新浪微博用户兴趣建模系统架构

原创：张俊林 布洛卡区 2016-05-18

/*作者注：这是2011年左右新浪微博个人兴趣模型的技术架构，所以你从中是看不到目前很多流行的NoSQL平台的，因为它们那时候还没出生呢，现在应该有了很大变化了，不过以新浪微博对技术的重视程度，说不定还是这套在运转也说不定@^@。*/

在微博环境下，构建微博用户的个人兴趣模型是非常重要的一项工作。首先，从可行性方面而言，微博是一个用户登录后才能正常使用的应用，而且用户登录后会有阅读/发布/关注等多种用户行为数据，所以微博环境是一个构建用户兴趣模型的非常理想的环境，因为围绕某个特定用户可以收集到诸多的个性化信息。另外，从用户兴趣建模的意义来说，如果能够根据用户的各项数据构建精准的个人兴趣模型，那么对于各种个性化的应用比如推荐、精准定位广告系统等都是一种非常有用的精准定位数据源，可以在此基础上构建各种个性化应用。

事实上，新浪微博在2011年前已经构建了一套比较完善的用户兴趣建模系统，目前这套系统挖掘出的个人兴趣模型数据已经应用在10多项各种应用中。对于每个微博用户，通过对用户发布内容以及社交关系挖掘，可以得出很多有益的数据，具体而言，每个微博用户的兴趣描述包含以下三个方面：用户兴趣标签、用户兴趣词和用户兴趣分类。

用户兴趣标签是通过微博用户的社交关系推导出的用户可能感兴趣的语义标签；用户兴趣词是通过对用户发布微博或转发微博等内容属性来挖掘用户潜在兴趣；用户兴趣分类则是在定义好的三级分类体系中，将用户的各种数据映射到分类体系结构中，比如某个用户可能对“体育/娱乐明星”这几个类别有明显兴趣点。以上三种个性化数据，用户兴趣标签和用户兴趣词是细粒度的用户兴趣描述，因为可以具体对应到实体标签一级，而用户兴趣分类则是一种粗粒度的用户兴趣模型。本文主要从体系结构角度来简介用户兴趣词以及用户兴趣分类这两类用户兴趣的挖掘系统架构。

|微博用户兴趣建模系统整体架构

微博用户兴趣建模系统整体架构如图1所示，其由实时系统和离线挖掘系统两个子系统构成。因为每时每刻都有大量微博用户发布新的微博，实时系统需要及时抽取兴趣词和用户兴趣分类，而离线挖掘系统的目的则是优化用户兴趣系统效果。

