第十章 语义搜索

1. 任务定义、目标和研究意义

随着 Internet 的爆炸性增长，万维网已经发展成为包含多种信息资源、站点遍布全球的巨大动态信息服务网络，为用户提供了一个极具价值的信息源。然而，传统搜索技术仍以关键词匹配、倒排索引和网页的链接结构为搜索依据，其查全率和查准率均无法满足用户日益提高的标准 [Arvind, et al., 2001] [Guo, et al., 2003] [Zhang, et al., 2007] 。与传统搜索技术不同，语义搜索是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句背后的真实意图，并依此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

语义搜索和传统的搜索系统有很大不同。以 Google 为例，Google 的传统搜索主要依据网站中是否存在关键词、有多少其他网站链向这个网站、用户的点击率等其他各种因素来决定呈现什么结果。Google 其实并不知道搜索词的含义。比如当你在 Google 中搜索“中国最大的城市”，Google 给你呈现的是一页包含这些关键词的链接，Google 并不知道这个问题的真正含义。相反，借助于语义网和知识图谱，语义搜索能直接给出这个问题的答案，而不是一页相关的链接。可以肯定的说，未来的搜索将会超越搜索词本身，进入由各种实体（entities）、实体的属性和实体的相互关系所组成的世界。语义搜索的目的即是借助于对实体的理解，它们之间的交互行为，用户对这些实体的理解获取准确的答案而不是一条条链接，通过利用语义技术，将推理结合到检索过程中，可以极大的提高当前的搜索效果，在语义 Web 环境下，可以更高效地发现信息资源。

2. 研究内容和关键科学问题

事实上，传统搜索技术提升的困难并不是因为目前的搜索技术本身不够先进，其根源存在于 Web 上现有的信息表达和组织方式过于简单。Web 上的网页使用的 HTML 语言或其改进版本，通过 DOM 树描述了网页的结构和格式，并引入包括图片、声音以及视频等多种媒体格式，使得信息的显示更加生动、形象。此外，相关文档之间可以采用超链接互相定向。然而，这种信息的表达和组织方法主要是为人们阅读服务的，对于计算机而言，缺失了 Web 页面所承载的语义信息。比如，某个 Web 页面中说明“小米 Note3 6GB 手机的价格是 2500 元人民币”。HTML 语言难以使得计算机理解：小米是一个移动设备的制造公司，Note3 是其生产的一款手机型号，6GB 是手机的内存容量，2650 的含义是手机的销售价格，其单位是人民币。

目前有不少针对自然语言理解的研究，希望通过自动的方式将自然语言的语义转换为计算机可以理解的表达方式，但是当前的研究进展无论是处理的效率还是结果的精确度都不能达到令人满意的程度。因此，现有的信息表达机制限制了计算机帮助人们处理、综合和分析信息的能力。为此，万维网发明人 Tim Berners-Lee 在 20 世纪九十年代末提出了语义网的构想。他指出，“语义网是现有万维网的扩展，在其中信息被赋予明确的、完善的语义，以使得计算机和人能够更好地进行协作” [Tim, et al., 2001] [Nigel, et al., 2006] 。为了实现在 Web 上表达语义的需求，包括万维网联盟(W3C) 和因特网工程技术组(IETF)在内的研究机构制定和开发了一系列技术规范。它们是在 Web 上进行语义表达和处理的技术基础，构成了一个层次化的技术框架。语义网是对万维网中信息表达方式的一次革新，它给出了一套技术框架使得 Web 上的信息可以方便地被计算机进行处理和理解。语义搜索是架构在语义网上的搜索引擎，将语义 Web 技术引入搜索引擎，为用户提供精准的检索结果。近两年来国外学者采用不同的方法和技术对该问题进行了深入的研究，并得出了不少有益的结论，也建立了相关的原型系统。但是一方面，由于语义 Web 处于发展阶段，另一方面也由于技术条件的限制，目前并不存在一个“通用”的解决方案，现有的语义搜索引擎系统也都处于起步研究阶段，离实用的商业化水平还相距甚远。

