第34卷 第15期 计 算 机 工 程 2008年8月
Vol.34 No.15 Computer Engineering August 2008
·软件技术与数据库· 文章编号：1000—3428(2008)15—0083—02 文献标识码：A 中图分类号：N945

基于LSA的Web信息采集和统计服务

李晓婷[1]，张 磊[2]，沈建京[2]
(1. 西安通信学院通信装备管理系，西安 710106；2. 解放军信息工程大学电子信息工程系，郑州 450001)

摘 要：在网络信息时代，传统的统计预测方法已经不完全适用，而对特定领域的信息采集和统计的需求日趋明显，使有效定向采集和统计特定领域信息并得到其相应的预测结果成为一个日益重要的研究方向。该文通过运用汉语分词、潜在语义分析和语义匹配等技术，构造了用户兴趣模型，并同时使用了面向服务的体系结构来设计该Web信息采集统计服务，通过具体的实验验证了对Web信息结构分析和未知信息相关性预测来控制信息采集统计的效果。

关键词：信息采集；潜在语义分析；面向服务的架构；Web服务

Web Information Collection and Statistics Services Based on LSA

LI Xiao-ting[1], ZHANG Lei[2], SHEN Jian-jing[2]
(1. Department of Communication Equipment Management, Xi'an Communication College, Xi'an 710106;
2. Department of Electrical Information Engineering, PLA Information Engineering University, Zhengzhou 450001)

【Abstract】In network information age, the traditional statistics and prediction methods have not been applicable to Web information collection and statistics anymore and owing to the requirements of information collection and statistics in special area are clearer than before, it makes the effectively directional information collection and statistics in special area and getting the corresponding predictive results become a more important research direction. This paper applies the technologies of Chinese word segmenting, Latent Semantic Analysis(LSA), semantic matching, and constructs a user interest model. In the mean time, it uses Service-Oriented Architecture(SOA) to design the Web information collection and statistics service, and validates the effect of the analysis of Web page gathering structure and unknown information, forecast for the relevance of Web page to control the information collection and statistics by concrete experiments.

【Key words】information collection; Latent Semantic Analysis(LSA); Service-Oriented Architecture(SOA); Web service

1 概述

在大量使用的基于关键词信息检索方法中，所考虑的仅仅是孤立的关键词，其他的语言成分并没有被利用，因而极易出现检索结果与用户的要求出现偏离。面对这些急需解决的问题，本文提出了基于LSA的Web信息采集统计服务的方法，该方法是基于SOA的潜在语义分析的定向采集相关信息，可避免无关信息的采集和处理，缩短采集时间、减少信息存储、加快检索时间和节约网络资源。

所采集到的信息质量直接关系到整个检索系统是否能够为该领域的用户提供良好的检索服务，该方法是特定领域Web信息检索统计分析和预测的基础。引入了潜在语义分析的文档自动标引与检索技术，用统计方法来捕捉并量化了标引词之间的语义关联关系；再将不同类别的向量项与用户查询词进行语义相关度分析，从近义词、关联词及区分词3个角度对初始查询进行扩展，这些扩展词可以从检索到的文档中获得，并且不断在检索过程和算法运行过程中根据用户的兴趣进行更新。在算法中对每次检索结果根据用户兴趣模型采用文本分类的聚类方法进行分析，找到聚类中心，把聚类中心向量项作为标引词，这样不仅极大提高了信息采集和预测的准确性，而且使信息采集统计服务系统在统计学习的基础上具有一定的自主性、智能性。

