

体系结构方案设计

—091250232 钟晓诚

目录

- 一：关注点2
- 二：体系结构需求定义3
 - 体系结构需求描述和设计约束3
 - 用例视图5
 - 非功能用例场景6
- 三：初始体系结构.....8
 - 逻辑视图8
 - 开发视图10
 - 进程视图12
 - 部署视图13
- 四：设计决策14
- 五：最终高层体系结构15
 - 逻辑视图：16
 - 开发视图16
 - 进程视图17
 - 部署视图18
- 六：风格19

一：关注点

系统的关注点见下表所示：

关注点	类型	描述	灵活性
网页爬取	功能需求	实现对网页的采集工作，用于站点资源的监视和资料库的更新	采集过程中，可以构造适当的启发策略，来指导机器人的路径选择和采集范围，减少文档采集的盲目性
内容处理	功能需求	对收集到的内容进行处理，提取特征元素	灵活性不大，要提取的特征元素可能发生变化
全文索引	功能需求	为收集到的内容建立索引以便于检索	灵活性不大，基本稳定
快速检索	功能需求	根据用户提供的检索条件实现快速的匹配	匹配算法有可能发生变化
排序	功能需求	将搜索结果按相关度进行排序，把最相关的结果放在最前面	排序算法有可能发生变化
用户接口	功能需求	为用户提供适当的交互界面，对用户输入词汇进行解析	解析算法可能发生变化
定时爬取	质量属性	网页爬取能够定期执行，定期更新储存库	时间可能发生变化
并发爬取	质量属性	爬取应该能够多机器（>=3）同时并发进行	并发机器数目有可能发生变化
可扩展性	质量属性	系统能够存储大容量数据，能够分布式使用多台机器的存储设备	能够在 2 小时内添加新的数据存储设备以扩充存储容量
安全性	质量属性	系统中储存的内容应该加密	加密算法可能发生变化
及时性	质量属性	系统应反应及时	能够在 10 秒内给出查询结果
可靠性	质量属性	系统应及时发现系统中的故障	能够在 1 分钟内发现各服务器及进程的故障
易用性	质量属性	系统要具有高易用性	在查询时，能够返回“非字符匹配”的相

			关结果
容错性	质量属性	系统可能发生故障，但必须拥有尽快修复故障的能力	系统应能够在 4 小时内能够恢复工作
可修改性	质量属性	系统的要求可能会发生变更	可能的变更点包括：爬取算法；对爬取网页的解析规则；加密算法；检索匹配算法；排序算法等等
法律规则	质量属性	系统应能够进行敏感词过滤	敏感词随时可以调整
商业规则	质量属性	系统能够实现竞价策略，可按照加权的方式对某些搜索结果的先后顺序进行调整	加权算法可能发生变化
人员技能	开发环境	团队成员对搜索引擎开发技术了解欠缺	灵活性不大，只有加强团队的学习能力
团队组织	开发环境	项目计划有时间限制，在学期结束前必须有系统原型交付	灵活性不大，项目交付时间基本不会变化
无	商业环境	无	无
软件环境	技术环境	不要求多平台、多浏览器的系统实现	灵活性变化不大
硬件环境	技术环境	系统应运行在至少八台机器上	拥有随时增加计算或存储设备的能力
支撑技术	技术环境	系统在开源的搜索引擎框架上修改完成	灵活性变化不大

二：体系结构需求定义

体系结构需求描述和设计约束

系统的体系结构需求描述和设计约束如下表所示：

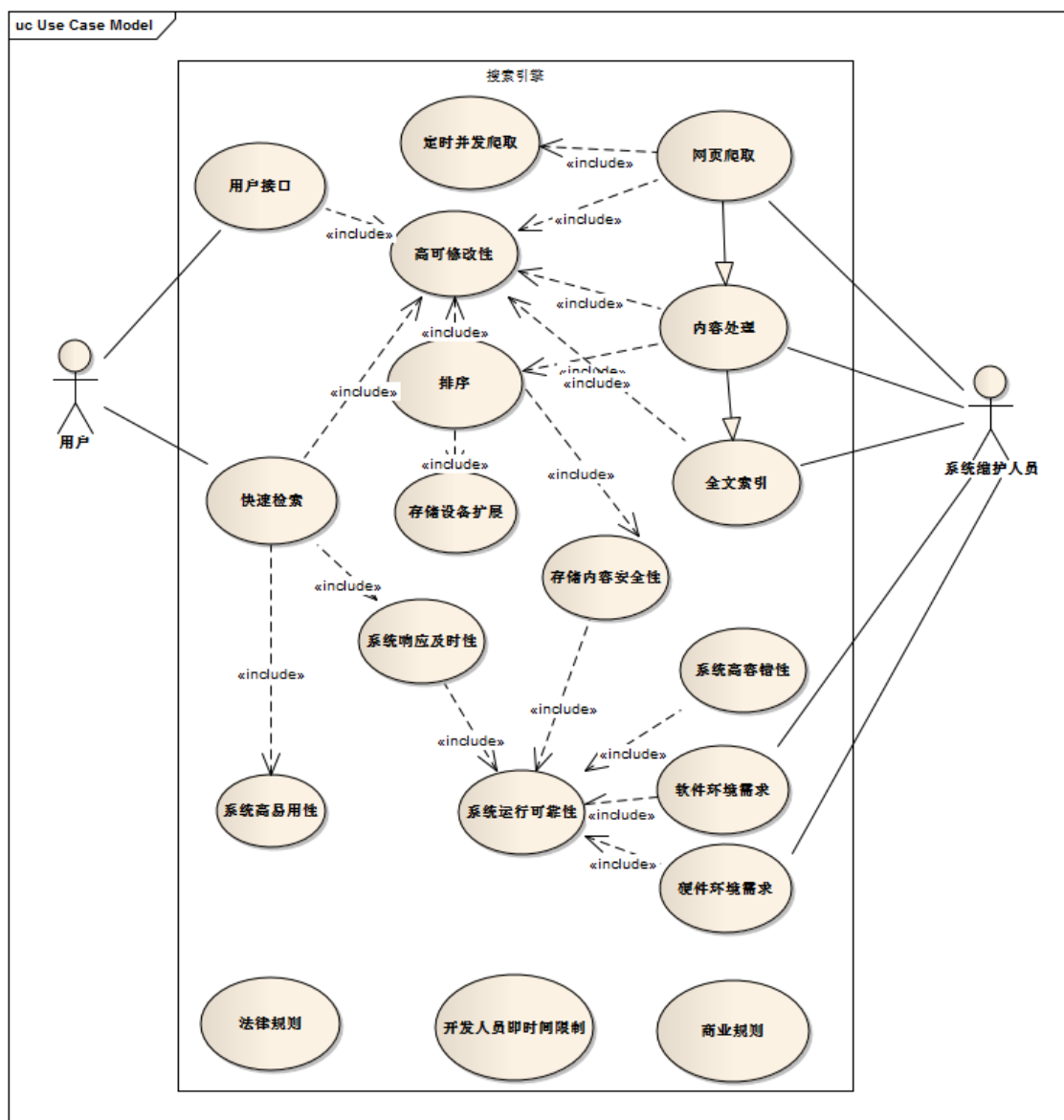
体系结构需求 ID	描述	设计约束	相关约束	优先级(小为高)
R1	网页爬取功能	C1 适当的启发策略，减少盲目性		1
R2	内容处理功能	C2 系统应能够		2

