

服务电信行业 打造生活搜索

## TRS 电信 114 企业搜索引擎解决方案

TRS 公司结合电信企业建设 114 搜索引擎的需求，应用多年来自主开发的 TRS Database Server 作为企业搜索引擎服务的平台，配套 TRS 文本挖掘基础件、TRS 网络信息雷达和数据内容分发服务模块，向电信行业推出智能、安全、跨平台、个性化的电信行业 114 企业搜索引擎解决方案。此方案已经在中国电信全国中心 114 企业搜索引擎项目和上海电信号码百事通搜索引擎项目中获得了成功应用。

## 目录

1. 概述.....	3
1.1 方案背景.....	3
1.2 企业搜索引擎与互联网搜索引擎的区别.....	4
1.3 电信 114 企业搜索引擎建设需求分析.....	8
1.4 电信 114 企业搜索引擎建设目标.....	11
2 TRS电信 114 企业搜索引擎解决方案架构.....	11
2.1 系统整合电信多种信息来源，支持内容实时增量索引.....	13
2.2 集群架构支撑大规模部署应用，支持内容海量安全管理.....	14
2.3 应用垂直搜索实现内容挖掘分析利用，支持电信企业开发特色搜索服务.....	16
2.4 搜索高效、准确，向用户提供智能个性搜索体验.....	17
3 方案特点和优势.....	18
3.1 “安全”的搜索引擎.....	18
3.2 更高的搜索准确性和智能性.....	18
3.3 个性化的搜索体验.....	19
3.4 强大的异构资源整合搜索.....	19
3.5 标准、开放的系统，提供强大的系统扩展能力.....	19
3.6 具有充分满足需求的自主核心技术和产品.....	20
3.7 采用先进的搜索引擎技术.....	20
3.8 基于内容的自动分类和聚类技术.....	21
3.9 基于内容的信息去重技术.....	22
3.10 优异的全文检索性能.....	23
3.11 成功的应用模式和丰富的应用经验.....	26
3.12 专注的服务.....	26
4 联系方式.....	28
5 版权声明.....	29

## 1. 概述

### 1.1 方案背景

2006 年，随着中国互联网搜索市场的迅猛发展，百度、Google、雅虎、搜狐、搜狗、新浪爱问、中国搜索等国内外搜索引擎在中国市场展开了空前激烈的竞争。在搜索市场的巨大市场价值的吸引下，国内电信运营商纷纷推出向综合信息服务提供商转型的战略举措，加入争夺搜索市场份额的行列。

2006 年 6 月，中国电信全面升级 1 1 4 查号业务，推出了全新的号码百事通业务，为将近 8 亿的电话用户提供方便、快捷的综合信息服务。号码百事通立足于百姓的衣、食、住、用、行、乐，着眼于生活的便利、便捷，致力于为用户提供综合信息服务。在中国电信瞄准这一巨大的潜力市场的同时，国内另一固网运营巨头中国网通，也开始在其北方 10 省推广类似的电话搜索引擎服务。重量级的电信运营商加入搜索引擎市场的争夺，预示着 2007 年搜索市场将产生风起云涌的变化。

目前用户对互联网服务的使用、获取方式的变化以及服务商为此进行的技术创新，是搜索形态变迁的关键，也是不同搜索服务商竞争的核心。作为互联网产业发展最重要的方向之一，以 Google、百度为代表的第二代搜索服务实际上与互联网服务整体发展的第二个阶段是相互对应的。第一阶段是网站呈现，目录分类；第二阶段是内容交互，网络搜索；第三阶段是任意聚合，个性搜索；第四阶段是自由交互，智能发布与搜索。第三代互联网搜索引擎将在个性化、语义智能分析、搜索结果优化等方面取得明显进步。

虽然电信运营商推出搜索引擎有其本身的品牌和资源优势，但是常规的互联网搜索模式已经被成熟的搜索引擎深度挖掘，百度、google 等搜索巨头长期积累的竞争优势却不是短时间就能赶超。那么电信运营商的新搜索引擎如何才能立足搜索市场并在局部胜出呢？其关键在于深入整合挖掘电信运营商的企业内

部网络资源和信息资源优势，侧重于特定关键领域提供特色搜索服务，为用户提供真正有价值的信息。

对于中国电信和中国网通而言，114 巨大的品牌资源、庞大的用户群、广泛的知名度是电信运营商实现信息服务平台的最佳载体。通过 114 平台的发展，刻意逐步将基于语音的增值服务嫁接到统一的平台上来，并为客户提供提供衣、食、住、用、行、乐等方面便利快捷的综合信息。目前，中国电信的号码百事通业务和中国网通的 114 电话导航业务，在国内各省区陆续展开。而作为该业务核心的 114 企业搜索引擎更是 2007 年各省市电信公司建设的重中之重。

北京拓尔思 (TRS) 信息技术有限公司是国内企业搜索引擎和内容管理软件的领导厂商，公司在企业搜索引擎领域占据着国内企业级搜索引擎市场的 70%。TRS 公司结合电信企业建设 114 搜索引擎的需求，应用多年来自主开发的 TRS Database Server 作为企业搜索引擎服务的平台，配套 TRS 文本挖掘基础件、TRS 网络信息雷达和数据内容分发服务模块，向电信行业推出智能、安全、跨平台、个性化的电信行业 114 企业搜索引擎解决方案。此方案已经在中国电信全国中心 114 企业搜索引擎项目和上海电信号码百事通搜索引擎项目中获得了成功应用。

## 1.2 企业搜索引擎与互联网搜索引擎的区别

搜索引擎的出现，整合了互联网上众多的网页资源，并提供信息导航和信息查询服务，使信息的价值得到了网民和厂商的普遍认可。一提到搜索引擎，就自然联想到互联网搜索引擎，再加上一些厂商刻意的推波助澜，造成了互联网搜索引擎取代所有搜索引擎的概念。而实际上我们可以看到不同搜索引擎之间的差别很大。

TRS 电信 114 搜索引擎是以 TRS 的企业级搜索引擎为基础的。TRS 所说的企业搜索引擎 (Enterprise Search Engine, 简称 ESE) 中的企业并非指单纯的企业，政府、教育、科研、媒体、医疗、军队、安全部门都有类似的应用需求，这里的“企业”可以理解为“企业级”，即企业级搜索引擎。那么，对于企业级搜

索，我们对“搜索”的诉求又是什么呢？和互联网搜索引擎相比，它又有哪些不同呢？

实际上，搜索引擎服务是内容管理技术的一个典型应用。我们不妨从内容管理的框架来看搜索引擎的各个环节，即从信息内容的采集，加工，管理，到服务，以至到信息内容的“发现”来比对一下企业级搜索引擎的不同。

	互联网搜索引擎	企业级搜索引擎
异构资源 搜索和整合	互联网通信协议  以HTTP传输协议为主获得的HTML和特殊格式文档（DOC、PPT、PDF、MP3、图片等）	企业环境下各种信息采集接口  HTML/XML（HTTP） RDBMS（API/SQL） 文件系统(NFS、FTP) Office/Lotus OA/Instant Communication Enterprise Application.....
数据 实时更新	更新周期长， 静态缓存的索引，周期切换	企业信息更新需要即时反映 动态更新索引，保证数据一致性
准确性 相关性	不可能查全 相关性排序以Page Rank、Title、Meta为主 面临SEO问题和商业性因素	更全面 精确计算，字、词混合索引；复合元数据查询（结构化特征） 更准确、排序更合理
安全性	公开信息，不存在安全问题	访问权限控制非常重要
管理、挖掘和 应用	找到信息后服务完成	需要完备整合和管理 智能挖掘分析（各种分类、聚类、提取手段） 安全开放接口支持其他应用系统 面向企业需求，个性化服务

