第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)06—0079—03 文献标识码：A 中图分类号： TP311.52

基于本体的论文复制检测系统

聂规划，付志超，陈冬林，刘平峰
(武汉理工大学电子商务研究所，武汉 430070)

摘 要：针对目前学术论文抄袭现象严重的问题，在分析目前学术论文复制检测研究方法的基础上，提出基于本体的论文复制检测系统框架模型，分别从本体访问层、本体表示层、本体映射层描述论文复制检测系统的框架。利用语义网本体技术，探讨论文本体的构建和论文相似度的计算。

关键词：复制检测；语义万维网；本体；领域本体

Ontology-based Thesis Copy Detection System

NIE Gui-hua, FU Zhi-chao, CHEN Dong-lin, LIU Ping-feng
(Institute of e-business, Wuhan University of Technology, Wuhan 430070)

【Abstract】 Aiming at serious thesis copying problems, this paper presents systematic frame model of ontology-based thesis copy detecting system which is based on analysis about foreign and domestic researching method and existing problems, and describes framework of thesis copy detecting system from three layers, such as ontology access layer, and ontology represent layer, ontology map layer. Semantic and ontology technology is utilized to discuss the build of paper ontology and calculation of paper similarity, which has certain theoretical meaning for the build of thesis copy detecting system.

【Key words】 copy detection; semantic Web; ontology; domain ontology

1 概述

电子学术资源获取的便利性为学术论文的抄袭、非法扩散等不道德行为提供了方便，必须进行学术论文非法复制的防止和检测。论文抄袭识别如仅靠人工来做，工作量巨大，效果也无法保证。

论文复制检测是针对学术论文的文本复制检测，其核心任务是判断论文文本之间的相似度。文本复制检测技术产生于 20 世纪末，已有很多典型的系统。根据其采用的算法，可分为 2 类：基于数字指纹的字符串匹配方法的系统和基于词频统计的相似度计算方法的系统[1-2]。前者有 SIF, COPS, KOALA, hingling, MDR 等系统，后者有 SCAM, DSCAM, CHECK, CDSDG[3]等系统[1, 4]。上述文本复制检测系统具有如下的特点：

(1)主要采用数字指纹、关键词匹配、基于语形的相似度计算模型等识别技术，大部分不能解决语义方面的相似度。

(2)在文本表示方面采用的方法不灵活。基于字符串匹配的系统中，如何选取文本快是让系统使用者很难把握的因素。基于词频统计算法的系统中，以向量空间模型(VSM)来表示文本，然而 VSM 的相似度方法缺乏考虑语义的相似性。

(3)文本复制检测系统的结构一般都是按 COPS 的体系结构[1, 4]，如何构建文本库以及如何集成异构的文本库等问题在这种体系结构中不能得到很好解决。

本文针对文本复制检测系统中存在的不足，提出基于本体的论文复制检测系统框架模型。

2 基于本体的论文复制检测系统框架模型

2.1 基本原理

由于 Internet 上的信息过量，在 Internet 上获取信息的查准率和查全率等问题一直难以解决。万维网创始人 Tim Berners-Lee 提出了语义万维网(semantic Web)，希望从根本上解决这个难题。本体是某个领域内不同主体(人、代理、机器等)之间进行交流(对话、互操作、共享等)的一种语义基础。因此，可构建一个学术论文的领域本体，然后通过该领域来集成异构的论文库，将该领域本体中包含的知识表示和语义关系应用到论文的相似度检测中去。

2.2 设计目标

针对目前文本复制检测系统的不足，本文提出基于本体的论文复制检测系统的 4 个基本设计目标：

(1)集成各种异构论文库，例如中国学术期刊网、万方数字化资源系统和重庆维普资源系统的论文库，并能适配多种格式的论文。

(2)复制检测要综合考虑语义相似性，有自然语言识别能力。在论文复制检测系统中加入本体技术，将论文领域本体作为语义理解的基础，可在一定程度上实现自然语言的识别。

(3)提供一个统一的查询模型。用户通过统一的查询模型，透明的从各个异构的论文库检索论文。

(4)由于系统是基于本体的，应有能力为用户提供友好的、准确的检索请求表达方式，而不是让用户去根据需求切

基金项目：国家自然科学基金资助项目(70572079)；国家科技支撑计划基金资助项目(2006BAH02A08)；湖北省自然科学基金资助项目(2006ABA303)

