**一、概述**

本课题组针对以上几个方面，在近年的**语义关联挖掘**[Zhang2012, Zhao2012, Wang2013b, Zhang2016a, Zhang2016c]、**语义推理**[Ma2011, Ma2013, Ma2014, Zhou2015, Zhou2016a, Zhou 2016b, Jabbour2016]、**语义数据摘要**[Zhang2007, Zhang2009, Jiang2012]、**大规模语义查询** [Jin2014, Jin2015, Jin2017]、**大规模链接数据挖掘**[Jiang2013, Wu2014a, Wu2014b, Wu2016a, Wu2016b]等工作的基础上，拟对知识图谱中的语义关联挖掘方法展开研究。本研究具有如下特点：

1. 研究**基于链接模式的挖掘方法。研究海量语义关联的挖掘方法，提出基于频繁子图的关联模式挖掘方法，并以此实现高效地语义关联挖掘算法。该方法充分利用前述语义关联检索模型带来的表达能力，提高了语义关联挖掘的目的性和效率。**
2. 研究语义关联检索模型，将语义关联背后的概念框架抽象为关联模式，将语义关联表示为关联模式和实例信息的综合，并以此为基础进一步提出语义关联检索模型。课题提出的模型将解决传统基于有向路径表示方法的局限性，能够刻画多对象之间的复杂语义关联，为语义关联检索打下坚实的理论基础。
3. 研究链接数据的时空模型，提出时空模型和建模方法。该模型能够刻画链接数据中对象以及对象间关系的时空信息，并且能够适用于语义关联的时空挖掘。该模型解决了现有RDF时空模型的局限性，将为本课题的研究提供合理的数据模型。
4. 研究海量语义关联的索引方法。充分利用关联模式为语义关联带来的灵活性，提出基于关联模式的差异性索引，该方法将不同语义关联之间的共性和特性分别进行索引，不仅能够降低索引时间开销和空间冗余、还将提高索引查询效率，为实时检索提供技术的保证。
5. 研究语义关联理解问题，实现语义关联摘要方法，将语义关联的复杂结构摘要为用户可快速理解的自然语言片段，帮助用户理解检索结果。 本研究能够在摘要中体现不同语义关联之间的相  
   关性和区别，对语义关联的快速准确的理解将起到极大的作用，将提高海量语义关联检索的易用性。
6. 研究基于描述逻辑的时空推理规则，提出在推理过程中引入外部知识库的方法。该方法能够在挖掘过程中发现语义关联间存在的隐含时空关系，同时能够解决挖掘过程中的时空约束满足问题；研究基于多维切片的语义关联时空挖掘方法。该方法能够有效地降低大规模链接数据挖掘的复杂度，提高挖掘效率。
7. 研究语义关联的时空演化，提出语义关联在时间维度的变点检测方法、周期性演变分析方法和空间差异分析方法。语义关联时空演化的研究将帮助用户深入理解语义关联的动态发展过程。
8. 研究大规模链接数据中语义关联的查询方法，提出语义关联在时空维度的索引方法，提出查询语言和对查询结果的摘要方法。该方法能够帮助用户在海量挖掘结果中提高查询效率，并且提高了挖掘结果对于用户的可理解性。

**二、国内外研究现状**

在语义网各研究领域和交叉学科领域中，语义关联一直受到极大的关注。语义关联的相关理论和方法已在国防、化工、生物医学、交通等领域得到了广泛的应用。例如Wang在文献[Wang2013]中提出利用语义关联的挖掘帮助复杂药物的研制； Ji在文献[Ji2015]中介绍了生物医学无结构化数据中如何通过语义关联的挖掘找到实体间隐含的关系。Sheth在社会关系网中识别语义关联用于国土安全的研究[Sheth2005]。但是传统语义关联挖掘没有处理时空维度的能力，忽略了链接数据中对象本质上具有的时空属性，这不仅浪费了链接数据丰富的时空信息，也限制了用户对链接数据的洞察能力。本课题的研究将为复杂网络分析、时空数据挖掘的研究提供宝贵的经验。课题的完成将极大的加强现有语义关联挖掘的能力，加深用户对链接数据的理解，为语义网的可持续发展提供有力的技术保障。在更深远的意义上，将促进人类的知识的分享和合理利用。综上可见，本课题的研究具有重要的理论价值和广泛的应用前景。

本部分将对课题所涉及的关键问题进行国内外发展现状的综述和分析。主要包括关系数据模型和图模型中的关联模式和挖掘方法、知识图谱中的语义关联模型和挖掘方法、时空关联挖掘等等。

**（一） 关系数据和图模型中的关联模型**

项集之间的逻辑蕴含关系被称为关联规则。关联规则揭示了数据之间的隐含相关性，但这种相关性通常不具备语义。Han针对经典Apriori算法[Agrawal1994]的固有缺陷，采用分治法，在不产生候选频繁模式的情况下寻找关联规则[Han2000]。文献[Hipp2000]对早期的关联规则挖掘方法进行了综述。近年来，关联规则挖掘的研究中开始越来越多的出现对时空维度关联规则的挖掘，例如文献[Lee2009b，Hai2011，Shaheen2013]。这也促使了本课题组思考语义关联的时空维度问题。此外，在图挖掘领域中，关联通常指的图中两个节点之间存在的路径。关联发现问题被建模为图搜索问题。A\*算法是图搜索中一个经典算法框架[Hart1968]。近年来，随着无标度网络的不断发现，针对无标度网络中的关联搜索也展开了相应的研究。Barabási对无标度网络中的关联搜索进行了综述[Barabási2009,Barabási2013]。图模型中的关联挖掘方法对生物医学、化学、交通、互联网、移动网络等领域产生了巨大的影响。例如，剑桥大学的研究[Bullmore2009]通过fMRI影像将人脑的结构和功能区域建模为复杂网络，并通过该网络的无尺度特性发现人脑功能区域间的关联。文献[Vespignani2012]中的研究涉及了社会技术系统中动态过程间形成的复杂关联。综上所述，关系数据模型和图模型中的关联模型和挖掘方法已经得到了充分的研究，其研究成果不胜枚举。这些研究不仅激发了面向链接数据的语义关联研究，同时也为语义关联的挖掘方法提供了巨大的参考价值。

### （二） 基于内存的子图挖掘方法

频繁子图挖掘可以在图集合中发现非常基本的模式，在大型图数据库中可以用它建立图索引并进行相似性搜索，区分不同的图组群，对图进行分类和聚类分析。目前图论中频繁子图挖掘算法主要有四种：

1. 基于贪心策略的算法：代表算法有SUBDUE和GBI等。
2. 基于归纳逻辑程序的算法：代表算法有WARMR。
3. 基于Apriori的算法:代表算法有AGM，FSG，AcGM等。
4. 基于模式增长的算法:代表算法有gSpan，FFSM，MoFa和Gaston等，这些算法能挖掘出所有的频繁子图，另一些算法是挖掘有限制的子图如CloseGrpah在图集上挖掘闭的频繁子图，MARGIN和SPIN挖掘最大频繁子图，CloseCut和Splat在一组关联图集上挖掘精确的信息密度大的频繁子图，TSMiner是在图集上挖掘大图等。

基于贪心的算法有SUBDUE和GBI。SUBDUE是一种基于最小描述长度和知识背景的不精确算法，它采用最小描述长度作为度量，并根据背景知识对最小描述长度的计算方法进行修改，采用一种启发式的束（beam）搜索方法来缩小搜索空间，提高了算法效率。特别适用于那些不要求精确计算支持度的领域，如社会人际关系网络。之后SUBDUE还扩展为图分类算法SubdueCL，SubdueCL不再采用最小描述长度，而是采用基于子图置信度的启发式方法。GBI类似SUBDUE，用一个顶点代替每个发现的子图来不断的压缩图，以得到具有最小规模的图。它可以处理有向的或无向的带环图，在文献[29]中GBI还作为特征构造器，用于构造图数据中的决策树分类器的特征。

基于归纳逻辑程序的算法：该算法的目标是归纳出一个可以正确的分类正样本集和负样本集的规则集。该类算法大多是贪心的，采用不同的启发式方法来剪枝查找空间。它们趋向于找出具有高支持度的子图，并可作为较好的识别器，但是并不能保证能够发现所有的频繁图。虽然WARMR可以发现所有的频繁子图，但是计算复杂度较高。