\*SEO：搜索引擎优化，利用工具或其他手法夺取较好的网络排名。

### 1、复杂结构数据的搜索

互联网上搜索的数据一般都是网页形式的，尽管这几年网上丰富起来的图片、MP3 等信息形式，但其组织形式仍是基于 HTML 组成的网页。而企业级用户需要搜索的数据既有互联网站上的，也有内部网站上的；既有网页形式的，又有各种数据库形式的，如 SQL Server、Oracle 数据库等；既有结构化数据，又更多的是各种电子文件格式的非结构化及半结构化数据，如 Word、Excel、Lotus Notes、PDF、XML 等；既有文本形式的数据，又有多媒体形式的数据；而且，同一机构的数据还可能分布在不同介质的载体上。

然而，不管数据的形式、来源、位置、平台如何不同，企业用户总是希望内外数据能无缝结合，用一个搜索工具和统一的界面，发出几个简单的检索请求就能对所有资源进行检索，并很快就能有满意的结果。

并且，互联网搜索内容对于用户来说都是未知的，而企业级搜索的对象基本上是已知信息源，其中包括企业资料库、目录、帮助文本、源代码信息库、新闻组等，在对这些信息进行索引时，用户需要按照内容而不是通过比较源链接来进行排列。

## 2、严格的安全搜索

在企业内部，安全的问题是无法回避的。因为企业内部的信息不象“人人平等”的互联网信息，其信息内容带有明显的“等级”安全特性。所以，当搜索技术变得无所不能，人们反而开始担心，如果搜索的结果泄漏了企业的机密怎么办？如果企业原有的安全架构对新的搜索技术失效了怎么办？这些疑问都让用户感到如鲠在喉，岌岌小心。

很多业内人士在谈到搜索安全的话题就忧心忡忡，他们普遍认为搜索环境并没有为企业级应用做好足够的准备，未来充满太多的变数。而在一些实际的应用中，我们看到，即便为数据定义了文档级和数据库级的双重安全保障，搜索引擎的“魔爪”还能透过授权的索引文档来“搜索”它们。

因此，针对企业网中不同的用户对不同的资源，其使用权限都可能不一样，需要企业搜索引擎能够对用户、资源、权限分级管理和控制，确保系统的安全。

## 3、高可靠的查全和查准

作为专业用户，企业用户需要查找的信息专业性强、概念复杂，而对查询的查全率和查准率有着非常高的要求。因此，需要利用各种手段来提高搜索引擎的查准率和查全率。

从查全率来看，互联网搜索引擎无从谈起查全率，因为互联网上的信息如此泛滥无边，任何一个搜索引擎服务商都无法穷尽互联网上的每个网页。而在企业级的某些应用中，是不允许有所遗漏的检索。必须对企业内部每个需要提供服务的信息进行索引。在检索机制上必须保障效率的前提下达到全面搜索的要求。

同样的道理，在互联网上因为信息自由的特点，决定了搜索只能通过“关键词匹配”这种核心检索手段去实现。而在企业内部，信息的组织复杂了许多。企业级搜索引擎有完善的信息分类体系，元数据，对象数据多层逻辑的组织形式，在查询上满足基于对象数据内容和元数据标引体系的精确查询要求。

#### 4、智能化的检索服务

企业内部的搜索服务，带有鲜明的业务特性，不像互联网搜索引擎仅提供信息参考。在企业内部的搜索结果将直接参与到企业的运营、决策中。所以，对于搜索的结果处理，搜索过程中采用相关智能技术以达到迅速、准确、全面定位目标信息非常重要。例如采用相关度分析技术，使相关度较高的结果排在结果列表的前面，相关度较低的结果排在后面，并屏蔽无用和错误的信息；构造强大的语义规则库，使系统能够正确地判断与检索词相关的同义词、近似词、上位词、下位词，帮助用户判断结果的相关度，并进行进一步的查询；支持完善的信息分类体系，对检索结果自动分类或者信息聚类；提供智能化的概念扩展查询等，都将有利于企业对信息资源的高效利用。

#### 5、企业搜索引擎通常都和企业其他的 IT 应用有机结合

以内容管理技术为框架，搜索技术为支撑，企业搜索引擎通常与数据管理、内容管理、记录管理、竞争情报、团队协同、过程管理、信息门户等知识管理的各个环节紧密结合，构成管理企业知识资产的完整而又灵活的体系。知识内容管理对搜索引擎技术提出了更高的要求，而先进的搜索引擎技术则为知识内容管理提供了工具和保障。在市场上我们也可以看到，国内外企业级搜索引擎厂商，有许多也是知识内容管理解决方案的提供商。

## 6、实时的信息搜索服务

正如前所叙，企业内部的搜索服务，具备业务特性，需要将搜索结果参与企业的运营和决策。所以通过搜索引擎提供的服务，必须能够动态地反应实际情况，即当内部的信息发生变化时，必须能够实时反应。在企业，不允许出现像互联网搜索引擎服务那样信息滞后更新的现象。

### 1.3 电信 114 企业搜索引擎建设需求分析

2007 年，要想在搜索市场占领先机，就需要使电信 114 企业搜索引擎具备差异化的竞争优势。如何培养用户使用习惯？如何聚集商业客户合作伙伴？如何实现平台搜索技术的改造升级？这些成为是电信公司建设 114 搜索引擎需要面对和解决的问题。目前各省市基本都建设了号码百事通和 114 电话导航的平台。

从客户需求角度分析，114 搜索引擎用户存在前向查询客户和后向被查询客户两种角色，这两种客户对于号码百事通业务有着不同的需求和价值：

从前向查询客户角度分析：

首先，前向查询客户拨打 114 是要获得能够解决衣食住行各类生活问题的相关线索，具体而言，就是提供各类服务组织的电话号码。

其次，用户获得信息线索后需要进行一定的选择比较，即客户为了最终解决某个问题或完成某件事情，需要把获得的信息、线索进行比较分析，找到最佳路径。

再次，进行订单交易，即客户通过比较选择后，确定了对象，有直接转接或者预订的需求。在现代的信息社会里，“预先确定”已经成为人们享受某项服务之前的必要环节。

最后，完成服务，即客户实现最终消费、解决问题或者完成服务的过程。在这个环节中，客户可能会用到电子支付，也会有服务质量反馈等后续事务。

由此可见，前向客户需求链的存在，为 114 搜索引擎业务提供了良好的发

展空间。正是为了满足前向客户需求、最大限度地为客户提供“一站式”便捷服务，号码百事通需要建立丰富的本地生活类信息数据库，为前向客户提供查询转接、短信播报等业务，从而获得广阔的号码信息增值服务新市场。

从后向被查询客户需求角度分析：

与前向查询客户的需求环节相对应，根据需求层次不同，后向被查询客户的需求分为三类：

第一类，后向客户需要将 114 平台作为信息发布的媒介。中国电信 114 有着广泛的客户群体，像使用电视、报刊、互联网这些主流媒体一样，政企客户需要尽可能地在 114 这一语音媒体上发布更多的信息。中国电信 114 能够以其诚信、高品质的品牌形象，为政企客户提供广传播、可信赖的语音信息发布平台，能够让更多的客户了解后向客户的服务能力和企业形象。