作者简介：聂规划(1957－)，男，教授、博士生导师、博士，主研方向：知识管理，人工智能，电子商务；付志超，硕士研究生；陈冬林、刘平峰，副教授、博士

收稿日期：2008-07-13 **E-mail：**esteem_84@163.com

分成独立的关键词。

2.3 体系架构

根据设计目标，本文提出基于本体的论文复制检测系统框架模型，该系统框架模型分为 3 个层次：本体映射层，本体表示层和本体访问层，如图 1 所示。

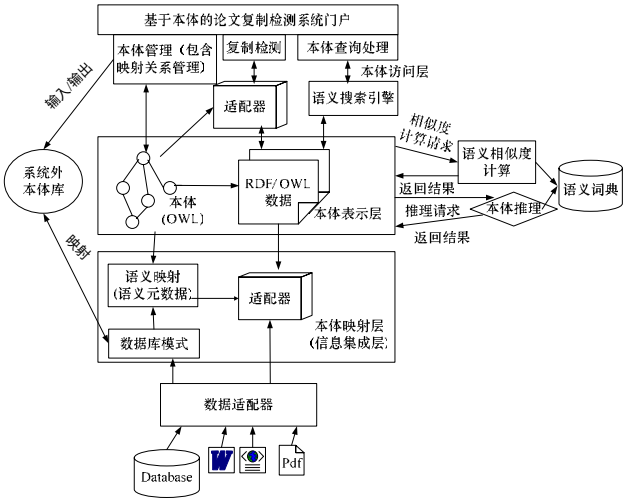

图 1　基于本体的论文复制检测系统框架模型

该系统的设计思想是首先建立异构论文库和本体的语义映射关系，然后生成 RDF/OWL 形式的语义数据，最后基于该语义数据进行相似度计算，提供对本体的语义查询，将计算或者查询结果返回给用户。

(1)本体映射层

要将数据库映射到本体，首先要建立数据库模式和本体之间的语义映射关系，该语义映射关系以语义元数据的形式来表示。由于数据库模式是二元模型，是由主键和外键建立表之间的关系，而本体属于层次模型，因此要建立语义映射，从数据库中抽取出本体是一个非常复杂的过程。本文借助 D2RQ 工具来实现数据库到本体的映射。D2RQ 是一个用来描述关系型数据库的 schemata 和 OWL/RDFS 本体之间映射的开源工具。它除了能将 RDF、RDB 统一起来，还实现了多个不同格式数据源查询的优化。

(2)本体表示层

该层主要负责存储本体，并进行语义相似度计算，响应推理请求。本体创建好后，如何持久化本体是一个非常重要的问题。用 RDF/OWL 进行本体化表示，形成元数据本体。由于这些元数据的值，即元数据本体的实例很多，如果直接把 RDF/OWL 数据当作 XML 存储，会存在如下问题：1)数据量大；2)无法响应并发操作；3)RDF/OWL 本身是一种描述框架，是以 XML 作为载体的，但由于 XML 模式不能区分 XML 属性和元素在含义上的不同，因此，对同样的信息内容，可将其映射成多种不同的 XML 结构，这样的后果是同一张用 RDF 描述的图映射成 XML 的结果可能并不是唯一的；4)基于 XML 的简单匹配很难满足 RDF/OWL 的灵活性，很多 RDF/OWL 的高级属性都不能匹配成功。

基于上述分析，本文采用关系数据库的方式存储本体。另外该层还要提供相似度计算和推理功能，这些都是建立在语义基础上的，因此，还要定义一个语义词典本体，用以描述论文所涉及到的各专业词汇之间的语义关系。

