《智能信息处理》课程考试

基于本体在电子商务检索中的研究

王悦

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 09 日

基于本体在电子商务检索中的研究

王悦

(大连海事大学 信息科学技术学院 大连 中国 116026)

摘 要 语义网的出现将会极大改善万维网时代网络信息低智能化、弱共享化的状况。它在电子商务检索方面的应用，将克服电子商务检索中信息难以有效准确识别的技术难题，从而加快电子商务进程。文章介绍了语义网，以及语义网与万维网的区别，着重阐述了语义网在电子商务检索中的应用。

关键词 电子商务检索；万维网；语义网；本体

中图法分类号 TP18 **DOI 号**：10.3724/SP.J.1016.2014.01229

Ontology-based Applications in E-commerce

WANG Yue

(Information Science and Technology College, Dalian Maritime University, Dalian 116026, China)

Abstract Emergence of the Semantic Web will greatly improve the World Wide Web network information era of low intelligence, weak share of the situation. It is used in retrieval of e-commerce, e-commerce will overcome the retrieval information is difficult to effectively and accurately identify technical problems, thus speeding up the process of electronic commerce. This paper introduces the Semantic Web, as well as the difference between the Semantic Web and the World Wide Web, the Semantic Web focuses on the application of e-commerce Retrieval.

Key words Retrieval of e-commerce; Web; Semantic Web; Ontology

1 引言

在互联网[1]日益普及的今天，人们充分体会到网络的巨大魅力。但现在所使用的互联网功能并不尽人意[2]，计算机不能理解网页内容的语义，以提供智能检索服务，更无法满足自然的“人-机-人”信息沟通与交互的需求，面对信息量呈几何级数膨胀的电子商务信息更是力不从心。新一代网络技术-语义互联网是一种智能化程度很高的网络技术，语义网[3]的开发与应用，可以打破电子商务信息检索的技术瓶颈，让用户通过网络进行的信息搜索，共享与交互更加便利。

2 语义网

2.1 语义网的介绍

为了克服目前互联网的缺陷，科学家们正在开展下一代能理解人类语言的智能网络-语义网的研究，语义互联网是对当前互联网的一种扩展，其目标是通过使用本体和标记语言，可扩展标记语言，资源描述框架等使互联网资源的内容能被机器理解，为用户提供智能检索，基于语义的内容检索和知识管理等智能服务。

简单的说，语义网或称语义 Web，是一种能理解人类语言的智能网络，语义网不但能够理解人类的语言，且还可以使人与电脑之间的交流变得像人与人之间交流一样轻松。语义网能使机器理解含语义的文档和数据。语义网就好比一个巨型的大脑，它由数据库智能化程度极高，协调能力非常强大的各个部分组成，可以解决各种难题。在语义网上连接的每一部电脑，都能分享人类历史上所有科学、商业和艺术等知识。它不但能够理解词语和概念，而且还能够理解它们之间的逻辑关系。在语义网中，网络不仅能够连接各个文件，而且还能够识别文件里所传递的信息，也就是说，它是一种聪明的

网络，可以干人所从事的工作。“语义网”是按照能表达网页内容的“词语”链接起来的全球信息网，是用机器很容易理解和处理的方式链接起来的全球数据库。语义网是对未来网络的一个设想，在这样的网络中，信息都被赋予了明确的含义，机器能够自动地处理和集成网上可用的信息。

2.2 语义网与物联网的区别

目前在万维网中，网页仅仅是一个单调的内容显示，电脑只负责将一个网页链接到另一个网页，网络不能按照用户的要求自动搜寻和检索网页，直至找到所需要的内容。而语义网则是希望计算机能“看懂”网页的内容，使计算机成为“智能”的导航工具。当然语义网还并不仅仅能完成这个功能，它比这还要“聪明”得多。语义网是对万维网本质的变革，它的主要开发任务是使数据更加便于电脑进行处理和查找。其最终目标是让用户能对因特网上的海量资源达到几乎无所不知的程度，计算机可以在这些资源中找到你所需要的信息，从而将万维网中一个现存的信息孤岛，发展成一个巨大的数据库。

语义网将使人类从搜索相关网页的繁重劳动中解放出来。因为网中的计算机能利用自己的智能软件，在搜索数以万计的网页时，通过“智能代理”从中筛选出相关的有用信息。而不像现在的万维网，只给你罗列出数以万计的无用搜索结果。“语义网”是由比现今成熟的网际搜索工具更加行之有效的、更加广泛意义的并且自动聚集和搜集信息的文档组成的。

