# 基于自然语言处理的容器画像可视化平台软件V1.0

# 说明书

**西安电子科技大学**

## 软件描述

随着软件工程的不断发展，软件已经渗入各行各业，软件从产业化时代已经走向了服务化、网络化的时代。在云计算领域，随着软件即服务（SaaS, Software as a Service）思想的落地，以软件作为服务进行交付的方式早已成为现实。在云环境下，云服务商提供进程级别的容器作为软件进行交付部署，用户可以直接启动容器镜像获得软件服务，或在容器的基础上配置自己的相关服务并再次打包为新的容器镜像。在服务化、网络化的软件时代，容器已成为软件的主要载体。

目前，容器仓库挖掘领域的相关研究主要对Dockerfile文件进行分析，该文件的结构化、同质化的特点使得容器知识易于发现，得到的容器知识被广泛应用于容器构建、构建配置，容器推荐和容器标注等多个方面。然而，由于Dockerfile文件信息底层、语义信息低维的局限性，研究无法从更加抽象、高维的语义角度对容器进行描述，容器缺乏简洁、高维的特征描述，导致容器在计算机和用户面前仍然是一个难以理解的黑箱。如何建模，分析挖掘容器知识成为了容器仓库挖掘的重点研究问题。

本软件实现基于自然语言处理的容器画像可视化展示，通过分析与容器相关的数据特点，以简洁、层次化的标签体系结构建模了容器画像。为挖掘容器知识，在容器画像设计的驱动下，本文结合了自然语言处理领域（NLP，Natural Language Processing）的相关技术，设计了容器画像标签生成机制支持容器标签生成，并实现了容器画像标签抽取工具，该工具针对文本数据，完成了数据采集、解析、预处理、标签挖掘和标签扩展生成工作，在多个语义层次为容器画像产生了标签。

## 软件功能

本软件主要完成容器标签生成和容器画像生成，主要包括：

* 数据采集功能：用于收集Docker Hub中的容器镜像数据信息。
* 标签生成功能：根据容器描述生成标签。
* 数据存储功能：负责容器相关数据在数据库中的增删改查操作。

## 3.软件的设计与实现

### 3.1 软件结构设计

可视化平台软件整体框架如图1所示。主要通过数据采集模块基于API请求模块将DockerHub中容器镜像数据信息保存到数据库中做持续化存储。然后通过标签生成模块进行文本预处理进而生成容器镜像的标签，最后通过Word2Vec等技术进行容器镜像的标签扩展，最后通过数据存储模块保存容器镜像并实现其画像的可视化。



图1 基于自然语言处理的容器画像可视化平台软件整体架构图

本系统主要基于C语言爬虫框架及Python自然语言处理框架，然后以Web的形式实现容器画像的可视化。

### 3.2 软件流程描述

本软件的使用可分为3个功能模块来介绍，即数据采集模块、标签生成模块和数据存储模块。

**数据采集模块：**包括API请求模块、数据解析模块和存储模块。API请求模块用于请求获得Docker Hub的容器数据信息，作为网络服务中的客户端，该模块采用curl库实现API对Docker Hub服务器的请求访问；数据解析模块解析容器数据信息，采用cJSON库实现对返回数据的解析；存储模块采用本地的MariaDB作为存储数据库，完成数据解析后的存储，如图2所示。



图2数据采集模块流程图

**标签生成模块：**，首先针对由Markdown格式的长文本描述信息，采用了基于C的MD4C工具库实现了Markdown文档的解析，通过匹配标题信息，找到容器的长文本描述信息。然后，根据文本预处理方法筛选单词，并对单词进行词汇归一化，如图3所示。



图3文本预处理流程图

**数据存储模块：**建立了容器信息表、容器发布者信息表和容器标签表。容器信息表中描述了大量有关容器的信息，本文提出的容器画像标签生成机制主要利用该表中的容器短描述和长描述信息作为标签生成的基础数据。

### 3.3 典型应用

本软件为用户提供了基于自然语言处理的容器画像可视化平台软件，可以广泛的应用于容器知识库建立。

现以容器知识库系统为例，介绍本软件的应用场景。

1. 用户可以通过按钮一键获取部署在DockerHub中的容器镜像信息，并可以通过可视化界面浏览。
2. 用户可以通过按钮一键对容器镜像信息进行自然语言处理，对容器镜像的标签进行提取。
3. 用户可以可视化观察生成的容器画像示意图

