1. For general questions relating to the features/functionality of HCM Extracts (rather than issues directly concerning any inaccuracies within this document itself) please use the HCM Reporting forum for a quicker response from the current infrastructure owners. http://myforums.oracle.com/jive3/forum.jspa?forumID=2632

技术设计文档

术语抽取系统

Terminologies Extraction System

文档创建者: Soul Mr. Zhao / 赵翔

创建日期: 2014/9/21

最近修改日期: 2014/9/21

项目地址: <https://github.com/soulzhao/termsDiscoverer>

当前版本: 2.0

状态: 编写中

模板版本: 1.51

1. Applications Release should be one of the following: Fusion v0, Fusion v1, Fusion v2 or Future
2. The Creation Date and Last Updated dates are field codes. To update the dates (and all other field codes in this document), press CTRL-A followed by F9.
3. Refer to the Release Planning Spreadsheet to get the corresponding RPS ID for this document.
4. Product Functional Area should be...
5. Technical Design Document should contain the name of the prerequisite TDD for this document
6. Enter 1.0 into the Version field the first time you publish the document. The version is incremented every time a change is published for the document. Only the digit after the decimal point is incremented: when you reach 1.9, continue to 1.10, 1.11, etc.
7. Edit the Status above. Choices are IN DRAFT, IN REVIEW and APPROVED.
8. **INFORMATION ABOUT USING THIS TEMPLATE**
9. This template includes special features to make document creation easier.  
     
   **Oracle Menu**  
     
   In the main Word menu bar you should see a menu called **Oracle**. This menu contains the following options:
10. Replace Document Variables: Displays a dialog where you can enter replacement values for the document variables marked in blue throughout this template.
11. Strip Yellow Notes: Permanently removes the yellow notes (such as this one) from your document.
12. Insert Formatted Table: Inserts a new table with standard Oracle formatting.
13. Format Current Table: When invoked when the cursor is inside a table, this function will format the table with standard formatting while preserving the contents.
14. Insert List Table Inserts a two-column, borderless table that is useful for a list of phrases with definitions. The first column is set to bold. Tab in the last cell to automatically add a new row.
15. **Keyboard Shortcuts**
16. Use the following keyboard shortcuts to format headings and paragraphs as you type:
17. Ctrl-Shift-1 Heading 1
18. Ctrl-Shift-2 Heading 2
19. Crtl-Shift-3 Heading 3
20. Ctrl-Shift-4 Heading 4
21. Ctrl-Shift-5 Heading 5
22. Ctrl-Shift-N Body Text (N = Normal)
23. Ctrl-Shift-B Bullet (use this for Oracle standard bullets)
24. Ctrl-Shift-L Number List
25. Ctrl-Shift-H Highlighted Variable (for creating new document variables)

文档版本控制

修改记录

1. The first change record should indicate that the document has been published for review. Keep track of all the changes made to the TUD after it has been published for review

| 修改日期 | 修改人 | 版本 | 修改内容 |
| --- | --- | --- | --- |
|  |  |  |  |
| 2014/9/21 | 赵翔 | 1.0 | 编写系统架构，目标平台，里程碑信息以及QA |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

审稿人

1. Identify all contributors to the TUD. You can include details about each person’s contribution in the Content Contribution field.

| 姓名 | 单位 | 职位 |
| --- | --- | --- |
|  |  |  |
| Soul, Mr. Zhao 赵翔 |  |  |
| Yingjie Zhang, 张英杰 |  |  |
|  |  |  |
|  |  |  |

参考文档

1. Include all references to documents that served as inputs into the TUD, or that are related to this TUD in the Document References table. For example, if the TUD refers to a Technical Solution Approach document, include the URL to this document here.

| 文档名 | 文档类型 | 文档地址 |
| --- | --- | --- |
|  |  |  |
| 术语抽取需求分析 | Local MS Word File | <requirements.docx> |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

