1、说明黑马头条技术架构与业务流

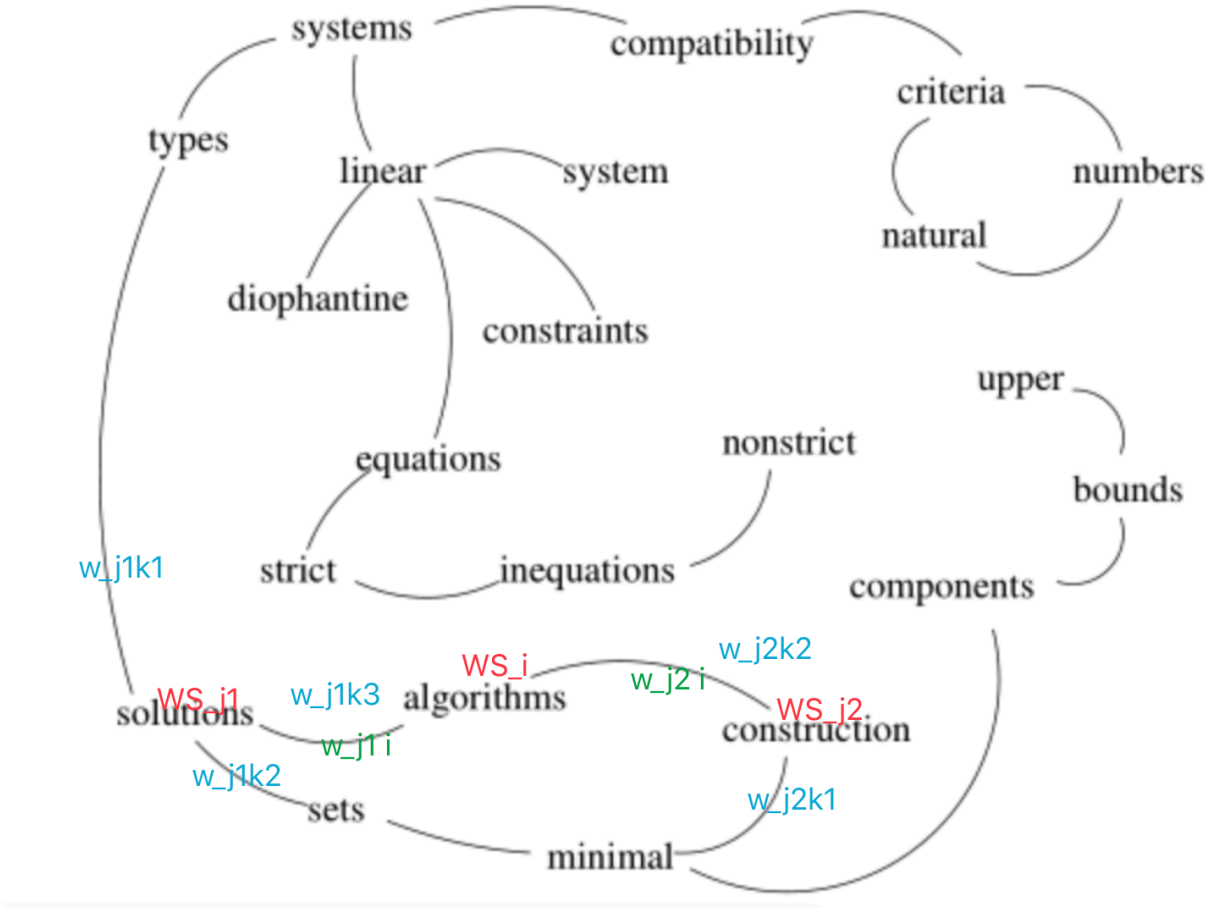
黑马头条推荐系统是使用lambda大数据实时和离线计算整体架构，利用黑马头条用户在APP上的点击行为、浏览行为、收藏行为等建立用户与文章之间的画像关系，通过机器学习推荐算法进行智能推荐。增加热门文章和新文章的推荐占比，达到千人千面的用户推荐效果。

- 1、基础数据层：包括业务数据和用户行为日志数据。业务数据主要包含用户数据和文章数据，用户数据即黑马头条注册用户的基础数据，文章数据在自媒体平台上传的文章的基本信息。用户行为日志数据主要通过flume进行收集到hdfs当中。主要技术组件flume,sqoop
- 2、基础画像计算：基于离线和实时数据，对各类基础数据计算成用户画像、文章画像。对用户、文章进行特征建立然后，计算用户的连召回集合存储, 排序即对候选集合中的文章进行用户相对的模型结果排序，生成一个排序列表。实时计算部分进行flume,kafka对接，实现在线计算结果存储hbase。主要技术als,内容召回，热门、新文章召回。排序部分：LR，wide&deep,离线计算更新使用supervisor,apscheduler实现定时更新
- 3、实时推荐：通过对外提供rpc接口来实现推荐业务的接入Feed流推荐，通过ABTest对接推荐中心，加入缓存与召回读取服务，进行实时排序逻辑。

2、请说明TFIDF原理、Textrank原理？它们的区别与联系？

TFIDF：文章中词的频率TF * IDF(词的逆文档频率)

TextRank 的迭代计算公式如下：



$$WS(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

$WS(V_i)$为词i的分数，$In(V_i)$为i词的相连词语，$WS(V_j)$为i词的相连词语的分数，$Out(V_j)$为相连词语的若干个连接权重，w_{ji}为**两个点之间的连接权重（TextRank 将共现次数作为无向图边的权值）**

- 解释：
  - 1 - d: 1 - 阻尼系数
  - d乘以某一部分为：求和{ (图中相连词语j的分数*对应边的权重) / (j的所有边的权重和)}
- 判断收敛条件
  - 临近两次误差较小：如0.0001
  - 迭代次数：100~200
  - 窗口：2~10

联系与区别：

- 联系：
  - 1、都依赖于分词的效果（业务场景的词典以及停用词的词典），大多数自然语言处理领域受影响都是分词效果
  - 2、都只是统计一些词的词频或者词的共现次数，没有基于语义方式去进行得到词的向量或者权重，没有词之间的联系
  - 通常做文本挖掘，两种方式都会做计算，
- 区别：
  - 1、TFIDF需要考虑完整的语料，语料越完全越多越好，算法简单计算量小，但是短文本应用没有textrank效果好
  - 2、TextRank虽然考虑到了词之间的关系，但是仍然倾向于将频繁词作为关键词。使用图模型迭代计算，计算量大。TextRank算法仅考虑文档内部的结构信息，导致一些在各个文档的出现频率均较高且不属于停止词的词语最总的得分较高。原因是没有考虑词语在整个语料库中的权重。
- 解决：综合TextRank与IDF
  - 1、因此在TextRank算法得到的每个词的得分基础上，乘上这个词在整个语料库的IDF值
    - IDF值是TF-IDF算法中的一个概念，该值越大，表示这个词在语料库中出现的次数越少，越能代表该文档。