## 4.运行安装说明

本软件的运行环境要求如下。

1、CPU主频不小于1.6GHZ；

2、内存容量不小于2G；

3、硬盘容量不小于10G；

4、操作系统Windows XP SP2或更高版本以及Linux；

基于自然语言处理的容器画像可视化平台软件是基于web系统的可视化软件，只需要执行shell脚本，即可完成安装。

## 5.使用说明

### 5.1软件主界面

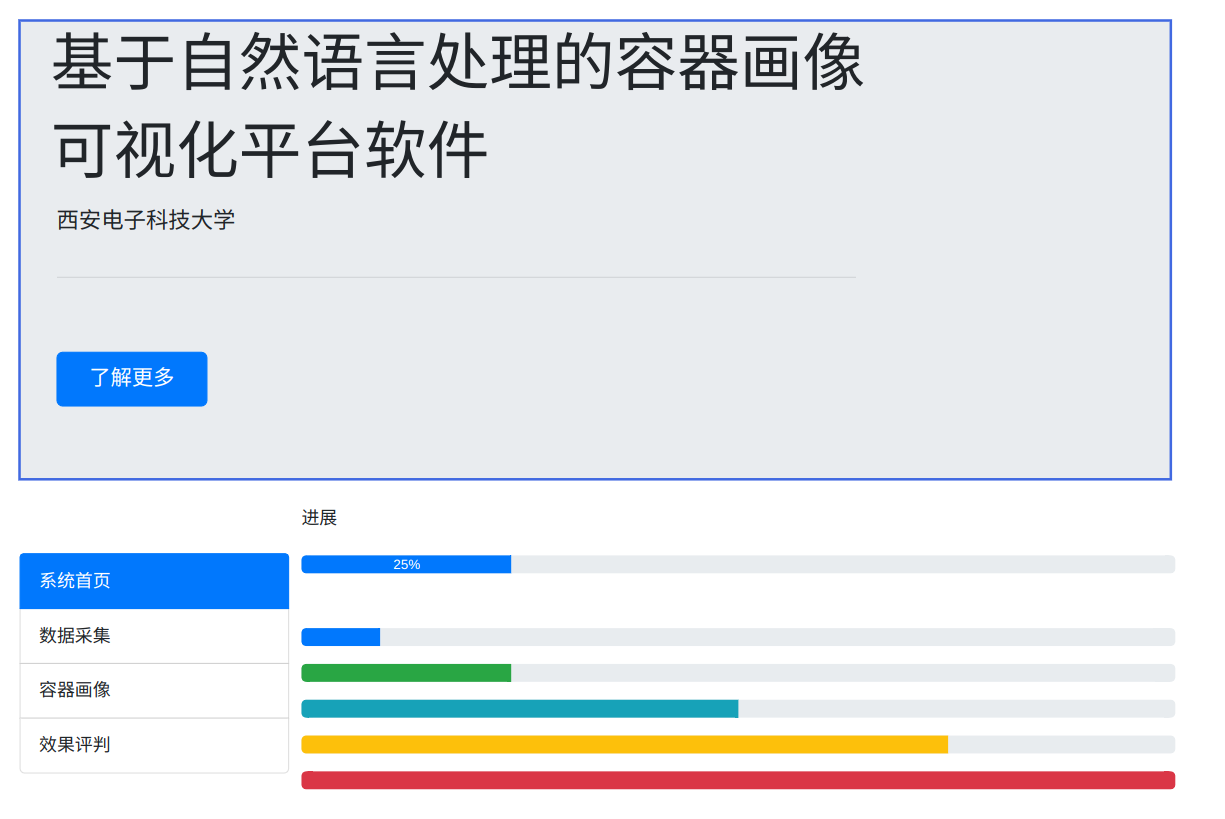
基于自然语言处理的容器画像可视化平台软件，其界面如图4所示。从图4可以看出软件提供了基本的功能供用户进行选择。