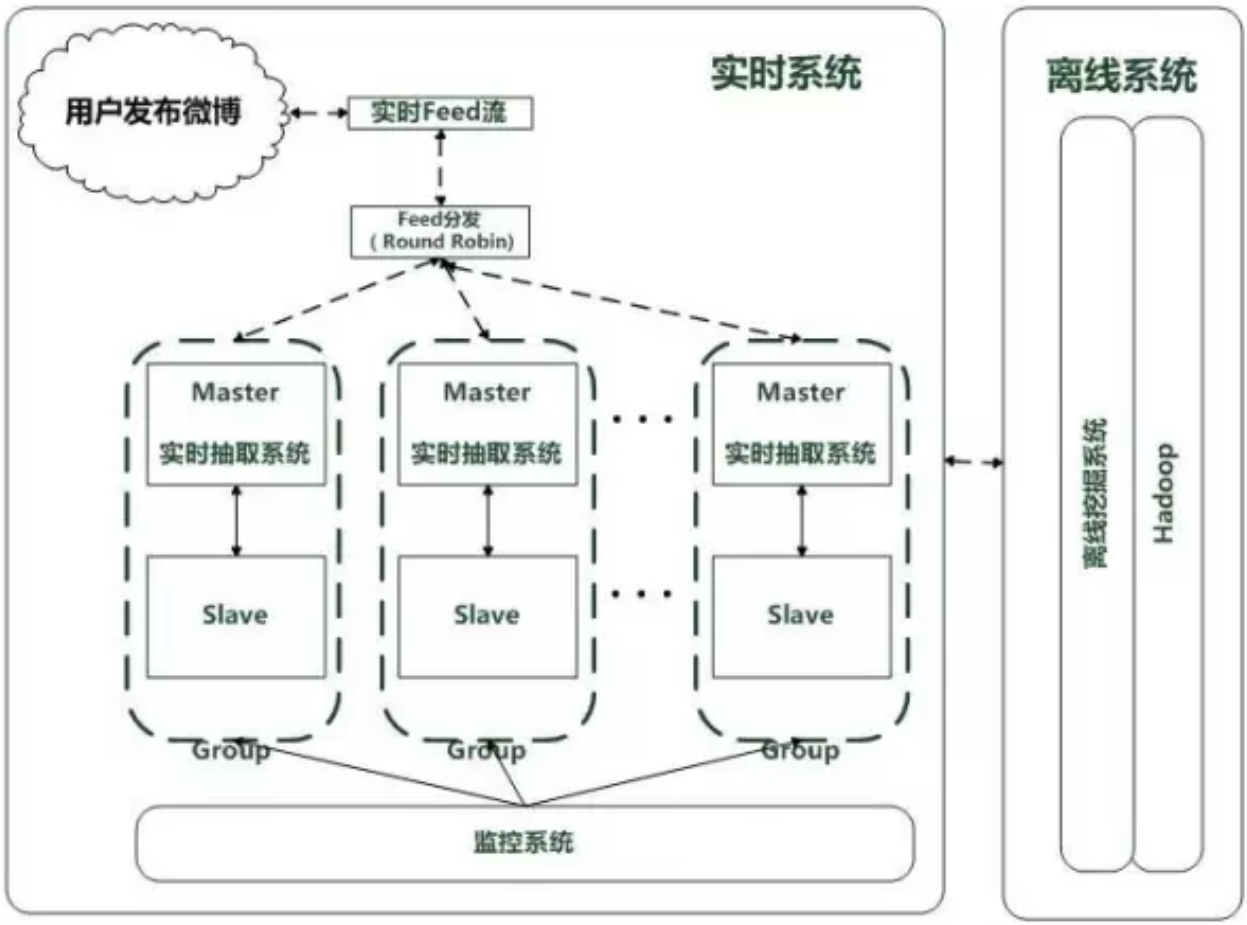

图1.微博用户兴趣建模系统整体架构统

每当有用户发布新的微博，则这条微博作为新信息进入实时Feed流队列，为了增加系统快速处理能力，实时系统由多台机器的分布式系统构成，通过Round Robin算法将实时Feed流队列中新发布的微博根据发布者的UID分发到分布式系统的不同机器中，为了保证

系统的容错性，由Master主机和Slave机器组成一个机器组，监控系统实时监控机器和服务的运行状态，一旦发现Master机器故障或者服务故障，则实时将服务切换到Slave机器，当故障机器恢复时，监控系统负责将服务切换回Master 机器。

离线挖掘系统是构建在Hadoop系统上的，通过MapReduce任务来执行挖掘算法，目标是优化用户兴趣词挖掘效果。

|实时抽取系统

对于实时抽取系统来说,每台服务器可以承载大约1亿用户的用户兴趣挖掘。当用户发布微博后，此信息实时进入原始Feed流队列中，语义处理单元针对每条微博快速进行语义计算，语义处理单元采取多任务结构，依次对微博进行分词，焦点词抽取以及微博分类计算。焦点词抽取与传统的关键词抽取有很大差异，因为微博比较短小，如果采取传统的TF.IDF框架抽取关键词效果并不好，所以我们提出了焦点词抽取的概念，不仅融合传统的TF.IDF等计算机制，也考虑了单词在句中出现位置，词性，是否命名实体，是否标题等十几种特征来精确抽取微博所涉及的主体内容，避免噪音词的出现。微博分类则通过统计分类机制将微博分到内部定义的多级分类体系中。