		处理项目过程中特征元素发生变化的情况		
R3	全文索引功能	无		2
R4	快速检索功能	C3 系统应能够根据用户提供的检索条件实现快速的匹配		1
R5	排序功能	无		2
R6	提供用户接口	C3 系统应能够根据用户提供的检索条件实现快速的匹配		2
R7	定时爬取	C4 爬取应定时进行，定期更新存储库	C1	3
R8	并发爬取	C5 系统应能够支持多处理器并发爬取功能	C1, C4	3
R9	存储设备的可扩展性	C6 系统应能够在 2 小时内添加新的数据存储设备		3
R10	存储内容的安全性	C7 系统应对存储数据提供加密算法		3
R11	系统响应的及时性	C8 系统应能够在 10 秒内给出查询结果	C3	3
R12	系统运行的可靠性	C9 系统应能够在 1 分钟内发现各服务器及进程的故障		3
R13	系统的高易用性	C10 系统在查询时应能够返回“非字符匹配”的相关结果	C3, C8	3
R14	系统的高容错性	C11 发生故障时系统能够在 4 小时内能够恢复工作	C9	3
R15	系统具有高可修	C11 系统的要	C1, C3, C4	3

	改性	求随时会发生变更		
R16	系统应遵守法律规则	C12 敏感词随时会变更	C11	4
R17	系统遵守商业规则	C13 加权算法变更	C11	4
R18	开发人员要求	C14 8-10 小组		5
R19	开发时间要求	C15 学期结束前		5
R20	软件环境要求	无		
R21	硬件环境需求	C16 系统应运行在至少八台机器上	C5, C6	3

用例视图

下面是系统的用例视图：



非功能用例场景

下面是对系统的质量属性需求定义可验证的场景描述：

项目	内容	
场景 ID	S1	
商业目标	扩大并发爬取机器数量	
相关需求和设计约束	R8	C5
场景内容	刺激	新的爬取机器
	刺激源	系统维护人员
	环境	爬取机器数量不足

	制品	搜索引擎爬取子系统
	响应	增加并发爬取网页的机器数目
	响应的度量	爬取机器数量>=3

项目	内容	
场景 ID	S2	
商业目标	扩展存储设备数量	
相关需求和设计约束	R9	C6
场景内容	刺激	新的存储机器
	刺激源	系统维护人员
	环境	数据存储机器数量不足
	制品	搜索引擎数据存储子系统
	响应	增加新的存储机器数目
	响应的度量	2 小时内添加新的数据存储设备

项目	内容	
场景 ID	S3	
商业目标	响应的及时性	
相关需求和设计约束	R11	C8
场景内容	刺激	新的用户查询请求
	刺激源	用户
	环境	查询环境
	制品	搜索引擎查询系统
	响应	快速响应用户的查询
	响应的度量	在 10 秒内给出查询结果

项目	内容	
场景 ID	S4	
商业目标	运行的可靠性	
相关需求和设计约束	R12	C9
场景内容	刺激	系统发生故障
	刺激源	系统
	环境	系统运行错误，发生故障
	制品	系统
	响应	系统及时检测出错误原因
	响应的度量	在 1 分钟内发现各服务器及

