基于语义的中文网页检索

余一骄[1] 刘 芹[2]
(华中师范大学语言学系 武汉 430079)[1] (武汉大学计算机学院 武汉 430072)[2]

摘 要 用户期望搜索引擎能提供基于语义的网页信息检索。基于本体、基于自然语言理解、基于文本统计分析的方法是实现中文网页语义检索的主要途径。分析了它们的实现方法、技术挑战和优、缺点;建议中文网页语义检索系统的开发应选择与普通用户联系紧密的应用领域,并以汉语词汇为索引单元,适量地采用中文信息处理技术。基于语义的中文网页检索应在以下方面加强研究:语义相关性评价方法、本体构建和实体抽取算法、基于语义的索引、大规模语义标注样本集开发等。

关键词 语义检索,中文网页,本体,分类,聚类,信息抽取
中图法分类号 TP391.3 **文献标识码** A

Semantic-based Chinese Web Page Retrieval

YU Yi-jiao[1] LIU Qin[2]
(Department of Linguistics, Central China Normal University, Wuhan 430079, China)[1]
(School of Computer, Wuhan University, Wuhan 430072, China)[2]

Abstract Semantic-based chinese Web page retrieval is a promsing application. The existing semantic retrieval mechanisms are categorized into three types, which are based on ontology, natural language understanding, and text classification and clustering respectively. The three technologies were reviewed and examined in detail. Semantic-based chinese Web page retrieval system should focus on popular fields to draw great attention from Web users. Moreover, Web pages should be indexed with words rather than chinese characters. Advanced chinese information processing technologies should be integrated into semantic retrieval systems. Some directions for future research were finally presented, including semantic relevance ranking, ontology definition and instance automatic extraction, semantic-based indexing, and large-scale semantic training collections construction.

Keywords Semantic information retrieval, Chinese Web page, Ontology, Classification, Clustering, Information extraction

1 前言

电子政务、电子商务、博客、微博等在我国发展十分迅猛,中文网页数量持续急剧增加。到2011年6月底,我国网络用户已达4.85亿,搜索引擎是受众最广、使用频率最高的互联网应用[1]。目前,通用搜索引擎主要提供基于字符串匹配的信息检索服务,其检索结果尚不能很好地满足用户的查询需求[2]。例如,2011年8月13日在百度中输入"人为"一词进行检索,得到的第2~4条检索结果分别是"人为什么活着"、"以人为本"、"犹太人为什么聪明"。从字符串匹配的角度来看,以上3条结果都是正确的。但它们却存在分词错误,并不符合查询者的检索意愿。如能根据语义,把待检索文本首先切分为词串,例如把"犹太人为什么聪明"切分为"犹太人"、"为什么"、"聪明",然后进行基于汉语词汇匹配的检索,则可以改进上述检索结果。由此例可见,未来有必要对中文网页语义检索技术进行更深入的研究。

中文网页中通常包括图像、视频、音频等多媒体数据。因基于内容的图像、视频、音频检索与文本检索差异较大,本文仅对中文网页文本的语义检索展开了讨论。中文网页文本检索与传统的中文全文检索存在不可忽视的差异。首先,中文网页内包含了〈title〉等HTML标记信息,利用这些标记信息有助于提高信息检索质量。其次,中文网页中除了有效文字内容外,还包括版权、导航、广告等无用的信息。信息检索系统获取中文网页后,须先对网页进行预处理,过滤噪声文本。再次,即使是经过预处理的中文网页文本,与全文文本还是有区别。全文检索中处理的大多是完整的句子,而中文网页中经常用不完整的句子表述重要信息。例如,大学教师的个人网页中,常用表格来描述教师的姓名、籍贯、学位、职称、办公地点等信息。除去HTML标记后,这些离散的词串并不能构成符合中文文法的句子。对不完整的句子进行词法分析、语

到稿日期:2011-09-28 返修日期:2011-11-22 本文受教育部人文社会科学研究项目(10YJA740120)、湖北省教育厅人文社会科学研究项目(2010b032)资助。

余一骄(1978-),男,博士,副教授,主要研究方向为信息检索、计算机网络,E-mail:yjyu@mail.ccnu.edu.cn;刘 芹(1978-),女,博士,讲师,主要研究方向为计算机网络。

法分析，难度会增加。因此，中文网页语义检索不仅需要传统的中文全文检索技术，还需针对中文网页的信息表述特征，研究特定的解决方案。

基于语义的网页检索与自然语言类型联系紧密，针对不同语种的语义检索研究进展差异很大。目前，英文网页语义检索研究最深入，中文网页语义检索研究还比较少。英语是屈折语，汉语是孤立语，二者的特征差异导致中、英文网页语义检索面临着不同的技术挑战[3]。已有的中文网页语义检索研究大多以追随英文网页语义检索技术为主，而对中文网页的特点以及汉语与英语的差异考虑得还不够充分。

本文致力于回顾中文网页语义检索成果，归纳中文网页语义检索的实现途径，并比较各种实现途径的优、缺点；还将讨论中文网页语义检索研究面临的 3 个重要问题，并展望将来应重点研究的方向。

本文第 2 节介绍中文网页语义检索的实现途径；第 3 节讨论中文网页语义检索面临的 3 个关键问题及解决方案；第 4 节展望未来的研究方向；最后总结全文。

2 中文网页语义检索的实现途径

基于语义的网页检索研究大致可以分为以下 3 类：基于本体、基于自然语言理解、基于网页文本统计分析。以下介绍 3 种途径的实现方法、面临的技术挑战，以及优、缺点。由于本文关注的焦点是中文网页语义检索的实现途径，因此对各类实现途径所涉及的中文信息处理技术、分类算法、聚类算法、机器学习技术等不作详细介绍。

2.1 基于本体的语义检索

本体(Ontology)是包含了概念、关系、公理、实体的四元组，它是描述领域知识的一种有效方式。早期基于本体的中文网页搜索，大多是依据本体实现查询扩展[4]。近年来，基于本体的中文网页语义检索则是把信息检索操作转换为针对特定问题进行逻辑推理求解的过程[5,6]。

2.1.1 实现方法

基于本体的语义检索，主要依赖领域知识，根据不同概念之间的关系、公理进行逻辑推理，将推理结果作为语义检索的输出结果。图 1 宏观地描述了基于本体技术实现中文网页语义检索的步骤。

步骤 1　针对特定应用领域定义本体；

步骤 2　从待检索的网页集中抽取有效信息，生成 RDF 三元组；

步骤 3　获取用户查询请求，对请求进行语义标注；

步骤 4　对 RDF 三元组进行逻辑推理；

步骤 5　将推理结果反馈给用户。

图 1　基于本体的语义检索

学术界已定义多种本体描述规范，如 RDFS(Resource Description Framework Schema)、SHOE(Simple HTML Ontology Extensions)、OWL(Web Ontology Language)等。本体定义、描述工具软件也较多，例如美国斯坦福大学开发的开源软件 Protég 在国内外语义网系统开发中应用得很广泛。尽管网页中发布的信息内容变化频繁，但是领域知识的变化却很缓慢。本体一旦定义，则长期不变。因此，图 1 中的步骤 1 通常只做一次。