第二类，后向客户需要 114 成为企业的一个营销渠道。高品质企业客户看重中国电信良好的品牌形象和 114 “一对一”信息传递的特性，希望 114 能够成为其强有力的营销渠道。传统的企业营销采用公共媒体广泛行销，近来越来越多的企业开始注重精准信息的分众传递，即根据目标客户群的年龄、职业、地域特征，选择信息投放范围和信息投放方式，以期把营销做得快速、准确、有力度。例如，“分众传媒”瞄准中国高中端商务人士做专业楼宇广告联播，在纳斯达克成功上市；

第三类，后向客户需要 114 成为其业务交易平台。企业使用 114 进行宣传、营销之后，需要进一步延伸服务，尽可能地促成交易。信用卡、电子支付的盛行，为电话支付、在线交易（bizon—line）提供了坚实的基础，114 延伸为业务交易平台，可以更大程度地为前后向客户提供服务。

可见，后向客户的需求同样为号码百事通业务提供了广阔的发展空间，号码百事通诚信、便捷、广泛服务的品牌形象，能够为优质政企客户提供良好的宣传、营销和交易平台。

业务的丰富也对 114 企业搜索引擎提出了更高的建设要求。大致需求有以下几点：

1. 对业务应用所需的企业信息库急需快速补充、整理和完善，才能为前端查询用户提供有价值的信息服务

在业务应用方面，114 企业搜索引擎的业务应用主要包括优先报号、实名查询、品牌查询、临时报号、查询转接、短信报号、话务呼转、企业名片、企业广告、指路服务、个人号簿、企业总机、签约客户分析、注册客户分析等 14 大类。

这些业务应用推广的关键是必须保证数据库中的信息必须准确、有效、及时。目前从全国范围看，各省电信公司的企业信息库还需快速补充和完善，以应对日益增长的外部电话和网络搜索的需求。

2. 对于庞大的信息资源库和数据库内的信息要进行特定领域的挖掘和搜索

由于基于 114 转型的电话搜索业务，主要是向前端用户提供准确、及时、深入的衣、食、住、行、乐等相关的综合信息服务。因此针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息和相关服务，才能真正获得目标用户的青睐。所以需要建立需要具有特色垂直搜索功能的 114 企业搜索引擎。

3. 跨地域、跨业务、跨平台的信息共享不到位，无法发挥规模效应

全国各省市的电信公司分布区域大，业务应用众多，公司内部存在各种数据库和信息平台。企业内部的数据和信息没有进行很好的共享和利用，所以无法发挥全国电信公司的规模效应；114 企业搜索引擎应实现对各省本地和跨区域的信息查询，并支持对互联网信息的查询。

3. 平台需要聚集并支持大规模的商业合作伙伴运营，后台商务信息的互动发布，竞价排名。

电信行业 114 企业搜索引擎，不仅需要满足前向查询用户查询需求，而且为后向被查询企业客户创造了优先接触目标消费者和发布商务广告的机会。

4. 各省电信公司都需要开发自己的特色业务，因此要求平台具有良好的扩展性。

为了解决上述问题，需要各地电信 114 企业数据信息库的数据模型统一并且进行升级改造，建设全国统一搜索引擎以实现全国 114 信息的共享和全国跨地域信息查询。

## 1.4 电信 114 企业搜索引擎建设目标

114 企业搜索引擎的建设将会延伸 114 的功能，丰富 114 的信息服务内容和形式，力争把 114 做成语音搜索领域的 Google。

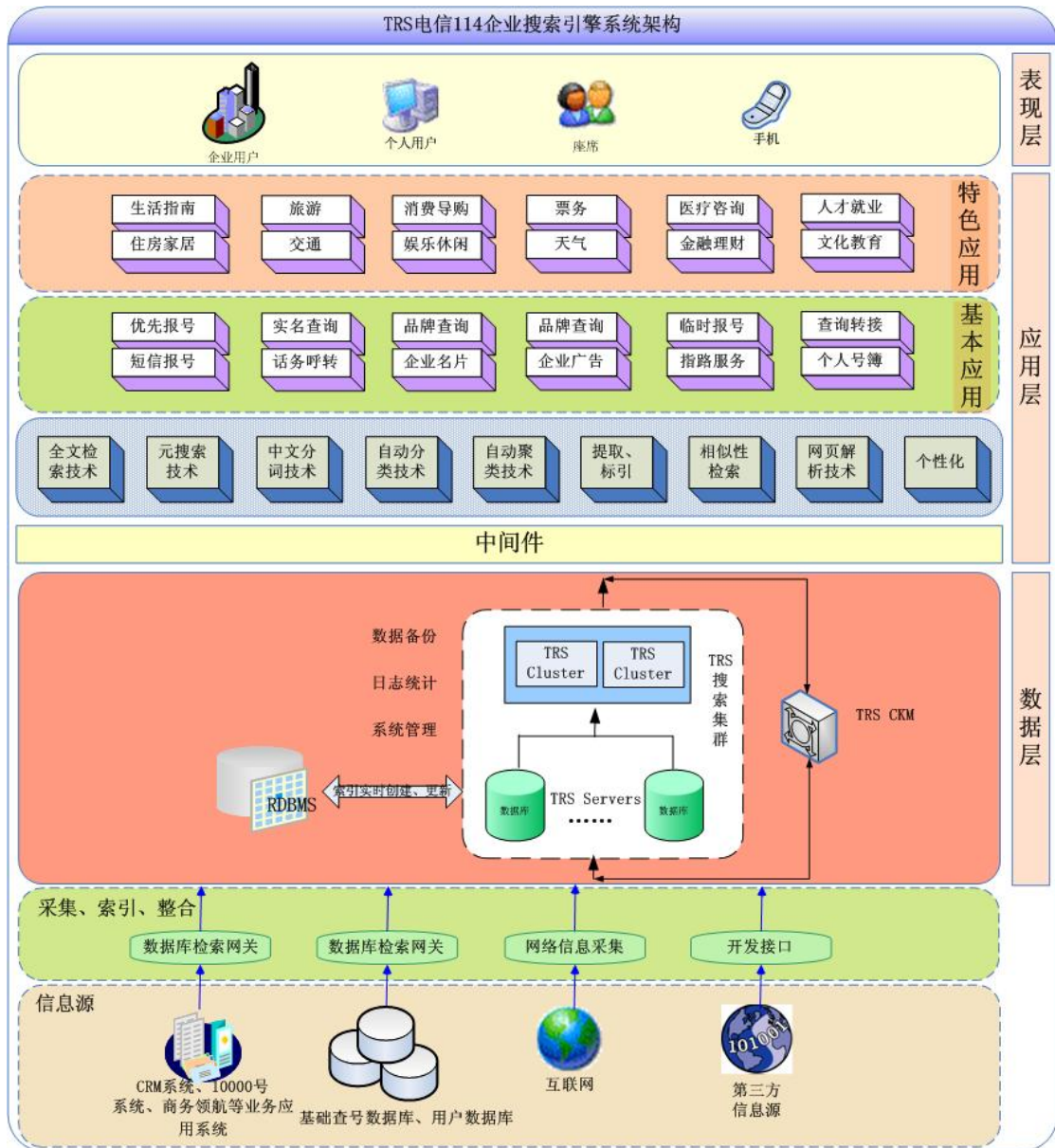
电信企业建设 114 搜索引擎建设工程将实现以下目标：

- (1) 帮助电信公司建设全国统一 114 企业搜索引擎，实现跨省查询业务，并为没有建设搜索引擎的省份提供省内 114 企业搜索引擎。
- (2) 支持对各省市电信公司 114 业务应用相关内部数据库和信息资源库进行补充完善，并进行个性化服务的信息挖掘和整合。系统支撑大规模用户跨地域和跨平台搜索。
- (3) 114 企业搜索引擎具备开发特定领域垂直搜索的能力，各地电信公司可以在此平台上进行二次开发。
- (4) 聚集商业客户及合作伙伴，满足企业客户竞价排名、信息发布、商机获取、资源共享等的双向需求。

