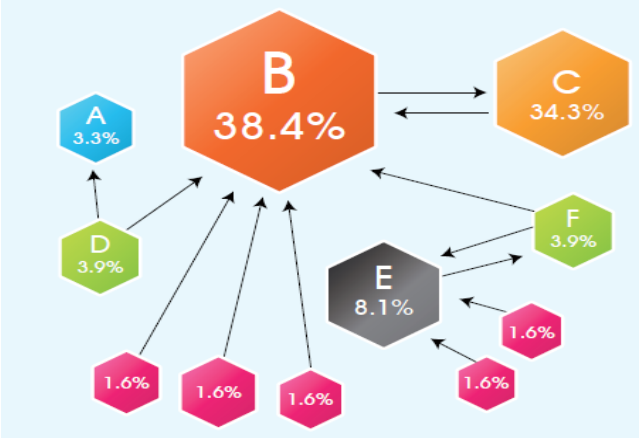
费马图计算助力搜狗提升百倍搜索精度 优化搜索体验

搜狗搜索是国内 TOP 级的搜索引擎，其搜索排序是基于 PageRank 算法实现（PR 值），数据体量达到万亿级别，常规 hadoop 分布式计算方法无法实现全部网页的检索排序。因此长期以来只能将“网页级”缩减到“网站级”或“目录级”，减少到 hadoop 能够处理的百亿级数据量。与费马科技合作后，依托费马图计算产品，搜狗搜索实现全量网页 PageRank 算法计算，从而为搜索用户提供更精准更高效的信息搜索体验。

面临挑战

搜狗搜索是搜狐旗下的子公司，负责搜索业务，成立于 2004 年，是全球首个百亿规模中文搜索引擎，收录网站达百亿规模。搜狗 2018 年其搜索相关广告营收为 10.23 亿美元，占比 91.01%，因此网页排序的质量对搜狗公司至关重要。PageRank 是网页排序的基础算法，重要性不言而喻。

由于网页数量非常多（中文网页数规模是千亿级别，链接数则达到了数十万亿），用 hadoop 处理万亿级别的网页 PageRank 是几乎不可能的挑战。因此，搜狗只能将网页级别，缩减到目录级别，或者进一步缩减到网站级别。缩到原始数据的 1%左右，降到 hadoop 可处理的规模。这种做法的代价是牺牲了 PageRank 的精度，造成网页质量的降低



然而随着自媒体数量的攀升、博客/问答等 UGC 内容的扩大，同一站点不同网页的权重差别相去甚远，举例来说，同一网站不同博主提供的内容搜索权重也理应不同。加之搜狗搜索对 UGC 内容的覆盖较为全面，不断巩固和增强搜狗搜索独家内容优势，在提供优质精准的全平台搜索服务时，其缩减的 PageRank 数据处理规模成为了瓶颈。

方案详述

PageRank 算法是由 Goole 创建人 Larry Page 提出的，它的基本假设是，对于任意一个网页：1）指向它的网页越多，那么它就越权威；2）指向它的那些网页越权威，那么它就越权威；3）它指向的网页越多，每个网页从它这里获得的权威值就越少。在实现中，将每个网页抽象成图中的一个顶点，网页之间的链接则抽象成顶点之间的有向边，然后 PageRank 算法就可以转化成一个在有向图上延着边传播“权威值”的迭代算法。由于数据量巨大，PageRank 算法往往需要大量的服务器配合进行分布式计算。而目前比较常见的，也是搜狗一直采用的做法，就是使用 Hadoop 的编写 MapReduce 程序来实现。

Hadoop 平台是一个通用的大数据计算平台，它使用 Java 编写，可以支持高达上千台机器的集群。但是 Hadoop 有个巨大的问题就是速度很慢，一方面是 MapReduce 模型本身对迭代计算不友好，另一方面它的实现也没有太多考虑效率。因此，对于全中文网页 PageRank 排序这种越大规模的计算问题，Hadoop 是无法满足的。

费马科技自主研发的 PandaGraph 是一个大规模图计算平台，它能够高效的处理超大规模图计算问题。开发 PandaGraph 的团队曾经开发出世界上最快的单机图计算系统 GridGraph，最快的分布式图计算系统 Gemini，以及规模最大的图计算系统 ShenTu。他们在 PandaGraph

中应用了大量前沿科研成果，使得 PandaGraph 成为目前市场上最快的，支持规模最大的图计算系统。同时 PandaGraph 还能与 Hadoop 平台无缝集成，直接利用 Hadoop 集群中的服务器进行计算。因此，PandaGraph 完美的契合了搜狗对中文网页全量 PageRank 的计算需求。

有了 PandaGraph，搜狗就不再需要对网页规模进行缩减，而可以直接在全量数据上进行 PageRank。在使用 100 台服务器规模的情况下，PandaGraph 在一天内完成了万亿级顶点，十万亿级边的超大规模数据的 PageRank。

搜狗搜索资深研究员田伟表示，“网页数据的量级是十分庞大的，目前业界只有通过图计算的方式能够实现全量数据的 PageRank，我们很荣幸的与业界顶尖的团队费马科技合作，不仅将网站级 PageRank 扩展到了网页级，大幅的节省了计算资源，同时其友好、强大的计算力，为搜狗在其他功能和业务的开拓方面，在底层支持系统层面带来了更多可能性，帮助我们不断创新。”