对比较简单的语义检索应用，例如，科技文献共享系统、大学学术信息搜索系统等，本体包含的概念、属性、关系、公理都很清晰，且数量较少。对这类应用场景，手工定义本体是高效的实现方式。然而即使是概念和属性都很少的本体，它们包含的实体(instance)数量也可能很大。例如“作者”是科技文献共享本体中的一个概念，待检索网页中所列出的论文作者可能高达几十万人[7]。

自动推理研究一般是针对结构化数据，而中文网页文本是非结构化的。因此，语义检索系统须从待检索网页中抽取出与特定应用相关的句子，然后把句子中的有效内容转化为语义推理引擎能够识别的结构化信息。从非结构中文网页中，抽取结构化信息的过程，称为中文网页的语义标注。RDF(Resource Description Framework)三元组是目前语义检索中使用最广泛的结构化信息存储方式。因此，基于本体的网页语义标注，也就是把无结构的网页文本信息转换为结构化的 RDF 三元组[5]。

一个面向应用的中文网页语义检索系统，其待检索网页数量至少数以万计；且随时间推移，还不断地有新网页加入到待检索文档集中。对海量的中文网页进行语义标注，需要全自动处理，并且算法的执行性能应尽可能优化。语义标注会产生大量的 RDF 三元组，其存储与管理也需要深入研究。

语义检索系统收到用户提交的查询请求后，依据本体中定义的关系和公理，针对 RDF 三元组，采用逻辑推理技术进行推理。鉴于自动逻辑推理难度高，目前的语义检索应用大多是针对较简单的应用场景进行推理的，例如大学校园信息查询[6]、基于语义的学术文献共享[8]等。语义网研究团体曾开发了一些软件包，如 Pellet，Jena 等。利用这些软件，可以降低语义检索系统实现逻辑推理的难度。

2.1.2 技术挑战

本体实体抽取的准确率决定逻辑推理的有效性，从而影响检索结果的质量。纯结构化的中、英文网页在因特网所占比率极低，因此语义网研究的一个重要任务就是将无结构的网页转变为结构化或半结构化信息。英文网页语义检索研究得最深入，然而即使是从英文网页中，自动抽取数据对象或本体实体，其准确率也难令人满意。在对象级的垂直检索系统 Libra 和 Windows Live Update 开发进程中，从英文网页中自动抽取 Web 对象及其属性，错误率较高[9]。

中文网页语义标注过程中，遗漏本体实体，可能导致语义检索系统召回率降低；增加错误的本体实体，可能导致语义检索系统准确率降低，以及逻辑推理链错误地中断。例如，在科技文献共享系统中，若不能准确抽取学者姓名，就无法将该学者关联到其论文发表信息，从而无法统计其论文被引用次数，最终不能客观地评价该学者的学术影响力。

文献[9]提出采用机器学习技术，以提高本体实体抽取正确率。文献[5]提出对中文文本进行必要的浅层句法分析，构造句子的语法依存树，然后根据依存树生成 RDF 三元组。英文网页的本体实体抽取技术，大多已应用到中文网页的语义标注[10]中。然而，中文网页语义标注还面临因汉语自动分词处理所带来的特殊挑战。

影响中文自动分词准确率的主要原因是未登录词和歧义，其中未登录词导致的分词错误率是歧义导致分词错误率的 5 倍以上[11]。不幸的是，语义检索系统需抽取的本体实体

大多是未登录词，例如人名、地名、机构名、商品名等。因此，中文本体实体抽取须重点考虑未登录词对抽取准确率的负面作用。

中文自动分词算法分为基于规则和基于统计的两类。基于规则的自动分词算法的分词精度，依赖电子词典的规模和质量。电子词典包含的词汇多，则分词正确率较高。但电子词典不可能包含网页中出现的全体人名、地名、商品名、机构名等专属名词，因此基于规则的自动分词算法不大适用于中文网页本体实体抽取。基于统计的分词算法可以自动标识出新词或未登录词，但对出现频率极低的专属名词，还是难以正确标识。例如，在基于语义的科技文献检索应用中，假设有一名叫"王胜利"的硕士研究生。他在读书期间发表了一篇论文，但该文被引用次数极少；毕业后他当了公务员，不再发表学术论文。从而，"王胜利"一词在待检索网页中仅出现一次或几次，但"王"和"胜利"在中文文本中出现频率却比较高。针对这种情况，基于统计的分词算法也可能出现本体概念"作者"实体抽取错误。

2.1.3 特点分析

利用本体实现基于语义的中文网页检索具有以下优点。

第一，应用逻辑推理技术，可以挖掘不同网页文本之间隐藏的数据关联关系。例如仅从一篇文献的发表信息，很难全面地判断这篇论文的学术影响(发表在高质量期刊上的论文，其被引用次数并不一定高)。但在基于语义的科技文献检索系统 SemreX 中，通过计算学术论文的被引用次数、所发表期刊的声誉等信息，可量化该论文影响力的大小[7]。

第二，语义网、知识工程和人工智能等领域，提出了许多理论，并提供了一些开源软件。只要找到一个较好的应用领域，定义好完备的本体，就可以较快地开展中文网页语义检索实验。信息检索是高度依赖实验的学科，能快速构建出检索系统，其对今后的理论研究，无疑有极大帮助。

该方案存在以下不足。

第一，由于本体定义局限，该方案难以提供覆盖面广泛的语义检索服务。检索系统覆盖面越广，本体的定义、维护就越困难。特别是多个领域对同一词汇的语义解释差异很大。例如，"网格"(Grid)一词，在分布式系统、电力系统、计算数学这3个学科中，其含义完全不同。

第二，当概念和属性的实体数量巨大时，逻辑推理难控制。用户期望即时获得检索结果，而逻辑推理很耗时。哪怕语义检索系统能提供很优秀的检索结果，但一旦检索时间过长，很多用户依然不能接受。此外，当本体中的公理数量巨大时，依据不同公理进行推理还可能产生相互矛盾的推理结果。特别是对错误的本体实体进行逻辑推理，会浪费大量的计算时间，却得到无价值的检索结果。

2.2 基于自然语言理解的语义检索

很长一段时间内，有研究者认为自然语言处理技术对信息检索的帮助很小，甚至会产生负面影响[12]。尽管如此，还是有研究者坚持开发基于自然语言理解的搜索引擎，例如 Textdigger，Hakia 等。2008 年 5 月发布的语义搜索引擎 PowerSet 采用帕洛阿尔托研究中心(PARC)研发的自然语言处理技术，以"读懂网络中的每个句子的含义"为理念，提供基于语义的网页检索服务[13]。以自然语言处理技术为核心的 PowerSet 被工业界认为是最有可能挑战 Google 的新型网页检索技术，2008 年 7 月它被微软公司收购，并将其技术融合到微软公司的网络搜索引擎中。PowerSet 的推出和被收购表明，基于自然语言理解的语义检索是一种不可忽视的语义检索途径。