总的来说，语义网背景下的语义搜索主要面临的问题有以下三点：

（1）与传统的 Web 文档相比，语义网文档的本质是 RDF Graph。给定一个 RDF Graph，可以采取多种语法格式对其进行序列化，如，RDF／XML，Natation3 等。采用不同的语法进行序列化，生成的语义网文档之间可能具有显著的差别，然而它们表达的语义却是一致的。有时，即使采用相同的语法，也会导致不同的结果文档，比如采用不同的 name space 前缀。因此，对于语义网文档的搜索而言，如何针对 RDF 数据模型的特点进行文档分析、索引建立和查询匹配即变得极为重要。

（2）理解一个 URI 所指称的实体对于判断语义网上的实体共指问题非常重要。实体共指是指客观世界的同一个对象，在语义网上（通常是被不同的信息发布者）使用不同的 URI 来指称。这种共指现象给语义网数据的整合和建立在其上的搜索均带来了困难。自动的共指消解技术能够帮助人们快速地找到可能的共指 URI 列表。要更好地解决实体共指问题，当前还是以人工参与为主。因此，提供一种快速、高效的办法理解一个 URI 所指称的实体，将能够很好地帮助人们做出共指判断，进而帮助人们理解所获取的信息的真实含义。

（3）在现有缺乏必要的手段形成语义网的背景下，如何利用语义网技术改进传统的 Web 信息检索系统对用户来说极为重要。传统 Web 是基于自然语言的方式进行组织的，而语义网提供的一系列的技术规范，包括语义的明确表达和语

义网数据查询，能够以一个特定领域的搜索系统为切入点，利用语义网技术帮助获取传统 Web 上的信息。

3. 技术方法和研究现状

语义搜索的研究涉及到多个领域，包括搜索引擎、语义 Web、数据挖掘和知识推理等。运用的主要方法可归纳为：（1）图理论；在语义网的技术框架中，RDF(Resource Description Framework)是一个非常基础、且又非常重要的数据模型。通过 RDF 数据模型可将语义网中的本体组织为图结构，图中的弧和由结点和弧组成的路径中都包含着信息，因此在语义搜索中应用到了不同形式的图遍历方法，如实例扩展及查询的形式化方法等；（2）匹配算法，在语义搜索中需进行概念与关键字或者实例与关键字的匹配，关键字提供了一种快速定位信息的入口，而关键字和概念的匹配方法是语义搜索中重要的一环；（3）逻辑特别是描述逻辑、模糊逻辑等。逻辑和推理已经被整合到未来的语义 Web 框架中。描述逻辑是知识的一种形式化表示方法[Baader, et al., 2003]，作为本体语言的基础为人们所熟知[Horrocks, et al., 2003]，如 OIL，DAML+OIL，OWL。语义搜索的目的是为了准确地理解用户的输入，因此必须要使计算机具有逻辑推理能力，即如果输入为“小米Note3 是 Note2 的升级版吗？价格是多少?”计算机要确切理解“小米”、“Note2”、“Note3”代表的含义，并且理解“Note2”和“Note3”之间的关系。

3.1.主流语义网搜索引擎

在新一代的语义搜索引擎中较为典型的有两个，且都是基于本体的语义搜索引擎，分别为：Swoogle 和 TUCUXI。其中，Swoogle 从搜索返回结果的 Web 文档中提取出本体，然后依据本体间的语义关联性确定出文档间的语义关系；TUCUXI 则通过所获得的本体在 Web 上以特定规则爬行，并通过语义处理找出最符合要求的网页。目前已开发出许多建立于本体上的语义搜索引擎，如，Congnition、Hakia、DeepDyve、Factbites、Kngine 等。

