



Anchiva 数据云 URL 过滤白皮书

内容概要

互联网内容的复杂、庞大与易访问性，促进社会进步与人类文明发展的同时，也给企业带来了生产力和资源的损失与浪费，这就需要一种技术来管理过滤互联网上各种各样的内容，于是诞生了URL过滤技术。URL过滤不仅有利于保障企业的生产效率，防止网络沉迷与滥用，而且是网络安全架构的一个重要组成部分。传统的URL过滤技术再也无法跟上当今web互联网数据的庞大规模、复杂性以及急速的增长速度。Anchiva的数据云URL过滤技术应运而生，面临日益多样化的web数据不仅最大程度的满足了实时分类的需求，而且真正做到了与语言、文化、地域无关，只与具体用户实际使用请求相关的URL过滤，将用户的需求和利益放在第一位，同时每一个用户也不是完全孤立的，通过云数据中心有效的结合在一起，共享互动成果，符合人类文明的进步，符合时代的进步。

URL 过滤需求

当今，互联网给我们带来了生活方式、生产方式上很多的便捷与好处。庞大的网络信息资源，使人们通过浏览器就能方便的浏览或获取到生活或生产过程中所需要的信息。但是，过分或不适当的网络资源访问不仅给企业带来了生产力和网络带宽的损失，还严重威胁着企业的网络安全，甚至网络上的不适当或非法内容还极大危害着个人的身心健康甚至给企业带来法律问题。具体如下：

1、非业务网站的过渡访问，带来生产力的损失

互联网内容的易访问性，使企业员工可以很方便的浏览到非商业的材料，从阅读新闻到网上购物，更有甚者，进行网络游戏、网络炒股等。显然会对正常的工作带来影响。同时，某些员工在上班时间浏览与工作无关的内容会给其他的同事带来非常不好的影响，容易导致公司其他员工的效仿，甚至引起整个团队士气的低落，这将给企业生产力带来极大的损失。

2、大带宽网络资源的下载浏览与使用，给企业带来带宽损失，甚至带来带宽使用极限

员工利用开放的网络可以很轻松的下载或观看音频和视频资料，尤其是P2P技术的广泛应用，极大的挤占着企业的网络带宽，甚至带来带宽使用上的极限，致使企业的正常网络运营与正常网络服务或访问出现问题，从而进一步影响企业的生产效率。

3、不经意恶意站点或URL的访问，威胁着企业的网络安全

如今，网站已经成为首要的病毒或恶意软件发布源。Gartner曾经报道，2008年第一季度，50%以上的合法网站曾经遭到过黑客的恶意行为。如此一来，员工不经意的web访问就有可能遭受到病毒或恶意软件的侵袭，进而对整个企业的网络安全带来威胁。

4、不适当的内容浏览，极大危害着个人的身心健康

许多网页包含不适当的内容，例如，色情、淫秽、邪恶、暴力等内容。不能自控的过多的访问或观看这些不适当的内容，将极大危害着个人的身心健康，甚至走上犯罪的道路。

5、非法内容的浏览，给企业带来法律上的麻烦

反社会观点和论坛、制造炸弹知识、极端政治文学、黑客工具等信息。当人们达到这些网页时，实际上色情或其他不适当的内容被浏览，在公司内，这容易招来公司的违法诉讼。



面临以上种种问题或麻烦，企业或单位该如何解决？全部关闭网络的大门不符合现代人类文明的发展与进步，也不利于合理商业业务的开展。单纯依靠使用者自我约束的网络使用制度已经无法执行。于是，诞生了专门针对 URL 进行过滤的技术，以此达到对员工上网浏览内容的控制与管理。

URL 过滤技术的发展

20 世纪 90 年代中期，URL 过滤解决方案依靠企业内部 IT 人员人工建立、更新与编辑站点黑白名单。这一做法的缺点是，所有分类由一个或少数几个人自由决定，对于这种资源密集型而且缺乏客观性的站点分类方法，不仅会使许多被认可的网站被封锁或被禁止的网站允许通过，而且随着 web 站点的快速增长与相关技术的日益复杂，这种方案很难实现客观、细粒度的 URL 分类，显然不能成为企业或单位有效的 URL 过滤方案。