## 2 TRS 电信 114 企业搜索引擎解决方案架构

TRS 公司作为国内企业级搜索引擎和内管理领域的领导软件厂商，一直在该领域拥有先进的理念、成熟的产品和先进的信息检索、内容管理和文本挖掘技术。TRS 电信 114 企业搜索引擎 2006 年在中国电信号码百事通全国中心搜索引擎项目，和上海电信号码百事通搜索引擎的基础上得到了成功的应用，积累了丰富的行业实践经验。这些项目实践对全国各地电信企业建立或升级 114 企业搜索引擎也具有良好的示范意义。

TRS 公司结合电信行业 114 业务应用需求，依托自身企业搜索引擎产品和中文信息处理技术，推出的电信 114 企业搜索引擎解决方案，能够全面整合索引搜索电信 114 业务应用的信息内容，并向用户提供高效的、准确的、安全的、个性化的搜索体验。其框架图如下：

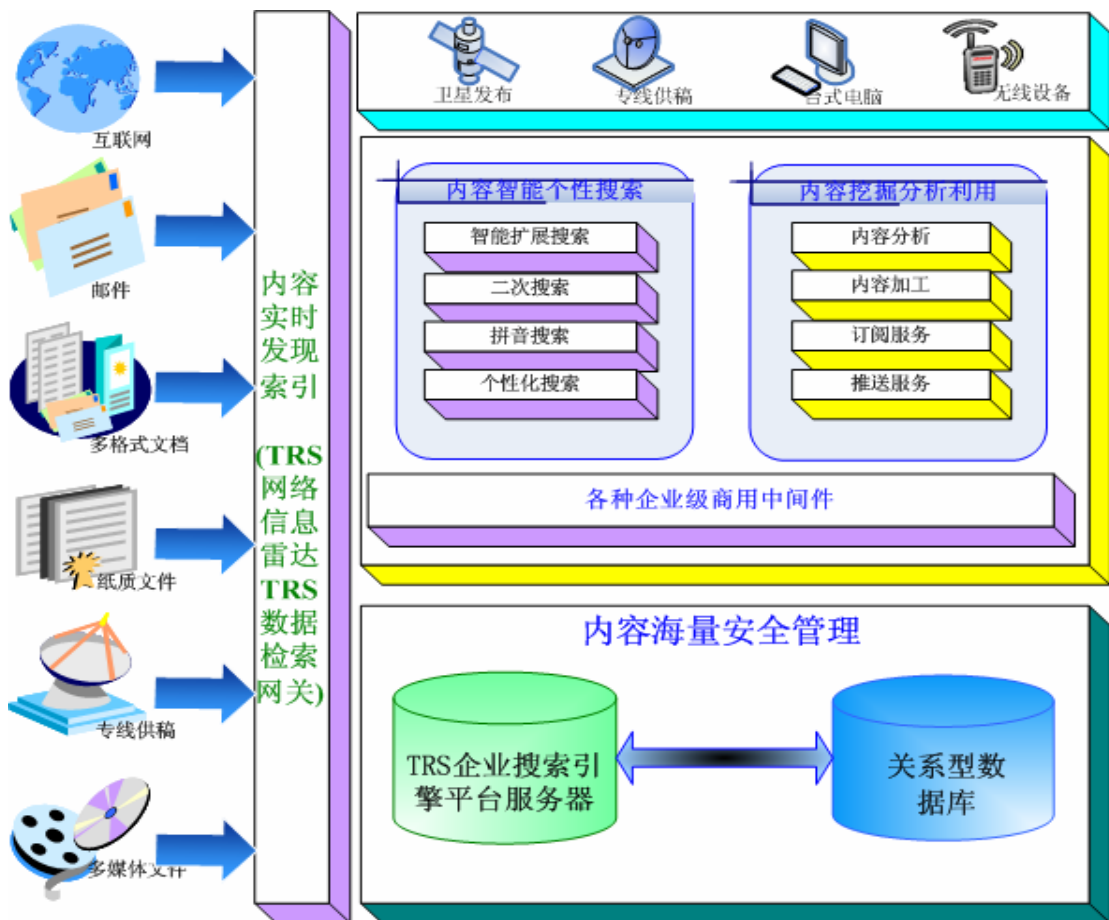


从图中可以看到,TRS 电信 114企业搜索引擎采用 TRS 高性能的 TRS Database Server 6.0 集群构作为索引和搜索基础平台; TRS 数据库网关作为跨平台数据导入工具,整合索引电信企业组织内部各类应用系统、数据库、外购信息库、自建信息库等多种信息资源; TRS 网络信息雷达作为网络信息采集工具,定向采集特定的网络信息资源; TRS CKM 作为文本挖掘的工具,应用文本自动分类、自动聚类、信息过滤等中文处理技术对业务应用信息技术深度挖掘分析; TRS 内容分发服务器作为内容服务模块,实现内容的个性化搜索服务。并且整个架构拥有很强的扩展性,对用户开放开发接口,电信企业可以自行开发具有特色的搜索服务。

## 2.1 系统整合电信多种信息来源，支持内容实时增量索引

在电信企业内部，许多信息内容的创建和生产都分散在各个应用系统中，而这些应用系统的数据存储基本上都是采用关系型数据库或者 NOTES 系统中。这些外部信息资源可能包括企业 CRM 系统、10000 号系统、帐务系统等应用系统，也可能包括查号数据库、企业信息库、用户信息库等业务应用数据库，也可能包括外购资源、自建数据库等其他信息源。

TRS 公司推出的企业搜索引擎解决方案，利用 TRS 公司多年来自主开发的 TRS Database Server 作为企业搜索引擎服务的平台。可以将电信企业内外部多种格式、多种介质形态、多种存储方式的内容信息，以实时的方式将这些信息进行索引。并且在索引的过程中力求能够做到准确，并配合以相关智能语言技术做信息的去重、自动标引等能力。其框架如图：