总之，语义网是一种更丰富多彩、更个性化的网络，你可以给予其高度信任，让它帮助你滤掉你所所不喜欢的内容，使得网络更像是你自己的网络。语义网最大的好处是可以让计算机具有对网络空间所储存的数据，进行智能评估的能力。这样，计算机就可以像人脑一样“理解”信息的含义，完成“智能代理”的功能。使用语义网搜索引擎搜索的结果也将比万维网更为精确。此外，由于大部分科技创新和突破，都是对已有知识的重新组合和更新，因此语义网也为新的科技创新提供了无尽的资源，它可以在很短的时间内，完成一个人甚至需要一辈子才能做出的组合结果。

2.3 语义网的层次结构

语义网的核心内容是建立一个明确的语义空间。其中，需要解决的关键问题就是语义的表达和结构化的信息集合及推理规则。XML(Extensible Markup Language)可扩展标记语言和 RDF(Resource Description Framework)资源描述框架技术解决了这一问题。XML 是一种用于定义标记语言的工具，其内容包括 XML 声明、定义语言语法的 DTD(Document Type Declaration 文档类型定义)、描述标记的详细说明以及文档本身，它提供了灵活、通用、丰富的结构化信息表示方式，是整个语义网的基石。RDF 提供语义信息和推理规则的表达方式，是语义网表达语义的关键。理论上构建的语义网有 7 个层次组成，如图 1 所示。

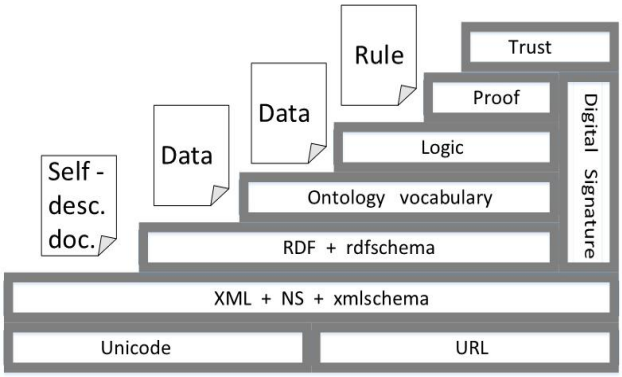

图 1 语义网结构图

Unicode 和 URI 层：保证用户使用国际化、通用化的字符集，同时也可以实现多国语言的混合存储和使用。

XML+NS+XML Schema 层：XML 是一种允许自定义标记的、描述 Web 文档和数据的结构化描述语言。命名空间 NS 为 XML 文档中的结构化标记提供上下文环境。XML Schema 在标记的使用上和文档结构上，为 XML 文档提供了明确的语义限制。

RDF/RDF Schema 层：RDF 定义一种用以描述资源及其相互关系的简单模型，是语义信息描述的有效手段。其基本数据模型包含 3 类对象:资源、属性和陈述；资源之间的关系通过属性和值来描述。RDF Schema 提供属性及属性间关系的表达机制，描述 RDF 的使用规则，定义领域字典，并用类型层次结构来组织该字典，构成完备的语义空间。

Ontology Vocabulary 层：在交流/通讯中扮演语义沟通的角色，用于描述的、概念化的显示说明。

Logic 层、Proof 层和 Trust 层：是语义表达的高级要求。其中，Logic 层提供推理规则的描述手段，Proof 层通过运用这些规则进行逻辑推理和求证，Trust 层则负责为应用程序提供一种机制，且对“是否信任”给出的论证，作出结论；Digital Signature 位于层次模型的右侧，并且贯穿于中间的

4 层。Digital Signature 是一种基于互联网的安全认证机制，通过它可以鉴别信息的来源和信息的安全性，用户据此决定是否接受该信息。

3 语义网在电子商务中的应用

3.1 基于语义网的电子商务信息检索模型

电子商务检索中，传统信息检索模型存在明显不足。在商品信息的组织与描述上，简单将关键词作为描述商品的基本元素。在检索操作上，通常是基于关键词的无结构查询，难以反映词语间各种语义联系，查询能力有限，误检率和漏检率很高，检索结果的真实相关度较低。为了解决这一问题，现在传统信息检索模型的基础上，引入语义网技术，以下就是基于语义网的电子商务信息检索模型见图 2。

