基于实时计算的Web分析系统的研究与开发

# 研究背景及意义

随着互联网的迅速发展，运营商面临的竞争日益激烈，挖掘用户行为产生的数据，可帮助运营商能更好的改善网站结构，做更好的业务和营销决策。现在很多公司每天都会产生数以TB级的大数据，如何对这些数据进行挖掘，分析成了很重要的课题。大数据时代，传统的web分析系统架构难以适应海量数据挖掘的要求。同时实时监控网站运行情况的需求也日益剧增。例如电子商务网站，需要处理并且挖掘用户行为产生的数据，产生推荐，从而带来更多的流量和收益，而每天处理海量的用户数据，需要一个低延时高可靠的实时流式分布式计算系统。再如社交网站，大家每天都会去社交网站是为了看看现在发生了什么，周围人在做什么，流式计算可以把用户关注的热点聚合，实时反馈给用户，从而达到一个圈子的聚合效果。

本课题研究工作是基于对公司网站，实现用户行为的采集，数据存储，以及实时分析。

# 与本课题相关的研究现状

## Web分析发展过程

美国网站分析协会WAA(Web Analytics Association)（现已改名为DAA:Digital Analytics Association）对网站分析的官方定义

“Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage.”

网站分析工具的鼻祖是[WebTrends](http://webtrends.com/)在1993年美国波特兰成立，基于日志分析用户行为。在1996年第一个网站计数器[Web-Counter](http://www.hitwebcounter.com/)诞生，之后在各大网站都开始出现访客计数器。 2005年Google收购Urchin之后推出[Google Analytics](http://www.google.com/analytics/)，2008年10月Yahoo!收购IndexTools后推出[Yahoo! Web Analytics](http://web.analytics.yahoo.com/" \t "_blank)，2011年Google宣布推出针对大客户的付费版Google Analytics Premium。同时国内的百度公司也推出了百度统计，并在2011年将百度统计开放给任何人使用，全部免费。

## Web分析流程

最早提出Web挖掘的是Oren Etzioni [2]，在他1996年的论文中声称，Web挖掘是利用数据挖掘技术从互联网的文档和服务中来自动发现可用信息。

--An Approach for Frequent Access Pattern Identification in Web Usage Mining

就挖掘的内容而言，web挖掘Web 挖掘可分为 Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘。

--- Formation of Algorithms Module and Dynamic Interface in Web Mining and Storage Process to Retrieve Metadata

Web使用记录挖掘，也就是本课题研究的Web分析，流程可分为源数据收集、数据存储、数据挖掘和模式分析 3 个阶段。

Web分析的关键指标包括：网站基础指标、流量质量指标、转化指标。

### 数据收集

早期数据收集主要跟服务器日志，而现在是使用用js脚本

### 数据存储

使用关系型数据库，到nosql数据库

### 数据处理

Batch定时任务，到hadoop,到实时计算。

Generally the web mining tasks can be classified into threecategories: web content mining, web structure mining and webusage mining. In addition, there are two other differentapproaches to categorize web mining. In both, the abovecategories are reduced from three to two: web content miningand web usage mining. In one, web structure is treated as partof web Content while in the other; web usage is treated as partof web Structure.

--- Formation of Algorithms Module and Dynamic Interface in Web Mining and Storage Process to Retrieve Metadata

Web usage mining includes three main steps: Data

Preprocessing, Knowledge Extraction and analysis of

extracted results.

Data Log fetch

Data clean

Pageview Identification

Computing the Reference Length

User Identification

Session Identification

**A NewClustering and Preprocessing for Web Log Mining**

Semantic Data Mining

Web Usage Mining A Review on Process, Methods and Techniques --- 匹配度很高

RTIC-C: A Big Data System for Massive Traffic Information Mining – 较高

Toward Scalable Systems for Big Data Analytics:

A Technology Tutorial 最牛逼的文章

现在很多公司每天都会产生数以TB级的大数据，如何对这些数据进行挖掘，分析成了很重要的课题。比如：

电子商务：需要处理并且挖掘用户行为产生的数据，产生推荐，从而带来更多的流量和收益。最理想的推荐就是根据兴趣推荐给用户本来不需要的东西！而每天处理海量的用户数据，需要一个低延时高可靠的实时流式分布式计算系统。

新闻聚合：新闻时效性非常重要，如果在一个重大事情发生后能够实时的推荐给用户，那么肯定能增大用户粘性，带来可观的流量。

社交网站：大家每天都会去社交网站是为了看看现在发生了什么，周围人在做什么。流式计算可以把用户关注的热点聚合，实时反馈给用户，从而达到一个圈子的聚合效果。

交通监管部门：每个城市的交通监管部门每天都要产生海量的视频数据，这些视频数据也是以流的形式源源不断的输系统中。实时流式计算系统需要以最快的速度来处理这些数据。

数据挖掘和机器学习：它们实际上是互联网公司内部使用的系统，主要为线上服务提供数据支撑。它们可以说是互联网公司的最核心的平台之一。系统的效率是挖掘的关键，理想条件下就是每天产生的海量数据都能得到有效处理，对于原来的数据进行全量更新。

大型集群的监控：自动化运维很重要，集群监控的实时预警机制也非常重要，而流式系统对于日志的实时处理，往往是监控系统的关键。

等等。

FROM <http://blog.csdn.net/anzhsoft/article/details/38168025>， Storm spark介绍。不错

试想下，如果，一个游戏新版本上线，有一个实时分析系统，收集游戏中的数据，运营或者开发者可以在上线后几秒钟得到持续不断更新的游戏监控报告和分析结 果，然后马上针对游戏的参数和平衡性进行调整。这样就能够大大缩短游戏迭代周期，加强游戏的生命力（实际上，zynga就是这么干的！虽然使用的不是 Storm……[Zynga研发之道探秘：用数据说话](http://www.csdn.net/article/2011-08-26/303631)）。

# 研究现状

## user traversals

## Frequent Pattern Tree Algorithms

# 研究目标

# 关键技术

数据不丢失

笔记：

Spark RDD可容错