2.2.1 实现方法

自然语言处理研究积累了大量的语言学知识资源和词法、句法分析软件。例如，描述英文词汇相互关系的有 WordNet，关于中文词汇之间关系的有 HowNet，关于中文自动分词系统的有 ICTCLAS 等。基于自然语言理解的信息检索，则利用这些已有的语言学知识库以及算法。

为了提高信息检索系统的召回率，20 世纪 90 年代初就有学者提出利用 WordNet 等语言学词典进行查询扩展，其操作步骤如图 2 所示。此后一直有研究者坚持利用语言学词典来改进查询扩展[14]。例如文献[15]把 WordNet 和常识知识库 ConceptNet 结合起来进行查询扩展，取得了较好的效果。

步骤 1 获取语言学知识库；

步骤 2 在语言学知识库中查询检索词，获取检索词的近义词、同义词等信息；

步骤 3 查询扩展；

步骤 4 利用新的查询，进行检索；

步骤 5 把检索结果反馈给用户。

图 2 基于语言学词典的语义检索

词典和本体在查询扩展研究中，都是使用频繁的知识资源，部分研究者甚至以电子词典为基础来构建本体。值得指出的是：尽管本体和电子词典都能描述近义词、同义词、反义词等信息，但二者还是有区别。语言学词典一般不描述上位词、下位词等信息，本体却常描述这些信息；词典解释义项，而本体不解释。基于电子词典的查询扩展，没有对被检索网页进行深入的语言学分析，因此还算不上真正意义上基于自然语言理解的语义检索。

深层次的语义检索，既需要对查询请求进行语言学分析，也需要对待检索网页文本进行语言学分析处理。词法分析、词性标注、词义标注的正确率已比较高，学术界对基于词义匹配的语义检索已有较多研究成果[16-18]，图 3 描述了基于词义匹配的语义检索的实现步骤。

步骤 1 对待检索文档实现自动分词、词性标注、词义标注；

步骤 2 对全体网页进行语义索引；

步骤 3 对查询请求作分词、词性标注、词义标注；

步骤 4 查询检索文件，实现词义匹配操作；

步骤 5 把检索结果反馈给用户。

图 3 基于词义匹配的语义检索

因词义标注中歧义消除比较困难，目前图 3 步骤 1 中的词义标注有不同的实现方式。文献[16]采用细粒度词义标注，标注出多义词在句子上下文中的义项。文献[17]根据 WordNet 词典、多义词的上下文信息，将名词词性的多义词标注为每个义项所对应的 WordNet 词典中的根词汇(WordNet 一共有 25 个根词汇)。这意味着，若某个名词具有多个义项，且每个义项所对应的根词汇是一致的，则上述方法无法区别该名词的多义性。粗粒度的词义标注可实践性较强，也能提高信息检索系统的性能。目前基于词义匹配的英文网页

语义检索较多地采用粗粒度的词义标注。

在中文信息处理中，词法分析、词性标注、词义标注等已取得实用性的成果，句法分析已有可用的成果，句义分析尚有较大的困难。因此，在现阶段最可行的基于自然语言理解的中文网页语义检索，就是基于词义匹配的检索。

用户输入的查询词中，不同词类所占比率差异很大。通常名词、动词等实词所占比率最高，而名词、动词中多义词比率很高。例如“编辑”具有名词、动词两种词性，有两个义项。而一些单字词的义项则更为丰富，例如“打”字在现代汉语词典(第五版)中，定义了 24 种义项。因此，词义排歧是基于词义匹配的中文网页检索中的核心问题之一。基于信息检索应用的排歧研究，与传统的词义标注中的排歧研究有区别。文献[19]指出，提高词义标注精确率后，未必能显著提高信息检索性能；相反，一些错误的词义标注结果，也没有显著降低检索效果。因此，排歧精度与检索性能之间的关系，还需要经过实验进一步研究。文献[20]通过统计搜狗的搜索日志文件发现，85%以上的中文查询请求不超过 3 个词。用户输入的检索词太少，缺乏必要的上下文信息，造成图 3 中步骤 3 难以正确地标注词汇义项。

2.2.2 技术挑战

消歧操作分布在自动分词、词性标注、词义标注、语法分析、语义分析等各阶段，高效率的消歧处理是自然语言处理研究中的长期挑战。通常消歧算法准确率越高，其计算复杂度就越高。当网页文本数量较少时，可采用复杂度较高的算法。但对信息检索系统而言，复杂度高的排歧算法并不合适。如何在排歧效率与复杂性之间选择合理的平衡点很关键。

词汇义项越多，排歧就越复杂。《现代汉语词典》(第五版)收录的词语中，多义词所占比率为 23%，其中两个义项的词汇为 18.4%，含 3 个或 3 个以上义项的多义词为 4.6%[21]。尽管多义词在中文词汇中所占比率不高，但它们在网页中的出现频率却很高，这为中文网页词义标注带来了诸多障碍。

信息检索领域比较关注中文句子中指代实体的自动识别问题[22]。自然语言中，相邻的句子或同一句子内部，常用代词指称某一特定实体。例如，“我看到小王来了，他手里拿着一本书。”中的“他”，显然是指“小王”。但“我看到小王和小李来了，他手里拿着一本书。”中的“他”到底是指“小王”，还是“小李”，或是其他人，这就不易确定。而“我看到小王和小李来了，他手里拿着一本书，小李后来把书给我看。”中的“他”，则可能是指的“小李”。从这 3 个例句可知，分析指代实体须要对上下文进行语义分析。

汉语中频繁使用的缩略词，其指代意义也需要根据上下文，甚至更多的信息来判定其指代实体。例如，国内大学的校名常有缩写形式。武汉地区把“华中师范大学”简称“华师”；上海把“华东师范大学”简称“华师”；广州把“华南师范大学”简称“华师”。当用户输入“华中师范大学”一词进行检索时，语义检索系统不容易判断是否该把包含“华师”一词的中文网页作为检索结果输出。

2.2.3 特点分析

基于自然语言理解的中文网页语义检索，其最大的特点就是接近人对网页的阅读、理解方式，因此最有可能取得与人工检索一致的检索结果。另外，自然语言词义变化比较缓慢，词汇之间的关系具有长期的稳定性。尽管不同应用领域的文本描述特征有差异，但对同一种自然语言，其语法规则是一致的。因此，只要对某一个应用领域实现了基于自然语言理解的信息检索，该语义检索系统所采用的技术就极有可能推广到其它领域。

