标签传播算法在微博用户兴趣图谱的应用

张俊林

新浪微博-搜索部-推荐组

2012-6-28

About me

- 中科院软件所　博士
- 《这就是搜索引擎：核心技术详解》作者
- 现任职新浪微博，从事语义分析，推荐系统，搜索技术，社交挖掘等方面研发工作

提纲

- 用户兴趣图谱的重要性
- 标签传播算法
- 使用标签传播算法计算微博用户兴趣图谱
- 大规模数据计算问题
- 算法效果示例

什么是用户兴趣图谱

- 用户兴趣图谱
  - 个性化概念
  - 根据用户的行为以及用户产生的内容等方方面面的数据来从中导出用户可能的兴趣点
  - 用户个性化兴趣建模

兴趣图谱的用途

- 个性化建模
- 推荐感兴趣的信息
  - 感兴趣的人
  - 感兴趣的微博
  - 感兴趣的新闻
  - 感兴趣的图片
  - 感兴趣的群组
  - 定向广告推送

兴趣图谱的用途

- 订阅微博重排序
  - 智能排序

构建用户兴趣图谱可利用的信息

- 微博环境下有很多可利用信息
  - 发表的微博内容
  - 转发评论的微博内容
  - 自标签
  - 参加的群组
  - 参加的投票
  - 我关注的人
  - 关注我的人
  - 社交行为
  - ……

构建用户兴趣图谱可利用的信息

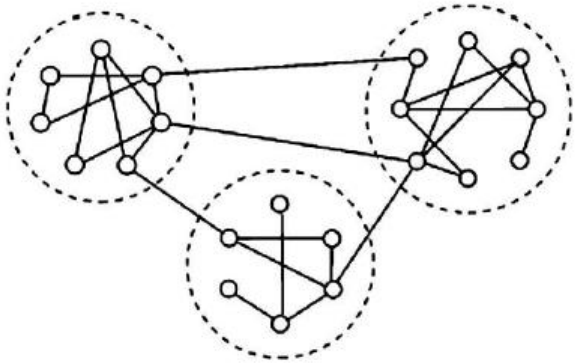
- 本讲座主要涉及到的信息
  - 发表的微博内容
  - 转发微博内容
  - 自标签
  - 社交行为
  - ……

提纲

- 用户兴趣图谱的重要性
- 标签传播算法
- 使用标签传播算法计算微博用户兴趣图谱
- 大规模数据计算问题
- 算法效果示例

标签传播算法

- 社交网络挖掘中很常用
  - 自动挖掘社交关系中的“团结构”



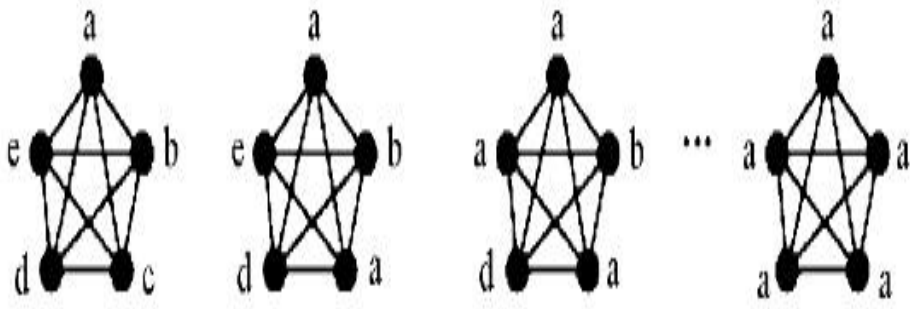
  - 优点
    - 思路非常直观和简洁，易理解易实现
    - 容易实现对大规模数据进行处理，具备很强的实用性