Swoogle 是由马里兰大学计算机科学和电气工程系于美国国家科学基金会（NSF）和美国国防部下署高级研究计划署（DARPA）的资助下所建立的。与那些传统意义上的语义网搜索引擎不同，Swoogle 在资源获取方面拥有一系列突出的解决方案，可自动发现语义网中 RDF 格式的文档，通过 Link-Following 和 Meta-Search 的方式识别出语义网文档（SWDs），通过语义分析不断发现新的语义网文档，并可对其中元数据建立相关索引提供高效率的查询服务，利用 Rational Random Surfing 模型提供高质量的排序结果[Ding, et al., 2004] [Ding and Finin, 2006]。Swoogle 的核心功能有：

- 提取语义网中的实例数据；

- 支持对语义网的浏览，提供语义网中文档的元数据；
- 搜寻语义网中的术语，譬如通过属性与类定义的 URIs 等；
- 搜索提取语义网中的本体，并使用独有的算法提供高质量的排序结果；
- 可存储各种类型的语义网文档。

Swoogle 与通常的本体存储器或本体标注系统相比，其最大的与众不同之处在于能够鉴别出异源本体，此外还具有语义网文档自动发现化制。

Congnition：目前可提供三个 Demo，Congnition Q&A，Medline Semantic Search 及 Wikipedia Semantic Search，涉及法律、医学与消费者信息等深度内容，且是首个真正实现人机对话界面的语义搜索引擎。

Hakia：由 Xerox 公司推出的 Hakia 搜索引擎通过理解用户查询，并利用本体进行查询扩展，将各种基于主题的相关信息汇总。其利用的技术包括：词形变换、同义词扩展、概念具体化、自然语言理解等，可为用户提供语义搜索范围内解决方案，能够满足用户对于低成本、高效率的搜索需求。其搜索范围包括新闻、网页、博客、维基词条、Pubmed 等，返回结果的呈现方式有深度语义（Galleries、Pubmed、可信站点）、表面语义（新闻、博客、网页）、常规搜索（Twitter 与图像）加结果页面链接。

Factbites：可依据事实进行回答，与结果链接相比，其更专注于内容分析，并可使搜索结果更有意义，到目前也只有简单搜索方式。其搜索结果呈现方式是从网页中所抽取出的有意义的、完整的语句清单加 URL。

DeepDyve：是深网或者隐形网络搜索引擎，可提供深度网络学术资源租赁服务与全文预览服务。其搜索范围可包括来自 Nature、IEEE、Elsevier、Wiley-Blackwel、Springer 等一流出版社的有关健康科学、生命科学、人文社会科学、物理科学与工程学等领域的权威评审期刊与专利等等深度网络学术资源，并同时可搜索 Wikipedia，现正慢慢扩展至更多的领域。其搜索结果主要为 PDF 文档，而搜索结果呈现方式是结果过滤项（主题、类型（可租用、仅供预览、免费）、时间、作者、期刊）加结果页链接。

Kngine：其可对任何主题进行搜索，能够支持移动端搜索，其语种包括英语、德语、西班牙语、阿拉伯语。以选项卡形式展现搜索结果，在选项卡下方可选择显示与每项相关的术语和网页，其搜索方式包括语音搜索和简单搜索。

3.2 技术要点及研究现状

随着计算机的普及和万维网技术的发展，万维网已经发展成为人类历史上最大的信息系统，也成为人们获取信息的重要来源。传统的搜索引擎引入的“关键词匹配导致难以理解用户意图”和“缺乏有效方法分析数据间关系”的问题无法保证返回用户满意的结果。基于此，结合语义信息的搜索引擎-语义搜索越发被

学者和工业界重视起来。事实上，语义搜索是传统搜索的进化，传统的搜索技术对于结合检索与推理的语义搜索有许多可借鉴的经验。因此，可在传统搜索引擎技术的基础上对语义搜索进行更深入的研究，建立实用性更强的语义搜索系统，改善当前的搜索效果，以期在更广泛的语义 Web 环境中发挥更大的作用。国外学者在近几年已经运用不同的方法对语义搜索领域进行了深入的学习与研究，并成功设计及实现了多个系统原型。但是一方面受制于语义网仍处于起步阶段，另一方面也由于目前技术水平的限制，至今还不存在一个既精准又高效的通用解决方案[Anuar, et al., 2016]。事实上，语义搜索研究目前仍处于探索阶段，现有的有关语义搜索的研究点主要有：