2 信息采集统计服务系统的结构设计

采用SOA的架构，以4层应用模型来实现主动式信息采集预测统计服务系统，分别是：表示层，业务外观层，业务层，数据服务层。该系统的结构如图1所示。4层应用模型实现了系统组件间的松散耦合，提高了系统的灵活性，增加了组件的可重用性，便于企业内系统及企业间系统的集成。它将业务功能包装成Web服务，使其与用户界面和数据访问相分离，使系统的维护变得简单，同时可以通过采用组件技术，降低企业服务器的负担，从而提高性能。在图1给出的系统结构中，企业内部端的普通用户、管理级用户和其他已有系统，以及企业外部端的普通用户和其他系统均处于表示层，并为其提供交互接口，使用于用户终端程序及其他应用服务可以方便地接入本系统。业务外观层主要提供标准化的服务接口及相关的业务逻辑和控制。业务层包含打印服务、描述服务、图生成服务等一系列的企业服务。业务外观层接受表示层提交的根据用户请求生成的用户兴趣模型后，调用位于业务层的相关服务模块，执行具体的事务逻辑，并向相应的服务发送请求，该服务向位于数据服务层的数据库服务器提出数据请求，并将结果返回至该服务，再将其传送给应用服务器，再由应用服务器将最后的请求响应传送给用户终端或其他应用服务。本系统在业务层，设计了7个服务模块。分别实现了打印服务、描述服务、图生成服务、图模板服务、报表服务、表模板服务、用户兴趣模型服务等功能。

基金项目：郑州市重大科技基金资助项目(052SGBG21076)
作者简介：李晓婷(1982－)，女，讲师、硕士，主研方向：分布式计算与应用；张 磊，硕士；沈建京，教授、博士生导师
收稿日期：2007-09-16 **E-mail：**christy_t@163.com

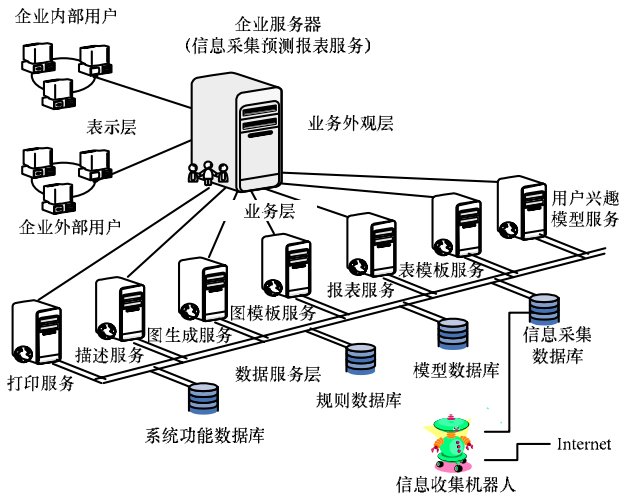

图 1　系统结构

3　系统框架描述

本文使用的是现有的较成熟的搜索引擎(如 Google, Yahoo 等)，作为海量 Web 资源库来代替大规模的资料库。由于自身没有文本库预处理的过程，相应地增加了搜索引擎选择模块，然后直接利用搜索引擎返回的结果进行相关文本的选择。

先简要介绍这个模型的每一个模块：

(1)需求分析：当用户把自己的需求描述提交到系统以后，系统开始对问题进行分析，并将这些需求做相应的标示，以供下一模块的输入。

表 1 列出了常见的需求类型。

表 1　需求类型

需求类型	特定词	例子
Who	谁	XP 都谁在使用
When	什么时候/何时/哪年	A 公司什么时候营业额过万亿
How	多少/几/多高/多长	快乐男生决赛当天短信票数有多少
What	是什么/什么是	SOA 架构的产品适合什么样的企业
Where	哪里/哪/什么地方	赈灾物资发往哪里
Why	为什么	为什么猪肉涨价
Other	……	……

针对不同需求类型制定相应的匹配信息抽取规则，以便在信息抽取阶段可按照相应的规则进行需求模式的匹配[1]。例如对于“对定义统计分析”类型的问题，其统计结果应该是，由查询到信息中句子的主动型的主语或被动型的宾语构成，然后对其结果利用 LSA 进行聚类操作。规则的制定要概括性强、简练，输出符合自然语法，具有较强的可读性。

(2)信息范围限定：根据需求信息的关键词的限定，对互联网的大量信息进行归类和提取，一般是利用搜索引擎得到相关文本信息后，再把文本解析成句子。在这个层次上，对句子进行分析和标示，然后根据其语义进行判断，是入库还是丢弃。