标签传播算法

- 问题：对于社交网络S，如何通过标签传播算法自动发现其中的密集连接子图？
- 基本思路
  - 初始阶段
    - 为图中每个节点赋予一个独一无二的标签L
  - 多轮迭代
    - 通过社交关系(即图的边）将标签向其它节点传播
    - 某个节点node将根据与其有边联系的其它节点的标签来决定自己此轮应该赋予哪个标签
      - 将其邻居节点的标签中出现次数最多的那个标签赋予自己
      - 如果邻居节点的标签数目一样多，无法找出最多个数标签，则随机赋予一个标签即可

标签传播算法

- 简单示例



提纲

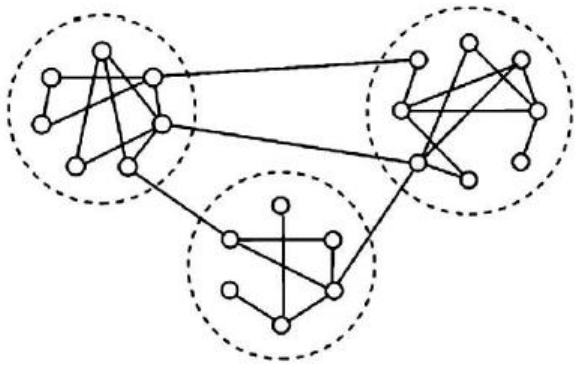
- 用户兴趣图谱的重要性
- 标签传播算法
- 使用标签传播算法计算微博用户兴趣图谱
- 大规模数据计算问题
- 算法效果示例

使用标签传播算法计算微博用户兴趣图谱

- Recap:利用到的信息
  - 发表的微博内容→兴趣词
  - 转发微博内容→兴趣词
  - 自标签
  - 社交行为

使用标签传播算法计算微博用户兴趣图谱

- 构建图结构
  - 图节点：用户ID
  - 节点之间的边：社交信息
    - 转发
    - 评论
    - @U
  - 初始标签
    - 内容兴趣词
    - 自标签
  - 基本假设
    - 如果两个用户之间的互动越频繁，那么两者之间的社交关系越紧密，而亲密的社交关系往往蕴含着潜在的兴趣关联或者较强的线下社交关系。



使用标签传播算法计算微博用户兴趣图谱

- 算法流程
  - 多轮迭代
    - Step1:找到与节点A有边联系的邻居节点，形成邻居节点集合S，S中的这些节点代表了与用户A有过互动行为的用户集合，同时把A也放入集合S中；
    - Step2:统计集合S所有用户中出现过的不同标签的频次，用户A本身初始的标签按照频次f参与计数，即给用户本身的标签一个较大的初始频次；
    - Step 3:将第二步获得的标签按照其频次和标签的IDF值使用类似于IF.IDF计算框架的方式计算其权重，并按照权重得分高低排序，取Top K标签作为节点A的新标签集合；
  - 迭代终止条件
    - 指定次数
    - 标签很少变化