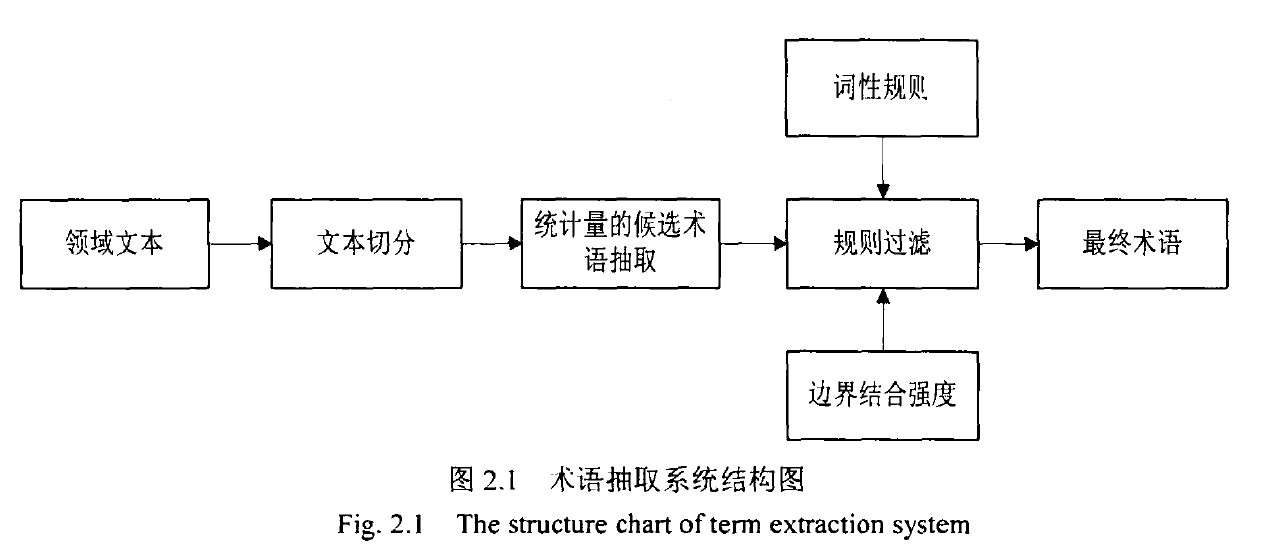
# 目录

# 项目简介

1. Author should include Uptake Summary, Expected Outcomes and Scope Approach. If applicable, include references to TDDs or other development artifacts that consuming teams should review. This section should not include background information; consumers should refer to related FSAs, FDDs, TSAs or TDDs for that information.

### 1.1 项目总览

术语抽取系统旨在对不同类型文章做处理，利用计算机自动抽识别和抽取出文章术语。术语抽取系统主要由三部分组成：文本切分，基于词频分布和信息熵的候选术语抽取，规则过滤。



### 1.2 进度安排

本项目采用里程碑管理，每个里程碑交付不同的成果。

项目分为四个阶段： 一. 分词阶段， 二. 候选词抽取阶段，三. 规则校验及过滤阶段，四，调试以及调整

| 阶段名 | 交付成果 | | | 预计日期 | |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | |  |
| **第一阶段：**  **分词阶段** | 输入： 文章文档  输出： Patricia Tree索引文件  词分布信息文件 | | | 2014/10/7 | |
| **第二阶段:**  **候选词抽取** | 输入：Patricia Tree索引文件（或直接内存索引）  词分布信息文件(或直接内存结构)  输出：候选词列表 | | | 2014/10/20 | |
| **第三阶段：**  **规则校验及过滤** | 输入：候选词列表  词分布信息文件(或直接内存结构)  输出：术语列表 | | | 2014/11/2 | |
| **第四阶段：**  **调试调整** | 为达到目标识别率，对规则等做调整 | | | 2014/11/08 | |

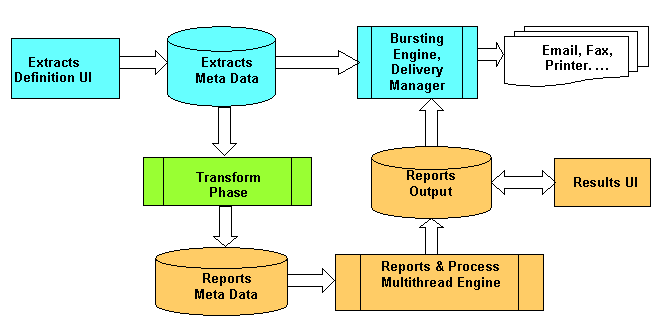
# 技术方案

1. Author should include Uptake Summary, Expected Outcomes and Scope Approach. If applicable, include references to TDDs or other development artifacts that consuming teams should review. This section should not include background information; consumers should refer to related FSAs, FDDs, TSAs or TDDs for that information.

### 2.1 目标平台技术

考虑到系统的

Extract Block/level defines the set of data that can be extracted from each logical entity, and block criteria

 **Process Flow Diagram of HCM Extracts**

#### Extract Definition

The complete Extract Definition will comprise Parameters, Blocks and Block Links, Filter Criteria, Records, Data Elements, Conditional Actions and Delivery options.

### 2.2 基本架构

HCM Extracts is a reporting or interface tool used for high volume complex reporting, extraction and archiving. It uses the Payroll Multi-threaded generic reporting engine for processing.

Extract Block/level defines the set of data that can be extracted from each logical entity, and block criteria (condition expression or fast formula) would define filter conditions on each logical entity for data extraction. Extract block links form a hierarchy of blocks that pass the context information during processing.

### 2.3 核心细节

HCM Extracts is a reporting or interface tool used for high volume complex reporting, extraction and archiving. It uses the Payroll Multi-threaded generic reporting engine for processing.

Extract Block/level defines the set of data that can be extracted from each logical entity, and block criteria (condition expression or fast formula) would define filter conditions on each logical entity for data extraction. Extract block links form a hierarchy of blocks that pass the context information during processing.

# Q & A 疑难点