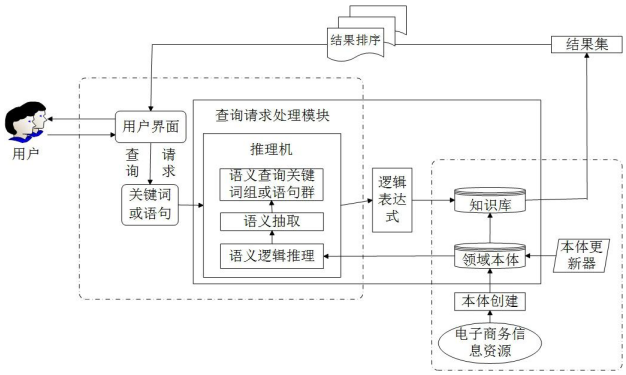

图 2 基于语义网的本体自动分类系统流程图

基于语义网的电子商务信息检索模型可分为 3 个模块：电子商务信息资源处理、用户接口及查询信息处理、检索匹配与输出。

3.2 电子商务信息资源处理模块

运用语义网技术，对电子商务信息资源进行有效处理，是本模型中的一个重要而关键的模块。利用语义网资源标注、概念检索技术以及电子商务领域中的分类体系和主题词表、语义字典等工具，构建能够充分描述电子商务信息资源领域知识的概念空间，建立本体模型，形成领域本体。在领域本体构建过程中，借助领域专家的帮助，充分运用专家知识和经验，捕捉相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出词汇和词汇间相互关系的明确定义。利用领域本体对文档进行标引。先从文档集中抽取出特征词汇，分析特征词汇，并建立与概念集之间的联系，从而达到使用领域本体对文档进行语义标引[4]。本体是共享的、形式化的、抽象的概念集合，概念及概念之间关系都已被精确地描述，通过语义标引方法，可以表示出商品信息中隐含的语义信息。商品的所属类别能够被更好地划分，并且概念之间有了明确的语义关系。本体构建的过程是一个持续的、不断修正的过程。领域本体初步建立好后，并不能一劳永逸。本体论是世界的反映，因此它必然随着现实的发展而变化[5]。在本模型中，采用本体更新器，根据信息资源的变化，对领域本体进行及时有效的扩充。本体更新器具备了根据网络信息的发展，及时更新领域本体中的本体知识的功能，如增加新的知识、修改不再适用的知识，并删除不再使用的知识。

在信息资源到领域本体构建的过程中，信息资源的元数据信息提取尤为重要。随着网络的发展，电子商务信息资源也在不断地丰富和扩充。在对信息资源进行合理整合组织的过程中，使用了元数据。元数据是描述资源属性、提供精确检索服务的结构化数据，能够被机器理解和处理，同时规范、标准的元数据也是构建信息共享平台的基础。由于数字化资源数量巨大，不可避免地造成了元数据标准难以统一，导致一些信息资源不精确甚至缺失、无法使用的情况。借助于 XML 组织文档，在无人工干预的情况下，依据原有电子商务信息检索中的元数据标准如 MARC、DC 等，提取用户感兴趣的元数据信息，并进行整合存储。首先，对数字化资源的文档等信息，去除在格式、内容、语言等方面存在问题或有严重缺失而影响使用的文档，即对信息资源进行初步整序，产生相对规整可用的文档信息，将不同格式的数字化文档转化为方便处理与解析的文本形式，存储在文档数据库中。采用 MARC、DC 等标准，根据数字化文档元数据的规范定义，产生提取元数据的各种应用模式，对文档数据库中的文档信息的元数据进行提取。为方便数据的共享与信息交流，一般采用 XML 将提取的元数据组织存储在元数据库中。而在元数据提取过程中，可以参考以下方法：区分文档各部分的重要标志，即对文档具有重要意义的关键词可认为是元数据提取过程的重要依据；对于许多文档中的普遍出现的元数据信息，可预先提取，对于符合某种共同模式的文档，采用统一模式。具体提取过程见图 3。

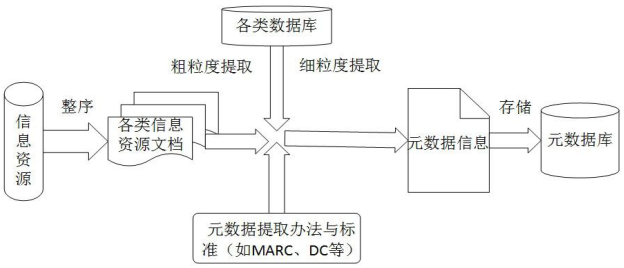

图 3 数字化文档元数据提取过程

在文档信息元数据提取过程中，将初步整序后的信息资源文档，利用 MARC、DC 等元数据提取办法与标准，以及各类数据库信息进行粗粒度提取，随后进一步细化，抽取出元数据信息存储于元数据库中。XML 不具备语义描述能力，而这一缺陷可以被 RDF 解决。RDF 定义一个简单的概念模型，指定相应的值，描述资源和资源之间的关系。RDF 以 XML 为语法基础，运用命名空间的思想，达到复用的目的，简化了程序，减少了创建元数据的工作量。在这些工作完成之后，利用元数据库中的信息、描述逻辑等以及电子商务领域中的分类体系和主题词表、语义字典等工具，在领域专家的知识和经验的帮助下，构建领域本体，最终存储在知识库中。信息资源领域本体构建流程见图 4。