当微博经过语义处理单元处理后，已经由原始的自然语言方式转换为由焦点词和分类构成的语义表示。每条微博有两个关键的Key:微博ID和用户ID，经过语义处理后，系统实时将微博插入“Feed语义表示Redis数据库”中，每条记录以微博ID为key，value则包含对应的UID以及焦点词向量和分类向量。考虑到每天每个用户可能会发布多条微博，为了能够有效控制“Feed语义表示Redis数据库”数据规模在一定范围，系统会监控“Feed语义表示Redis数据库”大小，当大小超出一定范围时，即将微博数据根据用户ID进行合并进入“User语义表示Redis数据库”。

图 2 单机实时抽取系统

在用户不活跃时段，系统会将“User语义表示Redis数据库”的内容和保存在Mysql中的用户历史兴趣信息进行合并，在合并时会考虑时间衰减因素，将当日微博用户新发表的内容和历史内容进行融合。为了增加系统效率，会设立一个历史信息缓存Redis数据库，首先将部分用户的历史数据读入内存，在内存完成合并后写入mysql进行数据更新。

|离线挖掘系统

出于精准定位用户兴趣的目的，在实时抽取系统已经通过“焦点词抽取”以及历史合并时采取一些特殊合并策略来优化算法，但是通过实际数据分析发现，有些用户的兴趣词向量还包含不少噪音，主要原因在于：微博用户在发布微博或者转发微博时有很大的随意性，并非每条用户发布的微博都能够表示用户的兴趣，比如用户转发一条“有奖转发”的微博，目的在于希望能够通过转发中奖，所以其微博内容并不能反映用户兴趣所在。为了能够更加

精准地从用户发布内容定位用户兴趣词，我们通过对实时系统累积的用户历史兴趣进行离线挖掘系统来进一步优化系统效果。

离线挖掘的基本逻辑是:微博用户发布的微博有些能够代表个人兴趣,有些不能代表个人兴趣。离线挖掘的基本目标是对实时系统累积的个人兴趣词进行判别，过滤掉不能代表个人兴趣的内容，只保留能够代表个人兴趣的兴趣词。我们假设如果用户具有某个兴趣点，那么他不会只发布一条与此相关的微博，一般会发布多条语义相近的微博，通过是否经常发布这个兴趣类别的微博可以作为过滤依据。比如假设某个用户是苹果产品忠实用户，那么他可能会经常发布苹果产品相关内容。

但是问题在于：如何知道两条微博是否语义相近？更具体而言，通过实时抽取系统累积的用户兴趣已经以若干兴趣词的表示方式存在，那么问题就转换成：如何知道两个单词是否语义相近？如何将语义相近的兴趣词进行聚类？如何判别聚类后的兴趣词哪些可以保留哪些需要过滤?

我们通过图挖掘算法来解决上述问题，将某个用户历史累计的兴趣词构建一个语义相似图，任意两个单词之间的语义相似性通过计算单词之间的上下文相似性来获得，如果两个单词上下文相似性高于一定值则在图中建立一条边。然后在这个图上运行Pagerank算法来不断迭代给单词节点打分，当迭代结束后，将得分较高的单词保留作为能够表达用户兴趣的兴趣词，而将其他单词作为噪音进行过滤。

图3 兴趣词语义相似图

图3是兴趣词语义相似图的一个具体示例，通过这个图可以看出，如果用户某个兴趣比较突出，则很容易形成一个连接密集的子图，通过在语义相似图上运行Pagerank算法，语

义相近的兴趣词会形成得分互相促进加强的作用，密集子图越大，其相互增强作用越明显，最后得分也会越高，所以通过这种方法可以有效识别噪音和真正的用户兴趣。

在具体实现时，因为每次运算都是在单个用户基础上，记录之间无耦合性，所以非常适合在hadoop平台下使用MapReduce来分布计算，加快运算效率。

小结

用户兴趣建模在微博环境下有着非常重要的作用，一个好的用户兴趣建模系统可以有效支持个性化推荐、搜索以及个性化广告推送系统。本文主要从体系结构角度，简介了微博用户兴趣建模分布式体系结构，并介绍了其中比较关键的数据挖掘算法。