该方案存在以下不足。

第一，文本语义分析复杂度高，系统对待检索网页的数量高度敏感。当被检索文档数量巨大时，几乎不可能进行深层次的语义分析。PowerSet 仅对维基百科全书中的网页进行检索，就有 200 多万个网页。而针对某些特定应用领域，例如国内中文新闻网页检索，其待检索网页数量则远不止 200 多万个。因此，目前并不是所有的应用领域都适合开展基于自然语言理解的中文网页语义检索。

第二，即使采用了先进的自然语言处理技术，有时也并不能显著提高信息检索性能。自然语言处理技术的应用程度与信息检索效率之间的关系，如今并不明确。对此，3.2 节将进行更详细的讨论。

2.3 基于网页文本统计分析的语义检索

通过对网页文本和查询请求的分类操作，可以确定待检索网页和用户提交的查询请求各自属于的语义子空间。通过对网页文本和查询请求的聚类操作，可以发现与查询请求语义近似度高，可能隶属于同一语义子空间的网页集合。信息检索系统在执行分类或聚类操作后，可将与查询请求隶属于同一语义子空间，或者是语义近似度在规定阈值范围内的中文网页作为检索结果输出。信息检索领域对中文网页分类、聚类操作进行了长期的研究[23-27]，新的分类、聚类特征不断被发现，改进的文本分类、聚类算法不断被提出。

2.3.1 实现方法

根据对分类和聚类算法的侧重程度、分类聚类操作执行对象的差异，基于网页文本统计分析的语义检索可细分为以分类待检索网页为主的语义检索，以及以聚类检索结果为主的语义检索。

图 4 描述了基于网页文本分类实现中文网页语义检索的实现步骤。在基于文本分类的中文网页语义检索中，既需要进行语义分类，也经常需要进行语义聚类操作。分类目录、分类算法和聚类算法，是实现图 4 所示途径的核心。

步骤 1 确定合理的语义类别；

步骤 2 建立训练样本集合，利用机器学习技术训练语义分类器；

步骤 3 将待检索网页进行分类，确定各网页隶属的语义类别；

步骤 4 获取查询请求，并对请求进行语义分类；

步骤 5 根据查询请求的语义类别，将同一类别的中文网页反馈给用户。

图 4 基于网页文本分类的语义检索

在对网页进行语义分类之前，首先要有明确的分类目录。目前有两种确定语义分类目录的方法。

第一，利用现有的、并已获得学术界和工业界广泛认同的分类目录，例如中图分类法、雅虎目录[28]等。雅虎目录是一种树状的分类，由专家人工定义，科学性较强。它包括艺术、商业与经济、计算机和因特网、教育等 14 大类，每个大类又包含若干个不同层次的子类。雅虎还根据不同语言、文化的差异，发布不同语种的雅虎目录，例如中文雅虎目录。为了把

Google广告内容合理地嵌入到与其语义相近的网页中，文献[29]将广告和待检索网页分类到雅虎目录中各子类，然后再计算同一分类中的广告和网页的语义相似度。

分类目录覆盖面广，符合人类的日常思维习惯，缺点是不一定适合待检索网页集合。例如，来自大学网站的网页大多关注学术、高等教育、科学研究等方面的信息，很少涉及娱乐、军事、体育等领域。利用雅虎目录对此进行分类，则多数类别的网页数量几乎为零，而教育类、学术类信息则分类过粗，难以满足用户的查询需求。

第二，先对待检索网页集合进行聚类，发现该网页集所包含的语义类别数量以及不同类别之间的差异。然后，把通过聚类获取的分类信息作为图4中步骤1的分类知识。由于聚类算法计算复杂度较高，信息检索系统中待检索网页数量巨大，因此对全体待检索网页进行聚类操作难度极大。但可以依据合理的采样原则，从待检索网页中抽取出较少的网页样本进行语义聚类[26]。根据聚类的结果，再确定适合该检索系统的分类目录。有了分类目录后，再训练语义分类器，然后对待检索网页以及信息检索系统运行过程中新增加的网页进行分类。

中文网页分类研究主要采用支持向量机(Support Vector Machine)。文献[24]先利用无监督聚类，对训练集中的正例和反例进行聚类，然后挑选例子训练支持向量机，从而获得分类器。文献[25]采用有指导的机器学习，训练中文网页分类器，并利用该分类器在“天网”搜索引擎上实现大规模中文网页的目录导航服务。文献[26]选择数据集的核心子集参与分类器训练，在特征选择阶段采用改进的基于词性的互信息特征选择模型，从而提高网页分类规模数据处理能力。

实施网页分类的难度，与分类目录所包含的子类别数量、待分类网页数量紧密相关。目前，网页分类算法在类别较少(如几个、几十个)时，可取得较好的效果。当子类别数量高达几百、几千时，其不仅分类器训练难度大，而且分类效果较差。此外，现有的分类、聚类算法，难以处理数百万计的网页文档[30]。有研究者提出不对全体待检索网页执行静态分类，而是针对每次检索所获结果进行语义聚类[31,32]。该方案的执行步骤如图5所示。

步骤1　利用现有的检索系统检索，获得检索结果；

步骤2　对检索结果进行聚类；

步骤3　根据聚类结果，重新进行相关性排序；

步骤4　将聚类排序后的结果，反馈给用户。

图5　对检索结果进行聚类的语义检索

图4与图5的根本差异是：第一，图4是对全体待检索网页进行分类，图5则只对每一次检索的结果进行聚类。显然，其查询所获结果包含的网页数量远小于全体待检索文档数量。第二，图4的分类操作在后台完成，无实时性要求，图5的聚类操作却需实时完成。因需要聚类的网页数量少，图5可以选择计算复杂性较高的聚类算法。网页聚类特征选择如今呈现多样化趋势，网页文本、超链接、网页的可视化结构等，都已得到高度重视[33]。

2.3.2　技术挑战

目前，中、英文网页分类、聚类操作，大多基于“词袋”(Bag of Words)模型。它使用词汇向量表示网页文本中不同词汇出现的频次特征，不考虑词汇出现的前后次序，更不能描述句子的句法结构。例如，“李明借给我一本书”和“我借给李明一本书”具有不同的语义，但其词袋模型表示结果则完全一致。用同一语义分类器来分类以上两个句子，其语义类别判定为一致。由此可见，词袋模型不能区分因句子内部结构差异所导致的语义差异。词袋模型缺乏对自然语言文本的深层次分析，这决定了网页文本分类、聚类、主题模型等不具备细粒度的语义区分能力。

两个不同的词在同一文本中出现的概率有差异。有些词汇共同出现的频率较大，例如“郭靖”和“黄蓉”；有些词汇之间共同出现的频率极低，例如“郭靖”和“坦克”(因为在《射雕英雄传》中不会出现坦克)；有些词汇的出现可能相互独立，例如“苹果”和“火山”。无论是中文，还是英文，词汇的数量都数以百万或千万计，统计全部词汇之间的共现概率在工程中几乎不可能完成。目前网页语义分类、聚类操作，大多假设任意两个词汇的出现是相互独立的。虽然这个假设与现实不符，但为了可实践性，也只能如此。