图4基于自然语言处理的容器画像可视化平台软件主页面

### 5.2 平台功能使用

5.2.1数据采集功能

数据采集模块首先通过API查询字符串将以查询语句的方式访问Docker Hub，通过生成查询字符串，然后利用curl库发送请求信息，获得查询结果，得到返回的多项容器数据，并再以API的方式分别访问各容器的详细信息，在得到以Json格式编码的数据信息后，对Json数据进行解析，最终借助数据存储模块将采集得到的容器数据存入数据库中。

如图5所示，当用户点击“开始数据采集”按钮，可以看到软件给出提示“正在从Docker Hub中进行数据搜集”。 因为爬虫需要一定的时间，待爬虫完毕后，将会给用户展示查询的结果。

图5 用户点击开始数据采集按钮软件收集数据示意图

如图6所示，当系统完成了爬虫，将生成两种样式：JSON文件和数据库文件，当点击特定容器镜像对应的JSON文件按钮时时，软件系统将会展示特定镜像的JSON格式，同时用户可以点击数据库可视化展示，将会对用户展示容器镜像信息存储在数据库中的信息。



图6 数据采集功能模块示意图

如图7所示，用户点击了JSON文件按钮后，图6展示了单个数据库的JSON文件解析示意图，图7展示了用户点击了数据可视化展示按钮后，将展示存储在数据库的容器镜像示意图。