3.2.1 引入推理和关联关系的语义搜索

在语义 Web 设计中，Web 中的资源用 URI 统一标示，并利用 RDF／OWL 标识资源的语义信息，由于数据间的语义明确便于计算机理解，基于此结构良好的数据的搜索克服了关键词查询的歧义性，同时在这些数据上还可以通过推理实现知识发现，推理出新的知识。随着语义 Web 研究的深入和应用的更新，Web 上的 RDF 资源对越来越多，基于推理的知识型语义搜索越来越得到关注，或许将成为未来语义搜索的主要方式。

Stanford 大学研制的 Triple 系统是一个基于逻辑程序设计的 RDF 查询系统，逻辑子句的问题求解能力使它能够解答较为复杂的问题,类似于“迈克尔杰克逊的《This Is It》专辑中有哪些歌曲？”这类推理型问题[Sintek and Decker, 2002]。马里兰大学设计的 HOWLIR 系统是基于 DAML 描述框架的语义 web 信息检索系统，它采用 DAML-JESSDB(一个基于 DAML 的推理系统)作为推理引擎。该系统自动产生并提取网页中的语义标签，同时也实现推理以产生更多关于网页的语义信息[Shah, et al., 2002]。搜索请求可以是针对语义信息的形式化查询，也可以是针对文本信息的关键字查询。文献[Dzbor and Motta, 2006]提出的 Swangler 系统则将语义标注转化为一般的文本查询关键字。清华大学提出了一种细粒度语义网检索模型，可对用户提供基于关键字的查询接口，检索系统以 RDF 图构建搜索策略，以 URI 资源为检索单位，查询结果是包含关键字在内的三元组集合[吴刚,等., 2005]。

资源间由关联关系引入的链接路径在某些特定领域比资源本身更具价值，比如在国家安全领域通常需要搜索资源之间的链接关系，这些关系可能意味着某些潜在的安全威胁。关联搜索中的主要问题在于如何定义链接的兴趣尺度，且这种定义方法不仅能够消除用户不感兴趣的关联关系，而且可以搜索到数据之间复杂的、隐藏的关联关系。文献[Anyanwu and Sheth, 2003]提出了一种大众化且简单的形式化计算方法，尝试发现资源间有价值的关联关系。语义搜索不仅要能够探索到资源之间的关联关系还需要获得合理的排序结果 [Aleman-Meza, et al., 2003]

[Anyanwu, et al., 2005]。知识库中实体之间关系的个数往往会超出实体本身，语义关联就是指实体之间的复杂关系。传统搜索引擎采用的排序方法只能对检索得到链接文本进行排序，无法对结构信息排序。为对语义搜索中获取的结构信息（多为 RDF 三元组）排序，目前多是将传统的结果排序算法做出改变以应用于语义搜索结果排序。文献[Bamba and Mukherjea, 2004]即利用语义 Web 资源的重要性对结果集进行排序。文献[Bai, et al., 2009]试图发现元数据上复杂的关系，提出了一种预测用户需求的排序方法来识别语义关联。

3.2.2 语义搜索中的查询扩展

传统的搜索引擎经常会因为词语含义的多样性而产生无意义的检索结果。产生词语多样性问题的根本原因在于，人们在现实生活中描述同样的对象或事件的用词存在着多样性。例如，单车和脚踏车都是对自行车这一概念的称谓。为解决这个问题，人们提出了基于概念的语义查询扩展（Semantic based QE），用概念来描述查询主旨，找到与查询语义相关的概念对查询进行扩展，因为概念是专门用来描述现实世界对象的。基于概念,可以消除现实世界中人们对同一真实对象的不同表达方式的理解差异。语义网的构建目标即是将网络中的概念构成网状结构，利用概念间的联系形成拓扑网络，而本体（语义网中的结点）则视为概念的具体表现形式。因此这种基于概念的查询扩展一直是语义搜索领域的研究重点。