大型的英文信息检索系统中往往包含了数百万个索引词汇[30]。随着新网页的增加，新词汇(如科技新名词、商业品牌、机构名等)还在不断涌现。增加一个索引词，词汇向量的维度就增加一维。尽管信息检索系统包含的索引词极多，但一个网页中出现的不同词汇数却较少。因此，词-文档矩阵是巨大的稀疏矩阵，其计算、存储都不方便。语义分类、聚类操作中，不得不对词-文档矩阵作降维处理。

隐含语义标引LSI(Latent Semantic Index)和概率隐含语义分析(Probabilistic Latent Semantic Analysis)在词-文档矩阵降维中应用广泛。LSI基于一个简单的假设：若某些词高频率地出现在同一文本中，则这些词具有语义近似性。例如，大连是众所周知的足球之城。“大连”和“足球”在同一网页中出现的频率很高。经过LSI算法处理之后，它会认为“大连”和“足球”之间存在语义关联。然而大连是城市名，足球是球类名，从语言学上看两个词汇既不是同义词也不是近义词。由此可知，LSI中的语义近似与语言学中近义词是完全不同的概念。LSI获得的语义近似，既包含词汇之间的词义近似性，也包含非同义词之间的强相关性。

2003年以来，LDA(Latent Dirichlet Allocation)模型、Labeled-LDA等改进的LDA模型在中文文本分类、聚类研究中应用频繁[34,35]。LDA模型包括词、主题和文档3层结构，是一个多层的产生式全概率生成模型。研究者认为LDA比LSI更适合网页文本分类，其理由是：LDA模型是全概率生成模型，具有清晰的内在结构，适合利用高效的概率推理算法进行计算；而LDA模型参数空间的规模与训练文档数量无关，因此更适合处理大规模网页文本分类[35]。

合理利用HTML标记中的文本内容，可以提高网页分类质量和速度。例如，位于标记〈title〉和〈/title〉之间的词串通常是网页制作者对该网页内容的人工标注。人工标注的主题分类，通常更准确。类似的HTML标记，还有〈b〉、〈/b〉、〈font size〉等。网页分类研究关注如何给不同标记中的词汇设置合理的权值[25,27]。

2.3.3　特点分析

利用网页文本统计分析实现中文网页语义检索，具有以

下优点。

第一，有利于找到那些利用不同词汇，表达同一语义信息的检索结果。信息检索召回率较低的一个不可忽视的原因就是，自然语言中，对同一种意思具有多种不同的词汇表达方式。利用LSI、LDA等，能够自动发现不同词汇表达之间的强相关性。

第二，检索结果与查询请求之间的语义相关度计算很便捷。在进行分类、聚类操作前，系统已经明确给出了待检索文本的语义计算标准；在分类、聚类过程中，定量计算了每个待检索文本所对应的语义向量。在查询过程中，只要计算出查询请求所对应的语义向量，就可以快速地找到语义近似的待检索文本。

该方案存在以下不足。

第一，分层分类(Hierarchical classification)和单层分类(Flat classification)难抉择。在分层分类中，一个网页通常属于一个语义子类别；在单层分类中，一个网页可以属于多个语义类别。分层分类执行效率高，但并不符合网页的语义分布状况；单层分类符合中文网页的语义分布状况，但其计算复杂度高，工程可实践性较低。是采用分层分类还是单层分类，不易确定。

第二，用户输入的中文查询词数量少，难以准确地确定查询请求的语义类别。国内用户在一次查询请求中，输入的检索词平均数为1.85个[20]。汉语常用词大多是多义词，输入词数少，用户又不希望进行更多的交互以确定查询范围，因此很难猜测用户的查询意图。

3 中文网页语义检索中的3个关键问题

以下3个问题在中文信息检索研究中，曾有过长期争论。在基于语义的中文网页检索研究中，它们依然是不可回避的问题。但由于检索目标的变化，对以上问题的考虑也与过去有所差异。

3.1 索引的基本单元

建立倒排序索引，是提高查询响应速度的关键。索引单元决定倒排序索引的创建、更新、管理机制，因此信息检索研究一直都很重视选择合理的索引单元。在英文中，词是表示语义的基本单位。布尔模型、向量空间模型、概率模型和语言模型在英文信息检索中应用时，绝大多数都是以单词或词干为基本单位进行检索的。在英文网页语义检索研究中，词是进行词义分析、本体实体抽取、词义匹配的基本单位。

学术界对中文信息检索是以单个汉字，还是以词或2字串等为索引单元，进行了长期的研究，研究者所持观点分歧很大[36]。文献[37]认为中文信息检索应以词为基本单位，并对自我指导的自动分词算法作了深入的研究。文献[38]则认为经过人工精确分词处理的中文检索系统，其效率不一定比基于单个汉字、2字串等的检索系统效率高。

尽管目前的中文信息检索系统大多不以词为索引单元，但这并不能否定基于词汇的索引在中文网页语义检索中的必要性。其理由是：第一，基于单个汉字、2字串的索引容易降低检索系统的查准率。“人为”一词的检索结果，清楚地反映了这一现象。第二，现代汉语词汇的平均长度为2.3个汉字[21]，有时对单个汉字不能进行有效的语义分析，如“葡”字。第三，多数2字串、3字串没有意义，例如“萄沟”、“萄沟美”等。因此，中文网页语义检索不能以单个汉字、2字串、3字串等为基本单位，而应以词为基本单位。

以词为基本单位的中文网页语义检索面临以下问题。

第一，汉语中词和短语的界限并不分明，导致难以建立公信力强的中文电子词典。现代汉语研究是根据频率高低来区分词汇和短语。例如“鸡蛋”出现频率高，认定其为词；“鹅蛋”出现频率较低，则认定其为偏正短语。对特定汉字串的出现频率，通过语料统计不难获得，然而频率高低之间的界限却并不明确。因此，即使获得了精确的频率信息，也难以确定它到底是词还是短语。

第二，汉语中词的切分粒度不明确。汉语中专属名词很多，例如人名、机构名、地址等，它们大多还可以细分为多个词，例如“华中师范大学”是专属名词，该词还可细分为“华中”、“师范”、“大学”3个更短的词。信息检索中，分词粒度与传统的中文自动分词算法研究存在差异。在自动分词软件中，“华中师范大学”作为一个词是合理的；而在信息检索中，分成“华中师范大学”则可能导致信息检索系统召回率偏低。文献[38]指出，在基于关键词匹配的信息检索中，将较长的词分为多个更短的词，有利于提高信息检索性能。但在基于语义的检索中，如果将机构名分得更小，则面临查准率降低的问题。中文网页语义检索中，自动分词粒度该如何选择，如今还没有明确的判断标准。