(3)本体访问层

本体访问是通过门户来实现的。门户是面向各类用户提供综合服务的窗口，门户是系统用户的统一入口。系统门户应该集成本体管理、复制检测、本体查询处理等功能。

本体管理从广义上来说主要是指本体的构建、编辑、验证、本体之间的映射以及本体进化，当然也应该包括映射关系管理。可见本体管理涉及的方面很多，并且也很复杂，不过随着语义网的发展，已经有一些专门的工具出现，例如 Protégé 可以很好地构建本体和编辑本体，前面提到的 D2RQ 工具，可完成对映射关系的管理。

复制检测是指将待检论文经过适配器预处理后得到论文本体的一个实例，然后将该实例与其所属分类的论文本体实例区匹配，经过语义相似度计算模块，计算出该论文与论文库中论文的相似度，从而来判断该论文是否是抄袭的。相似度计算流程如图 2 所示。

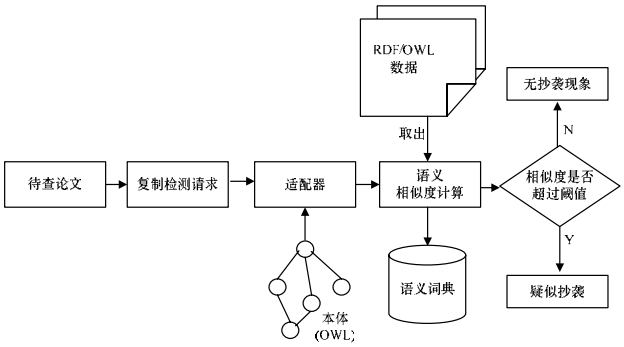

图 2　相似度计算流程

本体查询处理应包含传统的文献检索功能，另外还应该实现智能检索功能，即能根据用户的检索条件，进行语义推理，检索出与检索条件具有相同语义信息的文献。因此，对用户通过门户提出搜索请求，该功能模块需要对搜索条件作必要的语义推理、重写、分解以及优化检索条件，搜索语义数据。随着语义网的发展，W3C 推荐的语义搜索引擎的检索语言 SPARQL 已经成为标准。因此，系统将采用 SPARQL 语言来对本体进行查询。不管用户采用何种方式和门户的本体查询模块进行交互，系统都会将用户的搜索条件转换为 SPARQL 语言表示，对本体的实例(RDF/OWL 语义数据)进行查询，最后将查询的结果以规范的语义形式返回给用户。在用户看来，他能通过本体透明的对异构的关系数据库进行语义查询，而不必关心具体的数据模式。

3　基于本体的论文复制检测系统的关键技术

本文提出的基于本体的论文复制检系统框架的实现主要依赖于相关关键技术的实现，这些关键技术包括本体的构建、语义映射关系的建立、语义相似度的算法、语义推理、语义查询以及语义查询的转化等。

3.1　本体构建以及语义映射关系的建立

本体的构建过程是一个反反复复的叠加过程，须进行不断完善。构建本体的方式很多，可通过 Protégé 工具来手工构建本体，另外还可借助 D2RQ 工具半自动化生成本体。本系统中使用的本体属于领域本体，本文将采用手工和半自动化相结合的方法来构建该领域本体。

通过手工构建本体时，首先确定系统中本体的专业领域和范畴，然后通过和相关专家进行交流，确定该领域本体中的概念以及概念的属性、属性值，并确定概念间的关联关系。最后按确定的概念、属性等通过 Protégé 来构建本体，生成一个候选本体。

