基于关键词的语义网数据查询研究综述

李慧颖[1] 瞿裕忠[2]
(东南大学计算机科学与工程学院 南京 210096)[1] (南京大学计算机科学与技术系 南京 210093)[2]

摘 要 语义网数据的关键词查询是语义网研究的一个重要问题。首先给出语义网数据关键词查询的相关定义。根据研究目标不同,将已有解决方案分为混合型和非混合型的语义网数据关键词查询,后者又分为K-A和K-Q-A两种查询方法。调研了上述分类中当前常用的解决方案和研究进展。在此基础上,进一步介绍并比较了8个具有代表性的语义网数据关键词查询工作。最后讨论存在的挑战,并指出未来可能的研究方向。

关键词 关键词查询,RDF数据,Top-K,语义网
中图法分类号 TP18 **文献标识码** A

Keyword-based Search on Semantic Web Data: The State of the Art

LI Hui-ying[1] QU Yu-zhong[2]
(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]

Abstract The growth of the Semantic Web has seen a rapid increase in the amount of RDF data. Meanwhile, the demand for access to RDF data without detailed knowledge of RDF query languages is increasing. In this paper, some definitions in terms of keyword-based search on Semantic Web data were firstly presented. Then, a large number of popular existing solutions were surveyed according to hybrid and non-hybrid style(including K-A means and K-Q-A means). In addition, eight representative systems were introduced and compared with each other based on different facets, and their unique features were highlighted in details. Finally, some remaining challenges were discussed, and some future research directions were pointed out.

Keywords Keyword-based search, RDF data, Top-K, Semantic Web

1 引言

过去10年里,语义网[1](Semantic Web)在W3C的倡导下得到了快速的发展,其目标是提供一个通用的语义框架,实现数据在不同应用之间的共享和集成。RDF[2]作为语义网的基础,以三元组(triple)形式表示信息并交换万维网上的知识和数据。据统计[3],万维网上已经发现超过1000万个包含语义数据的文档,其中大多数是以RDF,RDFS[4],OWL[5]语言描述的。我们称这些由本体语言描述、以三元组形式组织起来的数据为语义网数据(也称为RDF数据)。

随着语义网数据的大量增加,普通用户直接对其进行查询的需求也在不断增加。目前已经有一些查询语言如SPARQL[6,7],SeRQL[8]等支持语义网数据查询,但它们对普通用户而言过于复杂。原因在于使用查询语言的前提是要求用户必须掌握其语法规则和待查询数据的模式信息。万维网搜索引擎(如Google等)中基于关键词的搜索技术得到广泛应用的事实表明,普通用户更倾向于简便的查询方式。因此,提供关键词查询方式对语义网数据的检索和重用成为一个重要问题。

目前,国内外已经有一些涉及语义网数据的关键词查询问题的研究。根据研究目标的不同,将已有解决方案分为两大类:一类是在万维网搜索引擎的基础上,通过查询网页的语义标注数据(RDF数据)来提高万维网搜索引擎的搜索效果[9-16]。这类方式多是通过结合语义网数据查询和万维网搜索来增强后者的搜索效果,我们称之为混合型语义网数据关键词查询方式;另外一类是直接在语义网数据中进行关键词查询以获取信息,我们称之为非混合型语义网数据关键词查询方式。对于第二类研究工作,根据处理方法的不同又分为由查询关键词直接构造查询结果(K-A)[17-23]和由查询关键词构造形式化查询语句再得到查询结果(K-Q-A)[24-29]两种方法。

虽然上述方案的具体实现有所不同,但都基本遵循图1所示的系统框架(加粗部分为混合型方式特有的模块),分为索引过程和查询过程。索引过程对语义网数据进行预处理,建立相关索引。查询过程根据用户输入的关键词查找用户需要的信息,按照一定的评分函数为候选结果排序,最终将前k个返回给用户。

到稿日期:2010-05-05 返修日期:2010-10-08 本文受国家自然科学基金(60973024),江苏省自然科学基金(BK2008290)资助。
李慧颖(1977—),女,博士生,讲师,主要研究方向为语义万维网,E-mail:huiyingli@seu.edu.cn;瞿裕忠 男,博士,教授,博士生导师,主要研究方向为万维网科学、语义万维网、计算机软件方法与技术。

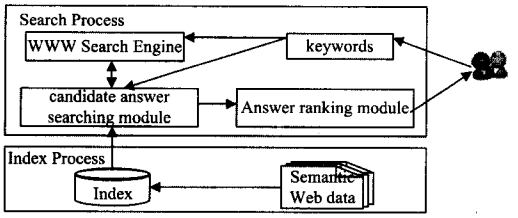

图 1　语义网数据关键词查询框架

本文第 2 节形式化地给出了语义网数据关键词查询问题的相关定义；第 3 节分别介绍两类不同方式的常用方法和研究进展，并指出其各自的研究难点；第 4 节进一步介绍和比较 8 种具有代表性的语义网数据关键词查询的系统；最后讨论目前存在的挑战以及未来的研究方向。

2　问题描述

本节形式化地给出语义网数据关键词查询的相关定义。语义网数据是指以 RDF 三元组的形式组织起来的数据，通常由 RDF，RDFS 或 OWL 语言描述，也称为 RDF 数据，可以通过有向图模型来表示，以 RDF 三元组的主体(Subject)和客体(Object)作为节点，以三元组的谓词(Predicate)作为从主体指向客体的有向边。依据文献[2]，简单地给出 RDF 数据图的一个定义。