使用标签传播算法计算微博用户兴趣图谱

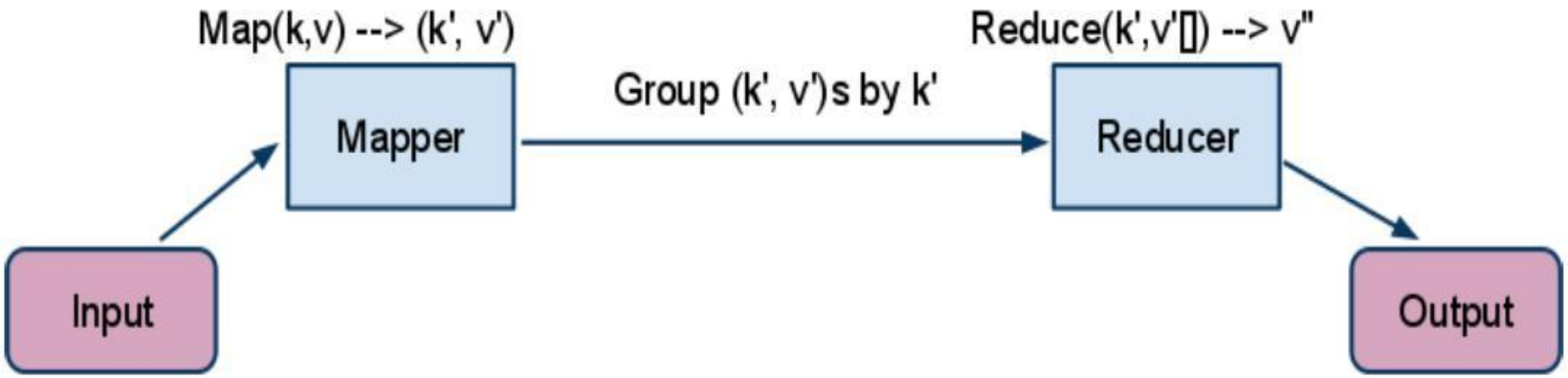
- 为何要引入类似IDF因子
  - 用户标签中包含一些非常通用的内容,信息含量低
    - 80后、音乐、网络营销 etc
  - 控制方法
    - 只保留信息含量高的
    - 计算权值公式中调节TF和IDF因子重要性
- 算法优点:
  - 结合内容因素和社交因素
    - 你所在的群体反映你的兴趣
  - 对发布内容少的用户也可以刻画其兴趣

提纲

- 用户兴趣图谱的重要性
- 标签传播算法
- 使用标签传播算法计算微博用户兴趣图谱
- 大规模数据计算问题
- 算法效果示例

大规模数据计算问题

- 微博用户超过3亿
- 分布式计算
  - Hadoop平台
  - MapReduce任务
- MapReduce



大规模数据计算问题

- WordCount

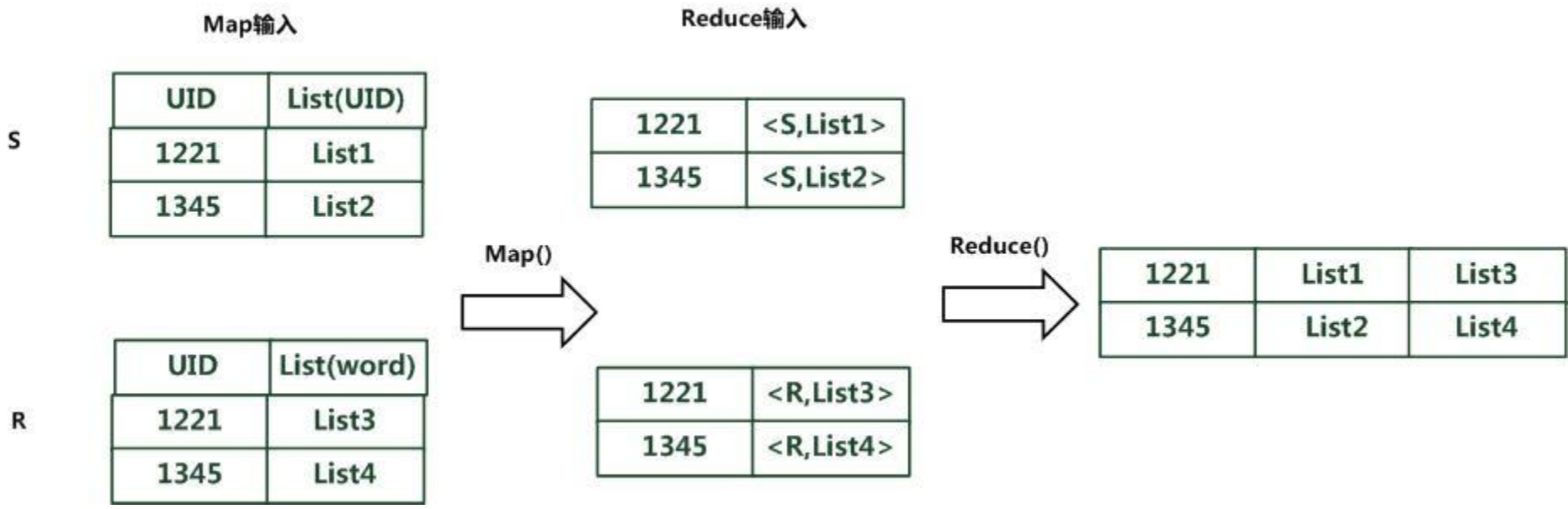
```
//Pseudo-code for "word counting"
map(String key, String value):
  // key: document name,
  // value: document contents
    for each word w in value:
      EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
    int word_count = 0;
    for each v in values:
      word_count += ParseInt(v);
    Emit(key, AsString(word_count));
```