目前语义搜索的研究侧重点围绕于查询语句或是文档中的语义发掘，注重发现目标资源间的关联，通过深度的查询理解而获得更高的查准率。文献[Jothilakshmi, et al., 2013]在领域模型的基础上提出了一种语义查询扩展方法，即结合概念级别（基于领域知识）、语法级别（基于 WordNet 的术语词汇）和随机模型 ME-HMM2（隐马尔可夫模型与最大熵模型相结合），取得了较好的效果。Pal 等在文献[Pal, et al., 2014]中提出了基于组合的概念映射查询扩展方法，考虑到每个候选扩展术语实用性的三个方面：其在相关文献和目标语料库中的分布、与查询术语的统计关联及术语在 WordNet 中的定义及其与查询术语间的语义关系，这种不同信息来源的组合能够对测试集合产生较好的效果。文献[Ngo and Cao, 2010]中提出了一种基于本体的广义向量空间模型进行文本的语义搜索，利用命名实体及其潜在的相关命名实体的本体特性获得文档和查询词的语义，并在此基础上构建了一个框架通过结合不同的本体，利用它们之间的互补优势进行语义标注和搜索。文献[Chauhan, et al., 2012]中提出了基于语义查询扩展的信息组织与检索系统，所提出的语义查询扩展方法包括一个基于领域本体的数学模型来计算概念之间的语义相似性和查询扩展算法，利用查询的概念及这些概念的同义词来执行查询扩展。Zhao 等在文献[Zhao, et al., 2015]中提出了一个基于物联网环境的、带有主题发现与语义感知功能的索引构建方案和检索系统 Acrost，通过以多主题为中心的搜集组合获得感兴趣信息的初始内容，基于聚合正则表达式和条件

随机域方法提取元数据，通过分析查询和对相关性内容排序进行语义感知检索。Bashar 等在文献[Bashar and Myaeng, 2014]中利用维基百科页面中的语义标注提出一种新颖的语义查询扩展方法用于对初始查询词消除歧义并丰富语义，然后将该方法应用于专利搜索、专利分类。

3.2.3 语义搜索中的索引构建

创建合理和有效的索引是保证搜索顺利进行的保证。传统的基于关键字的搜索引擎不能很好的解决一词多义,多词一义的问题。用户将花费很大的代价从搜索引擎返回的结果中寻找所需的结果或者换关键字重新查询。建立语义索引则是为搜索引擎解决以上问题提供了新方向。利用语义网中的本体去分析文档和查询语句的语义信息，从而为海量的无结构网页数据建立语义索引,查询时通过匹配用户意图和文档中以本体标识的概念的相关性给出结果。这种方法避免了基于关键词搜索的一词多义和多词一义问题。如文献[Ma, et al., 2007]提出了一种利用本体获取词间的语义关系,消除自然语言的多义性以标识文档的方法。文献[Mihalcea and Moldovan, 2000]介绍了基于 WordNet，用布尔模型添加词语语义信息到传统索引上从而建立了一个结合了语义和传统关键字的新型索引。文献[Buscaldi and Zargayouna, 2013]介绍了一种标注文档中概念的语义检索系统 YaSemIR，并且这个系统可配置以和不同的本体、不同的文档一起工作。文献[Setchi, et al., 2011]将文档中最有意义、最有代表性的词语找出,并且获取其语义以组成一个语义核心，以这些语义核心建立索引，查询匹配的时候仅就语义核心和查询语句进行匹配。文献[Roger, 2008]考虑了词语对的关联性，并依据关联性强度快速地构建了一个潜在语义索引分析系统。文献[Kokiopoulou and Saad, 2004]在改进的 K-近邻算法基础上，消除了传统潜在语义索引时间复杂度高的特点，并应用文档索引结果做基于反馈的文本过滤。

