当$C_{in} < C_{out}$时，λ的值越小，P-Rank方程越稳定
当$C_{in} = C_{out}$时，λ的取值与P-Rank方程的稳定无关

P-Rank的优化

- P-Rank的表示
  - P-Rank幂级数表示

$$S = \lambda C_{in} \cdot Q \cdot S \cdot Q^T + (1-\lambda) C_{out} \cdot P \cdot S \cdot P^T + I_n$$

  - P-Rank逆矩阵形式

$$vec(S) = [I_{n^2} - \lambda C_{in}(Q \otimes Q) - (1-\lambda) C_{out}(P \otimes P)]^{-1} \cdot vec(I_n)$$

随机游走提供理论基础

- P-Rank的随机算法与优化(时间复杂度$O(n)$)
  - P-Rank随机概率模型
  - 基于MonteCarlo的随机算法

利用低秩分解降低矩阵逆计算的复杂度

基于SimRank模型的可扩展相似性搜索

- SimRank相似性搜索：给定一个查询节点u，查找top-k个节点 v，其满足前k个最高的SimRank值$S(u,v)$
- 算法四大要素
  - 引入线性递归表达式---(加速计算)
  - 基于蒙特卡罗计算$S(u,v)$---(基于线性递归表达式+随机游走)
  - 只关注局部的节点信息---(SimRank值的距离衰减性)
  - 设置两个上界值---(修剪相似度搜索过程，加快运算)

总结

- 应用场景
  - 网页排名
  - 协同过滤
  - 网络图聚类
  - 近似查询
- 速度和精确度不可兼得

谢 谢