大规模数据计算问题

- 改造的标签传播算法
  - While(Condition is not satisfied)
    - 数据聚合MapReduce任务
    - 标签传播计算MapReduce任务
    - 数据聚合MapReduce任务
  - 与通常的MR任务不同，有两类输入数据
    - » 用户社交数据：<uid,list(uid)>的形式
    - » 用户的初始标签信息：<uid,list(word)>的形式
  - 我们希望的形式
    - » <uid,list(uid),list(word)>的数据形式

大规模数据计算问题

- 数据聚合MapReduce任务
  - 典型的One-2-One join操作，我们采取标准的Reduce-side join的方法



大规模数据计算问题

- 标签传播计算MapReduce任务
  - » 用户的数据经过聚合处理已经组织成为<uid,list(uid),list(word)>的形式
- Map操作

```
map(String key, String value):
    // key:   uid
    // value: list(uid)+list(word)

    /*解析value*/
    list uidList=parse(value)
    list wordList=parse(value)

    /*标签传播*/
    for each u in uidList:
       EmitIntermediate(u, AsString(wordList));
       If u==key://如果是节点自身，传播f次
          int i=0
          while i<f:
             EmitIntermediate(u, AsString(wordList));
             i+=1
```

大规模数据计算问题

- Reduce操作

```
reduce(String key, Iterator values):
    // key:    uid
    // values: list(word)列表

    /*统计标签频次*/
    map<string,float> tfCounter;
    for each v in values:
        list wordList=parse(v)
        for each w in wordList:
            tfCounter[w]+=1

    /*计算标签权值*/
    For each label in tfCounter.keys():
        float weight=TfIdf(label)
        tfCounter[label]=weight

    /*取Top K*/
    List topLabelList=GetTopK(tfCounter,K)

    Emit(AsString(topLabelList));
```

提纲

- 用户兴趣图谱的重要性
- 标签传播算法
- 使用标签传播算法计算微博用户兴趣图谱
- 大规模数据计算问题
- 算法效果示例

算法效果示例

微博用户名	Top 10 标签
张俊林say	自然语言处理/推荐系统/数据挖掘/机器学习/google/云计算/ Nosql/信息检索/计算语言学/搜索引擎
刘江CE	云计算/程序员/android/社会化媒体/产品设计/架构师/CSDN/研发管理/移动开发/ CTO
梁斌Penny	社交网络/机器学习/数据挖掘/自然语言处理/云计算/搜索引擎/数据分析/ python/信息检索/分布式计算
互联网的那点事	互联网/用户体验/新媒体/创业/天使投资/B2C/UED /创业者/交互设计/产品经理

算法效果示例

微博用户名	Top 10 标签
创业最前线	创业/天使投资/vc/创业家/商业模式/企业家/风险投资/PE/领导力/创业者
李开复	创新工场/知乎/社交网络/移动互联网/天使投资/lbs/创业家/智能手机/豌豆荚/ vc
周鸿祎	风险投资/奇虎/浏览器/互联网/360/杀毒/产品经理/创业/ipad/社交游戏
雷军	移动互联网/安卓/智能手机/米聊/天使投资/小米手机/Iphone/ 3G/创业/ miui

That's it,Thanks!