Apriori算法只能采用广度优先搜索，通过边或顶点的的逐层迭代来产生候选子图的。 代表算法AGM和FSG都采用邻接矩阵来表示图，不同的是AGM对图的顶点进行逐层构造来获得候选子图，最终求出的导出子图不一定连通，FSG则以边为迭代对象，挖出的频繁子图是连通的。由于FSG加强了候选子图的剪枝，算法效率比AGM要高。这类算法都是由两个大小为k的频繁子图连接产生大小为k+1的候选子图。该算法的缺点：如果k很大时，需要很长时间并且会产生许多冗余的k+1候选子图；在判断候选子图是否频繁时需要重复扫描数据库，这样会占用大量的内存空间和处理时间，降低了算法的整体效率。

基于模式增长的算法既可以采用深度优先搜索又可以采用广度优先搜索，这类算法都是通过逐步扩展边来生成候选子图的，该算法效率比Apriori算法效率高，系统开销小。由于基于模式增长的算法具有如此的优点，下面介绍一些基于模式增长的经典算法。

**1） gSpan算法**

gSpan算法是最早将模式增长思想用于频繁子图挖掘中的，该算法是在无向带标签的图集上挖掘所有频繁的子图。它主要采用了两种技术：DFS词典顺序（DFS Lexicographic Order）和最小DFS编码（Minimum DFS Code）。当对图进行深度优先遍历时可以得到几种不同的DFS树，相应的也会得到几种不同的DFS编码。根据DFS词典顺序可以从中找出最小的DFS编码，并把最小DFS编码作为图唯一的规范表示。该算法采用最右路径扩展，即新的顶点只能加在最右路径上，新的边只能加在最右顶点与最右路径上其它顶点之间，前者形成的边称为前向边，后者形成的边称为后向边。先进行后向边的扩展再进行前向边的扩展，边扩展边进行子图同构测试，这样既保证了挖掘结果的完整性又避免大量候选子图和冗余子图的产生。gSpan并不能避免子图同构测试，但是它为每个图保存了一个子图同构列表这样可以减少子图同构测试的次数，提高挖掘效率。在计算候选子图支持度的时候采用了只扫描该子图的超图进行扫描，而不是对整个图集进行扫描，这样避免了大量重复扫描图集的问题。gSpan算法中边扩展的时间复杂度是O(2n),子图同构测试的时间复杂度是O(2n)，总的时间复杂度达到O(2n \*2n)。gSpan具体算法如下：

**算法1 GraphSet\_projection(D,S)**

输入：图集D.

输出：频繁子图S.

1:在图集D上根据标签支持度的大小重新排列标签；

2:除去不频繁的顶点和标签；

3:对剩余的顶点和边赋予新的标签；

4:S1图集D中所有频繁的一条边；

5:按DFS词典顺序重新排列结合S1；

6:S S1；

7:对集合S1中的每条边e做：

8:用e初始化s, s.D是包含边e的图的集合；

9:Subgraph\_Mining(D,S,s)；

10:DD-e；

11:如果|D|<minsup；

12:break；

Subgraph\_Mining(D,S,s)

输入：图集D,频繁图s.

输出：频繁子图集合S

1:如果s不是最小DFS编码；

2:return；

3:否则把s加入到S中；

4:枚举包含s的图并记录s的孩子；

5:对每个s的孩子c做：

6:如果c的支持度>=minsup

7：sc；

8:Subgraph\_Mining(Ds,S,s)；

**2）CloseGraph算法**

gSpan算法是在图集上挖掘所有的频繁子图，由于子图数目呈指数增长，这给数据分析造成困难。为解决该问题X.Yan等人提出另一种挖掘算法CloseGraph。它是在gSpan的基础上改进能够在图集上挖掘无向带标签的闭频繁图。CloseGraph同gSpan的扩展方式和编码方式相同，不同的是CloseGraph还采用了另外两种技术：同等发生(Equivalent Occurrence)和提前终止(Early Termination)。利用这两种技术可以提前终止一些子图的生成与扩展，提高了挖掘效率。一个图g能够通过增加一条新边e来进行扩展，表示新生成的图。边e可能引进一个顶点到g中，也可能不引进顶点。如果g引进了一个新顶点，我们用来表示新生成的图。若没有引进新顶点则用来表示，表示在图集D上g出现的次数，表示在图集上可扩展为的次数。给定图和图集，如果那么就称g与同等发生。这就意味着在图集D中g出现的地方也会同时出现，这就是同等发生；另一种是提前终止，给定图集和两个同等发生的图g与，对任意图，，，如果（其中边e是g扩展为时所用的边），这时就可提前终止g的扩展只扩展即可。由于gSpan是按照严格的DFS词典顺序来挖掘频繁图的，直接用上面提前终止的条件会导致挖掘结果不完整。还需加另一条件即g的最小DFS编码要大于的最小DFS编码，这样才能保证挖掘结果的正确性。实验表明CloseGraph算法要优于gSpan算法，但边扩展和子图同构测试的时间复杂度仍是O(2n)，所以整体时间复杂度是O(2n \*2n)。

CloseGraph主要有三个递归步骤：（1）它产生一个频繁图；（2）判断该频繁图是不是闭的；（3）判断是否满足提前终止的条件来决定该频繁图是否需要被扩展。具体算法如下：

**算法2 CloseMining(D,min\_sup,S)**

输入：图集D和最小支持度min\_sup.

输出：闭频繁图集合S.

1:除去不频繁的顶点和边；

2:S0单顶点频繁图的编码；

3:S S0；

4:对S0中的每个编码s做

5:CloseGraph(s,NULL,D,min\_sup,S)；

CloseGraph(s,p,D,min\_sup,S)

输入：DFS编码s,它的父亲p,图集D和min\_sup.

输出：闭频繁图集合S.

1:如果s不是最小DFS编码则返回

2:如果并且且

不是提前终止失败的情况。则不需要扩展s，返回。

3:初始化C为空集合；

4:扫描图集D，扩展s得频繁图插入C中;

5:对s探测提前终止的情况；

6:如果在C中不存在使得support(s)= support()

则说明s是闭的。把s插入S中；

7:从C中删除不是最右扩展的到的频繁图；

8:对C按照DFS词典顺序进行排序；

9:对C中每个做

10:CloseGraph(,s,D,min\_sup,S)；

11:返回；

3）FFSM算法

FFSM（Fast Frequent Subgraph Mining）可用于有向图的挖掘，在图集上挖掘所有的频繁子图。该算法采用下三角邻接矩阵来表示图，对角线的位置存放顶点的标签，下三角的位置存放边的标签，如果对应位置不存在边则用0来表示，按照从左到右从上到下的顺序遍历该下三角邻接矩阵可得到该图的编码。一个含有n个顶点的图可能有n!个下三角邻接矩阵表示形式，相应的会得到n!中不同的编码，按照严格的词典顺序可以从中找出最大的编码，并把最大的编码作为该图唯一规范的表示。最大编码对应的下三角邻接矩阵称为规范邻接矩阵CAM(Canonical Adjacency Matrix)。该算法提出了两个高效的候选子图产生的操作FFSM-Join和FFSM-Extension，如果所有图都用CAM来表示，那么通过 FFSM-Join和FFSM-Extension操作产生的候选子图将是不完全的，于是该算法引入次最优CAM。在挖掘过程中会产生次最优CAM树，它可以保证所有频繁子图可以毫无歧义的枚举出来。该算法为每个频繁的子图保存一个嵌入集合（embeding set），该集合中存放的是频繁子图的顶点。在子图同构时把图的操作转化为集合的操作，提高了子图同构测试的效率。该算法是将结构图转化为标准邻接矩阵，把图的扩展转化为CAM的连接和扩展的操作。由于矩阵的操作比一般的序列操作要复杂的多，所以该算法的编码方式不适合大图的挖掘。该算法中子图同构测试的复杂度是O(m2)，边扩展的复杂度是O(m\*n\*2n)，总体复杂度是O(m4\*n\*2n)。其中m表示集合中顶点的数目，n表示频繁图边数。因为具有m个节点的完全图包含m(m-1)/2条边，所以总体复杂度可转化为O(n3\*2n)。实验表明该算法要优于gSpan算法。

FFSM具体算法如下：

**算法3 FFSM(D,min\_sup,S)**

输入:图集D,最小支持度min\_sup.

输出:频繁图集合S.

1:S频繁顶点和边的CAM；

2:P频繁边的CAM；

3:FFSM-Explore(P,S)；

FFSM-Explore(U,W)

输入:次最优CAM列表U,频繁子图的CAM集合W.