定义 1(RDF 数据图)　设 U,B,L 分别代表 URIref 集合、空白节点集合和文字集合。一个 $(subject, predicate, object)\in(U\cup B)\times U\times(U\cup B\cup L)$ 称为 RDF 三元组。其中谓词可以分为两个不相交的集合 R 和 A。R 表示关系(relation)集合；A 表示属性(attribute)集合。一个 RDF 数据图就是 RDF 三元组的集合。

对于 RDF 数据的关键词查询，根据事先给定的查询关键词，查询结果有多种不同的定义。文献[29]给出关键词查询结果的形式化定义。

定义 2(关键词查询结果)　给定查询关键词集合 $q=\{q_1,q_2,\cdots,q_m\}$ 和 RDF 数据图，文本信息中出现查询关键词的节点称为关键词节点，关键词查询结果是一个包含关键词节点的子图。

对于给定的查询通常会有多个查询结果，我们称之为候选查询结果，需要定义评分函数(ranking function)进行评分并返回 Top-K 个查询结果。本文对于 Top-K 关键词查询结果的定义如下。

定义 3(Top-K 关键词查询结果)　给定评分函数对候选查询结果进行评分，按评分结果降序排列，其中前 k 个就是 Top-K 关键词查询结果。

这里进一步解释语义网数据关键词查询问题的研究范畴。首先，该问题仅指对 RDF 数据的关键词查询，不包括数据库领域中关键字检索问题(请参阅文献[30,31])和 XML 数据的关键词检索问题(请参阅文献[32-34])。其次，这里的查询方式仅指关键词查询，不包括需要输入对象 URI 的语义关联搜索[35,36]。另外，语义网数据的关键词查询问题也不包括收集语义网文档及对语义网文档进行的检索(请参阅文献[37])。

3　分类

目前的研究工作根据研究目标的不同分为两类：一类是以提高万维网搜索引擎的搜索效果为目的的混合型语义网数据关键词查询方式。另外一类是以直接从语义网数据中获取信息为目的的非混合型语义网数据关键词查询方式。后者又根据处理方法的不同分为由查询关键词直接计算查询结果和由查询关键词计算形式化查询语句再得到查询结果两种方法。下面分别介绍它们的研究进展，并指出各自的研究难点。

3.1　混合型语义网数据关键词查询

混合型语义网数据的关键词查询最早于 2003 年由文献[9]提出，利用网页的语义标注数据和网页的文本数据，将语义网技术和万维网搜索引擎相结合，目的在于提高万维网搜索引擎的效果。在这种方式中，假设网页已经具有语义标准数据，当用户输入关键词时，系统会利用搜索引擎搜索网页，同时会向标注数据发起查询，两者的结果相互影响，最终返回给用户更合理的排序结果。对于混合型语义网数据的关键词查询可以从搜索引擎和 RDF 数据查询的交互方式及搜索引擎的效果从何处获得提高两个方面进行分类。

在交互方式方面，可以分为将万维网搜索引擎的结果作为 RDF 数据查询的输入[10]和利用本体和标注数据进行语义查询扩展[13-16]两种方式。在第一种方式中，网页的 URI 及其文本数据通常被建模为 RDF 数据图中的一个节点(网页节点)，其语义标注数据刻画了这个网页节点的属性或者和其他网页节点的关系。对于用户输入的查询关键词，系统先利用搜索引擎得到若干网页节点，将这些网页节点及其在搜索引擎中的评分分值作为 RDF 数据查询的输入，在此基础上利用激活扩散算法获得其它网页节点评分分值，最后将分值最高的若干网页节点返回。第二种方式要求事先定义领域本体和标注数据，搜索时先利用领域本体理解用户输入关键词的语义，通常是补充查询关键词之间的关系或者查询关键词的类名等，同时将标注数据作为网页内容的补充，接着将扩展后的查询通过万维网搜索引擎获得最终结果。

在搜索引擎的效果提高方面可以分为为搜索引擎结果提供论据(argument)、提高搜索引擎的查全率和查准率 3 个方面。为搜索引擎提供论据是在提供万维网搜索结果的同时，在标注数据中寻找与查询关键词相关的实体，将该实体及其属性信息提供给用户，以便用户更好地理解搜索引擎的结果。提高搜索引擎的查全率和查准率就是在现有搜索引擎的基础上利用对标注数据的查询来提高搜索结果的语义相关度，具体到不同的方法，有的侧重提高万维网搜索引擎的查全率，有的侧重提高万维网搜索引擎的查准率。

总体来说，混合型语义网数据的关键词查询是在领域本体和标准数据的支持下，利用对标注数据查询获得的信息来辅助万维网搜索引擎获得更好的效果。然而，大部分的领域本体需要手工创建和维护，基于本体的语义标准也需要手工参与，语义标注本身就是一个难点[38]。另外，这种方式中对标注数据的查询通常都是寻找与查询关键词相关的单个实体，很少考虑到查询关键词分布到多个实体的情况(只有文献[9]简单提到了这方面的问题)。而实际上，经常会出现要将多个实体(多个网页节点)组合在一起才能满足用户的查询需求的情况。如何处理上述情况是需要面对的问题。

3.2　非混合型语义网数据关键词查询

随着语义网数据的大量出现，尤其是很多大规模 RDF 数据的出现，直接对其进行关键词查询的需求也在不断增加。

非混合型语义网数据关键词查询从处理问题的不同方法上分为以下两种。

3.2.1 K-A 查询方法

由查询关键词直接计算查询结果的方法是以 RDF 数据图模型作为基础，定义查询结果模型(确定满足何种条件的子图作为查询结果)，建立相关索引以支持快速实时的查询响应，实现查询算法找到候选查询结果，对候选结果进行评分并将前 k 个返回给用户。对于 K-A 查询可以从查询结果的建模方式、索引的内容、评分函数 3 方面进行分类(具体参见图 2)。