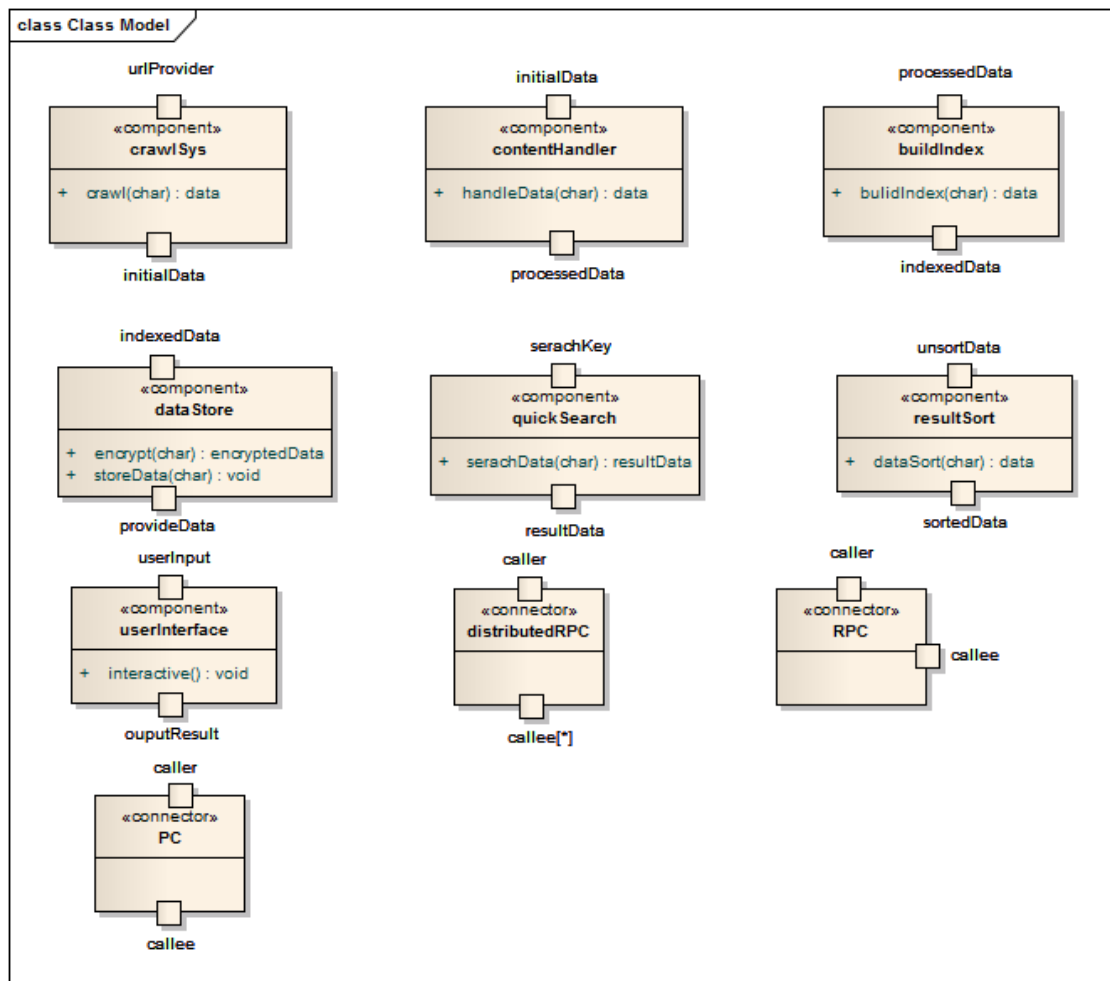
		进程的故障
项目	内容	
场景 ID	S5	
商业目标	系统的高容错性	
相关需求和设计约束	R14	C11
场景内容	刺激	系统发生故障
	刺激源	系统
	环境	系统运行错误，发生故障
	制品	系统
	响应	系统恢复正常运行
	响应的度量	发生故障时系统能够在 4 小时内能够恢复工作

三：初始体系结构

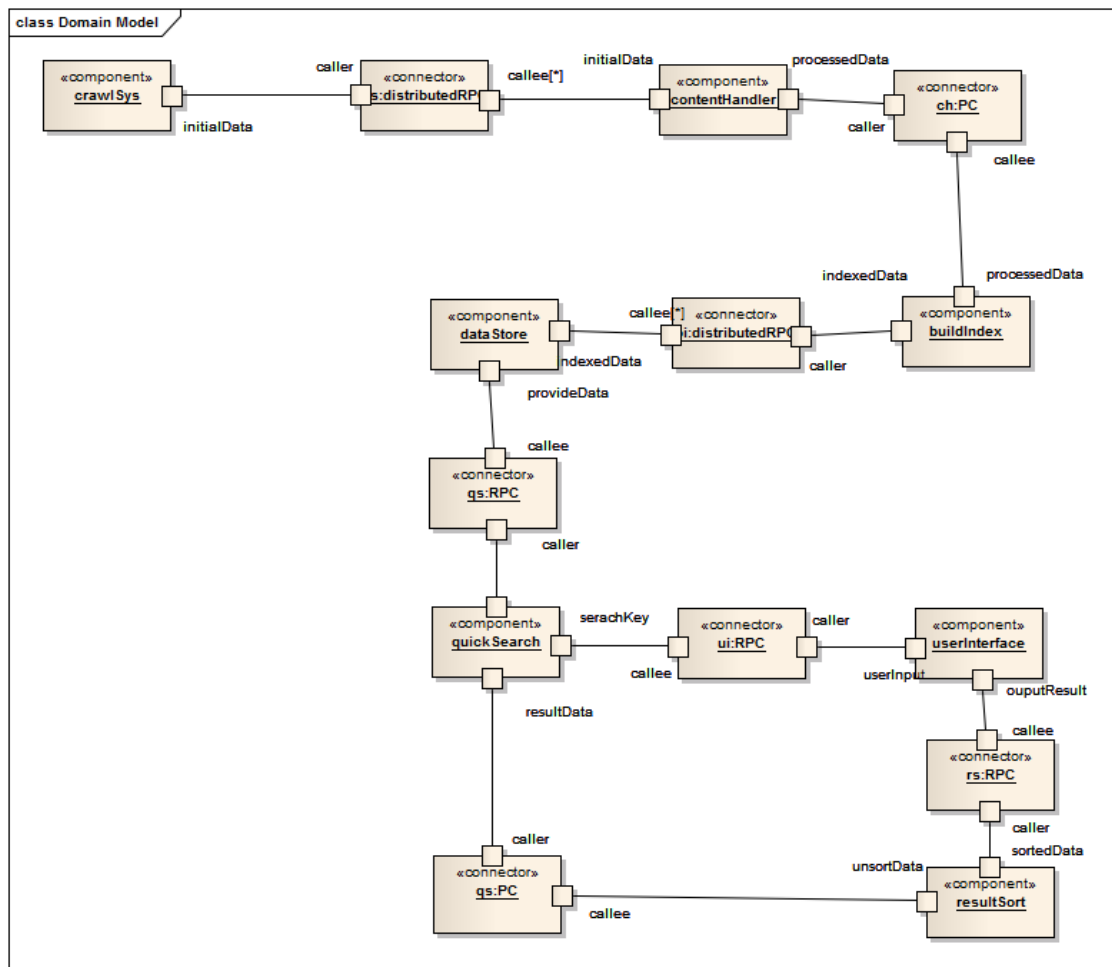
从空白开始建立初始软件体系结构，利用模块化、信息隐藏等基本的设计原则分析和处理概要功能需求，建立能够反映功能需求概要的体系结构雏形；根据体系结构关注点，依据经验与直觉，使用相关的设计策略，修改上一步的体系结构雏形，建立初始软件体系结构。下面分别给出初始体系结构的逻辑视图、开发视图、进程视图和部署视图。