输出:到目前为止包含所有频繁子图CAM的集合W.

1:For U中每个元素X做；

2: If X是CAM；

3: WW{X}，初始C为空；

4: For U中每个元素Y做；

5: CCFFSM-Join(X,Y)；

6: End For；

7: CCFFSM-Extension(X)；

8: 从C中除去既不频繁又不是次最优的CAM；

9: FFSM-Explore(C,W)；

10: End If

11:End For

**4） SPIN算法**

SPIN是在FFSM的基础上改进的，能够在无向带标签的图集上挖掘最大频繁图。主要策略是：先通过FFSM算法在图集上挖掘出所有的频繁树，然后在这些频繁树上构建频繁图。它把树挖掘和图挖掘统一到一个过程中。在SPIN算法中挖掘的树是无向带根节点的有序树，按照广度优先遍历该树会形成一个字符序列即该树的编码。一个树可以有几种不同的编码，根据词典顺序把编码最大的树作为该树唯一的规范表示。若一个树的编码是最大的并且包含图的所有顶点，那么就称该树是该图的规范生成树（Canonical Spanning Tree）。算法中主要采用两种技术来支持这种生成最大频繁图的策略，一是图的划分，把含有相同规范生成树的图划分为一等价类（Equivalence Class）中；另一是剪枝技术。下面分别介绍这两种技术。

图的划分分为两步：首先在图集上用FFSM算法挖掘出所有的频繁树，之后对于每个频繁树T，找出规范生成树与T同构的所有频繁图，这些频繁图就被分到同一个等价类中。

剪枝技术主要有三种：一种是从底向上剪枝（Bottom-Up Pruning）, 它采用动态排序（Dynamic Reordering）方法，首先剪枝掉一些不频繁的候选边来缩减查找空间，之后根据候选边的支持度按从小到大重新排序。留下的频繁候选边经过连接后仍可能是频繁的；另一种是尾部收缩（Tail Thrink）,如果一个树T有一个相关边（associative edge）集合S（所谓相关边就是在图集上T出现的地方边e也同时出现，其中边e不是T中的边，那么就称e是T的相关边）,那么就可以把S中的元素从T的候选边的集合中剪掉，不需要再对这些边一步步扩展，因为树T的超图也肯定会包含这些相关边。并且当这个相关边集合是致命的（表示树T加上这些相关边后形成图的规范生成树不再与T同构）那么就可以剪枝掉整个等价类。这样可以缩小查找空间。最后一个剪枝技术是外部边剪枝（Extern-Edge Pruning），这种剪枝技术在不知道候选边的情况下可以剪枝掉一个等价类。当对一个图引进一个新顶点时，该图与新引进的顶点之间形成的边就是外部边。如果一个图存在一个频繁的外部边，扩展该外部边后会形成一个新图，因为新图的规范生成树不再与原图的生成树同构，所以就可以把原图所在的等价类剪枝掉。这样同样可大大缩小查找空间。

实验表明SPIN算法要优于gSpan和FFSM算法。虽然SPIN算法采用了三种剪枝技术提高了挖掘效率但是在生成频繁树时采用FFSM算法，但是整体的时间复杂度还是O(n3\*2n)。

**5）频繁子图挖掘算法分析与比较**

除了gSpan和CloseGraph外，基于模式增长的算法还有FFSM以及SPIN算法。FFSM算法可用于有向图的挖掘，在图集上挖掘所有的频繁子图。SPIN是在FFSM的基础上改进的，能够在无向带标签的图集上挖掘最大频繁图。SPIN算法中采用了剪枝技术，时间效率要优于FFSM，但是整体时间复杂度两者都是O(n3\*2n)。其中子图同构测试的复杂度是O(m2)，边扩展的复杂度是O(m\*n\*2n)，总体复杂度是O(m4\*n\*2n)。m表示集合中顶点的数目，n表示频繁图边数。因为具有m个节点的完全图包含m(m-1)/2条边，所以总体复杂度可转化为O(n3\*2n)。根据以上对基于模式增长的挖掘算法的分析与讨论，表1列出了这些算法的特性。

表1：基于模式增长的挖掘算法的分析与比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 算法 | 最大频繁子图 | 图的编码 | 边扩展的复杂度 | 子图同构测试复杂度 | 总复杂度 | 搜索方法 | 算法特点 |
| gSpan | 否 | 最小DFS编码 | O(2n) | O(2n) | O(2n \*2n) | 深度优先 | 挖出所有频繁子图，不适合大型图集，频繁子图是连通的。 |
| CloseGraph | 是 | 最小DFS编码 | O(2n) | O(2n) | O(2n \*2n) | 深度优先 | 挖出闭的频繁子图，适合大型图集，频繁子图是连通的。 |
| FFSM | 否 | 邻接矩阵 | O(m\*n\*2n) | O(m2) | O(n3\*2n) | 广度优先 | 挖出所有频繁子图，不适合大型图集，频繁子图是连通的。 |
| SPIN | 是 | 邻接矩阵 | O(m\*n\*2n) | O(m2) | O(n3\*2n) | 广度优先 | 挖出最大频繁子图，适合大型图集，频繁子图是连通的。 |

### （三） 基于外存的子图挖掘方法

目前存在的挖掘算法大部分是基于内存的，而不是基于I/O的。这些算法主要采取有效的启发式方法来剪枝查找空间，挖掘频繁模式。很少有算法考虑到在大规模数据集上进行挖掘。之前的挖掘算法的研究已经取得了显著的进步，基于磁盘的图集挖掘仍然是个挑战。

在基于内存的挖掘算法中，设计的数据结构都是放进内存中的，如邻接链表，邻接矩阵等，这样在设计算法时这些结构是可以随机访问的。当邻接链表或邻接矩阵不能放进内存时，随机访问就会变的非常困难。对于基于外存的挖掘方法，要想能够随机访问图集，花费的代价将非常高。问题的瓶颈是随机访问图集，自然而然解决问题的方法就是对图集进行索引，设计高效的索引结构能够高效灵活构建索引，并且能够快速访问数据。

基于这些启发，ADIMINE[2]算法提出了ADI(Adjacency Index)，一种高效的图集索引结构，它使得大规模数据能够被随机访问，并且能够使挖掘过程中某些操作快速有效执行。论文中提出了ADI-Mine算法，它能够利用ADI结构进行挖掘，实验表明该算法效率优于gSpan，并且能够在较大数据集上进行挖掘。

由于该方法是基于磁盘的，在建立ADI索引时需要消耗很多时间。在动态数据集上该方法存在一定的局限性，数据的每一次更新都需要重新建立ADI，这样将花费很多时间，效率不高。但是在静态数据集上该方法具有一定的优势。

### （四） 基于划分的子图挖掘方法

目前存在的子图挖掘算法主要是基于内存的，这些算法假设数据集是相对静止的并且能够载入内存。但是当数据非常大不能放入内存时，就无法在这些数据集上进行子图挖掘。这种情况下可以采用划分合并的策略，分而治之。

基于划分的方法能够使挖掘过程更灵活，例如关联规则的挖掘[35]，数据分类[36,37]，数据聚类[38,39]等。文献[40]率先研究了划分图的方法并开发了METIS软件包。METIS算法主要基于多层次图划分方法并尽量减少顶点和边的断裂，把数据集划分成较小的数据集，但是该论文只是图的划分并未提到数据挖掘及结果合并的问题。基于划分的子图挖掘方法包括三个步骤：首先把数据集拆成小的数据集；其次从小的数据集中进行挖掘；最后合并挖掘结果，恢复出全局正确的子图。

**1）PartMiner**

论文[3]提出了一种基于划分的策略来进行频繁子图的挖掘算法PartMiner。该算法主要由三个阶段组成，在第一个阶段递归调用图划分算法把图划分成较小的子图，然后把这些子图分配到各个单元中；第二个阶段运用已存在的基于内存的挖掘方法在每个单元中挖掘频繁子图；最后一个阶段采用merge-join的方法找出数据集中全局的频繁子图。该算法同时能够处理动态数据，把变化的数据集孤立为一个单独的数据子集，当数据变化时，只是在变化的数据集上重新执行挖掘算法。实验表明该方法要优于ADIMINE算法。

在划分的过程中，为减少单元间的连通度，需要按照一定的规则把子图分配到各个单元中。如果我们随机划分数据集，无规则地分配到各个单元，这样会导致单元间连通性不清晰，当合并挖掘结果时，merge-join算法需要在每两个单元中进行一次执行，这样会导致算法效率的降低。PartMiner算法在划分的过程中采用两个标准，一是减少子图间的连通度，二是孤立频繁更新的顶点及边到单独的子图中。这样在合并挖掘结果时只需在相邻的单元中进行合并，只需在更新的数据中进行重新的挖掘。这样可以提高挖掘的效率。