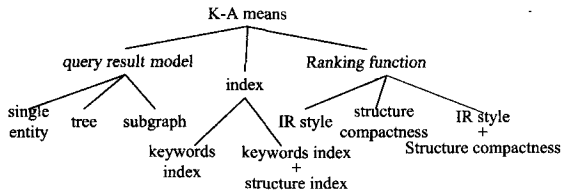

图 2 K-A 方法解决方案的分类

根据定义 1，RDF 数据可以建模为一个点和边都有标签的有向图。实际上，用户输入的查询关键词可能匹配到 RDF 数据图中的节点(关键词节点)或边(关键词边)，而用户想要获得的查询结果是将这些关键词节点和关键词边联系起来的信息。遗憾的是，现有的 K-A 解决方案不支持查询关键词匹配到边的情况。因此，查询结果通常是一个 RDF 数据图的子图，其中包括所有(或部分)的关键词节点。

目前 K-A 查询方法对结果的建模有 3 种：一种是将查询结果建模为单个实体(单个节点)，实际上就是寻找一个包含所有(或部分)查询关键词的节点，这是最简单的一种方式。一种是将查询结果建模为树，要求每个查询关键词都与一个树叶节点相关，并且每个树叶节点都是关键词节点，另外结果树中的其他节点都是为了关联关键词节点而引入的。这样，求最小(节点最少)的查询结果树就转化为图论中的斯坦纳树问题。最后一种是将查询结果建模为子图，要求查询结果是由关键词节点和关键词节点间路径上的点生成的子图。后两种建模方式的区别在于：从直观上理解，查询结果树模型是要找到一个将关键词节点连接起来的一个节点，称为连接节点(connecting node)，这个连接节点以及它到每个关键词节点的路径就构成了查询结果，但是寻找这样的最小查询结果是图论中的 NP 完全问题。现有的解决方案[19]多是利用启发式算法求近似解。而查询结果子图模型是将关键词节点间的所有路径都放入查询结果中，其优点是给用户提供的结果信息比较丰富，而且查询算法可以利用图论中一些已有的路径算法，比较易于实现。

在索引方面，可以分为仅建立关键词倒排索引和建立关键词倒排索引加上结构索引两种方式。关键词查询方式起源于信息检索，关键词倒排索引是必不可少的一部分。早期的工作通常只建立关键词索引，以辅助快速定位与查询关键词相关的节点或边。近期的工作为了支持大规模 RDF 数据的快速查询响应，会在关键词索引的基础上增加结构索引，其目的在于减轻实时查询的计算负担，将一些中间结果事先在预处理阶段计算好并建立索引。结构索引的具体内容主要取决于采用何种查询算法，例如为所有节点间的最短路径建立索引以实现查询结果树的构造算法[19]，或计算并索引 r-半径(r-radius)子图以实现查询结果子图的构造算法[22]。

在评分函数方面可以分为只考虑信息检索的文本相关评分策略、只考虑结构紧凑性的评分策略和将两者结合起来的评分策略 3 种方式。早期的工作通常仅考虑结构紧凑性的评分策略，例如将查询结果中的路径总长度的倒数作为评分结果，认为分数越高越贴合用户的需求。这是因为大家公认普通用户更倾向于紧凑的查询结果。后来，文献[27]提出将文本相关性作为评分策略，例如采用类似于 TF/IDF 的文本相似度计算公式进行评分。但是，区别于传统的信息检索技术，RDF 数据的关键词查询是在结构化的数据上进行查询，因此单纯考虑文本相似度也无法取得较好的效果。最后一种评分方式是综合考虑查询结果的结构紧凑性和文本相似度，取得了一定的效果。但是，实际上评分策略的选择应该依赖于最终面对的用户，比如普通用户可能更倾向于结构紧凑的查询结果，而试图在数据中挖掘线索的用户可能更倾向于发现一些不常见的关联。因此，与用户交互并进行查询精化的方式才能满足更多不同用户的需求。目前，在 RDF 数据查询中还没有这方面的工作。

另外，对 RDF 数据的查询问题还可以分为对单一数据源进行关键词查询和对多数据源进行关键词查询两种情况。

总体来说，K-A 查询方法的优势在于直接对 RDF 数据图进行查询，数据的粒度细，查询关键词匹配到数据图中的节点并直接在数据图中构造查询结果返回给用户。不需要背景知识，也不需要 RDF 模式(RDF schema)信息的支持。不足之处在于当面对大规模的 RDF 数据时，现有查询算法的效率有待提高。另外，随着数据规模的增加，建立索引所需的时间和空间也会大量增加。

3.2.2 K-Q-A 查询方法

K-Q-A 查询方法是将查询关键词匹配到 RDF 数据图的节点或边，在查询模板或模式信息的辅助下构造与用户查询关键词相关的候选形式化查询，将这些候选查询排序并返回用户，用户可以利用它们向 RDF 数据库发起查询并获得最终查询结果。K-Q-A 查询可以从候选查询的构造方法、候选查询使用的查询描述语言、评分函数 3 个方面进行分类(具体参见图 3)。