逻辑视图

下面是系统的逻辑视图：

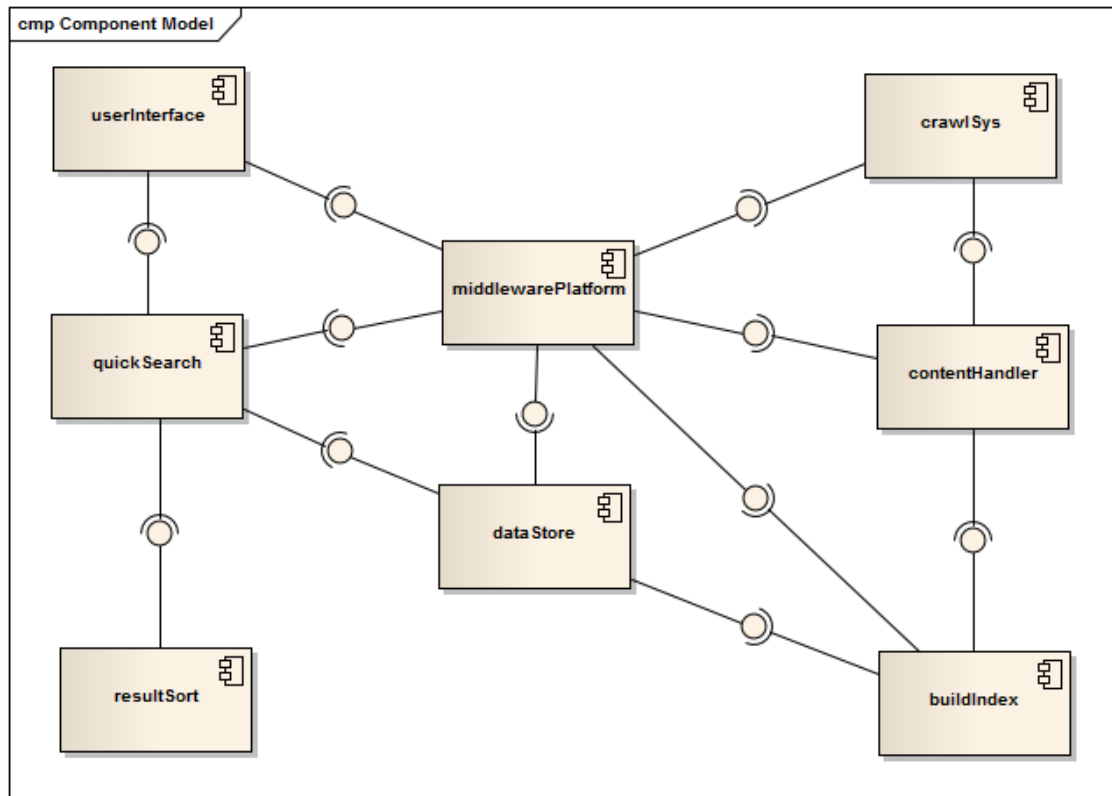


下面用对象图描述系统配置:

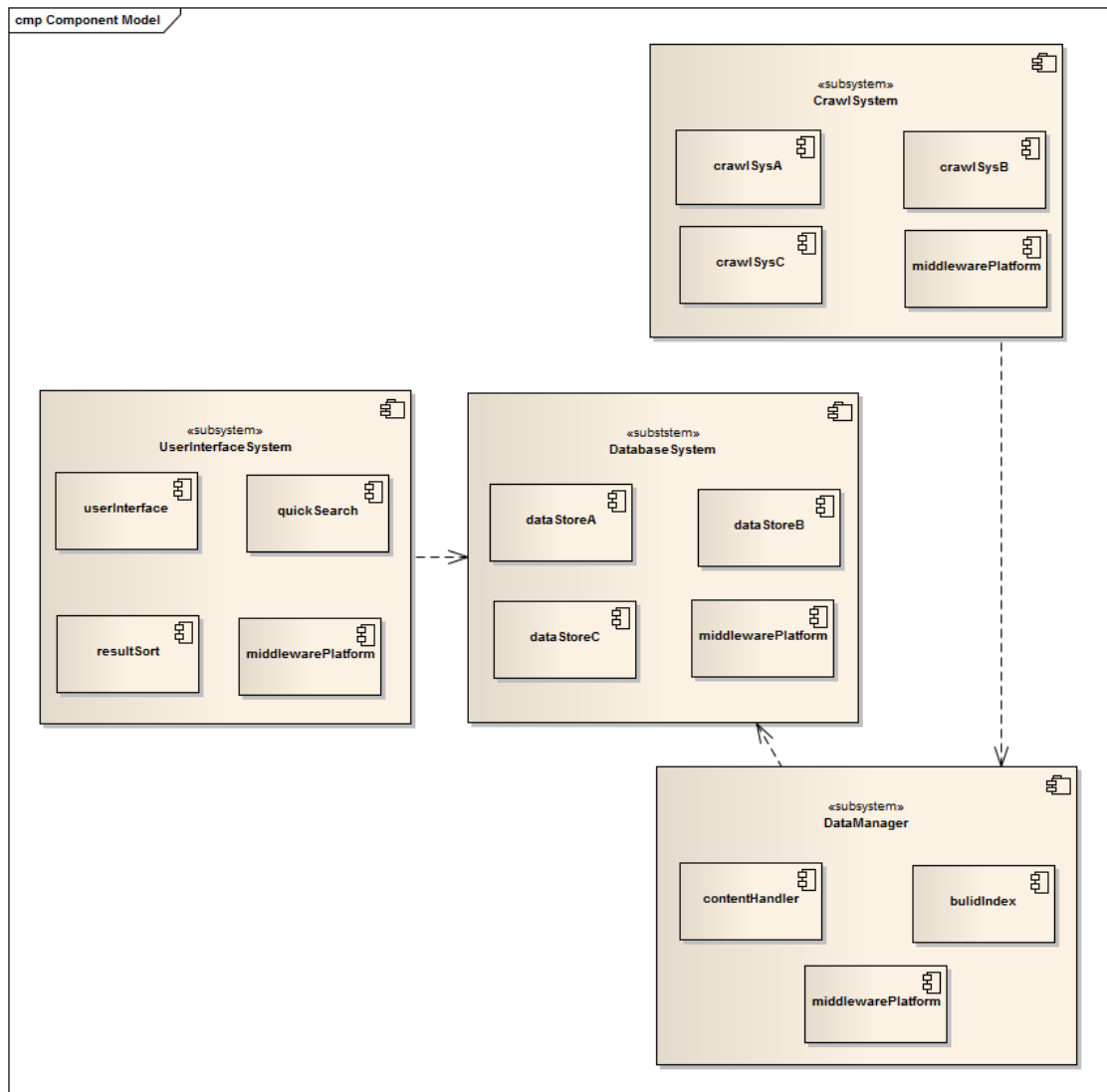


开发视图

下面是系统体系结构实现模块的组织：

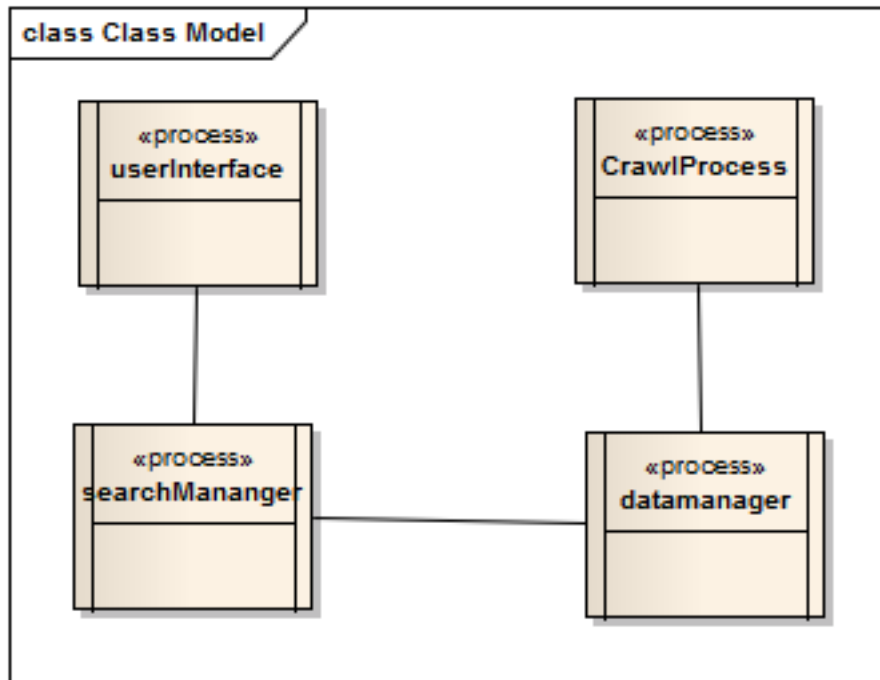


下面是系统模块的子系统组织和分层组织：

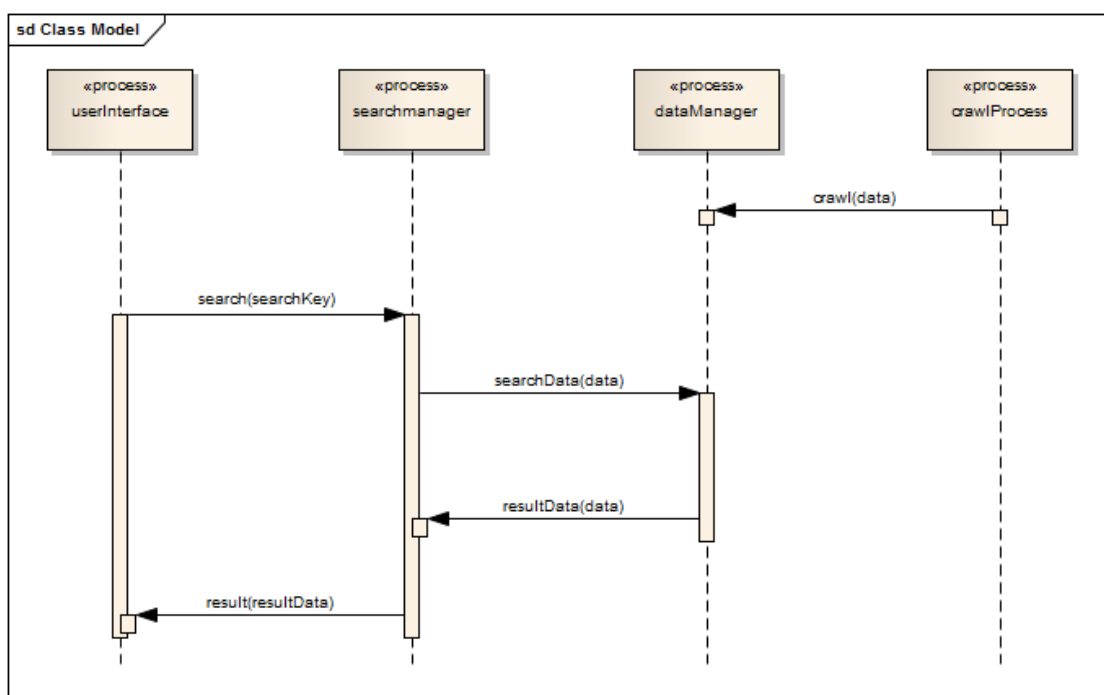


进程视图

下面是系统的进程视图：

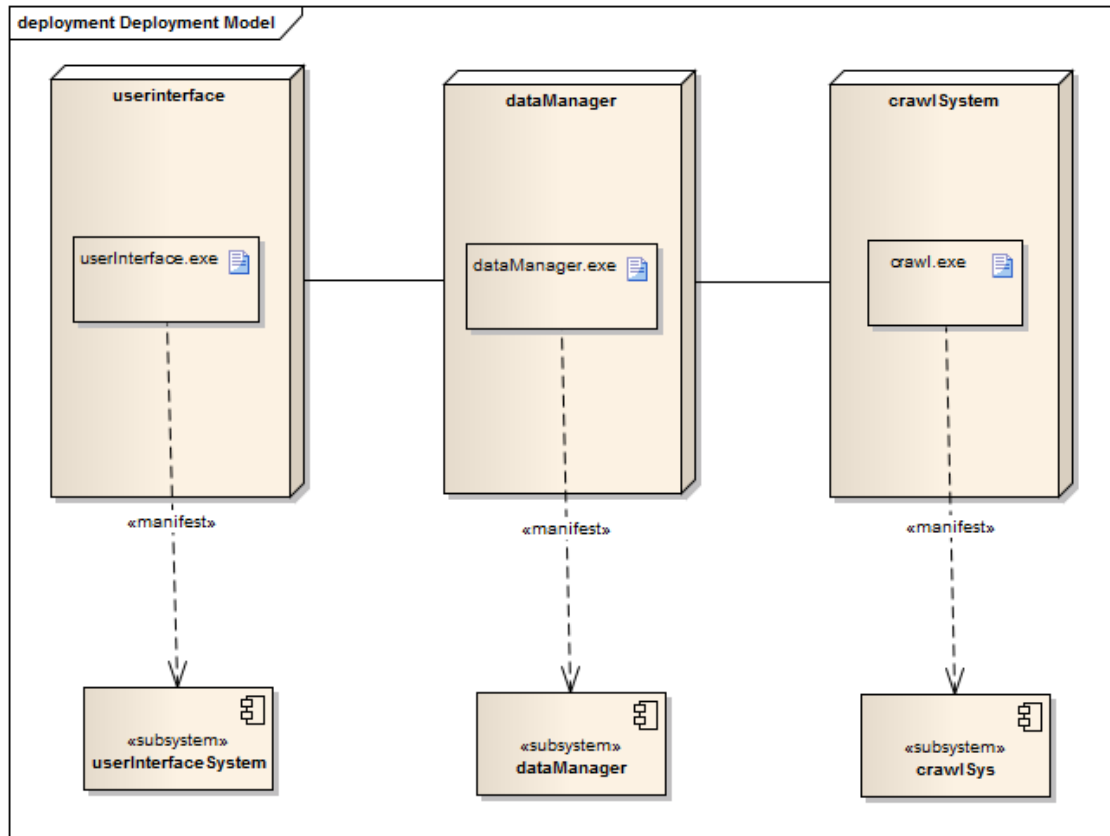


下面的顺序图表示的是系统进程之间的通信机制：



部署视图

下面是系统的部署视图：



四：设计决策

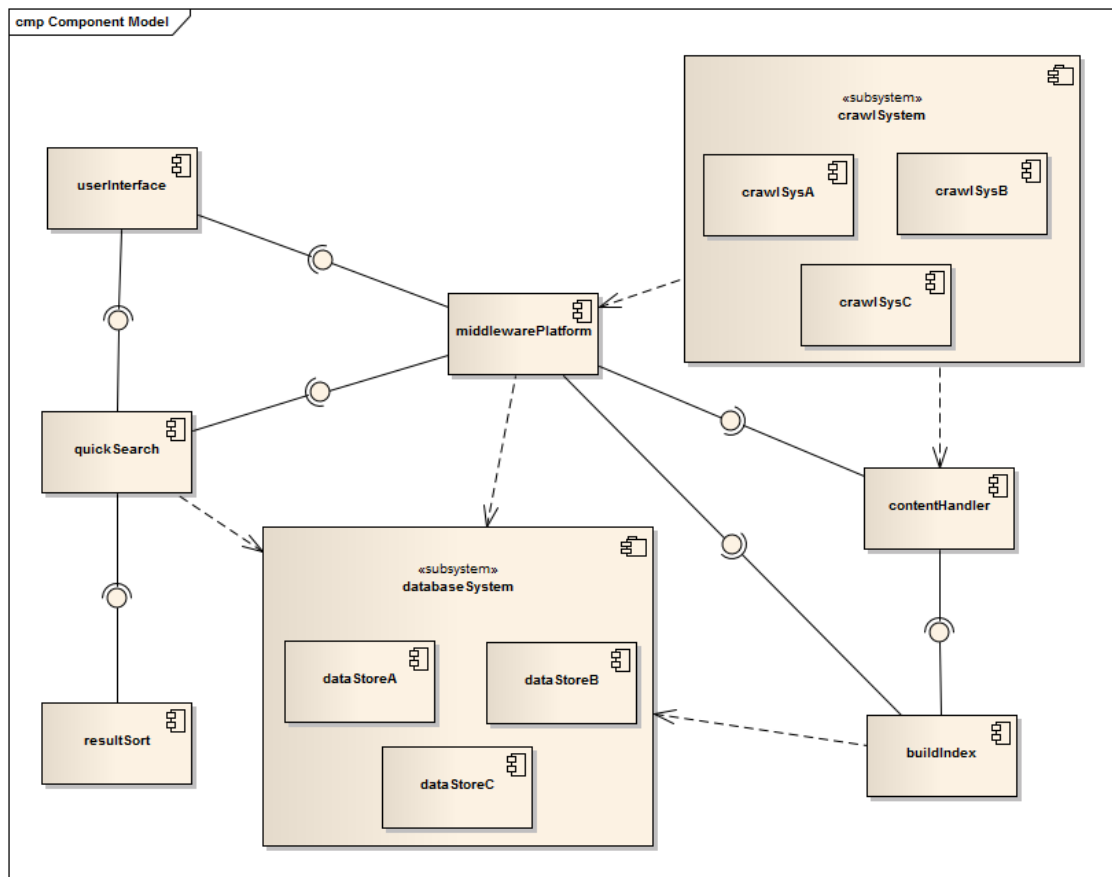
下面是针对系统的质量属性需求和约束定义的设计决策：

- R7 定时爬取
 - 编号：D1
 - C4 爬取应定时进行，定期更新存储库