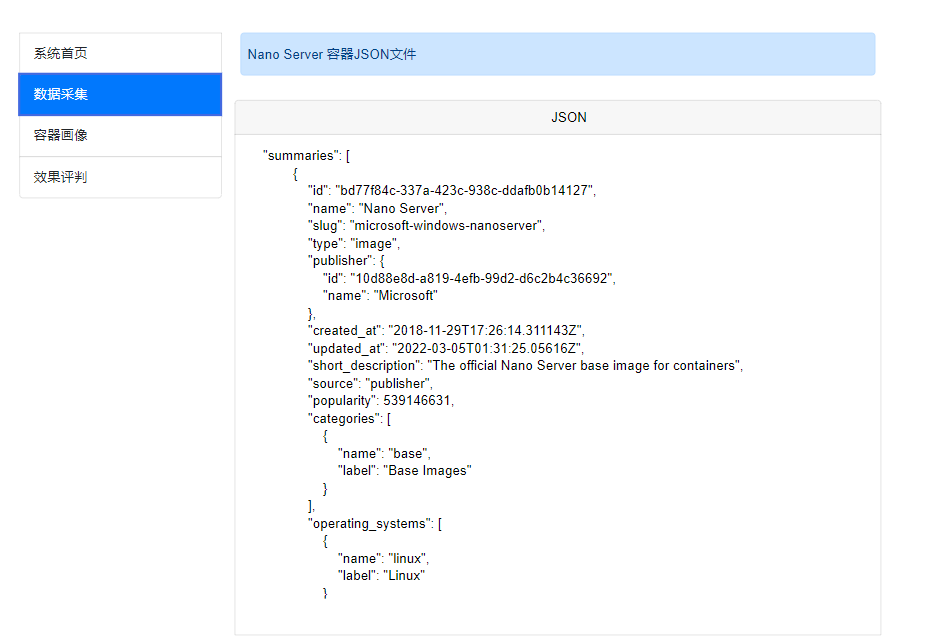


图6 Nano Server容器镜像JSON文件结果示意图

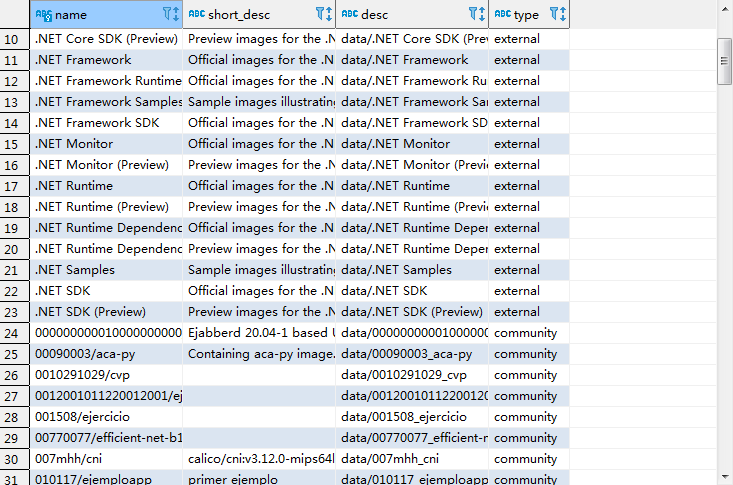


图7数据库可视化示意图

5.2.2 容器画像可视化功能

容器画像可视化功能主要包括对容器镜像的标签短描述和长描述进行自然语言处理。该过程需要大量的数据处理时间。当用户点击开始画像生成按钮后，处理完成后，将会给用户展示容器的画像，如图8所示。

当生成画像后，将会展示容器镜像的列表，用户可以选择特定的容器镜像的画像生成图，如图9所示。

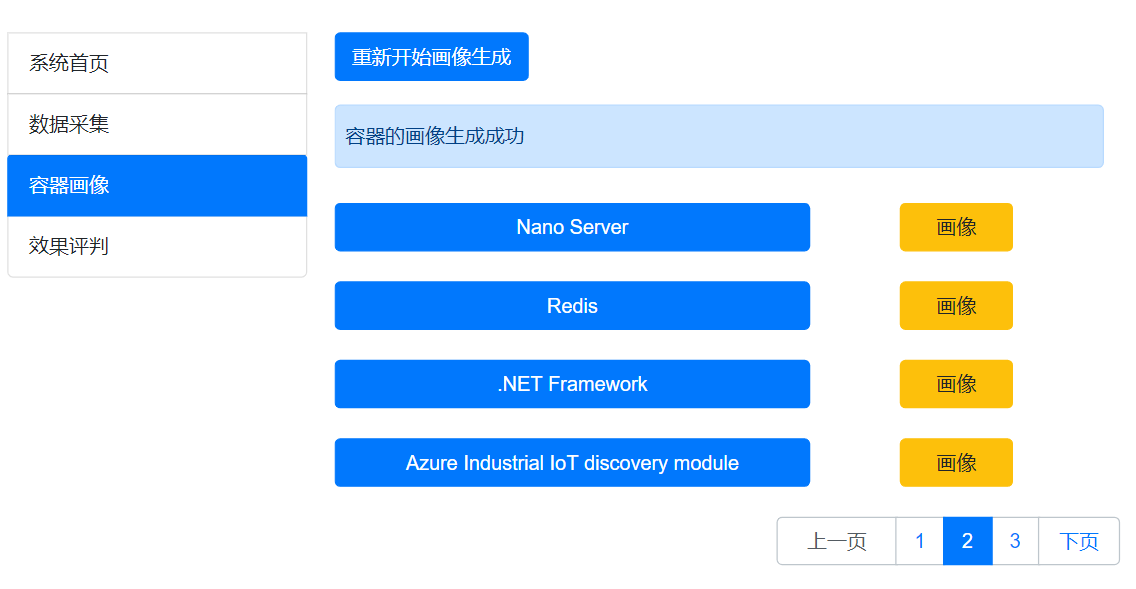


图9 容器画像可视化功能模块

本软件说明书选择.NET Framework容器Azure Industrial IoT discovery module容器为例子进行展示，其结果示意图如图10和图11所示。