(3)信息库预处理：对于待访问的文本库，在形成查询检索之前，需将信息库处理成系统可以理解的形式，如进行分词、词性标注和语段标识等。

(4)候选资料选择分析：从信息库中抽取一些可能含有所需信息的文本，并对这些候选文本进行句子拆分、词性标注和语段分析。

(5)相关资料匹配及排序正则：得到了相关信息文本后，系统一般是利用需求分析模块中所选取的问题关键字、用户兴趣模型与相关文本中的句子结合相应规则进行匹配来得到候选结果，通常要分析这些候选信息的类型，如人物/组织、日期型等，并按关键字出现的权重对这些候选答案进行统计和排序。

(6)用户兴趣模型：通过样本信息集 LSA 矩阵，可以构建一个等效 m 维向量——用户兴趣模型 M 来表示用户所感兴趣的信息领域，通过用户兴趣模型简单快速地获取这些领域的相关信息。页面文档的相关性是决定是否需要采集索引该页面的重要指标。

返回统计结果，并将最终的结果套用模板库的格式返回给用户。

4　系统关键技术

4.1　SOA 和 Web services[2]

SOA 是一个基于组件模型的分布式软件架构，它将应用程序的不同功能单元(称为服务)通过这些服务之间定义良好的接口和契约整合起来。基于 SOA 构建的企业级架构将独立于实现服务的硬件平台、操作系统和编程语言。通过面向服务的架构所提供的方法，可以构建企业级应用系统，将应用程序功能作为服务提供给终端用户应用程序或其他服务。而基于它的应用程序服务具有松散耦合、位置透明、协议独立等特点。SOA 的目标在于让 IT 变得更有弹性，以便更灵活、更快地响应不断改变的企业业务需求及用户需求。

SOA 的具体实现有很多，包括 Web services, CORBA, JINI 等。但由于 Web services 已经比较成熟，且具有许多成熟的技术规范的支持，并且越来越受到软件开发人员的重视，因此成为构建 SOA 的主要技术。本系统的实现就是采用 Web services 技术作为支撑。

Web services 是封装成单个实体发布到网上并提供 API 以供其他程序使用的功能集合。它建立于普遍使用的 http 协议之上，采用跨平台的 XML 语言作为统一的数据描述格式。

Web services 核心技术都可以很好地支持 XML，包括：Web service 描述语言(Web Service Description Language, WSDL，用于进行服务的统一描述、发现和集成规范)、UDDI(Universal Description, Discovery and Integration，用于服务的描述、发布和集成)、简单对象访问协议 (Simple Object Access Protocol, SOAP，用于服务调用)。

4.2　潜在语义分析(LSA)[3]

LSA 是一种基于潜概念索引的检索技术，是利用大型文本语料库及统计计算方法提取和表征词汇语境意义的一种理论和方法。

LSA 的第一步是构造一个文档-项矩阵，这是一个大型稀疏矩阵，矩阵的行对应文档，列对应项(一般为出现在文档中的字或主干词)。矩阵元素的值一般为：出现在文档中的项的频率数(第 j 个项出现在第 i 个文档中的次数)。其中，不常出现的项对于它们关联的重要性是无足轻重的。

然后，通过奇异值分解(SVD)方法，求出文档-项矩阵的一个降秩的近似矩阵。在 LSA 中，降秩的过程称为塌陷奇异值分解，其目的是使每一个文档和项由一个词汇表中全体词的维数低得多的向量来表示。这样，当进行语义分析时，处理的文档就会映射到一个低维空间，并与在这一空间中的文档进行比较。文档之间的相关度可以用它们所对应的向量之间的余弦值来表示。由于对于文档和项采用了降维表示，这种表示使空间包含了与空间维数相同的潜概念，LSA 的本质是，用这些潜概念的线性组合来描述文档和项。因此，它能

(下转第 88 页)