通过半自动化方式来构建本体时，主要通过考察数据库中的表、属性、主外键和包含依赖关系,定义一组从关系模型到本体的映射规则。基于这些规则能直接得到一个候选本体。

综合以上 2 个候选本体，进一步对该候选本体进行评价和精炼，生成最终的本体。综合考虑学术论的篇章结构[5]，确定的本体的具体结构如图 3 所示。

图 3 中定义了 3 类资源对象，分别是论文(Paper)、作者(Author)和段落(Paragraph)。在 3 类资源对象的基础上，还定义了 2 种对象属性(ObjectProperty)，其中，create 描述了作者和论文之间的写作关系，其定义域为 Author 类，值域为 Paper 类；对象属性 hasParagrap 描述了论文和段落之间拥有关系，其定义域为 Paper 类，值域为 Paragraph 类。

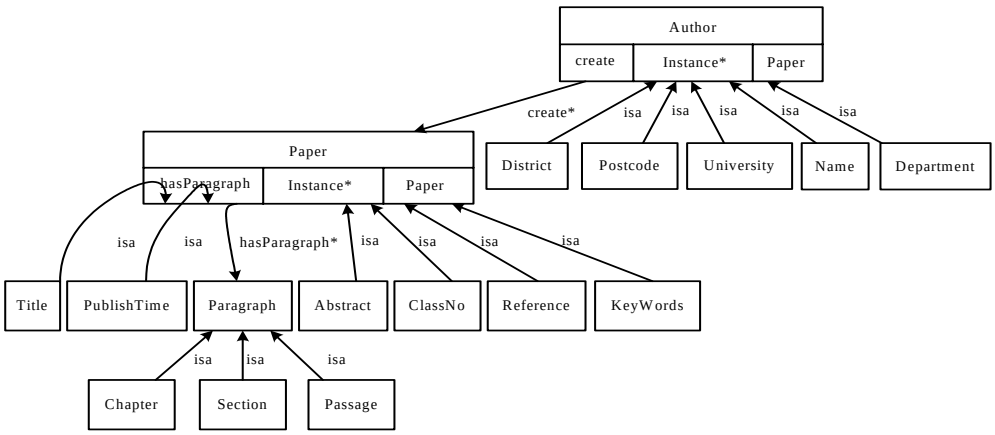

图 3　论文本体结构

数据库和本体之间的语义映射关系的建立关系到本体的构建和异构数据库的集成。由于关系数据库模型和本体模型不同，建立数据库和本体的语义映射关系是一个复杂的过程。可采用前面提到的工具 D2RQ 来实现语义映射关系。建立数据库于本体之间的语义映射不是要把数据库中所有数据语义化，其原则是只有数据库中的记录与本体中概念和属性具有真正意义上的语义联系，才会关注联系的类型以及怎样实现该语义联系，并将数据库的关系表转化为本体类，将数据库中的数据转化为实例。在建立语义映射关系的时候一般遵循如下规则：

(1)1 个表对应 1 个概念或 1 个表对应多个概念，以表中的某列(或多列)形成概念实例的标识。

(2)多个表对应 1 个概念，则该概念实例的标识需要这几个表联合起来，以每个表中的单列或多列通过连接才能形成该概念的实例标识。

(3)1 个表只是一个概念的一部分，将表做为这个概念实例属性的一部分即可。

(4)如果 1 个列对应 1 个属性，直接可以转化。

(5)如果 1 个列对应多个属性，即这个列的内容包含了 1 个概念的多个属性的内容。此时可编写专用的处理程序。

(6)多个列对应一个属性，则这些列必须合并才能从语义上构成这个属性的值。

3.2　语义相似度的计算方法

本文提出的复制监测系统模型是基于本体的，因此，在计算相似度方面，将充分利用本体的优势，综合考虑语义的相似性来计算论文的相似度。由于建立专业的语义词典需要大量的时间，本系统将采用知网(HowNet)作为语义词典。在综合考虑文献[6-7]中计算方法的优缺点后，按文献[6-7]中的计算方法，将论文相似度的计算划分为 3 个层次：词语层次，段落层次和篇章层次。