20 世纪 90 年代末，出现了专门对 URL 进行收集、分类的厂商。URL 过滤技术开始采用本地数据库分类引擎。URL 及其内容在根据预先定义的类别下通过分类引擎进行相应的关键字查找分析与分类（如赌博、色情及网上购物等），分类好的 URL 存储在一个集中的主数据库中，然后通过更新复制一份副本移交到客户本地数据库中。这种 URL 过滤方案的缺点是，随着网页数量的激增，由于一刀切的关键字分类技术和本地分类数据库的限制，无法实现更高、更准确的覆盖率和更广泛的 URL 分类。

2000 年初，URL 过滤解决方案试图采用启发式内容分析的方法，这种动态的分类技术，通过智能分析网站标题和网页 html 主体中相关内容的概率来确定 URL 类别。从理论上讲，相比前两类 URL 分类方案，这是一种很好的分类方法，然而在实际中它本身却存在问题，很多基于启发式的 Web 内容分析结果没有相关的配套技术实时地发送给终端用户，而且采用的仍然是传统的本地数据库进行存储。但是当今 web2.0 时代，web 数据是一个不定数据且日益多样化的集合体，而每个用户的需求却独特且具体，基于本地 URL 数据库的过滤技术，只能过滤存储本地用户需要的数据，不能存储所有相关及最新数据，以执行快速和准确的监测，因此这种传统的 URL 过滤技术也无法应对高度复杂且快速发展壮大的 web2.0 网络。

据Google调查，互联网上的网页数量以每天一亿的数量急速增长。以上三种URL 分类方法已经不能够准确有效的收集、分类所有的URL类别。数据存储和处理要求也已经远远超出了本地数据库能力。于是在2009年，业界出现了数据云的URL过滤技术，这种数据云URL过滤机制，基于云技术的URL收集、分类处理及发放策略，并不依赖于本地数据库有限的资源进行分析与检测，也不依赖于数据库更新最新的URL分类，利用的是专门的分类服务器群，根据实际网络的使用与普及方式对网页内容及语义进行全面分析后的分类。与传统的云不同的是，真正做到了云的客户端自动主动地去云的服务器端获取所需的数据，而不是单纯的基于云服务端的定时推送更新方法。下面我们对几种URL过滤技术做以下对比总结：

	90年代中期 自分类黑/白名单	90年代末 本地黑/白名单	2000年初 启发式检测分类	2009年 基于云的URL过滤技术
分类技术	企业IT人员人工分类	关键字查询分类引擎	启发式的关键字概率分类技术	完整的web内容及语义分析技术
存储方法	黑白名单文档	本地数据库	本地数据库/云端服务器群	云端服务器群/本地缓存
更新方法	人工编辑更新黑白名单	复制数据库副本更新方法	服务器定时推送更新	客户端随时获取URL分类更新
准确性	差	一般	较好	极好
覆盖范围	差	一般	较好	极好
总结	缺乏客观性的分类方法，资源集中，不准确。	误报、漏报率高，互联网的增长速度远远超过了本地数据库存储能力。	服务器不能实时推送准确的分类，没有客户端随时获取技术，客户端使用本地数据库存储，容量有限。	无处理性能和本地数据库存储限制，先进的本地缓存自动学习机制，能够满足每个客户独特且具体的需求。