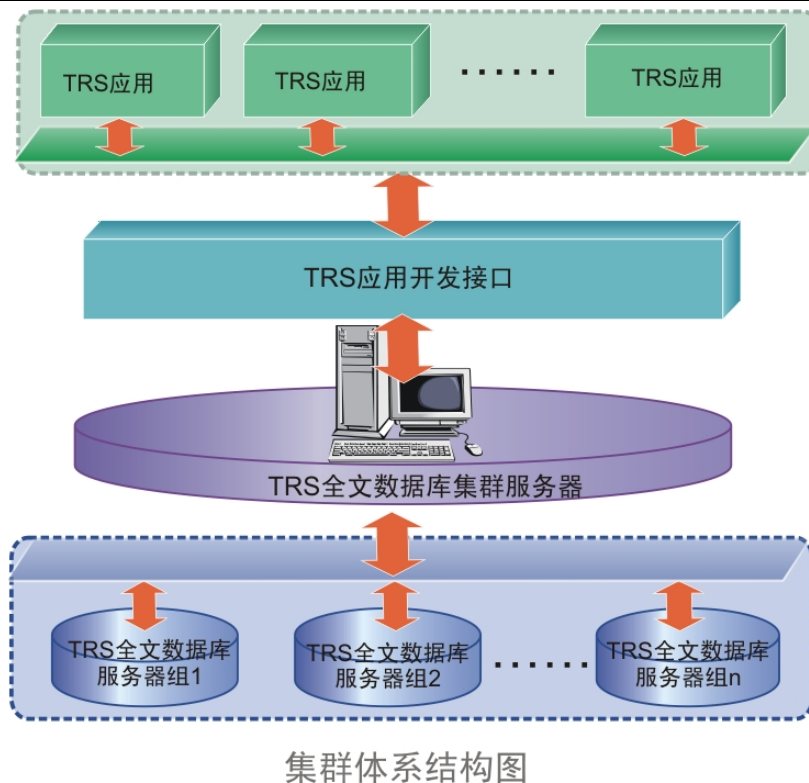


TRS 支持包括 Oracle、SQL Server、DB2、Sybase 和 Mysql 等业界主流关系数据库系统，既可以实现数据库数据一次性向 TRS Server 的迁移，实现历史数据的采集入库到数据管理平台；也可以实现增量动态运行方式，无缝透明支持各种应用数据的数据库采集。将应用中的数据实时地反映到 114 企业搜索引擎平台服务器中（TRS Database Server）。

## 2.2 集群架构支撑大规模部署应用，支持内容海量安全管理

在电信 114 企业搜索引擎应用中，海量信息的管理完全不同于互联网搜索引擎。首先在信息总量上，因为 114 企业搜索引擎应用是分散在各个电信企业内部，并且每个地域、每个应用系统的数据量都是海量的。随着业务应用的增加，数据量将会成倍增长。TRS 电信 114 企业搜索引擎平台服务器在系统架构上，应用集群技术，支持分布式部署，实现在单台或者数台机器中实现电信企业内部海量数据的管理。从信息安全的角度，电信 114 企业搜索引擎服务保障业务信息是按秩序、按组织规则授权方式的搜索。而不象互联网搜索引擎上的人人搜索。

TRS 全文数据集群服务器，是架构在多个物理 TRS 全文数据库服务器之上的分布式管理系统，它支持数据分布及负载均衡两种方式，并支持两种方式的组合运用，满足用户海量数据和高并发环境下的分布式检索、检索性能和可靠性要求。



TRS 全文数据库集群系统结构示意图

### “TRS 全文数据库服务器组” 内的数据库服务器之间负载均衡

组内的数据库服务器由集群服务器统一调度，一个服务请求只发往组内的一个数据库服务器，一个“TRS 全文数据库服务器组”至少包含一个数据库服务器。

### “TRS 全文数据库服务器组” 之间实现分布式检索

用户一个检索请求需要根据其所包含目标对象的分布情况，发往其中部分或全部的数据库服务器组，TRS 全文数据库集群服务器对检索结果集归并处理后返回给用户。

采用 TRS 全文数据库系统 V6 的集群架构可实现以下目标：

- 海量数据按需扩展和分布检索
- 大规模用户高并发条件下保证高性能
- 实现无单点故障的高可靠性应用

TRS 企业搜索引擎平台服务器（TRS Database Server）除了采用得到业界广泛使用全文检索的全部功能和性能，针对企业信息内容搜索引擎服务的管理和资源建设的新需求，发展了包括 Native XML，集群，Unicode，自然语言处理及智能检索等众多新功能，结合 TRS 领先的结构化和非结构化联合查询技术，从而满足了用户对电信 114 搜索引擎的广泛需求。更为重要的是，TRS 电信 114 企业搜索引擎平台服务器提供了多种安全机制的管理，涉及到系统安全和内容安全各个方面。为安全的搜索奠定坚实基础。

- Native XML：能够为更精确的检索提供存储和检索手段。
- 集群：满足海量信息处理和负载均衡的苛刻需求。
- Unicode：以中文为主，提供多语言支持，实现了国际化。
- 自然语言及智能检索：实现更加人性化和达到更好的检索效果。
- 异构：结构化和非结构化异构信息联合查询。

## 2.3 应用垂直搜索实现内容挖掘分析利用，支持电信企业开发特色搜索服务

搜索引擎的出现，整合了互联网上众多的网页资源，并提供信息导航和信息查询服务，使信息的价值得到了网民和厂商的普遍认可。但是，搜索引擎的发展格局是多方面的，市场需求的多元化也导致了搜索引擎的行业化和细分化，从而“垂直搜索引擎”成为了搜索引擎发展的必然趋势之一。

TRS 认为：垂直搜索引擎是针对某一个行业或组织，满足行业专业需求、或者组织某项业务需求的专业搜索引擎，是搜索引擎的细分和延伸，是对某类网页资源和结构化资源的深度整合，并为用户提供符合专业用户操作行为的信息服务方式。比如：用户搜索广州天河区的可带宠物就餐的川菜馆的电话、菜单价格、交通指路等这就是一种垂直搜索。

TRS 作为国内垂直搜索引擎的领军企业，全面支持垂直搜索的相关技术，并率先在国内进行了垂直搜索引擎的实践，为用户成功实施了包括中央政府门户网站政府搜索引擎、公安部搜索引擎、专利搜索引擎、企业经济情报预警搜索引擎

等等成功案例。

在电信行业建设 114 企业搜索引擎的背景下，虽然 TRS 作为搜索引擎厂商提供了成熟的基于企业级搜索的搜索引擎产品，但是各地电信企业对信息内容进行再组织、再开发，特别是应用智能的知识挖掘技术进行内容的挖掘和分析，并根据业务需求开发业务排序和展示，从而为用户开发特色搜索服务。如生活搜索、购物搜索、旅游搜索、票务搜索、教育搜索、行路搜索等等特色搜索服务。

## 2.4 搜索高效、准确，向用户提供智能个性搜索体验

在搜索性能方面，TRS 全文数据库在普通 PC 服务器环境下，在千万级记录的数据库上，也能获得亚秒级查询速度。集群架构的 TRS 电信 114 搜索引擎可以根据不同用户的规模，满足大规模座席的查询速度和并发数量的要求。

在要满足高效搜索的同时，TRS 企业搜索引擎平台服务器熔炼了 TRS 公司多年在中文智能处理方面的研究成果，并结合十多年来的企业及搜索引擎的应用经验。多种中文智能处理技术的应用，如智能分词，字词索引结合、主题词表概念扩展等技术的应用，同时 TRS 搜索引擎内嵌中文自动分词系统，使得查全率和查准率都得到极大的保障。

在 TRS 电信 114 企业搜索引擎解决方案中。采用了模块化的内容分发服务模块。让用户可以方便地通过页面设计模板封装等方式来实现个性化的搜索提交，结果个性呈现。并且系统结合多种信息分发机制，将搜索、浏览、订阅等功能有机集成。而对于信息发现和评估，系统也提供了很好的支持，并可以根据搜索的统计，来评估信息内容的使用情况及信息用户的搜索习惯。电信企业也可以根据自己的需求，开发符合本地用户习惯的搜索页面和结果排序页面。