第三，汉字组词能力强，新词层出不穷，导致索引困难。词汇数量太大，对信息检索系统开发是障碍。英文信息检索系统中，其词汇数量为数百万个[30]。“信息交换用汉字编码字符集基本集”包括的高频汉字有6763个[39]，仅由这些汉字组成的不同2字串就可达45738169个。另外，汉语中还有3字词、4字词等，如“黑龙江”、“信口开河”等。过多的词汇，导致索引、词-文档矩阵降维困难。

3.2 采用中文信息处理技术的程度

不对网页文本进行自然语言分析，可以显著降低信息检索系统开发的难度以及检索系统所需的硬件设备条件。早期基于字符串匹配的中文网页检索系统，大多避免使用中文信息处理技术。但从“人为”一词检索结果的分析可知，对中文网页文本进行适度的语言学分析，有助于提高查询的精确率以及用户的满意程度。

中文信息处理技术在中文网页语义检索中是否需要？如果需要，应该对待检索文本进行何种程度的语言分析处理？既然词是中文网页语义处理的基本单元，则中文词典、中文自动分词等中文信息处理资源和技术显然是必需的。因此，对第一个问题，答案已很清晰。对第二个问题，还需继续研究。

文献[40]在保持其它条件不变的前提下，不断调整中文分词结果的精度，发现信息检索效率与中文自动分词精度之间并不是简单的单调递增关系。当分词准确率低于75%时，信息检索性能随分词精度的提高而提高；此后检索性能随着分词精度的提高反而降低。一般认为，对网页分词越精确，越有利于提高检索的精确率。然而高精度分词在提高精确率的同时，却会降低其召回率，最终降低系统检索效率的总体评价。文献[40]的实验结果表明，高精度的中文信息处理结果未必能提高信息检索的效率，有时甚至还带来负面的影响。

文献[23]研究中文网页分类，它先用机器翻译软件将中文文本翻译为英文，然后利用英文文本分类技术进行分类。汉译英机器翻译技术错误率较高，尚未进入到实用化阶段。然而即使是这样的处理，也得到了较好的中文网页分类效果。虽然文本在机器翻译过程中，其错误率较高，但由于英文文本分类是基于“词袋”模型，“词袋”模型本来就忽略了句法结构等对文本语义的影响，因此，只要机器翻译软件能把中文网页中的关键词汇翻译正确，利用“词袋”模型依然可以取得较好的网页分类结果。文献[23]的实验结果表明，仅做简单的中文信息处理，也能显著地改进信息检索的效果。

语义信息检索中，对文本的语义分析应该做到何种程度，目前尚无共识。基于本体、词义匹配、网页文本统计分析的语义信息检索，主要是在词的层面进行处理，不涉及对语法、句子意义的分析。文献[41]在句意理解层面对文本进行了分析处理。到底哪种分析层次最适合中文网页语义检索，还需要进行更深入的研究。目前可以肯定的是，中文信息处理技术将与中文网页语义检索技术结合得越来越紧密。

3.3 选择合适可行的应用场景

中文网页语义检索拥有巨大的应用前景，但目前它还处于萌芽时期。有巨大应用潜力的技术，要想取得成功，必须尽快开发出具有示范性的应用系统。另外，新型的中文网页语义检索技术也需要在大规模应用中进行测试和验证。根据英文网页语义检索、中文信息处理的发展历程，可以预测基于语义的中文网页检索短期内难以在关键技术上取得重大突破。因此，中文网页语义检索实验有必要注意以下问题。

第一，选择一个能引起广大在线用户关注的应用场景，以提高用户数量。目前的语义网检索系统开发，大多针对科技文献搜索、专利搜索等与普通在线用户关联度较低的行业。因此，系统的用户数少，社会影响小。广大民众关注的焦点是饮食、住房、交通、体育、娱乐、财经、教育等方面的信息。如果能让用户在语义搜索中，亲身体会到通用搜索引擎不能提供的优质检索服务，这才能引起用户和工业界对中文网页语义搜索的真正关注。

第二，待检索网页数量不宜过大，但应尽可能完备。语义标注、自然语言理解、文本统计分析等所需时间和空间开销均很大。国内多数研究团体拥有的硬件设备有限，过高的设备要求将不利于广泛开展语义检索研究。为了显示语义检索的有效性，待检索网页集应尽可能覆盖某一个行业领域。例如，高考、研究生入学考试是能引起众多考生、考生家长关注的主题。在填报志愿过程中，考生和家长需要高校的教学、科研等信息。国内只有1500多所全日制大学、学院，且各大学官方网站的网页文本信息总容量不算太大。对其中文网页进行语义检索，则网页数量适中，且覆盖面较广。

4 未来的研究方向

基于语义的中文网页检索有必要从以下几个方向进行更深入的研究。

4.1 中文网页的语义相关性评价

来自不同领域的学者，对语义的理解存在差异；即使是来自同一领域的专家，对检索结果的语义近似性排序也不尽一致。如何定义检索结果与查询之间的语义相关性，一直都是很具挑战性的工作。

本文第2节所列的3类中文网页语义检索途径，对检索结果相关性的评价模式并不一致。基于本体的中文网页语义检索，其结果是根据逻辑推理所得的，因此其结果与传统的向量空间模型、概率模型、语言模型所计算出的语义近似性难比较。基于“词袋”模型计算出的语义近似性，完全没有考虑不同句法结构差异所导致的语义近似度差异。基于自然语言理解的语义检索，目前很少有文献明确地说明其语义相关性的定量计算方法。

3类语义检索途径，如今都能解决一部分问题，但又都存在不足，它们之间互补性强。未来的中文网页语义检索系统极大可能采用两种或多种语义检索途径，以提高语义检索的质量。对于混合型的语义检索系统其语义相关性的定义不再是单一的评价标准。根据多种语义相关性评价标准，建立组合型的语义相关性评价模型是可能的解决方案。

4.2 本体构建与本体实体抽取

基于本体的中文网页语义搜索是目前最常见的实现途径。定义本体是表达知识和开展逻辑推理的前提，未来中文网页语义检索必定需要开发更多的本体。目前本体自动生成技术还不成熟，本体开发主要是手工或半自动方式。未来有必要不断减少领域专家参与，以提高本体自动完善的能力。

从中文网页中高精度地抽取本体实体有待继续研究。中文本体实体抽取的重要前提是能自动识别专属名词。基于模板的信息抽取、基于统计的信息抽取、基于语法依存树的自动抽取算法都已应用到中文信息检索中。由于待检索网页太多，因此信息检索很难采用计算复杂度过高的本体实体抽取算法。如何在本体实体抽取精度与计算复杂性之间取得合理的平衡，还需通过大规模实验来进行探究。

在基于本体技术的语义检索中，中文文本自动分词、本体实体抽取、逻辑推理是多个分离的操作。前一阶段的错误会降低后继步骤的效率。当本体实体抽取错误率积累超过一定阈值时，后继操作可能会失去意义。因此，有必要提出一种本体实体正确率检测机制。当本体实体错误率超过预定阈值时，能自动报警，防止针对大量错误的本体实体进行逻辑推理，消耗了大量计算资源，却得到无意义的结果。

