灵玖软件：NLPIR文本聚类智能挖掘文本数据

产品名称	灵玖软件：NLPIR文本聚类智能挖掘文本数据
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

在当今时代信息爆炸的情况下，一方面网络信息的规模急剧膨胀，另一方面信息又凌乱无序，有价值的信息湮没在大量冗余信息中，对其发现和管理变得越来越困难。为此，以文本聚类信息进行数据挖掘引起了越来越多人的兴趣。

文本聚类就是从很多文档中把一些内容相似的文档聚为一类。文本聚类主要是依据著名的聚类假设：同类的文本相似度较大，而不同类的文本相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文本手工标注类别，因此具有一定的灵活性和较高的自动化处理能力，已经成为对文本信息进行有效地组织、摘要和导航的重要手段，为越来越多的研究人员所关注。

聚类分析是一种无指导的机器学习方法，在机器学习、统计分析、模式识别、数据挖掘、生物学等许多领域得到了广泛的研究与应用。聚类的基本目的是将数据对象按照一定的标准分成若干个簇，使得同一个簇中的对象之间相似度较大，不同簇之间的对象相似度较小。文档的聚类分析与一般的聚类分析类似，往往包括如下5个步骤：

(1). 模式表示，往往包括特征抽取和特征选择，把数据对象表示成适合于算法可计算的形式;

(2). 根据领域知识定义模式之间的距离测度公式;

(3). 聚类或者分组;

(4). 数据抽象表达(如果需要);

(5). 评价输出结果(如果需要)。

一个文本表现为一个由文字和标点符号组成的字符串,由字或字符组成词,由词组成短语,进而形成句、段、节、章、 篇的结构。要使计算机能够高效地处理真是文本，就必须找到一种理想的形式化表示方法,这种表示一方面要能够真实地反应文档的内容(主题、领域或结构等),
另一方面，要有对不同文档的区分能力。NLPIR文本聚类是目前反应使用效果比较好的文本聚类系统。

NLPIR文本搜索与挖掘系统针对互联网内容处理的需要，融合了自然语言理解、网络搜索和文本挖掘的技术，提供了用于技术二次开发的基础工具集。

NLPIR文本聚类模块是基于相似性算法的自动聚类技术，自动对大量无类别的文档进行归类，把内容相近的文档归为一类，并自动为该类生成标题和主题词。适用于自动生成热点舆论专题、重大新闻事件追踪、情报的可视化分析等诸多应用。

灵玖基于文章集合核心语义理解技术，不仅聚类速度快，而且准确率高，并能自动得到类别间的演化趋势。

文本信息挖掘是数据挖掘中重要的研究领域之一，文本数据是一种无机构或半结构化的数据，对文本信息的处理需要自然语言的支持，计算机不能很好的处理文本数据中的多义、歧义等问题，若要获得深层次信息需要深入探索研究。文本聚类作为文本信息挖掘的重要方式之一，对Intemet上搜索引擎和信息架构的构建等起到了重要的作用。