划分之后，我们可以在每个单元上进行数据挖掘。假设数据集被划分成两个单元和，表示在上挖掘的频繁子图，表示在上挖掘的频繁子图，我们希望能够恢复出原数据集上的频繁子图。首先根据频繁子图边个数的多少对每个单元中的频繁子图进行排序，用表示在单元中k条边的频繁子图。接着考虑在单元和中一条边的模式，由于一条边的模式没有共同的连接的边，所以得到。之后合并两条边的子图，基于相同的连接边合并两条边的模式得到三条边的候选模式，然后去掉其中不频繁的子图和同构图最终得到我们想要的三条边的频繁模式。当k>2时，，然后通过merge-join算法来获取k+1条边的候选子图，合并方法具有三个步骤：1）合并和来获得；2) 合并和来获得；3）合并和来获得。最终得到，除去其中不频繁的图和同构图最终可得到我们想要的k+1条边的频繁子图。

由于该方法在挖掘的过程中把支持度降为（k是数据集划分单元个数）所以会把不频繁的子图挖掘出来，挖掘的结果不正确，该方法不能直接应用到语义关联的挖掘中。

**2）PartGraphMining**

PartGraphMining[4]在划分时采用递增的聚类算法思想，主要是基于传统的K-means算法，只需要遍历一遍数据集，就可以把数据集划分成多个片段。每个片段的大小也就是片段中图的数目要保持平衡，这样在挖掘过程中可以保持参数的统一。为优化挖掘结果，提高挖掘效率，提出了质心(Centeroid)和相似度(Similarity)的概念。

**质心（Centeroid）：**片段的质心表示为，其中表示在中一条边的模式。在中每个元素都有相应的权重即，其中表示在片段中出现的次数。

**相似度（Similarity）:**图与片段质心间的相似度可以表示为SimGraph()。，若则SimGraph()=0;若，其相应权重为则SimGraph()=。

划分算法主要有三步：1) 把前k个图依次放入k个片段中，并初始化这些片段的质心；2) 考虑接下来k个图，把这k个图按一定规则放入不同的片段中，规则如下：(i) 保证图与合适片段间有最小的相似度；(ii) 保持片段大小的平衡；3) 重复第二步知道整个图集被划分完毕。

划分完毕后会得到k个大小相似的片段，在每个片段上采用基于内存的挖掘算法Gaston或者gSpan进行挖掘，支持度大小设为。我们可以得到每个片段中局部频繁子图，其中表示在片段上挖掘出的频繁子图。计算所有的并集；计算所有的交集，一定是数据集D上的频繁子图。重新遍历一遍数据集，判断候选子图是否频繁，最终输出全局频繁的子图。

当单个图规模较大无法放进内存时，该方法仍不能解决问题，而语义网中单个类型对象图可能包含亿个三元组无法直接放进内存，所以无法用该方法解决问题。我们仍需继续探索新的划分方法。

**（五） 知识图谱中的语义关联模型和挖掘方法**

在语义关联研究的早期，语义关联被定义为两个对象在RDF图中直接或间接的关系[Aleman-Meza2003]。对语义关联的建模通常沿用图论中有向路径的方式。Anyanwu在2003年形式化了属性序列的概念以及其上的操作符[Anyanwu2003]，并在此基础上定义了对象间的四类语义关联。路径模型虽然仅能刻画有限的语义关联，但其直观性和算法的简洁性影响了之后的大量研究工作。例如，文献[Zhou2011]中提出的Conkar系统将语义关联的路径模型进一步定义为对象之间的无环路径，与此类似的工作有RelFinder 路径模型[Heim2009, Heim2010]、基于learning-to-rank的路径模型[Chen2012]以及Explass路径模型[Cheng2014]。这些模型在路径的长度、关系类型、方向性等限制条件上存在差异。在文献[Fang2011]中提出的REX系统中，语义关联是两个对象间的、对特定受限图模式的实例化。虽然REX的关联模型涉及比路径更为复杂的子图结构，但其本质是由路径合并得到的，依然属于路径模型。类似的模型还有RECAP中提到组合路径模型[Pirrò2015]等等。与传统路径模型不同的是子图模型，后者通常被定义为多个对象之间存在的子图结构。文献[De2006]中将语义关联定义为覆盖中心节点和附近特定节点的特殊子图。而文献[Chen2011]也提出了类似的模型。在文献[Kasneci2009]提出的MING系统中，语义关联被建模为斯坦纳树问题，特定的子图被抽取后用于解释一组特定对象间的关系。该模型与本课题模型的最大的区别在于：前者是查询相关的，在获得语义关联前需要首先查询并确定一组特定的对象，而本课题中的语义关联是查询无关的，是对整个链接数据的建模。在最近的研究成果中，有两个研究工作与本文中的模型类似。文献[Yang2014]和[Cheng2016]分别提出了语义关联的树结构模型和子图结构模型，但这两个工作都是将多对象间的语义关联问题转化为对象两两之间的路径问题，所以本质上属于路径模型。本课题组在语义关联建模和挖掘方面已取得了丰富的研究成果，于2012年提出了语义关联的图模型[Zhang2012]，突破了以往基于路径模型的局限性，并且在[Zhang2012b]中进一步提出借鉴图挖掘中的研究成果，通过频繁子图挖掘获得链接数据中具有典型性的语义关联。申请人提出了语义关联挖掘的优化方法[Zhao2012, Zhang2016c]，针对语义关联挖掘任务，进一步提出了链接数据的图压缩方法和聚类方法以提高挖掘效率[Jiang2013, Zhang2016b]；在语义关联挖掘结果的搜索和排序方法上，提出了基于虚拟文档的搜索模型[Wang2013b]和多目标语义关联搜索模型[Zhang2016a]。基于图模型的语义关联挖掘已经逐渐成熟，是本课题的理论基础。语义关联挖掘的研究目前已广泛应用于生物制药[Chen2009, Wild2012]、生物医学[Nebot2012, Liu2013]、基因分析[Makita2013]、国土安全[Sheth2005]、情报学[Yuan2013]等领域。

**（六） 传统领域中的时空模型和时空推理**

在地理信息系统和时空数据库中，时空模型的研究已较为成熟。对本课题具有借鉴意义的时空模型通常可以分为以下类：(1) 时空立方体模型：该模型将空间二维结构和时间一维结构组合为立方体，其中每个截面为现实世界的一个状态；(2)序列快照模型：通过时间切片方法将连续变化的空间现象转换为离散部分，时间信息作为空间信息的时间戳。(3)基态修正模型：对序列快照模型进行了修正，通过存储相邻快照的变化来减少数据冗余；(4) 时空复合模型：将空间分隔成具有相同时空过程的最大的公共时空单元；(5) 基于事件的模型：通过事件序列和时间轴来表达一些过程的时空变化；(6) 面向对象的模型：以面向对象的基本思想组织地理时空，每个时空对象中封装了对象的时态性、空间特性和属性。以上时空模型各有优势。链接数据和GIS数据在模型上存在较大差距，因此以上模型并不能直接应用于语义关联挖掘。但这些模型其中基于事件的模型对时态RDF的提出起到了重要作用；面向对象的模型可以为本课题所参考；而序列快照模型中的时间切片法也会被本课题所借鉴。在空间推理方面，空间推理包含多种空间关系的推理，如拓扑、方位等。Randell[Randell1992]提出的RCC拓扑关系推理，用于对空间位置的拓扑连接关系进行推理；Andrew[Andrew1991]提出了基于圆锥和基于投影的方位模型，用于对空间对象的相互方位进行推理；其中基于投影的方位模型又称为主方位模型；Goyal[Goyal2001]提出了方位关系矩阵推理法，使用九宫格对方位进行推理。在时间推理方面，影响最大的工作是Allen在[Allen1983]中提出的时态关系代数，用于对13种互不相交且联合完备的时态关系进行推理。本课题将对上述时空推理模型进行彻底研究，使其能够用于语义关联挖掘中的时空推理。