4. 技术展望与发展趋势

国内外主流的搜索引擎厂商对于语义搜索的前景极为看好，普遍认为其是机遇与挑战并存的新领域。为了改进搜索效果和提升竞争力，处于主流地位的传统搜索引擎巨头也开始尝试语义搜索技术。2008 年微软收购了语义搜索引擎 Powerset[27]，希望以此提高 Bing 的语义功能和认知度以缩小与 Google 在搜索质量上的差距。百度则在 2009 年即开始涉足语义搜索领域，与哈尔滨工业大学建立合作研究实验室，专门对语义搜索中的关键技术-自然语言处理进行研发，并推出一款基于语义的“框计算”应用，专门用于对中文中的生僻字进行查询[Yang, el al., 2000]。Google 也于 2012 年 5 月推出知识图谱，将其应用于搜索引擎中以增强搜索结果，标志着大规模知识图谱在互联网语义搜索中的成功应用，视其为

[27] https://www.cnet.com/news/report-microsoft-to-buy-powerset/

下一代语义搜索的第一步。

虽然各大互联网公司都试图在语义搜索上有所突破，但目前国内外科研机构对语义搜索的研究还处于初步探索阶段，并未形成一种通用的框架和方案。虽然提出了多种系统，但受限于语义网技术的尚未普及，并未有一套实用的语义搜索系统。已提出的系统有的只是对传统的信息检索功能进行补充和完善，有的只能提供形式化的查询，有的仅仅是有限利用了本体中的结构数据，并不存在能紧密结合两者功能的系统，实现的推理功能尚处于初步尝试过程中，目前也不存在较为成熟的基于语义的结果排序方法。未来的语义搜索研究方向可沿以下几点展开：

① 语义搜索概念模型。语义模型能改善当前搜索引擎的搜索效果，未来可扩展成为构建在语义Web上的新一代搜索引擎。

② 语义搜索本体知识库的构建、维护与进化。研究垂直领域的本体知识库构建方法、本体知识库设计方法和本体知识库查询方法，构建完备的领域本体知识库，探索本体知识库的维护方案，随着领域本体知识库的丰富还要研究并解决多领域异构的本体知识库的融合问题，提供本体相容性冲突检测方案。

③ 语义搜索的推理机制。结合领域本体，研究语义搜索中基于描述逻辑及模糊逻辑的推理问题，提高基于描述逻辑的本体推理技术的推理效率，扩大其推理算法的适用范围，结合文本信息获取用户的查询语义，提高处理用户查询需求的准确度。

④ 语义搜索的结果排序。传统搜索引擎采用的排序方法只能对文本信息进行排序，不能对实体之间的复杂关系排序，无法实现语义搜索结果的排序，因此需研究基于语义的结果排序方法，实现本体知识库中实体及实体之间关系的排序，提高返回结果的相关性。

⑤ 语义搜索的原型系统实现。基于以上研究，实现语义搜索引擎系统原型，在应用环境中进行测试并实现性能优化。

参考文献

[Arvind, et al., 2001] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. ACM Transactions on Internet Technology, 2001, 1(1): 2-43.

[Aleman-Meza, et al., 2003] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth. Context-aware semantic association ranking. In Proceedings of the First International Conference on Semantic Web and Databases, 2003, 24-41.

[Anyanwu and Sheth, 2003] Kemafor Anyanwu and Amit Sheth. P-Queries: Enabling querying for semantic associations on the semantic web. In Proceedings of the

12th International Conference on World Wide Web, 2003, 690-699.

[Anyanwu, et al., 2005] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. SemRank: Ranking complex relationship search results on the semantic web. In Proceedings of the 14th International Conference on World Wide Web, 2005, 117-127.

[Anuar, et al., 2016] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. Semantic retrieval of trademarks based on conceptual similarity. IEEE Transactions on Systems, Man, and Cybernetics, 2016, 46(2): 220-233.