## 3 方案特点和优势

### 3.1 “安全”的搜索引擎

TRS 搜索引擎技术支持内容安全性控制，可以通过域、IP 段、URL 等广域网范围的控制，实现授权搜索采集，不乱采集。同时，TRS 对查询内容进行分级控制，特定的人只能搜索和查询特定的内容。

在 TRS 搜索引擎技术中提供了信息智能过滤和禁用词典设置，通过这些技术，保障搜索引擎在提供便捷的搜索服务的同时，也保证对不良信息搜索的过滤。

另一方面，TRS 搜索引擎技术在安全模块设计上提供了对 PKI/PMI 体系支持的开放接口，在未来，很容易将本系统整合到信息安全保障体系之中。

### 3.2 更高的搜索准确性和智能性

TRS 搜索引擎技术支持按词索引、按字索引、按关键词索引，字词混合索引，适应不同应用环境的需求，同时 TRS 搜索引擎内嵌中文自动分词系统—检索“北大”，检索不出“东北大学”。

内嵌歧义处理实例规则库，正确识别歧义片断，提高分词准确性分词系统要达到一定的准确率，需要和人一样不断积累知识，也就是不断积累分词规则。TRS 公司从 80 年代末就开始积累分词规则，这些规则是需要从大量的语料中统计产生，如果语料的数量不够则产生的规则往往带有片面性，TRS 积累了 20~30GB 的文本语料，且这些语料能反应现中文语言的特点。如果一个语句切分时有歧义片段，有适合的规则则按规则切分歧义片段，提高查准率；

在查询方面，TRS 提供了基于词典的智能扩展查询，可以按同义词、主题词等词典进行智能扩展查询，例如，在查询“锐器”时，系统将自动将包括“匕首”“刀”内容的结果提供给用户参考。

### 3.3 个性化的搜索体验

TRS 搜索引擎内容分发服务模块充分考虑了信息搜索过程中工作繁忙、对信息的时效性要求高等工作特点。提供了任务定制查询、专栏预设查询、个性化排序等功能。例如：使用者可以定制查询任务，比如“专项斗争”、“专题文件”等单项任务，又如可以定制查询更新时间，查询系统将根据定制的任务，定期进行相关信息查询，定期将查询结果推送到用户的工作界面，方便信息需求者。

另外，系统还提供了个人检索历史记录、个性化界面设置等等功能，不同的使用者可以选择适合自己的工作查询界面，提升系统的易用性和灵活性。

### 3.4 强大的异构资源整合搜索

TRS 搜索引擎技术不但能搜索网页内容，而且能搜索各种 RDBMS，文件系统等多种异构资源数据进行整合搜索，为用户提供更全面的信息搜索应用。在未来，可以在当前搜索引擎系统上不断扩展新的搜索应用。

### 3.5 标准、开放的系统，提供强大的系统扩展能力

标准、开放是一个应用系统得以发展和壮大的基础，通过标准开放的模式，可以保证用户更多地采用先进的技术搭建个性化的应用。

随着技术的发展，各个软件供应商越来越在某一领域具有专利或优势技术，但是用户的需求是全方位的，因此，最好的解决方案就是采用统一规范标准的接口进行应用集成，这也是国际化软件发展趋势。

TRS 公司设计的建设方案在很多方面为系统应用集成提供了保证，如支持系统三层体系结构，支持 J2EE 标准中间件，支持 XML 数据交换规范，提供底层数据库的各种平台的完善的开发接口，提供模块组件，支持二次开发，开放底层数据存储格式等等。

本系统在架构设计方面，不但满足了现有的需要，而且为系统未来发展进行了考虑。首先，数据层采用了 TRS 集群服务器，实现了 TRS 全文检索数据库的集

群和负载均衡应用，在应用层实现了应用服务器的集群和负载均衡设计，在采集方面利用分布式采集和任务集中控制的模式可以进行大规模采集应用，在未来可以通过增加硬件的方式，就能提供系统的处理能力。

另外，随着未来负载和访问量的增加，可以分步建立镜像中心，满足大规模应用需要。

### 3.6 具有充分满足需求的自主核心技术和产品

TRS 公司在本项目所涉及的众多方面具有全方位的核心技术和产品，并且公司发展的战略定位和本项目的需求完全吻合。

本方案涉及底层数据库系统、中文知识挖掘、信息搜索等多种技术，是一个大型和复杂的信息系统，TRS 信息技术有限公司在信息检索、内容管理和知识管理方面具有领先的产品和技术优势，致力于成为中文内容管理领域的领导者。TRS 全文信息检索系统已经在超过 1 千家用户的多个系统和应用中得到成功应用；TRS 中文知识管理和自然语言处理方面的研究成果，是业界第一个实用化的相关产品，其中包括中文自动分类系统、自动聚类系统、网页内容过滤、内容去重等。这些研究成果来源于公司相关的研究机构-中文信息处理研究中心-承担的国家自然科学基金、国家 863 计划等国家级研究项目。TRS 公司致力于成为中国信息检索和内容管理市场的技术和市场领导者，并且把行业化应用作为我们的既定战略，因此在技术研发的持续投入上符合用户对信息系统持续发展的需求。

### 3.7 采用先进的搜索引擎技术

近年来搜索引擎技术得到较大发展，为本系统的实现提供了技术手段。本系统所涉及的搜索引擎相关技术包括网页自动采集和更新、网页自动分析技术。

网页自动采集和更新

为保证本系统要求功能的顺利实现，所采用的搜索引擎技术具备以下功能：

- 1) 支持增量更新的策略，每次采集只采集上次更新后新生成的网页，而不是全部再采集一遍，从而保证信息更新的效率。增量更新策略是对信息

采集非常重要的方法，也是网络上搜索引擎普遍存在的缺点。

- 2) 支持灵活的采集策略，包括可以指定采集的目录和层次，以及使用检索逻辑来定位内容，比如可以采用“自行车-比赛”这样的检索逻辑来控制抓取“有关自行车，但并不是关于自行车比赛”的网页。除了在采集模块提供检索逻辑定位内容的方式，我们还在发布模块提供专题服务的方式实现同样的功能，而且我们建议应采用以专题服务为主实现网页内容定位。具体内容参见设计方案中的专题服务部分。

#### 网页自动分析技术

采集到的网页，为了满足本系统的应用，必须经过以下加工处理：

- 1) 正文内容提取：剔除广告、导航信息、版权等无用信息，只保留正文内容以及必要的图表；
- 2) 格式自动转换：自动将 HTML 格式转换为 TEXT 文件，方便再加工；
- 3) 属性自动标引：对有条件分析出标题、版次、日期、作者、栏目、分类等属性的网页，分析并标注这些属性信息（元数据自动提取）；
- 4) 属性自动提取：自动搜索、记录网页中的单位名称、系统名称等标识网页属性的信息。

### 3.8 基于内容的自动分类和聚类技术

为了对采集到的大量网页信息进行标注分类，必须采用适当的机器自动分类方法，尽量减少需要人工参与的环节。但是，必要的人工干预能够提供分类的准确率。

在本方案系统应用设计中所提出的机检分类和自动分类，分别代表了语义规则分类方法和统计原理分类方法两种典型的分类技术，为了描述方便，我们分别称它们为基于语义规则的自动分类和基于统计原理的自动分类。