**（七） 基于RDF模型的时空模型和时空推理**

受到时空数据库和时空挖掘的影响，语义网的研究者在2004年之后逐渐开始研究基于RDF模型的时空模型。最早提出的相关工作是2004年Hobbs提出的时间本体[Hobbs2004]。该工作提出了一个时间本体，使用OWL公理来描述时间点和时间区间，并且形式化了时间点和时间区间的推理规则。Gutierrez提出了时态RDF[Gutierrez2005, Gutierrez2007]，它通过为RDF图加上时间标签来描述RDF图的演化，并进行版本控制。该研究还提出了一套面向RDF的时态查询语言。针对描述逻辑的时态扩展和时间推理，在2005至2008年诞生了一系列工作[Artale2005, Welty2006, Kim2008, Milea2008]。Pugliese 在[Pugliese2008]中提出了时态RDF数据库的概念，并提出使用tGRIN结构对时态RDF数据库进行索引。该工作将对本课题的时空索引提供参考。Sheth在[Sheth2008]中描绘了一个称为EventWeb的系统，该系统能够抽取语义数据中事件的时间、空间和主题信息，并能够提供结构化查询。EventWeb系统的思路和本课题有类似之处，但其模型基于事件，无法满足语义关联的挖掘需求。Tappolet在[Tappolet2009]中提出使用命名图来刻画RDF图中的时间信息，并且将SPARQL查询语言扩展为-SPARQL，用于对RDF图中的时间信息进行查询。Perry在[Perry2008,Perry2011]中提出了SPARQL-st查询语言，能够支持对RDF的时空查询。Rula在2012年对当时的大规模链接数据进行了深入研究[Rula2012]，分析了其中时间信息的数量、分布和采用的时间模型。该工作对本课题实验数据集的选取具有重要指导意义。Batsakis在[Batsakis2011]中提出SOWL这一时空本体，该本体基于OWL2.0语法和SWRL规则语言，能够以四维流体模型或N元关系模型对时空信息进行表达和推理。SOWL实现了RCC8和基于圆锥的空间推理。著名语义网知识库YAGO2[Hoffart2013]从维基百科中抽取事件以及相应的时空信息，将事件的时空信息作为单独的维度进行刻画。YAGO2的体系结构和时空信息抽取方法将为本课题提供参考。

**（八） 时空演化问题**

在语义网的现有研究中，时空演化的研究主要集中于本体演化的研究，用于在本体工程中观察和规范本体随着时间变化的行为。本体演化的研究来源于数据库模式演化的研究，通常仅考虑演化的时间维度而忽略空间维度。Stojanovic在[Stojanovic2002]中提出了本体演化的研究问题，认为用户的需求随着时间发生变化，因此建立的本体也应当随时间发生演化。该工作提出了一个六步骤的演化过程用于帮助本体开发者更好地维护本体的动态性变化。Noy在[Noy2004]中深入讨论了本体演化与数据库模式演化的区别，并且分析了本体开发者的行为模式对本体演化造成的影响。Noy进一步在[Noy2006]中讨论了开发本体过程中的协作行为对本体演化的影响问题。Plessers在[Plessers2007]中提出了一个帮助用户理解本体演化的方法。当本体开发者不断地对本体进行修改时，该工作能够自动探测到演化的发生，并且将本体不同类型的变化定义为不同类型的行为，使得观察本体演化的用户能够理解演化背后的动机。这项工作为本课题带来了启示：时空演化的分析能够帮助用户发现并理解本体或链接数据中的动态性变化。Javed在[Javed2009]中进一步在不同层次中定义了基于模式的演化操作。Grau在[Grau2012]中形式化地定义了本体演化中的细粒度动态性，而Zablith为本体演化中的动态性做了完整的综述[Zablith2015]。随着链接数据的发展，链接数据的演化也开始得到了研究。Auer在[Auer2013]中讨论了链接数据在生命周期中的动态性问题，提出了链接数据动态性的研究对用户合理利用链接数据的重要性。Käfer2013在[Käfer2013]中详细讨论了链接数据的演化研究。通过对已有链接数据的可用性、失效率、变化率、三元组级别变化、术语级别变化等特点的长期跟踪分析，该工作能够帮助用户了解链接数据的动态性并做出合理的预测。本课题组相信，语义关联挖掘后的时空演化分析将为链接数据的合理利用提供重要的帮助。

**（九） 小结**

通过语义关联挖掘充分利用链接数据的价值已成为了语义网研究领域的主要课题之一。对国内外研究现状的分析表明：现有的语义关联建模、分析和挖掘方法忽略了链接数据中丰富的时空信息，限制了用户对链接数据的有效洞察。随着链接数据规模的不断增长，其影响已渗透入社会各行各业，用户对大规模链接数据中蕴含的丰富价值报以巨大的期望。因此，亟待提出一套语义关联时空挖掘方法，能够有效地建立起链接数据的时空维度、发现并帮助用户理解语义关联的动态性，并能够适应大规模链接数据。课题的完成将极大地加强现有语义关联挖掘的能力，加深用户对链接数据的理解，为语义网的可持续发展提供有力的技术保障。在更深远的意义上，将促进人类知识的分享和合理利用。综上可见，本课题的研究具有重要的理论价值和广泛的应用前景。

**参考文献：**

[Agrawal1994] Agrawal R, Srikant R. Fast algorithms for mining association rules. In Proceedings of 20th international conference on very large data bases(VLDB1994), 1994, 487-499.

[Allen1983] Allen J F. Maintaining knowledge about temporal intervals. Communications of the ACM, 1983, 26(11): 832-843.

[Aleman-Meza2003] Boanerges Aleman-Meza, Christian Halaschek-Wiener, Ismailcem Budak Arpinar, Amit P. Sheth. Context-aware Semantic Association Ranking. In Proceedings of the 1st International Workshop on Semantic Web and Databases, 2003, 33-50.

[Andrew1991] Andrew U F. Qualitative spatial reasoning with cardinal directions. In Proceedings of the 7th Austrian Conference on Artificial Intelligence, 1991, 157-167.

[Anyanwu2003] Anyanwu K, Sheth A. p-Queries: Enabling Querying for Semantic Associations on the Semantic Web. In Proceedings of the 12th International World Wide Web Conference, 2003, 690-699.

[Artale2005] Artale A, Franconi E: Temporal description logics. Handbook of Time and Temporal Reasoning in Artificial Intelligence. Elsevier, Amsterdam, 2005.

[Auer2007] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conferenec, 2007, 11-15.

[Auer2009] Auer S, Lehmann J, Hellmann S. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. In Proceedings of the 8th International Semantic Web Conference, 2009, 731-746.

[Auer2013] Auer S, Lehmann J, Ngomo A C N, et al. Introduction to linked data and its lifecycle on the web. Reasoning Web. Semantic Technologies for Intelligent Data Access. Springer Berlin Heidelberg, 2013: 1-90.

[Aydin2009] Aydin I, Karakose M, Akin E. The prediction algorithm based on fuzzy logic using time series data mining method. World Academy of Science, Engineering and Technology, 2009, 51(27):91-98.

[Barabási2009] Barabási A L. Scale-free networks: a decade and beyond. Science, 2009, 325(5939):412.

[Barabási2013] Barabási A L. Network science. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2013, 371(1987):20120375.

[Batsakis2011] Batsakis S, Petrakis E G M. SOWL: A Framework for Handling Spatio-temporal Information in OWL 2.0. In Proceedings of the 5th international conference on Rule-based reasoning, programming, and applications, 2011, 242-249.

[Berners-Lee2001] Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific american, 2001, 284(5): 28-37.

[Berners-Lee2006] Berners-Lee T. Linked Data. Design Issues. WWW Consortium. 2006.

[Bizer2009] Bizer C, Heath T, Berners-Lee T. Linked data - the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, 2009, 205-227.

[Bullmore2009] Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 2009, 10(3):186-198.

[Chen2009] Chen H, Ding L, Wu Z, et al. Semantic web for integrated network analysis in biomedicine. Briefings in bioinformatics, 2009, 10(2):177-192.

[Chen2011] Chen C, Wang G, Liu H, Xin J, Yuan Y. SISP: A New Framework for Searching the Informative Subgraph Based on PSO. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011, 453-462.

[Chen2012] Chen N, Prasanna V. Learning to Rank Complex Semantic Relationships. International Journal on Semantic Web and Information Systems, 2012, 8(4), 1-19.

[Cheng2014] Cheng G, Zhang Y, Qu Y. Explass: Exploring Associations between Entities via Top-K Ontological Patterns and Facets. In Proceedings of the 13th International Semantic Web Conference. 2014, 422-437.