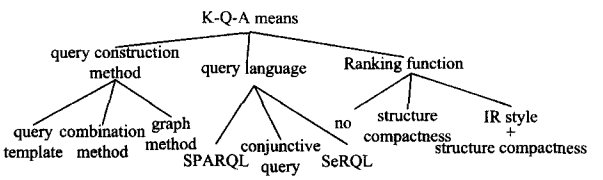

图 3 K-Q-A 方法解决方案的分类

由关键词到形式化查询再到查询结果的方法目前主要的研究工作集中在从关键词构造形式化查询部分，因为已经有一些比较成熟的工作帮助用户发起形式化查询并最终获得查询结果。因此，如何从用户输入的关键词构造符合其要求的形式化查询语句成为关键问题。从用户的角度看，他并不了解数据的模式信息和形式化查询语言的语法规则，只是利用查询关键词描述其想要得到的信息。所以，K-Q-A 查询方法的工作是依靠模式信息或背景知识找到查询关键词之间的关联，确定用户想查询的对象并构造符合语法规则的形式化查询语句。

目前形式化查询的构造方法有 3 种：一种是采用事先定

义查询模板的方法[28]，即利用模式信息在查询关键词比较少的情况下事先定义几种可能的查询模板，当具体查询出现时，找到适合的模板构造查询语句并输出查询结果。一种是基于组合的方式[25]，即确定每个关键词对应的节点集合，利用组合的方式确定这些关键词节点之间所有可能的关联，对每种组合方式构造候选查询图。第3种是利用图算法构造候选查询[29]，通过模式信息或背景知识将数据图进行摘要(summary)，在此基础上使用图算法(如图搜索算法等)得到候选查询图并转化为相应的查询语句返回用户。前两种方法的不足之处在于，随着查询关键词数目的增加，查询模板或组合数目会极大地增加，系统可扩展性能不高。最后一种方法是在摘要图上构造候选查询，因此能够实现快速响应，近期的K-Q-A工作主要采用这类方法。

候选查询使用的查询语言根据不同的应用可以分为3种：一种是SPARQL查询语言，它是目前比较主流的RDF数据查询语言，是W3C的候选推荐标准。一种是SeRQL查询语言，它是Sesame(一个支持RDF数据存储、形式化查询和推理的开源项目)提出的一种RDF查询语言。一种是将候选查询用合取查询的形式表达，合取查询的抽象程度高于具体的查询语言，用户可以根据不同的应用将合取查询再转换为SPARQL查询语句或SQL查询语句等。因此，第3种方法的灵活度比较高。

对于评分函数同样分为3种。在K-Q-A查询方法中，这里的评分函数主要是对候选形式化查询进行评分。早期的工作基本上没有考虑到评分函数。近期基于图算法的工作对候选查询的评分策略是基于结构紧凑性和文本相关性两个方面考虑的，类似于K-Q方法对最终查询结果的评分策略。

总体来说，K-Q-A查询方法(主要针对近期基于图算法的工作)的优势在于，将RDF数据抽象为摘要图并在此基础上构造候选查询，通常摘要图的规模要远小于数据图，因此候选查询构造算法的效率高。不足之处在于，抽象后的摘要数据粒度粗，在一些特殊情况下无法得到令人满意的候选查询。

4 语义网数据关键词查询系统

上一节已经从两个不同角度对部分已有工作进行了分类和归纳。在本节中，首先对8个较具代表性的系统工具做简要介绍，接下来对它们做进一步的比较和分析。

4.1 系统简介

TAP[9]是由美国斯坦福大学和IBM Almaden研究中心等于2003年共同开发的一个语义搜索系统。它的特点是在提供万维网搜索结果的同时提供相关语义信息，将查询关键词匹配到RDF数据中的实体，为用户展示与查询关键词最相关的实体以及该实体的属性和关系，目的是为万维网搜索结果提供论据。TAP是混合型语义网数据关键词查询研究的早期代表性工作。

SemSearch[24,28]是由英国开放大学于2006年设计实现的一个基于关键词的搜索系统。该系统支持用户对单源的语义网数据进行关键词查询，它采用的技术是将用户输入的关键词通过事先定义好的模板转换为SeRQL查询语句，向后台数据库发起查询并最终将结果返回给用户。实验显示，当查询关键词的数目是2时，需要事先定义7个查询模板，而随着用户输入的关键词数目的增多，事先要定义的模板数目将极大地增加。SemSearch是非混合型语义Web数据关键词查询早期研究工作的代表，它提出了一个通用的语义Web数据的关键词查询框架，许多后续工作都继承了它的系统框架。

SPARK[25]是由国内上海交通大学2007实现的一个关键词搜索系统。在定位与查询关键词相关的RDF资源过程中，它除了采用自然语言处理的取词根(steming)、计算编辑距离等技术，还利用了WordNet查找同义词的方法，使得查询关键词可以匹配到更多的RDF资源(关键词节点)。每个查询关键词(设共m个)对应的关键词节点构成一个集合，从m个集合中挑出不同的关键词节点进行组合，使得每种组合方式中恰好包含m个关键词节点，从这些节点出发利用克鲁斯卡尔算法得到最小生成树作为候选查询图。将候选查询图转化为SPARQL查询语句，用概率论方法为它们评分并返回给用户。由于使用组合的方式来确定可能的候选查询，在查询关键词数目增多或匹配到的RDF资源很多的情况下，系统的性能会受到很大的影响。另外，系统不支持用户将属性名或关系名作为查询关键词输入。