#### 基于语义规则的自动分类（机检分类）

基于语义规则的自动分类是利用人工定义的语义规则对信息进行分类，人们通过维护一个规则表来控制分类的效果。

- 优点：原理简单，容易实现，控制效果明显。

- 缺点：语义规则的制定和维护需要大量的人工参与，不能利用语料库的知识资源；不能有效解决对多语言的支持，对多语言需要分别建立对应的规则表，工作量大；人工制定的语义规则不可能完全反映分类的内在规律。

基于语义规则的自动分类方法比较适合通过简单的规则即可明确判定的分类，比如按地区分类，按事件分类等。

基于统计原理的自动分类（自动分类）

基于统计原理的自动分类方法是建立在统计学习理论和机器学习方法之上的根据内容进行自动分类的方法，其基本原理是利用概率统计学原理，采用机器在大量语料库上自动学习的方式，分析出各个分类的内在特征，然后通过对比未知对象与各个分类特征的相关程度来判定其类别归属。

基于统计原理的计算方法在近年来得到普及的开发和应用，并在诸如语音识别、汉字识别、拼音输入法等领域的应用中表现出良好的实用价值，比基于规则推理、语义分析等语言学知识的方法表现出更强的灵活性和适应性。

- 优点：学习过程由机器自动进行，不需人工干预；人们对分类质量的控制转换成提供语料库的方式，更加符合信息管理员的工作特点；在给定语料库的前提下，机器对分类特征的提取不会产生遗漏或误差，计算结果稳定。

- 缺点：基于统计原理的自动分类适合于对内容进行自动分类，而不适合地区、事件、来源等类型的分类。

在此系统设计、实现中，不但可以先按内容、地区、来源等多种方式快速标引网页，而且可以基于内容对采集信息进行自动、准确的分类，这两种分类方法有机结合为搜索引擎系统提供全面的、准确的、快速的、智能的分类服务。

### 3.9 基于内容的信息去重技术

在该项目中，采用了信息去重、相似性检索技术，主要应用在网页的排重过程分析中。虽然简单的规则判断提供了一种可选择的方式，但合理的方案应是基于网页内容本身的判断，基于网页内容的判断应该是排重的主要手段。因此我们

建议采用基于内容的、成熟的信息去重相似性检索技术实现内容的排重判断，在排重判断的过程中，TRS 公司设计、实现的系统将不但处理文字内容，而且要对文中的数字内容进行判断。

### 3.10 优异的全文检索性能

TRS 全文检索系统在行业里具有领先的性能，是中文全文检索的事实上的标准，众多的全文检索厂商都以 TRS 的性能指标作为自己软件评测的标准和系统发展方向。

目前国内唯一的商用千万级数据库——新华社多媒体数据库就是采用 TRS 作为底层检索平台，目前，该系统已有将近 16T 数据量，检索（包括简单检索和复杂检索）的平均响应时间是秒级。并且，TRS 的检索性能随着数据的增加不会呈线性下降，可以在一个非常广的数据规模范围内保证用户的实际应用。

## TRS 全文数据库系统 V6 的性能指标

产品	测试环境
<b>TRS Database Server</b>	Dell 2850: Inter(TM)2.8G×2; 4G Memory; RAID0 Linux AS4.0 update2 (64 位) Windows2003 SP1 (64 位)
<b>Cluster</b>	Dell 2600: Inter(Xeon)2.4G×2; 2G Memory; RAID0 Linux AS4.0 (32 位) Windows2003 sp1 (64 位)
<b>数据类型:</b>	<b>新闻类数据</b>

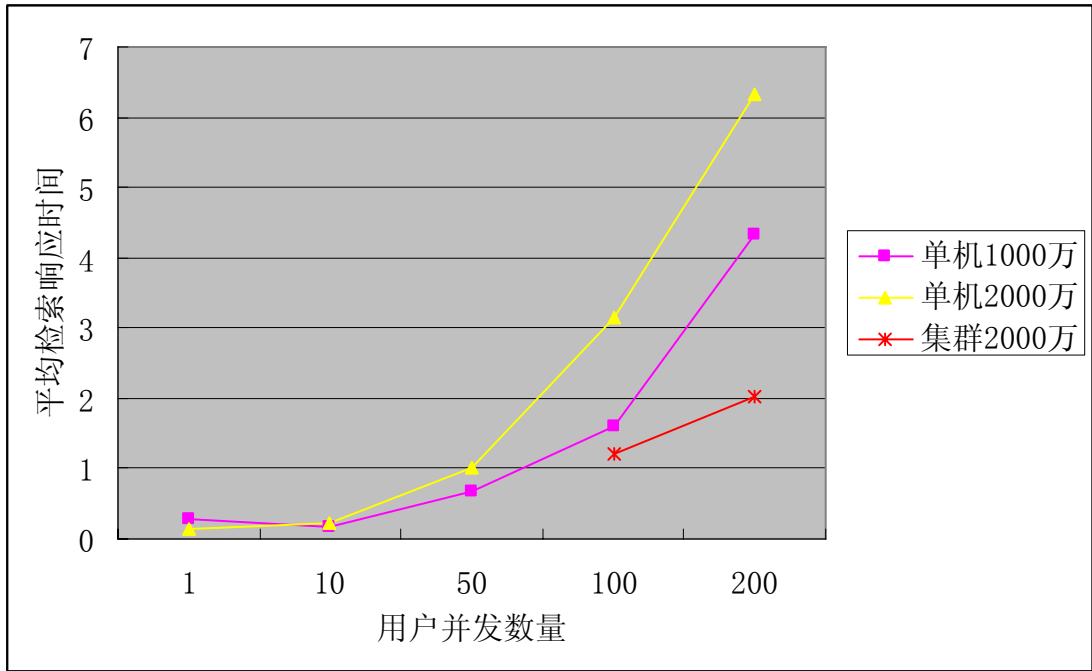
单机环境下 TRS 全文数据库系统的性能指标:

数据量	并发用户数量	平均检索响应时间(秒)
1000万	1	<b>0.27</b>
	10	<b>0.17</b>
	50	<b>0.682</b>
	100	<b>1.603</b>
	200	<b>4.334</b>
2000万	1	<b>0.14</b>
	10	<b>0.23</b>
	50	<b>1.015</b>
	100	<b>3.148</b>
	200	<b>6.336</b>

集群负载均衡模式下 TRS 全文数据库系统的性能指标:

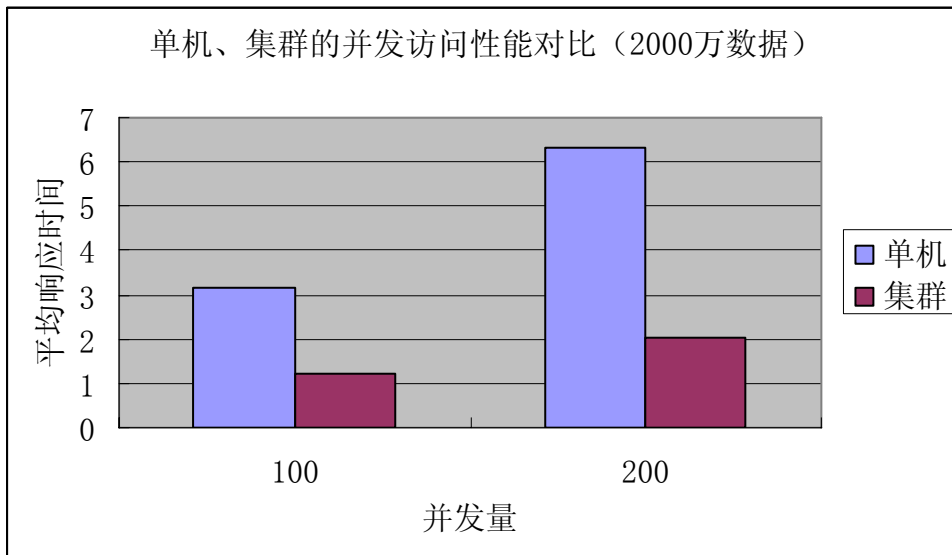
数据量	TRS Database Server 数量	并发用户数量	平均检索响应时间(秒)
2000万	2	100	<b>1.214</b>
	4	200	<b>2.018</b>