[Cheng2016] Cheng G, Liu D, Qu Y. Efficient Algorithms for Association Finding and Frequent Association Pattern Mining. In Proceedings of The 15th International Semantic Web Conference. 2016, 119-134.

[Cudre-Mauroux2010] Cudre-Mauroux P, Wu E, Madden S. Trajstore: An adaptive storage system for very large trajectory data sets. In Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE2010), 2010, 109-120.

[De2006] De C, Faloutsos C. Center-Piece Subgraphs: Problem Definition and Fast Solutions. In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining. 2006, 404-413.

[Eliassi-Rad2004] Eliassi-Rad T, Chow E. Using Ontological Information to Accelerate Search in Large Semantic Graphs: A Probabilistic Approach. LLNL Technical Report, 2004.

[Fang2011] Fang L, Sarma A, Yu C, Bohannon P. REX: Explaining Relationships Between Entity Pairs. In Proceedings of the VLDB Endowment (PVLDB), 2011, 241-252.

[Freksa1992] Freksa C. Using orientation information for qualitative spatial reasoning. In Proceedings of the International Conference GIS - From Space to Territory, 1992, 162-178.

[Goyal2001] Goyal R K, Egenhofer M J. Similarity of cardinal directions. In Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases, 2001, 36-55.

[Grau2012] Grau B C, Jiménez-Ruiz E, Kharlamov E, et al. Ontology Evolution Under Semantic Constraints. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, 2012, 137-147.

[Gutierrez2005] Gutierrez C, Hurtado C, Vaisman A. Temporal RDF. In Proceeding of the 2nd European conference on the Semantic Web, 2005, 93-107.

[Gutierrez2007] Gutierrez C, Hurtado C A, Vaisman A. Introducing time into RDF. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2):207-218.

[Hai2011] Hai Y, Li X. A General Temporal Association Rule Frequent Itemsets Mining Algorithm. International Journal of Advancements in Computing Technology, 2011, 3(11):63-71.

[Han2000] Han J, Pei J, Yin Y. Mining Frequent Patterns Without Candidate Generation. In Proceedings of the 2000 ACM SIGMOD International Conference Management of Data, 2000, 1-12.

[Hart1968] Hart P E, Nilsson N J, Raphael B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics, 1968, 4(2):100-107.

[Heim2009] Heim P, Hellmann S, Lehmann J, Lohmann S. RelFinder . Revealing Relationships in RDF Knowledge Bases. In Proceedings of the International Conference on Semantic Multimedia, 2009, 182-187.

[Heim2010] Heim P, Lehmann S, Stegemann T, Aroyo L, Antoniou G, Hyvönen E. Interactive Relationship Discovery via the Semantic Web. In Proceedings of the 7th Extended Semantic Web Conference, 2010, 303-317.

[Hendler2012] Hendler J, Holm J, Musialek C, et al. US Government Linked Open Data: Semantic.data.gov. IEEE Intelligent Systems, 2012, 27(3):25-31.

[Hipp2000] Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining—a general survey and comparison. ACM SIGKDD Explorations Newsletter, 2000, 2(1):58-64.

[Hobbs2004] Hobbs J R, Pan F. An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing, 2009, 3(1):66-85.

[Hoffart2013] Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 2013, 194:28-61.

[Jabbour2016] S, Ma Y, Raddaoui B, Sais L, Salhi Y. A MIS Partition Based Framework for Measuring Inconsistency. In Proceedings of the 15th International Conference on the Principles of Knowledge Representation and Reasoning (KR'16). 2016, 84-93.

[Janik2005] Janik M, Kochut K. BRAHMS: A Workbench RDF Store and High Performance Memory System for Semantic Association Discovery. In Proceedings of the 4th International Semantic Web Conference, 2005, 431-445.

[Javed2009] Javed M, Abgaz Y M, Pahl C. A pattern-based framework of change operators for ontology evolution. In Proceedings of On the move to meaningful internet systems: OTM 2009 Workshops, 2009, 544-553.

[Ji2015] Ji M, He Q, Han J, et al. Mining strong relevance between heterogeneous entities from unstructured biomedical data. Data Mining and Knowledge Discovery, 2015, 29(4):976-998.

[Jiang2012] Jiang X, Zhang X, Gui W, et al. Summarizing Semantic Associations Based on Focused Association Graph. In Proceedings of the 8th International Conference on Advanced Data Mining and Applications (ADMA2012), 2012, 564-576.

[Jiang2013] Jiang X, Zhang X, Gao F, et al. Graph Compression Strategies for Instance-focused Semantic Mining. In Proceedings of the 7th China Semantic Web Symposium & 2nd Web Science Conference (CSWS 2013), 2013, 50-61.

[Jin2014] Jin J, Khemmarat S, Gao L, Luo J: A Distributed Approach for Top-k Star Queries on Massive Information Networks. In Proceedings of the 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS2014). 2014, 9-16.

[Jin2015] Jin J, Khemmarat S, Gao L, Luo J: Querying Web-Scale Information Networks Through Bounding Matching Scores. In Proceedings of the 4th International World Wide Web Conference (WWW2015). 2015, 527-537.

[Jin2017] Jin J, Luo J, Khemmarat S, Gao L: Querying Web-Scale Knowledge Graphs Through Effective Pruning of Search Space. IEEE Transactions on Parallel and Distributed Systems (TPDS), 2017, Accepted.

[Kantz2004] Kantz H, Schreiber T. Nonlinear time series analysis. Cambridge University Press, 2004.

[Karli2009] Karli S, Saygin Y. Mining periodic patterns in spatio-temporal sequences at different time granularities. Intelligent Data Analysis, 2009, 13(2):301-335.

[Kasneci2009] Kasneci G, Elbassuoni S, Weikum G. MING: Mining Informative Entity Relationship Subgraphs. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, 1653-1656.

[Kim2008] Kim S K, Song M Y, Kim C, Jang H C. Temporal ontology language for representing and reasoning interval-based temporal knowledge. In Proceedings of the 3rd Asian Semantic Web Conference, 2008, 31-45.

[Klyne2004] Klyne G, Carroll J. Resource Description Framework (RDF): Concepts and Abstract Syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. 2004.

[Lee2009] Lee M, Kim W. Semantic association search and rank method based on spreading activation for the Semantic Web. In Proceedings of the 2009 IEEE International Conference on Industrial Engineering and Engineering Management, 2009, 1523-1527.

[Lee2009b]Lee A J T, Chen Y A, Ip W C. Mining frequent trajectory patterns in spatial–temporal databases. Information Sciences, 2009, 179(13):2218-2231.

[Liu2013]Liu H, Dou D, Jin R, et al. Mining biomedical ontologies and data using rdf hypergraphs. In Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA2013), 2013, 141-146.

[Ma2011] Ma Y, Qi G, Hitzler P. Computing inconsistency measure based on paraconsistent semantics. Journal of Logic and Computation. 2011, 21(6):1257-1281.

[Ma2013] Ma Y, Distel F. Concept Adjustment for Description Logics, In Proceedings of the 7th International Conference on Knowledge Capture (KCAP'13). 2013, 65-72.

[Ma2014] Ma Y, Pe?aloza R. Towards Parallel Repair: An Ontology Decomposition-based Approach. In Proceedings of Description Logics 2014. 2014, 633-645.

[Makita2013] Makita Y, Kobayashi N, Yoshida Y, et al. PosMed: ranking genes and bioresources based on Semantic Web Association Study. Nucleic acids research, 2013.

[Milea2008] Milea V, Frasincar F, Kaymak U. Knowledge engineering in a temporal semantic web context. In Proceedings of the 8th International Conference on Web Engineering, 2008, 65-74.

[Mukherjea2004] Mukherjea S, Bamba B. BioPatentMiner: An Information Retrieval System for BioMedical Patents. In Proceedings of Proceedings of the 30th International Conference on Very Large Data Bases, 2004, 1066-1077.

[Nebot2012] Nebot V, Berlanga R. Finding association rules in semantic web data. Knowledge-Based Systems, 2012, 25(1): 51-62.

[Noy2004] Noy N F, Klein M. Ontology evolution: Not the same as schema evolution. Knowledge and information systems, 2004, 6(4):428-440.

[Noy2006] Noy N F, Chugh A, Liu W, et al. A framework for ontology evolution in collaborative environments. In Proceedings of the 5th International Semantic Web Conference, 2006, 544-558.

[Pérez2006] Pérez J, Arenas M, Gutierrez C. Semantics and Complexity of SPARQL. In Proceedings of the 5th International Semantic Web Conference, 2006, 30-43.