 - 把爬取进程设置为定时任务
 - 影响：进程视图，部署视图
 - 详细设计约束：对爬取进程进行修改，变为定时执行任务
- R8 并发爬取
 - 编号：D2
 - C6 系统应能够支持多处理器并发爬取功能，要求至少 3 个以上处理器并发
 - 把爬取过程作为一个独立的进程，增加并发爬取网页的机器数目
 - 影响：进程视图，部署视图
 - 新的体系结构约束 CA1：一致性更新与 Cluster 访问
- R9 存储设备的可扩展性

- 编号：D3
 - C6 系统应能够在 2 小时内添加新的数据存储设备
 - 封装数据存储过程，提高存储设备的可扩展性
 - 影响：逻辑视图；开发视图；进程视图；部署视图
 - 详细体系结构约束 CA1：一致性更新与 Cluster 访问
- R11 系统响应的及时性
- 编号：D4
 - C8 系统应能够在 10 秒内给出查询结果
 - 设计算法进行查询优化，设计用户接口和存储数据端的通信 3 间隔
 - 影响：开发视图
 - 详细体系结构约束 CA2：通信规则
- R12 系统运行的可靠性
- 编号：D5
 - C9 系统应能够在 1 分钟内发现各服务器及进程的故障
 - 使用 Ping/Echo 方法检测服务器故障
 - 影响：所有 4 个视图
 - 详细设计约束：Ping/Echo 规则
- R14 系统的高容错性
- 编号：D6
 - C11 发生故障时系统能够在 4 小时内能够恢复工作
 - 使用冗余服务器提高可靠性，发生故障时使用冗余服务器
 - 影响：所有 4 个视图
 - 详细设计约束：一致性更新与 Cluster 访问
- R19 开发时间要求
- 编号：D7
 - C15 学期结束前至少能提交项目的 beta 版本
 - 使用分层式结构，方便并行开发
 - 影响：开发视图