OntoLook[13]是由国内华中科技大学于2007年开发的一个基于关系的搜索系统。该系统假定已有领域本体并对现有网页进行了标注，在标注的RDF数据和网页文本的基础上实现关键词检索。用户需要输入查询关键词及其对应的类(关键词类)。系统利用领域本体找到关键词类之间的直接关系，构建一个概念关系图，计算出概念关系图所有的生成子图，每个生成子图可以得到一组属性-关键词对(查询关键词+关系+查询关键词的形式)集合，利用这样的属性-关键词集合向后台数据库发起查询，得到相关的网页。该系统的优势在于有了领域本体支持，可以得到查询关键词之间的关系，将其加入到查询中并在标注数据的支持下得到更符合用户要求的网页；但是，系统只能发现关键词之间的直接关系，如果两个关键词之间通过两个或多个关系间接关联起来，系统是无法处理的。

BLINKS[19]是美国杜克大学和IBM T. J. Watson研究中心于2007年共同开发的基于二级索引的关键词查询系统。系统采用的是K-A查询方法，支持以对点标签的有向图进行关键词查询。为了处理大规模有向图，BLINKS首先采用两种算法进行图划分：一种是基于广度优先搜索算法，一种是基于METIS算法。通过划分将大规模数据图分割为若干节点数目固定的块，块之间通过门户节点关联起来。接着，系统建立索引，除了关键词索引外，为了实现快速实时查询，BLINKS建立了所有节点对之间的最短路径索引(结构索引)。为了获得处于不同块的节点间的路径信息，系统除了在每个块内索引节点对之间的最短路径以及节点和门户节点之间的路径信息外，还在块和块之间索引了门户节点的信息。确定查询关键词节点集合后，基于结构索引，利用量化均衡和等距启发式规则寻找候选查询结果树，最后以候选结果树中路径长度和的倒数作为评分函数将结果排序输出。该系统的优势在于采用空间换取时间的思路，事先进行图划分并建立了结构索引，利用启发式算法寻找候选查询结果，效率较高；但是，系统的算法不支持对边标签的图进行关键词查询，即无法处理用户将属性名或关系名作为查询关键词输入的情况。另外，系统建立结构索引的时间和空间开销也是比较大的。

EASE[22]是由国内清华大学和新加坡国立大学于2008

年共同开发的适用于非结构、半结构和结构化数据的关键词查询方法。首先将上述数据建模为点标签的无向图，建立关键词索引和结构索引。这里的结构索引是通过图邻接矩阵的幂次计算得到的 r 半径图（r 是系统变量，实验中 r 设为 2）。接着，确定查询关键词对应的 r 半径图，并将其中的非斯坦纳节点（指非关键词节点和非关键词节点路径上的点）删去，得到 r 半径斯坦纳图，作为候选查询结果。为候选查询结果评分并返回给用户。EASE 的优势在于利用图邻接矩阵的幂次计算 r 半径图并作为结构索引，算法简单有效；不足之处在于系统变量 r 的选择如果较大，则计算结构索引的时间和空间代价会很大。如果 r 值选得较小，则查询关键词可能不会同时出现在一个 r 半径图中，即查询结果不能包含所有的查询关键词（只能包含部分）。

Falcons[20] 是由国内东南大学于 2008 年开发的一个在语义网中发现和浏览实体的系统。该系统从语义网中抓取了约 700 万个 RDF 文档，以关键词查询的方式为用户提供 RDF 数据中实体的检索。系统为每个实体建立虚拟文档[39]，在此基础上建立倒排索引，以保证可以由关键词快速定位到相关实体，最后为用户返回一组排序后的实体。Falcons 的优势在于支持在语义网范围内查询实体，可以将多个数据源对同一个实体的描述融合在一起，来建立实体的虚拟文档；不足之处在于 Falcons 目前关注的还是单个实体的查询，不支持将多个实体联系在一起，以满足用户的查询需求。

TWRC09[1)][29] 是由德国卡尔斯鲁厄大学和国内上海交通大学于 2009 年共同开发的 RDF 数据的关键词搜索系统。系统在 RDF 数据的基础上计算摘要图，摘要图与模式信息类似，但它在模式信息的基础上从数据图中抽取出一些信息补充进来。用户输入关键词查询时，系统实时计算带参数的摘要图，即在摘要图上加入关键词节点和关键词边等。接着，在带参数的摘要图上进行类似广度优先搜索算法的搜索，得到候选查询图。最后将候选查询图排序并由此生成合取查询返回给用户。系统的优势在于通过将数据图进行摘要的方式极大地缩小了查询算法的搜索范围，提高了系统性能；不足之处在于抽象后的摘要图数据粒度粗，在此基础上构造查询图时，有些用户查询可能无法返回令人满意的结果。

4.2 系统的比较分析

除了上述系统外，还有许多各具特色的系统工具，如 RSS[21]，QuizRDF[11]，Swoogle[40]，Q2Semantic[27] 等，参见文献[41-43]等的工作。由于目前还没有统一的评价标准，很难对它们进行定量的评价。一般来说，系统之间的主要区别在于：(1)查询方式；(2)构建索引的方法；(3)评分函数以及(4)候选查询使用的查询语言（针对 K-Q-A 方法）。根据以上几个方面，现将上述 8 个 RDF 数据的关键词查询系统进行总结（见表 1）。