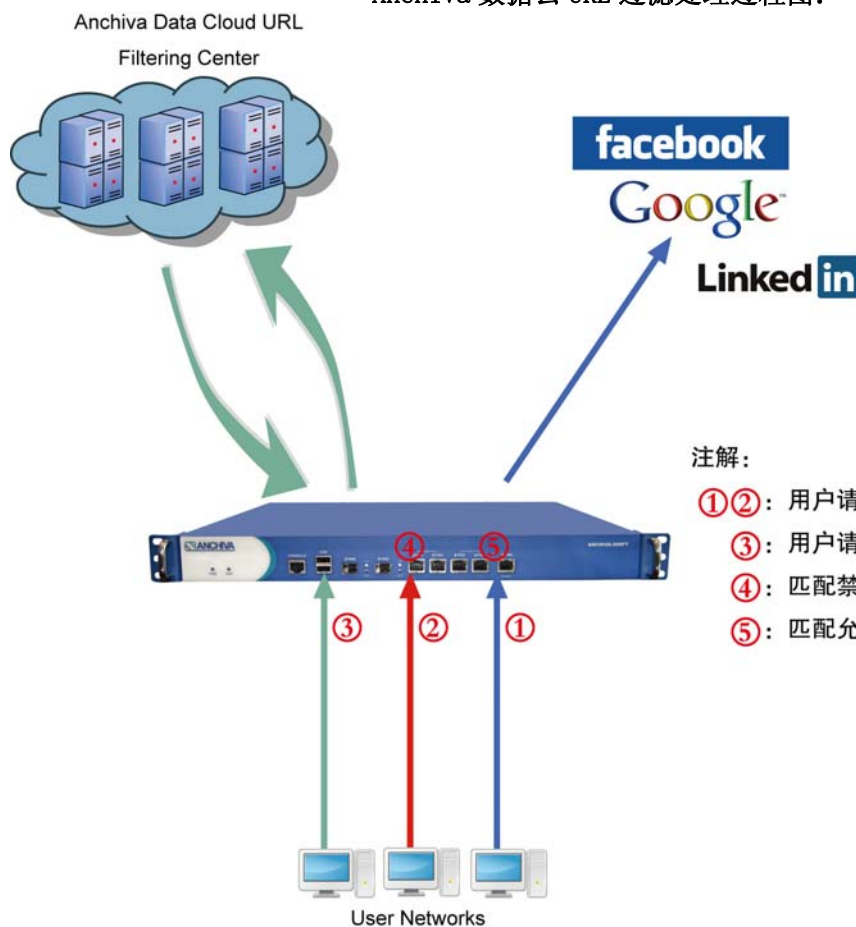
Anchiva 数据云 URL 过滤运行机制

Anchiva 数据云的 URL 分析过滤技术由两部分组成：部署在企业网络边界处的 Anchiva web 安全网关（SWG）和 Anchiva 基于云的 URL 分类中心。网关设备和 URL 分类中心实时通信获取最新的 URL 分类。不仅打破了传统本地数据库的限制，而且与其他基于云的技术不同的是 Anchiva web 安全网关中具有 URL 缓存技术，这个本地的缓存为每一个独立的用户存储最为相关的 URL，这些设备本地的 URL 类别，有效地确保了 URL 匹配的最佳性能，并且采用的是一种本地缓存自动学习的机制，随着企业用户使用时间的增长，这个本地缓存的 URL 库会更加的贴近每一个用户的实际需求，将能给客户更加精准的 URL 过滤。

以下是 Anchiva web 安全网关 URL 过滤对 HTTP-GET 请求的处理过程：

- 1、Anchiva web 安全网关的 URL 匹配处理引擎接受来自本地用户的 HTTP-GET 请求。
- 2、URL 匹配处理引擎首先从设备上的本地缓存中查找相关的 URL 分类。
- 3、如果 URL 匹配处理引擎从本地缓存中查找到了正确的 URL 分类，那么将该 HTTP-GET 请求根据客户设定好的相关过滤策略进行允许或阻止的操作。
- 4、如果没有在本地缓存中找到正确的 URL 分类，URL 匹配处理引擎会自动将该 HTTP-GET 请求发送到 Anchiva 数据云 URL 分类中心。
- 5、Anchiva 数据云 URL 分类中心将自动查询并返回正确的分类给设备的 URL 匹配处理引擎。
- 6、URL 匹配处理引擎根据 URL 分类中心返回的分类对该 HTTP-GET 请求按照客户设定好的相关过滤策略进行允许或阻止的操作，并在设备本地缓存的 URL 类别中添加相应的 URL 分类。

Anchiva 数据云 URL 过滤处理过程图：



注解：

- ①②：用户请求访问本地缓存中存在的URL
- ③：用户请求访问本地缓存中不存在的URL
- ④：匹配禁止策略被阻断
- ⑤：匹配允许策略，正常访问