**1 简介**

本文档描述采集系统Uyint的总体设计。

1.1**采集系统设计背景**

网络的迅猛发展带来的是互联网信息的爆炸式增长，如何有效的获取并利用这些数据就成为了一个巨大的挑战。对于一些传统的需求，人们可以通过搜索引擎来进行数据的检索和搜集。一定程度上，搜索引擎也成为了许多普通人进入互联网的入口。然而对于商业上的一些应用，搜索引擎并不能完全满足对于数据的需求。搜索引擎的局限主要有一下几点。

1. 不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎所返回

的结果包含大量用户不关心的结果。

1. 通用搜索引擎的服务器资源有限，对于特定源的覆盖率并不能保证。
2. 互联网信息的数据结构多样化，搜索引擎对于复杂数据结构无法做到通用解析。
3. 通用搜索引擎基于关键词检索，对于特定的语义检索支持较差。

为了解决上述问题，定向的采集系统就应运而生了。定向采集系统是针对业务所需要的特定数据而进行一定的业务数据采集。例如业务需求某个特定门户频道下的新闻信息；某个社区内部的帖子信息；社交网站的评论信息等等。这些特定的数据类型都是搜索引擎所不能覆盖到的。而定向的采集系统则能够根据其需求所要网页不同的数据结构进行定制化的解析，从而将业务所需数据进行采集。

2．采集总体说明

2.1 采集系统概述

2.2 采集需求