[Perry2008] Perry M S. A framework to support spatial, temporal and thematic analytics over semantic web data. Doctoral Dissertation. Wright State University, 2008.

[Perry2011] Perry M, Jain P, Sheth A P. Sparql-st: Extending sparql to support spatiotemporal queries. Geospatial semantics and the semantic web. Springer, 2011, 61-86.

[Pirrò2015] Pirrò G. Explaining and suggesting relatedness in knowledge graphs. In Proceedings of the 14th International Semantic Web Conference, 2015, 622-639.

[Plessers2007] Plessers P, De Troyer O, Casteleyn S. Understanding ontology evolution: A change detection approach. Web Semantics: Science, Services and Agents on the World Wide Web, 2007, 5(1):39-49.

[Pugliese2008] Pugliese A, Udrea O, Subrahmanian V S. Scaling RDF with time. In Proceedings of the 17th international conference on World Wide Web, 2008, 605-614.

[Pustejovsky2003] Pustejovsky J, Casta?o J M, Ingria R, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. New Directions in Question Answering, AAAI Press, 2003, 28-34.

[Randell1992] Randell D A, Cui Z, Cohn A G. A spatial logic based on regions and connection. In Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning, 1992,165-176.

[Rula2012] Rula A, Palmonari M, Harth A, et al. On the diversity and availability of temporal information in linked open data. In Proceedings of the 11th International Semantic Web Conference, 2012, 492-507.

[Shaheen2013] Shaheen M, Shahbaz M, Guergachi A. Context based positive and negative spatio-temporal association rule mining. Knowledge-based systems, 2013, 37:261-273.

[Sheth2005] Sheth A, Aleman-Meza B, Arpinar I B, et al. Semantic association identification and knowledge discovery for national security applications. Journal of Database Management, 2005, 16(1):33.

[Sheth2008] Sheth A, Perry M. Traveling the semantic web through space, time, and theme. Internet Computing, IEEE, 2008, 12(2):81-86.

[Stojanovic2002] Stojanovic L, Maedche A, Motik B, et al. User-driven ontology evolution management. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, 2002, 285-300.

[Stumme2006] Stumme G, Hotho A, Berendt B. Semantic web mining: State of the art and future directions. Web semantics: Science, services and agents on the world wide web, 2006, 4(2):124-143.

[Tappolet2009] Tappolet J, Bernstein A. Applied temporal RDF: Efficient temporal querying of RDF data with SPARQL. In Proceedings of the 6th European Semantic Web Conference, 2009, 308-322.

[Vespignani2012] Vespignani A. Modelling dynamical processes in complex socio-technical systems. Nature Physics, 2012, 8(1):32-39.

[Viswanathan2012] Viswanathan V, Krishnamurthi I. Finding relevant semantic association paths through user-specific intermediate entities. Human-centric Computing and Information Sciences, 2012, 2(1):1-11.

[Wang2013] Wang C C N, Hecht D A, Sheu P C Y, et al. Semantic computing and drug discovery - a preliminary report. In Proceedings of the 7th International Conference on Semantic Computing, 2013, 453-458.

[Wang2013b] Wang C, Zhang X, Lv Y, et al. Searching Semantic Associations Based on Virtual Document. In Proceedings of the 7th China Semantic Web Symposium & 2nd Web Science Conference (CSWS 2013), 2013, 62-75.

[Welty2006] Welty C, Fikes R. A Reusable Ontology for Fluents in OWL. In Proceedings of 4th International Conference on Formal Ontology in Information Systems, 2006, 226-236.

[Wild2012] Wild D J, Ding Y, Sheth A P, et al. Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. Drug discovery today, 2012, 17(9):469-474.

[Wu2014a] Wu T, Ling S, Qi G, Wang H. Mining Type Information from Chinese Online Encyclopedias. In Proceedings of the 4th Joint International Semantic Technology Conference (JIST'14). 2014, 213-229.

[Wu2014b] Tianxing Wu, Guilin Qi, Haofen Wang: Zhishi.schema Explorer: A Platform for Exploring Chinese Linked Open Schema. In Proceedings of the 8th Chinese Conference of Semantic Web and Web Science (CSWS'14). 2014, 174-181.

[Wu2016a] Wu T, Qi G, Wang H, Xu K, Cui X. Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic Model. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16). 2016, 287-293.

[Wu2016b] Wu T, Yan S, Piao Z, Xu L, Wang R, Qi G. Entity Linking in Web Tables with Multiple Linked Knowledge Bases. In Proceedings of the 6th Joint International Semantic Technology Conference (JIST'16). 2016, 239-253.

[Yang2014] Yang M, Ding B, Chaudhuri S. Finding patterns in a knowledge base using keywords to compose table answers. In Proceedings of the VLDB Endowment, 2014, 7(14), 1809–1820.

[Yuan2013] Yuan P, Jin H, Li Y, et al. SemreX: A Semantic Association-Based Scientific Literature Sharing System. Semantic Web and Web Science, 2013, 161-167.

[Zablith2015] Zablith F, Antoniou G, d'Aquin M, et al. Ontology evolution: a process-centric survey. The Knowledge Engineering Review, 2015, 30(1):45-75.

[Zhang2007] Zhang X, Cheng G, Qu Y. Ontology Summarization Based on RDF Sentence Graph. In Proceedings of the 16th International World Wide Web Conference(WWW2007), 2007, 707-716.

[Zhang2009] Zhang X, Cheng G, Ge W Y, et al. Summarizing Vocabularies in the Global Semantic Web. Journal of Computer Science and Technology(JCST), 2009, 24(1):165-174.

[Zhang2012] Zhang X, Zhao C, Wang P, et al. Mining Link Patterns in Linked Data. In Proceedings of the 13th International Conference on Web-Age Information Management (WAIM2012), 2012, 83-94.

[Zhang2012b] Zhang X, Zhao C, Zhang S, et al. A Partition-Based Approach to Mining Link Patterns. In Proceedings of the 2nd International Conference on Computer Science and Service System (CSSS2012), 2012, 2165-2168.

[Zhang2016a] Zhang X, Lv Y. Towards Multi-Target Search of Semantic Association. In Proceedings of the 6th Joint International Semantic Technology Conference (JIST2016). 2016, 236-244.

[Zhang2016b] Zhang X, Lv Y, Lin E. Object Clustering in Linked Data using Centrality. In Proceedings of China Conference on Knowledge Graph and Semantic Computing (CCKS2016). 2016, 1-12.

[Zhang2016c] Zhang X, Cheng W. Pattern Mining in Linked Data by Edge-labeling. Journal of Tsinghua Science and Technology. 2016, 21(2):168-175.

[Zhao2012] Zhao C, Zhang X, Wang P. A Label-based Partitioning Strategy for Mining Link Patterns. In Proceedings of 7th International Conference on Knowledge, Information and Creativity Support Systems (KICSS2012), 2012, 203-206.

[Zhou2011] Zhou M, Pan Y, Wu Y. Conkar: Constraint Keyword-based Association Discovery. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, 2553-2556.

[Zhou2015] Zhou Z, Qi G, Wu Z, Shi J: A Platform-Independent Approach for Parallel Reasoning with OWL EL Ontologies Using Graph Representation. In Proceedings of the 27th IEEE International Conference on Tools with Artificial Intelligence. 2015, 80-87.

[Zhou2016a] Zhou Z, Qi G, Liu C, Mutharaju R, Hitzler P: Reasoning with Large Scale OWL 2 EL Ontologies Based on MapReduce. In Proceedings of the 18th Asia Pacific Web Conference (APweb2016). 2016, 429-433.

[Zhou2016b] Zhou Z, Qi G, Glimm B: Exploring Parallel Tractability of Ontology Materialization. In Proceedings of the 22nd European Conference on Artificial Intelligence. 2016, 73-81.