表 1 表明，首先，较早的语义网关键词查询工作都是以万维网搜索引擎作为基础的（混合型），因为当时的语义网数据并不丰富，只有少量的对网页进行标注的语义数据。因此，当时的工作主要是利用语义数据进行查询扩展等，以提高万维网搜索引擎的效果。随着近几年语义网的快速发展，语义网数据大量涌现，直接对语义网数据进行关键词查询的工作（非混合型）也在不断增多。在非混合型工作中，早期比较常见的是 K-Q-A 查询方法，从利用模板的方法到近期利用图算法的方法。而 K-A 查询方法的系统是最近两年出现的，通常只针对点标签的有向或无向图。这两种方式的利弊已经在第 3 节详细分析过，它们的共同特点是近期的工作都采用图论的算法（如广度优先搜索等）实现候选结果或候选查询图的构造，系统效率主要取决于索引和算法的选择。

表 1 关键词查询系统概述

Keyword search system	Query means	Index	Ranking function	Formal query language
TAP (2003)	Hybrid	Keyword index	IR style	——
SeamSearch (2006)	K-Q-A	Keyword index	no	SeRQL
SPARK (2007)	K-Q-A	Keyword index	Probabilistic ranking model	SPARQL
OntoLook (2007)	Hybrid	Keyword index	no	——
BLINKS (2007)	K-A	Keyword index & path index	Path length	——
EASE (2008)	K-A	Keyword index & r-radius graph index	Path length & IR style	——
Falcons (2008)	K-A	Keyword index & virtual document	IR style	——
TWRC09 (2009)	K-Q-A	Keyword index & summary graph	Path length & IR style	Conjunctive query

其次，对于索引的建立经历了从早期只建立关键词索引到后来的采用关键词索引加结构索引的方式。随着语义网数据量的不断增加，能够在大规模数据上进行关键词查询成为重要的研究问题，因此近期的工作都利用了结构索引来支持高效的查询算法。

另外，评分函数也是经历了从早期的不支持评分到利用结构紧凑性评分再到利用结构紧凑性加文本相似度评分的过程，说明关键词查询系统越来越重视查询结果的相关性评价。

最后，K-Q-A 查询方式使用的形式化查询语言也发生了从具体的查询语言（如 SeRQL，SPARQL 等）到比较通用的查询描述方式（如合取查询）的转变，因为合取查询可以较容易地转化为其他的形式化查询，为后面发起查询得到最终查询结果争取了更大的灵活性。

结束语 本文形式化地给出了语义网数据关键词查询的相关定义；从研究目标和查询方法的不同，阐述了语义网数据关键词查询问题的常用解决方案和研究进展；介绍并比较了 8 个现有工作。语义网数据关键词查询是一个新兴的研究课题，虽然有一些相关领域的研究成果可供借鉴，例如数据库领域关键词查询和 XML 数据关键词查询的一些方法，但是由于语义网数据关键词查询问题具有自身的特殊性，该领域仍然面临许多有待解决的挑战。总结起来有以下方面。

· 对关键词查询方法的改进：目前虽然已经提出了不少

1) 由于该系统的作者尚未给此系统起名，本文组合作者的姓氏开头字母及发布年代，暂作为该系统的名称。

语义网数据关键词查询的方法，但它们都有各自的局限性。混合型语义网数据关键词查询利用对标注数据的查询来提高万维网搜索引擎的效果，难点在于领域本体和标准数据难以获得，也不支持将标注数据中的节点联系起来，共同响应用户的查询。另外，对于非混合型的 K-A 查询方法来说，它的优势在于直接针对数据图构造查询结果，数据粒度细，不需要数据模式信息的支持。不足之处在于面对大规模的数据图，算法的效率有待提高。而对于 K-Q-A 查询方法来说，它的优势在于可以利用数据的模式信息生成形式化查询或利用抽象的摘要图构造查询图，算法效率高。不足之处在于抽象后的数据粒度粗，有些查询无法获得满意的结果。所以，在今后的研究中有必要提出一些新的方法来融合它们的优势，同时对现有的方法做进一步的改进。

• 对评分函数的改进：目前对候选结果的评分基本上考虑的都是结构紧凑性和文本相似度两个因素。但实际上，不同的用户侧重的评分策略会有所不同，比如普通用户可能更倾向于结构紧凑的查询结果，而试图在数据中挖掘线索的用户可能更倾向于发现一些不常见的关联。因此，如何与用户交互并进行查询精化是需要改进的地方。

• 支持多数据源的关键词查询：目前大部分的工作只支持单数据源的关键词查询问题。虽然 Falcons 可以支持多数据源的查询问题，但它目前关注的还只是单个实体的查询。而在拥有大量 RDF 数据的语义网中，可能很多情况下会出现用户的查询关键词出现在多个数据源中的多个实体上。这时不但需要将多个数据源融合起来，还需要将多个关键词实体联系起来，才能响应用户的查询，即多数据源的关键词查询问题。多数据源的关键词查询问题显然更加复杂，涉及到很多具体的问题，比如针对用户的关键词查询、如何选择数据源、选择后的数据如何进行融合、如何建立索引使得数据融合后索引能够保证不丢失信息等等。这些都是今后工作需要考虑的问题。

总之，语义网数据关键词查询是语义网研究中的一个重要问题。国际上在这方面的研究很活跃，并已有多个原型系统。目前国内也有一些相关研究，并且取得了一定的进展。可以预见，随着语义网数据的不断增多和普通用户对语义网数据查询需求的增多，将会涌现更多支持语义网数据关键词查询的相关方法和工具。

参 考 文 献

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web[J]. Scientific American, 2001, 284(5): 34-43

[2] Hayes P. RDF Semantics [S]. W3C Recommendation 10 February 2004

[3] Ding Li, Finin T. Characterizing the Semantic Web on the Web [C] // Proc of International Semantic Web Conference. LNCS 4273. Athens, GA, USA, 2006: 242-257