TRS 全文数据库系统的性能图示:



海量数据高并发环境下单机、集群访问性能对比图示：

（排版时，下面的图可由上面的图引导而来，例如：在上图下面两根线的位置加一个箭头）



由上图我们可以看到，在海量数据高并发环境下，TRS 全文数据库服务器集群负载均衡模式可以成倍提高访问性能。

### 3.11 成功的应用模式和丰富的应用经验

TRS 公司多年来服务于众多的关键信息系统建设, 在应用集成能力和服务能力上得到了用户的认可。

如国务院新闻办三网一库的核心数据库建设、国家计委纵向网、卫生部信息发布网站, 国家统计局综合网站信息服务系统, 北京市劳动和社会保障局网站、外交部新闻监控采集系统、新华社多媒体数据库平台、中央人民广播电台网站内容管理平台、央视在线主持系统、人民日报资料库、中粮、中国五矿等等, 我们对资源库建设、入库、编辑、生成、展示等整个信息的生产和服务的应用需求有着深刻的理解, 这是领先其他公司的显著优势, 通过这些服务, 我们与客户建立了良好的合作关系, 赢得了客户的信任。特别是我们为新华社多媒体数据库进行的提速工作, 再次展现了公司在信息检索技术领域国际化的领先水平, 我们通过并行检索、Bi-Gram 索引、服务器群集和数据库智能化自我管理以及多层次 Cache 技术等, 使得多媒体数据库在千万级数据库记录的综合查询性能获得了成倍的提高, 这是对关键业务的关键性突破。

### 3.12 专注的服务

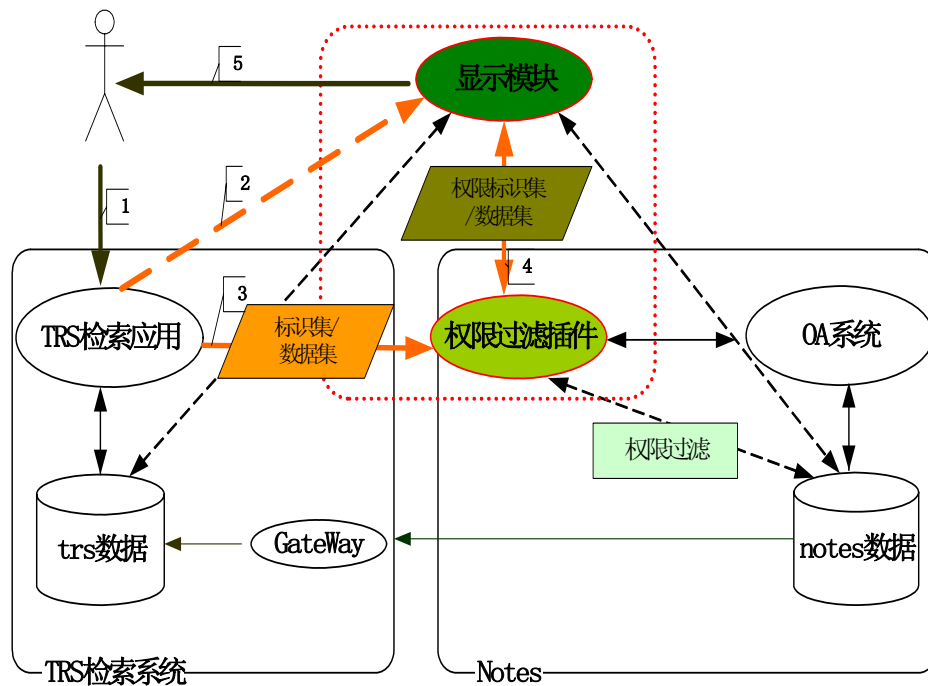
TRS 专注于咨询和开发服务, 采用开放的体系结构、技术和产品, 确保用户的利益, 保护现有投资。

TRS 公司专注于软件产品的提供和技术服务, 在咨询和服务能力上居于领先地位, 在本项目中, 我们在关键性平台上采用了 TRS 成熟产品, 并且保证提供开放的技术体系, 提供完整和可扩充性, 提供应用集成能力, 同时根据用户需求, 推荐针对应用的行业内具有优势的相关产品, 总体设计方案强调系统组件的最佳组合, 选择合适的技术和产品。

#### 企业·广东移动搜索引擎服务

在广东移动搜索引擎服务的系统建设中, TRS 重点解决了资源信息的整合和安全问题。

资源信息的整合包括了对办公自动化数据的整合、对论坛数据的整合、对CM系统数据的整合，通过整合，实现统一的检索入口，实现统一的Portal服务。此外，各种数据对原有系统中的权限（主要是Notes系统中的权限）得到了很好的继承，只有相关的授权用户才能对自己权限范围内的信息进行浏览和检索。



广东移动搜索引擎从技术上可以分为三个部分：Notes 业务平台，TRs 检索系统，权限过滤和显示模块。用户进行全文检索首先在 TRs 中命中相应检索记录，然后通过权限过滤命中相应的数据集合，最终通过显示模块展示给相关人员经过安全验证的信息内容。

通过这种方式，Notes 中的全部业务权限得到了充分的继承，有效体现了 TRs 企业搜索引擎的开放性和可集成能力，充分体现了 TRs 安全检索的概念。

通过共享资源库的建设，为内部工作人员提供了一个题材丰富的信息资源系统，采用 TRs 企业搜索引擎系统，将各部门信息资源整理加工，形成集公文、政务信息、行业规章、地方特色信息等资源的数据库，为各级领导提供决策服务参考，为相关工作人员提供信息快速查询平台。

安全检索和资源整服务，使内部的信息得到有序的共享，并进一步优化了管理流程，实现“一站式”信息发布和办公服务模式。

## TRS 联系方式

**TRS总部营销服务中心：**

北京市朝阳区安翔北里 11 号

北京创业大厦B座 1008 室

邮编: 100101

电话: (010)64848899

传真: (010)64889088

Email:[info@trs.com.cn](mailto:info@trs.com.cn)

**TRS总部研发中心：**

北京 北四环中路 35 号健翔桥

北京信息科技大学图书馆三层

邮编: 100101

电话: (010)64859900

传真: (010)64879084

Email:[trs@trs.com.cn](mailto:trs@trs.com.cn)

**上海分公司**

上海市陕西北路 66 号科恩国际中心 1505A

邮编: 200041

电话: (021)51168966/67/68

传真: (021)51168968 转 1024

Email:[trs.sh@trs.com.cn](mailto:trs.sh@trs.com.cn)

## 版权声明

TRS®是北京拓尔思信息技术有限公司的注册商标。本文中涉及的各种产品和服务的名称可能是拓尔思公司的商标,其他所有提及的产品和服务名称可能是各自持有者的商标。

版权所有 © 2007 北京拓尔思信息技术有限公司  
保留所有权利