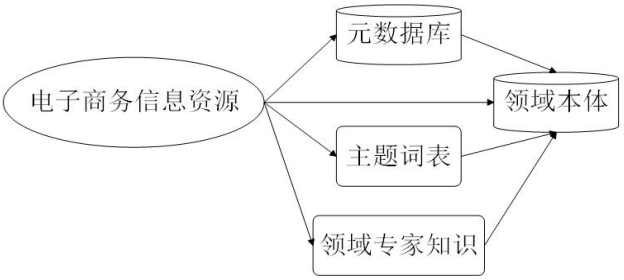

图 4 信息资源领域本体构建流程示意图

3.3 用户接口及查询信息处理模块

传统信息检索是直接使用用户输入的关键字来进行检索查询，但是其效果不尽如人意，主要原因在于容易对信息需求的理解失真，无法灵活更改信息需求，又难以表达检索出结果的相似性程度。同时，用户真正的检索意图很难用几个关键字表达清楚。人机交互开始受到重视与关注，成为信息检索效率提高所要研究的一个方面。用户接口的人机交互是建立在语义的基础上。在本模型的用户接口及查询信息处理模块中，当用户在用户界面采用自然语言输入查询请求，一般是关键词或语句，推理机首先对查询请求进行预处理，负责将用户提交的自然语言查询词语或语句转换成合适的本体查询词或语句。推理机是指系统中实现基于知识推理的部件，是基于知识的推理在计算机中的实现，主要包括推理和控制两个方面，是知识系统中不可缺少的重要组成部分。推理机在电子商务信息资源领域本体的基础上，利用本体领域内的知识和一些基本的自然语言理解技术对关键词或语句进行分析，通过语义相似度的计算，进行语义推理，从领域本体中抽取出与用户查询关键词或语句具有语义相似度的本体，最终得到用户真正的检索意图。

这一过程，仍然需要借助领域本体，将领域本体知识导入推理机中，推理机根据领域本体概念以及概念之间的关系对关键词或语句进行逻辑推理，找到相关的语义关键词或语句；利用语义相似度，进行抽取和推理，形成语句查询关键词组或语句群，或是以用户对关键词组或语句群选择后的最终结果，或是以系统分析选择后默认的关键词或语句，代替用户输入的关键词或语句，形成逻辑表达式，提交至检索匹配与输出模块。在这一模块中，构建的领域本体和语义相似度是语义查询请求提取的基础。这样用户只需一次查询就可以通过不同的语义关系进行检索，获得不同的商品，不需要反复地检索和检索扩展，提高了信息资源的利用率，也可以避免因为用户描述不当而带来的搜索误差，使信息检索变得快速而且准确。

3.4 检索匹配与输出模块

本模块的主要功能是采用处理后的关键词组或语句群，利用构建的领域本体，在知识库中搜索用户真正需要检索的信息。当用户输入查询请求，推理机对输入的关键词或语句，在领域本体的基础上，进行语义推理，抽取出语义关键词组或语句群，形成逻辑表达式，提交至检索系统。在检索时，系统能够对知识库中采用 RDF、RDFS 等描述的实例进行有效推理，推理过程需要依据具体的推理规则，而系统程序员则可根据具体情况创建合适的推理规则。然后，系统从电子商务信息资源中搜索出符合该语义词或句的所有相关商品资源，即结果集。这时，对得到的结果集进行整合，就要依据在查询请求处理子模块时，分析得到的语义关键词组或语句群，或结合用户对返回的语义关键词组或语句群的选择，与用户的原始查询请求对比，并根据与语义查询关键词组或语句的相关程度以及语义分析得到的用户检索意图，根据领域本体中的信息，依据一定语义相似度算法，综合相关信息和形成相关度数值，得出具体的语义相似度数值。然后以此进行排序，相似度越高，排名越靠前。将最终

经过排序的结果集通过用户界面返回用户。

4 结束语

虽然，目前语义网还处于初期发展阶段，从制定应用方案，构建整个网络体系，到服务于实践中还需要很长时间。但是未来，随着语义网技术的日趋成熟，语义网技术将更加全面地为电子商务检索向智能化的方向迈进，提供有力的支持。

参 考 文 献

1 中国互联网信息中心．第 28 次中国互联网络发展状况统计报告[R]．北京：中国互联网信息中心，2011

2 RenFu-ji.AdvancedInformationRetrieval[J].ElectronicNotesinTheoreticalComputerScience.2009,255:303-317

3 凌海云.基于语义网的智能搜索技术的研究与实现［D］.成都：电子科技大学，2004：4

4 何绍华,宫兆晖.基于语义网的网络信息检索相关性研究［J］.情报杂志，2007（2）:120－121

5 王文峰,赵莉.语义网中的本体对象研究及应用［J］.枣庄学院学报,2007（2）:52－53

6 周义刚, 姜赢. 语义网下动态知识组织模型构建研究[J]. 图书馆理论与实践, 2019.