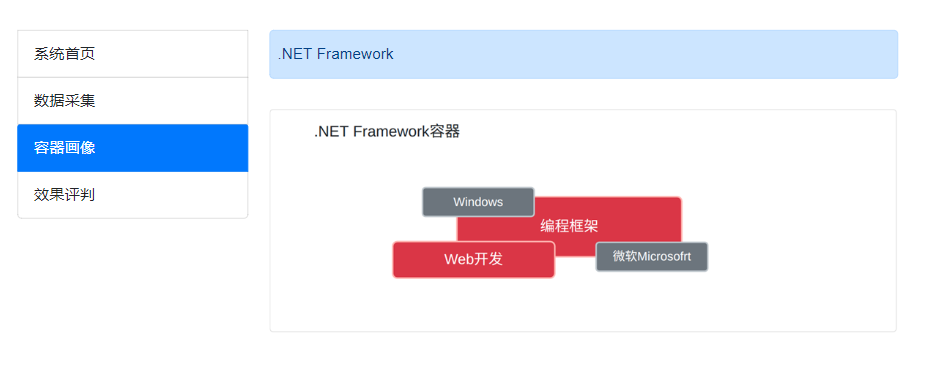


图10 .NET Framework容器画像

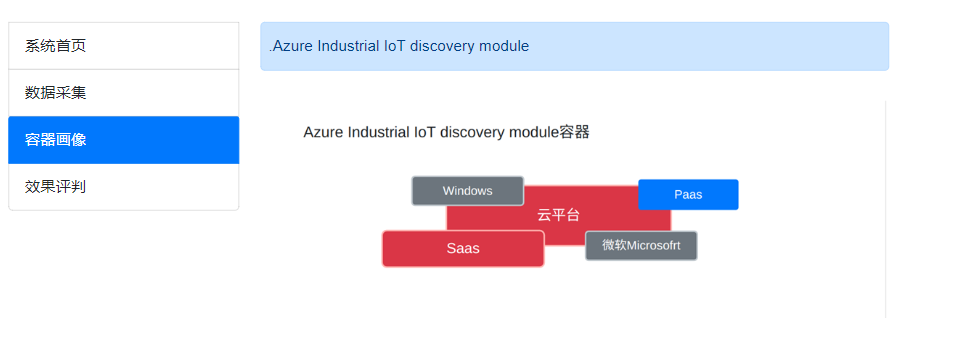


图11 Azure Industrial IoT discovery module容器画像

5.2.3 效果评判功能

效果评判功能是对容器技术的画像生成的效果进行评判，是本软件提供的一种辅助功能，用于表征容器画像生成的效果，如图12所示。

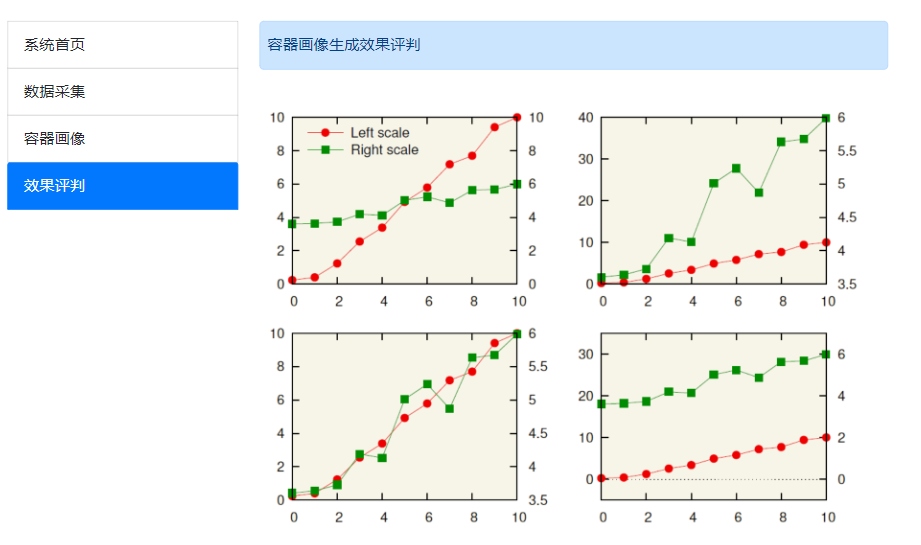


图12 效果评判结果示意图