(1)词语层次，主要包括词语与词语、词语与句子、词语与段落之间的相似度计算。由于在汉语中实词才是表达文章意义的关键词汇，因此在相似度算法中省略了虚词部分的相似度计算，这样可在保证计算准确性的前提下提高计算效率。文献[6-8]给出计算 2 个实词语义相似度的详尽方法，计算公式为

$$SimWS(s_1,s_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}simWP_j(p_1,p_2) \tag{1}$$

其中，$\beta_i(1\le i\le 4)$ 是可以调节的参数，并有

$$\beta_1+\beta_2+\beta_3+\beta_4=1,\ \beta_1\ge\beta_2\ge\beta_3\ge\beta_4,\ simWP_j(p_1,p_2)(1\le j\le 4)$$

分别表示 2 个实词第一独立义原、其他独立义原、关系义原、符号义原的相似度，这 4 种义原相似度的计算方法在文献[6, 8]中都有详尽阐述。

文献[6]中考虑到除实词的第一义原相似度之外，其他义原相似度是相对独立的，本文在文献[8]的基础上将实词语义相似度算法进行了改进，改进的计算公式为

$$SimWS(s_1,s_2)=\beta_1 simWP_1(p_1,p_2)+\sum_{i=2}^{4}\beta_1 simWP_1(p_1,p_2)\times\beta_i simWP_i(p_1,p_2) \tag{2}$$

(2)段落层次，主要是段落与段落之间的相似度计算。本文在论文本体中，定义了 Paragraph 类，Paragraph 包含子类 Chapter、Section、Passage，分别表示章、节、段，该结构描述了论文的段落信息，抽取出论文的各个段落。借助向量空间模型(VSM)，将每段映射成为一个特征向量，则段落 d_i 可表示为

$$V(d_i)=\{(t_{i1},w_{i1}(d_i));(t_{i2},w_{i2}(d_i));\cdots(t_{ij},w_{ij}(d_i));\cdots\}$$

其中，$t_{ij}(i=1,\ 2\cdots,n;\ j=1,\ 2\cdots,\ m)$ 为一列互不雷同的词条；$\omega_{ij}(d_i)$ 表示 t_{ij} 在 d_i 中的权值。参考文献[7]中的完备二部图的最大权匹配算法，得到段落相似度的计算公式为

$$simPD(d_i,d_j)=\frac{x_1y_1\times\sqrt{\omega_{i1}\times\omega_{j1}}+x_2y_3\times\sqrt{\omega_{i2}\times\omega_{j3}}+x_3y_n\times\sqrt{\omega_{i3}\times\omega_{jn}}+\cdots+x_ny_2\times\sqrt{\omega_{in}\times\omega_{j2}}}{\min(n,m)} \tag{3}$$

其中 $x_l,\ y_k$,分别表示段落 $d_i,\ d_j$ 对应特征向量中的第 l 个和第 k 个词条

$$x_l y_k=SimWS(x_l,y_k),\ 1\le l\le n,\ 1\le k\le m$$

其中，n, m 分别表示段落 d_i, d_j 对应特征向量中词条的个数。

(3)篇章层次，主要依据式(3)得到的段落之间的相似度计算整篇文章的相似度。设待计算相似度的 2 篇论文为 $p_1=(d_{11},\ d_{12},\ \cdots,\ d_{1i},\cdots,\ d_{1m})$，$p_2=(d_{21},\ d_{22},\ \cdots,\ d_{2i},\cdots,\ d_{1n})$

其中，d_{1i}, d_{2i} 分别表示 p_1, p_2 的第 i 段和第 j 段。则文章 p_1, p_2 相似度的特征矩阵为

$$\boldsymbol{P}_{12}=\boldsymbol{p}_1\times\boldsymbol{p}_2^{\mathrm{T}}=\begin{bmatrix} d_{11}d_{21} & \cdots & d_{1m}d_{21} \\ \vdots & & \vdots \\ d_{11}d_{2n} & \cdots & d_{1m}d_{2n} \end{bmatrix} \tag{4}$$

其中，$d_{1i}d_{2j}$=sim$PD(d_{1i}, d_{2j})$。

(下转第 84 页)