4.3 语义索引

索引是影响信息检索系统查询响应时间的关键。对本体实体进行索引，与传统的倒排序索引有很大差异。在近10年的语义网研究中，对本体实体的管理有的是通过RDF数据库，如Sesame来完成[8]；有的是通过关系数据库管理系统，如SQLServer来完成[6]。以Sesame为例，我们观察到当RDF三元组数量超过200万条以上时，语义数据库管理系统效率较低，很难满足用户的实时访问需求。关系数据库虽然可以管理数百万的RDF三元组，但当RDF三元组数量更大时，访问速度依然缓慢。针对特定领域的中文网页语义检索，检索系统开发者难以预计和控制RDF三元组的数量。未来有必要对大规模RDF三元组的高效管理和索引进行深入研究。

采用基于词义匹配的中文网页语义检索，其倒排序索引的构建也比较困难。基于词义的倒排序索引表有两种可能的实现途径：第一，以汉语词汇为索引词，在索引列表(Posting list)中加入词汇在文档上下文中的义项信息；第二，以汉语词

汇义项为索引单位，来建立倒排序索引，在索引列表中只存储网页的文档编号。哪种方法在工程中更可行、高效，还需通过大规模实验进行比较。

4.4 经语义标注的大规模中文网页样本集建设

信息检索实验离不开大规模的网页测试集。机器学习技术在信息检索中应用频繁，如今在本体实体自动抽取、网页文本分类器训练，以及词义标注等各个领域，几乎都会使用机器学习技术。广泛地使用机器学习技术，就需要大规模的中文网页样本集。缺乏语义标注信息的中文网页样本，就不能直接将其应用于有指导的中文网页语义标注、网页分类训练，也难以实现语义聚类以及语义相关性评价。对大规模中文网页进行语义标注，需要大量的时间和人力，国内许多研究机构和研究者还不能独立完成。未来需要大规模、开放、经过了语义标注的中文网页样本集。

TREC 和 NTCIR 是目前国外最具影响的中文网页测试集，它们最初都只提供了繁体中文测试集，后来才逐渐加入简体中文网页测试集。NTCIR8 虽包括简体中文文本，却是针对跨语言检索研究，而并不是针对中文网页语义检索研究。TREC 和 NTCIR 的简体中文样本集主要来自新华社发表的新闻报道，从网页来源来看，二者都有待丰富。另外，TREC 和 NTCIR 都没有对中文网页进行语义标注。

北京大学网络实验室发布的 CWT 是国内使用较广泛的简体中文网页测试集。CWT 还是国内搜索引擎和网页挖掘会议测评比赛的待检索文档集[42]。CWT 网页来源比较广，且具有较强的时效性。未来，可以优先选择 CWT、NTCIR 等具有较大学术影响，且数据格式比较规范的中文网页样本集进行语义标注。上述组织长期发布中文网页测试集的经验，可以保证训练集、测试集的质量；同时还可以借助 CWT、NTCIR 的传统学术影响力，快速地将经过语义标注处理的训练样本集进行推广。

结束语 本文首先将基于语义的中文网页检索研究分成基于本体、基于自然语言理解和基于网页文本统计分析的 3 种途径；然后分析它们的实现方法、面临的技术挑战、以及优、缺点；最后指出，3 类方法都存在计算复杂度高、可扩展性较弱等不足。

语义检索的基本单元、自然语言分析的程度、应用场景的选择等是中文网页语义搜索研究必须面对的问题。文中认为基于语义的中文网页检索应以汉语词汇为基本单位，并尽量采用计算复杂度较低的中文信息处理技术，来提高本体实体抽取、词义标注的精度。中文网页语义检索系统开发，则应尽量选择用户数量众多的应用，以扩大中文网页语义检索对工业界的吸引力。最后指出中文网页语义检索应在基于语义的相关性评价、本体构建和本体实体自动抽取、语义索引方法以及大规模语义标注样本集建设方面展开更深入的研究。

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] Ren Fu-ji. Advanced Information Retrieval [J]. Electronic Notes in Theoretical Computer Science, 2009, 225: 303-317

[3] Li Zhi-han, Xu Yue, Geva S. A hybrid Chinese information retrieval model [C]// Proceedings of the 6th International Conference on Active Media Technology. Berlin: Springer-Verlag, 2010: 267-276

[4] Yang Ling-peng, Ji Dong-hong, Tang Li. Chinese Information Retrieval Based on Terms and Ontology [C]// Proceedings of NTCIR-4. Tokyo: National Institute of Informatics, 2004: 1-8

[5] 荆涛，左万利，孙吉贵，等. 中文网页语义标注：由句子到 RDF 表示[J]. 计算机研究与发展，2008，45(7)：1221-1231

[6] Wen Kun-mei, Lu Zheng-ding, Li Rui-xuan, et al. Design and mplementation of Semantic Search Engine Smartch [J]. Journal of Southeast University(English Edition), 2007, 23(3): 317-321

[7] Ning Xiao-min, Jin Hai, Wu Hao. RSS: A Framework Enabling Ranked Search on the Semantic Web [J]. Information Processing and Management, 2008, 44(2): 893-909

[8] Jin Hai, Yu Yi-jiao. SemreX: a Semantic Peer-to-Peer Scientific References Sharing System [C]// Proceedings of International Conference on Internet and Web Applications and Services 2006. Washington: IEEE Computer Society, 2006: 97-102

[9] Nie Zai-qing, Wen Ji-rong, Ma Wei-ying. Object-Level Vertical Search [C]// Proceedings of Biennial Conference on Innovative Data Systems Research 2007. Asilomar, 2007: 235-246

[10] 林鸿飞，杨志豪，赵晶. 中文文本的信息自动抽取和相似信息检索机制[J]. 小型微型计算机系统，2007，28(11)：2074-2079

[11] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007，21(3)：8-19

[12] Baeza-Yates R. Challenges in the Interaction of Information Retrieval and Natural Language Processing [C]// Proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistics 2004. Berlin: Springer-Verlag, 2004: 445-456

[13] PowerSet [OB/OL]. http://en. wikipedia. org/wiki/ Power_set

[14] Liu Shuang, Liu Fang, Yu C, et al. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases [C]// Proceedings of SIGIR 2004. New York: ACM Press, 2004: 266-272

[15] Hsu M-H, Tsai M, Chen H. Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach [C]// Proceedings of 4th Asia Information Retrieval Symposium. Berlin: Springer-Verlag, 2008: 213-224

[16] Christopher S, Oakes M P, John T. Word Sense Disambiguation in Information Retrieval Revisited [C]// Proceedings of SIGIR 2003. New York: ACM Press, 2004: 258-265

[17] Bum K S, Cheol S H, Chang R H. Information Retrieval using Word Senses: Root Sense Tagging Approach [C]// Proceedings of SIGIR 2004. New York: ACM Press, 2004: 258-265