[4] Brickley D, Guha R V. RDF vocabulary description language 1.0: RDF schema[S]. W3C Recommendation. February 2004

[5] Patel-Schneider P F, Hayes P, Horrocks I. OWL web ontology language semantics and abstract syntax[S]. W3C Recommendation. February 2004

[6] Perez J, Arenas M, Gutierrez C. Semantics and Complexity of SPARQL[C] // Proc of International Semantic Web Conference. LNCS 4273. Athens, GA, USA, 2006: 30-43

[7] Prud'hommeaux E, Seaborne A. SPARQL Query Language for RDF[S]. W3C Candidate Recemendation 2006. Latest Version: http://www.w3.org/TR/rdf-sparql-query/

[8] Broekstra J, Kampman A. SeRQL: A Second Generation RDF query language [C] // SWAD-Europe Workshop on Semantic Web Storage and Retrieval. Amsterdam, Netherlands, 2003: 13-14

[9] Guha R, McCool R, Miller E. Semantic search[C] // Proc of the 12th International Conference on World Wide Web. Budapest, Hungary, 2003: 700-709

[10] Rocha C, Schwabe D, Aragao M. A hybrid approach for searching in the semantic Web [C] // Proc of the 13th International Conference on World Wide Web. New York, USA, 2004: 374-383

[11] Davies J, Weeks R. QuizRDF: Search Technology for the Semantic Web[C] // Proc of the 37th Hawaii International Conference on System Sciences. 2004: 1805-1812

[12] Zhang Lei, Yu Yong, Yang Yin, et al. An enhanced model for searching in semantic portals[C] // Proc of the 14th International Conference on World Wide Web. New York, 2005: 453-462

[13] Li Yu-fei, Wang Yuan, Huang Xiao-tao. A Relation-based Search Engine in Semantic Web[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2): 273-282

[14] 黄瑞，史忠植. 一种新的 Web 异构语义信息搜索方法 [J]. 计算机研究与发展，2008，45(8)：1338-1345

[15] 田萱，杜小勇，李海华. 语义查询扩展中词语-概念相关度的计算 [J]. 软件学报，2008，19(8)：2043-2053

[16] Lamberti F, Sanna A, Demartini C. A Relation-based Page Rank Algorithm for Semantic Web Search Engines[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(1): 123-136

[17] Ding Li, Finin T, Joshi A, et al. Swoogle: a search and metadata engine for the semantic Web[C] // Proc. of the Thirteenth ACM Conference on Information and Knowledge Management. Washington D. C., U. S. A, 2004: 652-659

[18] Ding Li, Pan R, Finin T, et al. Finding and ranking knowledge on the semantic Web[C] // Proc of 4th International Semantic Web Conference. LNCS 3729. Galway, Ireland, 2005: 156-170

[19] He Hao, Wang Hai-xun, Yang Jun, et al. Blinks: Ranked Keyword Searches on Graphs[C] // Proc of the ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007: 305-316

[20] Cheng Gong, Ge Wei-yi, Qu Yu-zhong. Falcons: searching and browsing entities on the semantic Web[C] // Proc. of the 17th International Conference on World Wide Web. Beijing China, 2008: 1101-1102

[21] 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

[22] Li Guo-liang, Ooi B, Feng Jian-hua, et al. EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data[C] // Proc of International Conference on Management of Data / Principles of Database Systems. Vancouver, BC, Canada, 2008: 903-914

[23] Ning Xiao-min, Jin Hai, Jia Wei-jia, et al. Practical and effective IR-style keyword search over semantic web [J]. Information Processing and Management, 2009, 45(2): 263-271

[24] Lei Yuan-gui, Uren V, Motta E. Semsearch: A Search Engine for the Semantic Web[C] // Proc of the EKAW. LNAI 4248. Podebrady, Czech Republic, 2006: 238-245

（下转第 50 页）

简单，运行稳定，并且有较高精度。通过把节点定位问题转换成约束优化问题，再运用粒子群算法进行求解的方法，压缩了搜索空间，加快了收敛速度，可较快地得到较优解。仿真实验表明，CPSO 定位算法与最小二乘法相比，在不同测距误差、不同测距半径、不同锚节点数和不同节点数的情况下，都能得到更高精度的解。这说明 CPSO 定位算法具有更强的抗误差性、更好的收敛性和更少的硬件设备投入等优点。

参 考 文 献

[1] Kim K, Lee W. MBAL: a mobile beacon-assisted localization scheme for wireless sensor networks[C]// Proc. of 16th International Conference on Computer Communications and Networks. Hawaii USA, 2007: 57-62

[2] Jose A C, Patwari N, Hero A. Distributed weighted-multidimensional scaling for node localization in sensor networks[J]. ACM Transactions on Sensor Networks, 2006, 2(1): 39-64

[3] He T, Huang C D, Blum B M, et al. Range-free localization schemes in large scale sensor networks [C]// Proc. of the 9th Annual International Conference on Mobile Computing and Networking(MobiCom). San Diego, California, USA: ACM Press, 2003: 81-95

[4] 林金朝，陈晓冰，刘海波. 基于平均跳距修正的无线传感器网络节点迭代定位算法[J]. 通信学报，2009，30(10)：107-113

[5] 孟令军，王宏涛，夏善红. WSN 节点声测距 TOA 值频域估计方法[J]. 电子与信息学报，2010，32(4)：993-997

[6] Reynet O, Jaulin L, Chabert G. Robust TDOA passive location using interval analysis and contractor programming[C]// Accepted to Radar 2009. 2009

