基于流失计算的Web分析系统的研究与开发

# 研究背景及意义

Oren Etzioni [2]，首次提出了Web挖掘的时候在1996年在他的论文中。他声称，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。

通过这个网络服务器的分析数据，我们可以有不同的信息，如用户的上网行为这是企业营销最重要的方面，可以帮助我们的用户分析，网站设计meliorate和做更好的业务和营销决策我们网站更受欢迎和用户友好。

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

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