[Baader, et al., 2003] Franz Baader, Diego Calvanese, Deborah L Mcguinness, Daniele Nardi, and Peter F. Patel-Schneider. The description logic handbook: Theory, implementation and applications. Cambridge University Press. 2007.

[Bamba and Mukherjea, 2004] Bhuvan Bamba and Sougata Mukherjea. Utilizing resource importance for ranking semantic web query results. In Proceedings of the Second International Conference on Semantic Web and Databases, 2004, 185-198.

[Bai, et al., 2009] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Supervised semantic indexing. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, 187-196.

[Buscaldi and Zargayouna, 2013] Davide Buscaldi and Haïfa Zargayouna. YaSemIR: Yet another semantic information retrieval system. In Proceedings of the 6th International Workshop on Exploiting Semantic Annotations in Information Retrieval, 2013, 13-16.

[Bashar and Myaeng, 2014] A. S. Bashar and S. H. Myaeng. Wikipedia-based query phrase expansion in patent class search. Information Retrieval, 2014, 17(5): 430-451.

[Chauhan, et al., 2012] R. Chauhan, R. Goudar, R. Rathore, P. Singh, and S. Rao. Ontology based automatic query expansion for semantic information retrieval in sports domain. In Proceedings of the International Conference on Eco-friendly Computing and Communication systems, 2012, 422-433.

[Ding, et al, 2004] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In Proceedings of the 13th ACM International Conference on Information and Knowledge Management, 2004, 652-659.

[Ding and Finin, 2006] Li Ding and Tim Finin. Characterizing the semantic web on the

web. In Proceedings of the 5th International Semantic Web Conference, 2006: 242-257.

[Dzbor and Motta, 2006] Martin Dzbor and Enrico Motta. Study on integrating semantic applications with magpie. In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2006, 66-76.

[Guo, et al., 2003] Lin Guo, Feng Shao, Chavdar Botev, and Jayavel Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 2003, 16-27.

[Horrocks, et al., 2003] Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From SHIQ and RDF to OWL: The making of a Web Ontology Language. Web Semantics: Science, Services and Agents on the World Wide Web, 2003, 1(1): 7-26.

[Jothilakshmi, et al., 2013] R. Jothilakshmi, N. Shanthi, and R. Babisraraswthi. An approach for semantic query expansion based on maximum entropy-hidden markov model. In Proceedings of the 4th International Conference on Computing, Communication and Networking Technologies, 2013, 1-5.

[Kokiopoulou and Saad, 2004] E. Kokiopoulou and Y. Saad. Polynomial filtering in latent semantic indexing for information retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, 104-111.

[Mihalcea and Moldovan, 2000] Rada Mihalcea and Dan Moldovan. Semantic indexing using WordNet senses. In Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational, 2000, 35-45.

[Ma, et al., 2007] Wenhui Ma, Wenbin Fang, Gang Wang, and Jing Liu. Concept index for document retrieval with peer-to-peer network. In Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/ Distributed Computing, 2007, 1119-1123.

[Nigel, et al., 2006] Shadbolt Nigel, Hall Wendy, and Berners-Lee Tim. The semantic web revisited. IEEE Intelligent Systems, 2006, 21(3): 96-101.

[Ngo and Cao, 2010] V. M. Ngo, and T. H. Cao. Ontology-based query expansion with latently related named entities for semantic text search. Advances in Intelligent

Information and Database Systems, 2010, 283: 41-52.

[Pal, et al., 2014] D. Pal, M. Mitra, K. Datta. Improving query expansion with latently related named entities for semantic text search. In Proceedings of the 2nd Asian Conference on Intelligent Information and Database Systems, 2010, 41-45.

[Roger, 2008] Bradford B. Roger. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008, 153-162.

[Shah, et al., 2002] Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, and James Matfield. Information retrieval on the semantic web. In Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, 461-468.

