中文分词nlpir发布lucene分词支持包

产品名称	中文分词nlpir发布lucene分词支持包
公司名称	灵玖中科软件（北京）有限公司
价格	.00/个
规格参数	
公司地址	北京市海淀区苏州街49-3号5层11号
联系电话	010-62648216

产品详情

近日，在北京理工大学大数据搜索与挖掘实验室团队成员的不懈努力下，为了方便大家使用，NLPIR-ICTCLAS发布了lucene/solr的分词支持包功能。系统通过此次升级进一步提升了功能效果，为数据的提取提供了保障。

NLPIR-ICTCLAS分词系统前身为2000年发布的ICTCLAS词法分析系统，由张华平博士在多年研究工作积累的基础上研发出来，从2009年开始，为了和以前工作进行大的区隔，并推广NLPIR自然语言处理与信息检索共享平台，调整命名为NLPIR分词系统，主要功能包括中文分词;英文分词;词性标注;命名实体识别;新词识别;关键词提取;支持用户专业词典与微博分析。

词法分析是自然语言处理的基础与关键。在中文自然语言处理中，词是最小的能够独立活动的有意义的语言成分。汉语是以字为基本书写单位，词语之间没有明显的区分标记，因此进行中文自然语言处理通常是先将汉语文本中的字符串切分成合理的词语序列，然后再在此基础上进行其它分析处理。中文分词是中文信息处理的一个基础环节，已被广泛应用于中文文本处理、信息提取、文本挖掘等应用中。

NLPIR-ICTCLAS系统支持多种编码(GBK编码、UTF8编码、BIG5编码)、多种操作系统(Windows, Linux, FreeBSD等所有主流操作系统)、多种开发语言与平台(包括：C/C++/C#,Java,Python,Hadoop等。这次升级可以直接把lucene/solr支持包功能内嵌到分词系统里，增强了系统的搜索速度和分词的准确率，是为了适应系统需求，提高客户使用的效率。

Lucene是一个开放源代码的全文检索引擎工具包，即它不是一个完整的全文检索引擎，而是一个全文检索引擎的架构，提供了完整的查询引擎和索引引擎，部分文本分析引擎。Lucene的目的是为软件开发人员提供一个简单易用的工具包，以方便的在目标系统中实现全文检索的功能，或者是以此为基础建立起完整的全文检索引擎.

Solr是一个高性能，采用Java5开发，基于Lucene的全文搜索服务器。同时对其进行了扩展，提供了比Lucene更为丰富的查询语言，同时实现了可配置、可扩展并对查询性能进行了优化，并且提供了一个完善的功能管理界面，是一款非常优秀的全文搜索引擎。它对外提供类似于Web-service的API接口。用户可以通过http请求，向搜索引擎服务器提交一定格式的XML文件，生成索引;也可以通过Http Solr Get操作提出查找请求，并得到XML格式的返回结果。

NLPIR-ICTCLAS系统在长时间的实验和总结中，以满足客户的需求为基础，不断的提高系统的流畅性和准确率，为广大使用者提供一个安全、高效的使用环境。