[7] Niculescu D, Nath B D. Positioning in ad hoc networks [J]. Journal of Telecommunication System, 2003, 22: 267-280

[8] Dragos N, Badri N. Ad hoc positioning system using AOA[C]// Proc. of the IEEE INFOCOM 2003. San Francisco, 2003

[9] 邱岩，赵冲冲，戴桂兰，等. 无线传感器网络节点定位技术研究[J]. 计算机科学，2008，35(5)：47-50，63

[10] Parsopoulos K E, Vrahatis M N. Particle swarm optimization method for constrained optimization problems[C]// Proc. of the Euro-International Symposium on Computational Intelligence 2002. 2002

[11] Laskari E C, Parsopoulos K E, Vrahatis M N. Particle swarm optimization for integer programming[C]// Proc. of the IEEE Congress on Evolutionary Computation(CEC 2002). Honolulu, Hawaii, USA, 2002: 1582-1587

[12] Shi Y, Eberhart R C. Particle swarm optimization: developments, applications and resources[C]// Proc. of Congress on Evolutionary Computation 2001. Piscataway, NJ: IEEE Press, 2002: 86-86

[13] 杨轻云，孙吉贵，张居阳. 最大度二元约束满足问题粒子群算法[J]. 计算机研究与发展，2006，43(3)，436-441

[14] 胡一波，王宇平. 解约束优化问题的一种新的罚函数模型[J]. 计算机科学，2009，36(7)，240-243

[15] 王建刚，王福豹，段渭军. 加权最小二乘估计在无线传感器网络定位中的应用[J]. 计算机应用研究，2006(9)：41-43

(上接第 23 页)

[25] Zhou Qi, Wang Chong, Xiong Miao, et al. SPARK: Adapting Keyword Query to Semantic Search[C]// Proc of the ISWC. LNCS 4825. Busan, Korea, 2007: 649-707

[26] Tran T, Cimiano P, Rudolph S, et al. Ontology-based Interpretation of Keywords for Semantic Search[C]// Proc of the ISWC. LNCS 4825. Busan, Korea, 2007: 523-536

[27] Wang Hao-fen, Zhang Kang, Liu Qiao-ling, et al. Q2Semantic: A Lightweight Keyword Interface to Semantic Search[C]// Proc of 5th European Semantic Web Conference. LNCS 5021. Tenerife, Spain, 2008: 584-598

[28] Uren V, Lei Yuan-gui, Motta E. SemSearch: Refining semantic search[C]// Proc of 5th European Semantic Web Conference. LNCS 5021. Temerife Spatin, 2008: 874-878

[29] Tran T, Wang Hao-fen, Rudolph S, et al. Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped(RDF) Data[C]// Proc of IEEE International Conference on Data Engineering. Shanghai, China, 2009: 405-416

[30] Hristidis V, Gravano L, Papakonstantinou Y. Efficient IR-style Keyword Search over Relational Databases[C]// Proc of the VLDB. Berlin, Germany, 2003: 850-861

[31] Bhalotia G, Nakhe C, Hulgeri A, et al. Keyword searching and browsing in databases using BANKS[C]// Proc of the ICDE. San Jose, CA, 2002: 431-440

[32] Xu Yu, Papakonstantinou Y. Efficient keyword search for smallest LCAs in XML databases[C]// Proc of the ACM SIGMOD International Conference on Management of Data. Baltimore, 2005: 537-538

[33] Guo Lin, Shao Feng, Botev C, et al. XRANK: Ranked keyword search over XML documents[C]// Proc of the ACM SIGMOD International Conference on Management of Data. San Diego, 2003: 16-27

[34] 孔令波，唐世渭，杨冬青，等. XML 数据的查询技术[J]. 软件学报，2007，18(6)：1400-1418

[35] Anyanwu K, Sheth A. ρ-Queries: enabling querying for Semantic Associations on the Semantic Web[C]// Proc. of International Conference on World Wide Web. Budapest, Hungary. 2003: 690-699

[36] Anyanwu K, Maduko A, Sheth A. SemRank: Ranking Complex Relationship Search Results on the Semantic Web[C]// Proc of International Conference on World Wide Web. Chiba, 2005: 117-127

[37] Aquin M, Baldassarre C, Gridinoc L, et al. Characterizing knowledge on the semantic web with watson[C]// Proc. of the 5th International Workshop on Evaluation of Ontologies and Ontology-based Tools. Busan, Korea, 2007: 1-10

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

[39] Qu Yu-zhong, Hu Wei, Cheng Gong. Constructing virtual documents for ontology matching[C]// Proc. of International Conference on World Wide Web. Edinburgh, Scotland, 2006: 23-31

[40] Ding Li, Finin T, Joshi A, et al. Search on the Semantic Web[J]. Computer, 2005, 38(10): 62-69

[41] Uren V, Lei Yuan-gui, Lopez V, et al. The usability of semantic search tools: a review[J]. The Knowledge Engineering Review, 2007, 22: 361-377

[42] Priebe T, Schlager C, Pernul G. A Search Engine for RDF Metadata[C]// Proc of 15th International Workshop Database and Expert Systems Applications. Zaragoza, Spain, 2004: 168-172

[43] Wu Gang, Tang Jie, Li Juan-zi, et al. Fine-grained semantic web retrieval[J]. Journal of Tsinghua University: Science and Technology, 2005, 45(1): 1865-1872