[Sintek and Decker, 2002] Michael Sintek and Stefan Decker. TRIPLE-A query, inference, and transformation language for the semantic web. In Proceedings of the First International Semantic Web Conference on the Semantic Web, 2002, 364-378.

[Setchi, et al., 2011] Rossi Setchi, Qiao Tang, and Ivan Stankov. Semantic-based information retrieval in support of concept design. Advanced Engineering Informatics, 2011, 25(2): 131-146.

[Tim, et al., 2001] Berners-Lee Tim, Hendler James, and Lassila Ora. The semantic web. Scientific American: Feature Article, 2001.

[吴刚, 等., 2005] 吴刚, 唐杰, 李涓子, 王克宏. 细粒度语义网检索. 清华大学学报（自然科学版）, 2005, 45(1): 139-146.

[Yang, el al., 2000] Qiang Yang, Hai-Feng Wang, Ji-Rong Wen, and H. M. Zhang. Towards a next-generation search engine. In Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, 2000, 1-12.

[Zhang, et al., 2007] Lei Zhang, Qiaoling Liu, Jie Zhang, Haofen Wang, Yue Pan, and Yong Yu. Semplore: An IR approach to scalable hybrid query of semantic[C]. In Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007, 652-665.

[Zhao, et al., 2015] F. Zhao, Z. Sun, and H. Jin. Topic-centric and semantic-aware retrieval system for internet of things. Information Fusion, 2015, (23): 33-42.

第十一章 基于知识的问答

1. 任务定义、目标和研究意义

问答系统（Question Answering, QA）是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。华盛顿大学图灵中心主任 Etzioni 教授 2011 年曾在 Nature 上发表文章《Search Needs a Shake-Up》，其中明确指出：“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”[Etzioni O., 2011]。因此，问答系统被看做是未来信息服务的颠覆性技术之一，被认为是机器具备语言理解能力的主要验证手段之一。因此，对其开展研究具有非常重要的学术和实际意义。特别是近些年，随着人工智能热潮到来，无论是学术界还是产业界，都给予其极大关注和投入。

纵观问答系统的技术演进，其一直伴随的人工智能技术的发展而发展。近些年，问答系统更是取得一系列倍受关注的成果。2011 年，IBM Watson 自动问答机器人在美国智力竞赛节目 Jeopardy 中战胜人类选手，在业内引起了巨大的轰动。随着人工智能技术的突飞猛进，各大 IT 巨头更是相继推出以问答系统为核心技术的产品和服务，如移动生活助手（Siri、Google Now、Cortana、小冰等）、智能音箱（HomePod、Alexa、叮咚音箱等、公子小白等）等，这似乎让人们看到了黎明前的阳光，甚至认为现有的问答技术已经十分成熟。

尽管 IBM Watson 系统在 Jeopardy 中战胜了人类选手，但是其核心技术并没有突破传统基于“检索+抽取”的问答模式，缺乏对于文本语义深层次的分析和处理，难以实现知识的深层逻辑推理，无法达到人工智能的高级目标。Watson 的成功也已经被证明仅仅局限于限定领域、特定类型的问题，离语义的深度理解以及智能问答还有很大的距离，其他问答系统，如 Siri 等，也存在同样的问题。因此，面对已有问答模式的不足，为了提升信息服务的准确性与智能性，研究者近些年逐步把目光投向知识图谱（Knowledge Graph）。其意图是通过信息抽取、关联、融合等手段，将互联网文本转化为结构化的知识，利用实体以及实体间语义关系对于整个互联网文本内容进行描述和表示，从数据源头对于信息进行深度的挖掘和理解。同时，互联网中已经有一些可以获取的大规模知识图谱，例如 DBpedia[Lehmann et al., 2014]、Freebase[Bollacker, 2008]、YAGO[Suchanek et al., 2007]等。这些知识图谱多是以实体、关系为基本单元所组成的图结构。

基于这样的结构化的知识，分析用户自然语言问题的语义，进而在已构建的结构化知识图谱中通过检索、匹配或推理等手段，获取正确答案，这一任务称之