[18] Pierpaolo B, Annalina C, Giovanni S. Integrating Sense Discrimination in a Semantic Information Retrieval System [C]// Proceedings of International Conference on Web Intelligence. Berlin: Springer-Verlag, 2010: 249-265

[19] Jacques G, Gilles F, Said R, et al. Analysis of Word Sense Disambiguation-Based Information Retrieval [C]// Proceedings of CLEF 2008. Berlin: Springer-Verlag, 2008: 146-154

[20] 余慧佳，刘奕群，张敏，等. 基于大规模日志分析的搜索引擎用户

行为分析[J]. 中文信息学报,2007,21(1):109-114
[21] 王惠. 词义·词长·词频—《现代汉语词典》(第5版)多义词计量分析[J]. 中国语文,2009,329:120-130
[22] Li Wen-jie, Qian Dong-lei, Lu Qin, et al. Detecting, Categorizing and Clustering Entity Mentions in Chinese Text [C]// Proceedings of SIGIR 2007. New York: ACM Press, 2007: 647-654
[23] Ling Xiao, Xue Gui-rong, Dai Wen-yuan, et al. Can Chinese Web Pages be Classified with English Data Source? [C]// Proceedings of WWW 2008. New York: ACM Press, 2008: 969-978
[24] 李晓黎,刘继敏,史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器[J]. 计算机学报,2001,24(1):62-68
[25] 冯是聪,单松巍,龚笔宏,等. "天网"目录导航服务研究[J]. 计算机研究与发展,2004,41(4):653-659
[26] 傅向华,刘国,陈冬剑. 一种核心子集选择训练的大规模中文网页分类方法[J]. 小型微型计算机系统,2011,32(8):1608-1612
[27] 段军峰,黄维通,陆玉昌. 中文网页分类研究与系统实现[J]. 计算机科学,2007,34(6):210-213
[28] Yahoo Directory[OB/OL]. http://dir.yahoo.com
[29] Andrei B, Marcus F, Vanja J, et al. A Semantic Approach to Contextual Advertising [C]// Proceedings of SIGIR 2007. New York: ACM Press, 2007: 559-566
[30] Bruce C W, Donald M, Trevor S. Search Engines: Information Retrieval in Practice [M]. Beiging: China Machine Press, 2009: 154, 291
[31] Zhang D, Dong Yi-sheng. Semantic, Hierarchical, Online Clustering of Web Search Results [C]// Proceedings of APWeb 2004. Berlin: Springer-Verlag, 2004: 69-78
[32] Claudio C, Stanisiaw O, Giovanni R, et al. A Survey of Web Clustering Engines [J]. ACM Computing Surveys, 2009, 41(3): 1-38
[33] Paul B. Visual structure-based web page clustering and retrieval [C]// Proceedings of WWW 2010. New York: ACM Press, 2010: 1067-1068
[34] 李文波,孙乐,张大鲲. 基于Labeled-LDA模型的文本分类新算法[J]. 计算机学报,2008,31(4):620-627
[35] 刘振鹿,王大玲,冯时,等. 一种基于LDA的潜在语义区划分及Web文档聚类算法[J]. 中文信息学报,2011,25(1):60-65,70
[36] Nie Jian-yun, Ren Fu-ji. Chinese Information Retrieval: Using Characters or Words? [J]. Information Processing & Management, 1999, 35(4): 443-462
[37] Peng Fu-chun, Huang Xiang-ji, Dale S, et al. Using Self-supervised Word Segmentation in Chinese Information Retrieval [C]// Proceedings of SIGIR 2002. New York: ACM Press, 2002: 349-350
[38] Schubert F, Li Hui. Chinese Word Segmentation and Its Effect on Information Retrieval [J]. Information Processing and Management, 2004, 40(1): 161-190
[39] GB2312-80. 信息交换用汉字编码字符集基本集[S]. 北京:国家标准总局,1981
[40] Peng Fu-chun, Huang Xiang-ji, Dale S, et al. Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR [C]// Proceedings of COLING 2002. Stroudsburg: Association for Computational Linguistics, 2002: 1-7
[41] Wang Ding-ding, Li Tao, Zhu Sheng-huo, et al. Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization [C]// Proceedings of SIGIR 2008. New York: ACM Press, 2008: 307-314
[42] 李静静,闫宏飞. 中文网页信息测试检索测试集的构建、分析及应用[J]. 中文信息学报,2008,22(1):30-36

(上接第61页)

结束语 本文通过对Web服务通信安全的分析,提出了解决Web服务安全问题的重要方法:Web服务架构优化与SOAP协议扩展相结合。Web服务架构通过设置业务网关实现身份验证和授权功能;加入安全性内容的SOAP协议实现机密性、完整性和不可否认性,从而全面、较好地实现了Web服务通信安全的5大需求。下一阶段的工作主要集中在综合使用其他的安全手段和措施进一步加强通信安全,如在Web服务器端通过对RequestSoapContext.Current值的判断,防止未通过Token验证的直接访问;在Web服务端禁止非SOAP协议的连接(例如HttpPost和HttpGet方式)请求等。这些都是下一步对于Web服务通信安全研究工作的主要方向。

参考文献

[1] Hardjono T, Weis B. The Multicast Group Security Architecture [Z]. Internet Draft, draft-ietf-msec-arch-05.txt, Internet Engineering Task Force, 2004-01
[2] 王晓峻,周晓峰,王志坚,等. 基于PKI/PMI的Web服务安全框架[J]. 计算机科学,2008,35(4):48
[3] W3C. WS-Policy(1.5) Framework[EB/OL]. http://www.w3.org/TR/2007/REC-ws-poliey-20070904,2007
[4] Diego Z G, Maria B F. Ontology-based Security Policies for Supporting the Management of Web Service Business Processes[C]// The IEEE International Conference on Semantic Computing. 2008
[5] 岳昆,王晓玲,周傲英. Web服务核心支撑技术:研究综述[J]. 软件学报,2004,15(3):429
[6] Boyens C, Günther O. Trust is not enough: Privacy and security in ASP and Web service environments[C]// Manolopoulos Y, et al, eds. Proc. of the 6th East European Conf. on Advances in Databases and Information Systems. Bratislava: Springer-Verlag, 2002: 8-22
[7] Thelin J, Murray P J. A public Web services security framework based on current and future usage scenarios[C]// Arabnia H, eds. Proc. of the Int'l Conf. on Internet Computing (IC2002). Las Vegas: CSREA Press, 2001: 825-833
[8] 杨怀洲,李增智. 基于Web Services的安全业务体系结构的设计[J]. 计算机工程,2005,31(20):146
[9] 王茜,吴黎明. 单点登录在Web服务安全中的应用[J]. 计算机工程,2008,36(8):179
[10] 贺正求,吴礼发,洪征,等. Web服务安全问题研究[J]. 计算机科学,2010,37(8):32
[11] 刘志都,贾橙涪,詹仕华. SOAP协议安全性的研究与应用[J]. 计算机工程,2008,34(5):142