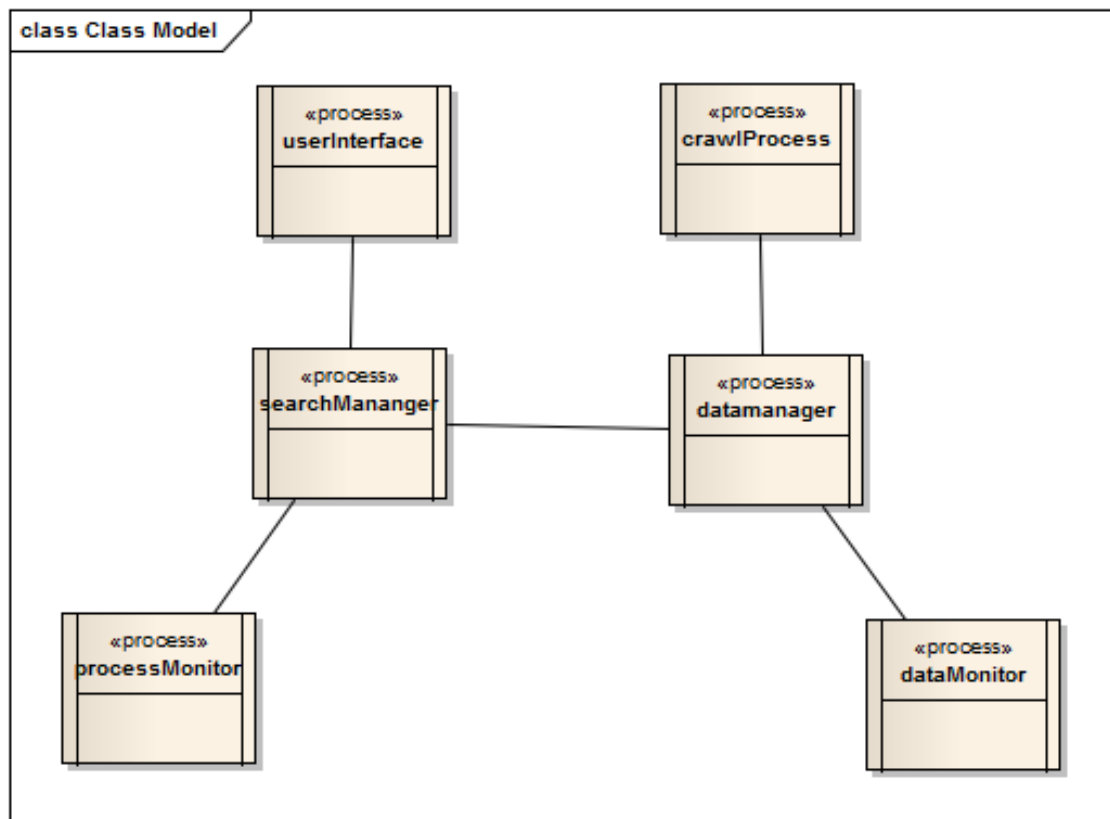
五：最终高层体系结构

使用 UML 表示法和 4+1 模型描述最终的高层结构，下面分别给出系统最终高层体系结构的逻辑视图、开发视图、进程视图和部署视图。



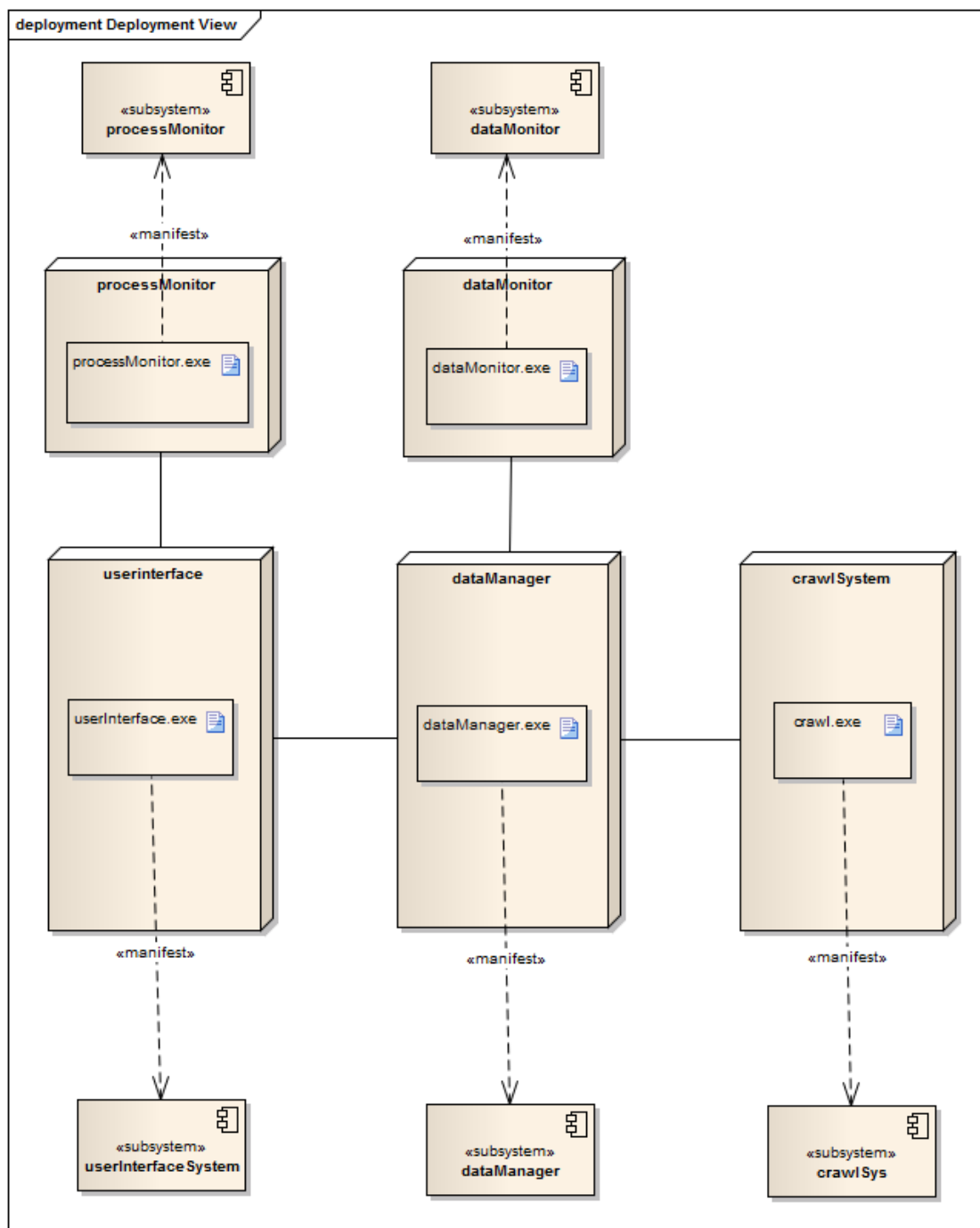
进程视图

下面是系统最终高层体系结构的进程视图：



部署视图

下面是系统最终高层体系结构的部署视图：



六：风格

风格编号	S1	风格类型	Layered Style
层次	Module-Level	相关功能	快速检索/结果排序 模块可以分层设计； 网页爬取/内容处理/

			全文索引模块可以分层设计
质量要求	R4, R5; R1, R2, R3;		
决策依据	C3; C1, C2;		
影响视图	逻辑视图, 开发视图		

风格编号	S2	风格类型	Object-Oriented Style
层次	Module-Level	相关功能	快速检索, 结果排序 网页爬取, 内容处理, 全文索引模块都可以运用面向对象风格进行设计
质量要求	R1, R2, R3, R4, R5;		
决策依据	C3, C1, C2;		
影响视图	逻辑视图, 开发视图		

风格编号	S3	风格类型	Point-to-Point Architecture Style
层次	Process-Level	相关功能	系统中各个模块之间的通信
质量要求	R9, R11;		
决策依据	C6, C8;		
影响视图	逻辑视图, 开发视图, 进程视图, 部署视图		

风格编号	S4	风格类型	Publish-Subscribe Architecture Style
层次	Process-Level	相关功能	系统中各个数据存储机器与数据管理模块之间的通信
质量要求	R9;		
决策依据	C6;		
影响视图	逻辑视图, 开发视图, 进程视图, 部署视图		

风格编号	S5	风格类型	Blackboard Architecture Style
层次	Process-Level	相关功能	系统中各个网页爬取机器与网页爬取管理模块之间的通信
质量要求	R7, R8;		
决策依据	C4, C5;		

影响视图	逻辑视图，开发视图，进程视图，部署视图
------	---------------------

风格编号	S6	风格类型	Distributed Architecture Style
层次	Physical Unit-Level	相关功能	由于系统中的不同模块需要部署在不同的机器上，所以采用分布式系统体系风格
质量要求	R8, R9, R12, R14, R21;		
决策依据	C5, C6, C9, C11, C16;		
影响视图	逻辑视图，开发视图，进程视图，部署视图		