**三. 关联挖掘技术路线**

**（一） 概述**

关联挖掘的技术路线分为：语义关联建模、基于模式增长的关联挖掘方法、大规模知识图谱中的关联挖掘、多目标关联检索等若干主要部分。

1. **基于链接模式的语义关联建模研究**

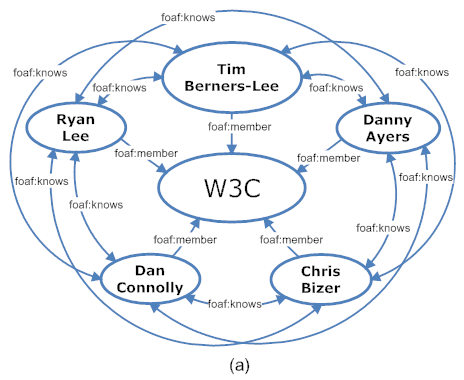
传统基于语义路径的语义关联模型存在两个局限：1）语义路径只能刻画对象两两之间的简单关系，无法刻画多个对象之间的复杂关联。如图1所示，从真实的语义数据中我们可以发现，万维网发明人Tim Berners-Lee与其他四位好友共同就职于W3C机构。传统语义路径模型需要将图1中的结构分解对象两两之间的20个语义关联，且每个关联互相独立，无法统一。实际上，这种多位好友共事关联大量的存在于真实语义数据中，应当统一为一个整体的语义关联；2）语义路径模型并没有考虑语义关联的典型性，即具有语义关联的两个对象之间的语义路径是否也出现在其它语义关联中。在很多情况下，平凡的、不重要的语义路径仅仅表示了两个对象在RDF图中的连通性，但并不表示两个对象之间存在具有意义的语义关联。

图1 多个对象构成的复杂语义关联

为了解决语义路径建模中的局限，我们提出了采用图结构对语义关联进行建模的方法。首先，我们对RDF图中的RDF三元组（RDF Triple）进行过滤，仅考虑描述有名对象间关联的三元组；其次，我们将RDF三元组扩展为链接五元组（Link Quintuple），每一个链接五元组中不仅包含主语、谓语和宾语的URI，同时包含了主语对象和宾语对象的类型信息；接着我们利用链接五元组将RDF图转化为类型对象图（Typed Object Graph），简称为TOG图。RDF图中资源的类型为隐式信息，需要通过RDF语义（RDF semantics）推理得到，而TOG图中对象的类型信息为显式信息；在TOG图的基础上，我们进一步提出了链接模式（Link Pattern）的概念，链接模式是TOG图中的频繁子图。作为语义关联背后的模式，链接模式确保了每一个挖掘到的语义关联在模式上均具有一定的典型性；最终，基于链接模式的定义，我们提出了语义关联的图模型。在给定的RDF图中，一个链接模式可能被该RDF图的多个子图实例化，每次实例化构成了一组对象间的语义关联。

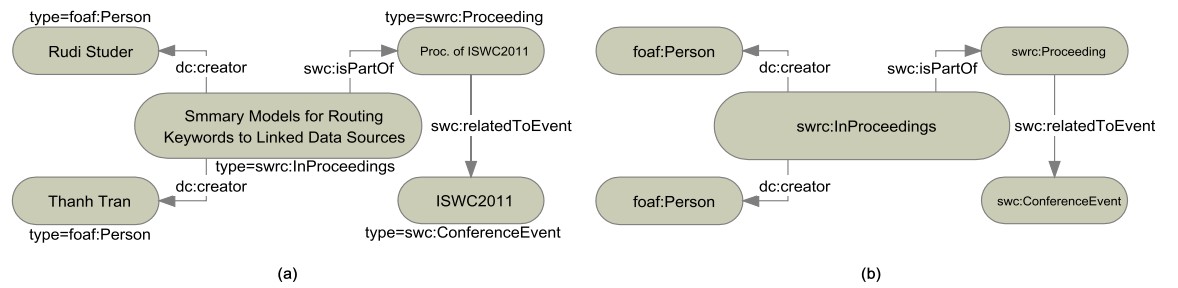
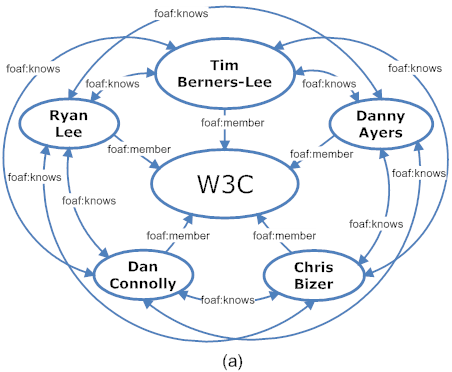
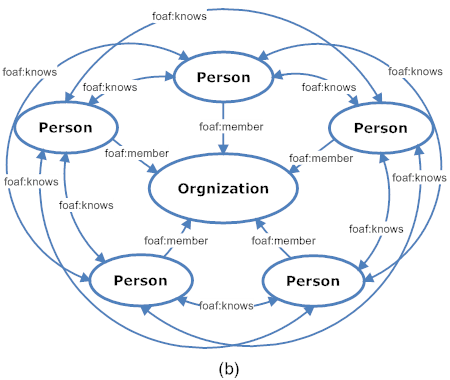
一个具体的TOG图和链接模式的例子如图2所示。其中图2（a）中的TOG图来源于DBLP数据集中抽取出的RDF片段，并在RDF片段中加上了语义对象的类型信息，该片段描述了发表在ISWC2011会议上的一篇论文的详细信息，包括具体的合著者和会议信息等等。图2（b）将图2（a）中的TOG去除了实例信息，仅保留类型信息。假如图2（b）中的结构在DBLP数据集中频繁出现，则将其称之为DBLP数据集中的链接模式，且进而认定对其进行实例化的所有TOG图均为语义关联。

图2（a）一个TOG图片段 （b）对应的链接模式

在研究的过程中，我们把语义关联背后的概念框架抽象为链接模式（Link Pattern），将语义关联表示为链接模式和实例信息的综合，并以此为基础进一步提出语义关联模型，即类型对象图模型（Typed Object Graph简称TOG）。该模型能够刻画多个对象之间的复杂模式和语义关联，为语义关联挖掘打下了坚实的基础。在介绍链接模式和语义关联的概念之前，我们以图3作为例子使读者能够对语义关联和链接模式有个整体的认识。

图3：(a) 语义关联 (b)链接模式

在语义网中，通常将语义数据分为两个层次，分别是实例层（Instance Level）和概念层（Concept Level）。图3-1(a)图为实例层的语义数据，(b)图为抽出的概念层的数据。(a)称为语义关联，(b)若频繁出现则称为链接模式。概念层上的Uri(Uniform resource identifier)通常称为术语（Term），它主要包括类（Class）和属性（Property）。实例层上的Uri或空白节点称之为对象，而实体既包含概念层上的术语又包含了实例层上的对象。例如图(b)中的foaf:Person是术语，图(a)中的Ryan Lee与之相应的实例层上的对象，foaf:knows和foaf:member都是属性。

语义关联并不能直接从RDF图中进行挖掘，这是因为对象的类型信息是语义关联挖掘中的主要元素，而这些类型信息是隐藏在RDF图中的。我们需要根据RDF的语义信息来推理出对象的类型。为解决这些问题，我们提出了一个新的模型即TOG模型，在该模型中可直接体现对象的类型信息，这使得在语义数据集中挖掘语义关联成为可能。下面我们将给出与语义关联挖掘相关的模型和定义。

**定义1 链接五元组（Link Quintuple）**：给定一个RDF文档，和一个三元组，在中链接五元组是对的一个扩展，其中和是对象节点，是对象属性，type(,) 和 type(,) 分别表示主语和宾语在文档中定义的类型。

**定义2 类型对象图（Typed Object Graph简称TOG）:**一个类型对象图包含一个链接五元组的集合：，这些链接五元组是从RDF图中的三元组扩展而来的，包含一个RDF三元组的集合，其中 and , 是从 中扩展得到的。

**定义3 RDF2Pattern子图同构（RDF2Pattern Isomorphism）：**给定一个类型对象图的子图和一个图,其中是从RDF文档中导出的，RDF2Pattern子图同构就是一个映射函数f: , 使得以下两个条件成立：(1) , (2) 。

**定义4链接模式（Link Pattern）：**给定一个RDF文档的数据集，一个链接模式是一个频繁出现的图结构，满足如下条件：

1. 是一个有向图，其中每个顶点是用户自定义的类，每条边是一个对象属性。
2. 在中存在一个RDF2Pattern的子图同构，使得。是包含p的子图同构图的个数，是用户给定的支持度的阈值。

**定义5 链接模式实例化（Instantiation of Link Pattern）:**给定语义数据集，是从中导出的RDF图。中存在链接模式，若能够用实例化当且仅当以下条件成立：(1) ，是一个对象，(2)并且，(3)且。

**定义6语义关联（Semantic Association）:**是语义数据集中的对象集合，是从中导出的RDF图，若是语义关联当且仅当是中链接模式的实例化，和分别是语义关联的顶点集合和边的集合。

1. **基于模式增长的语义关联挖掘方法的研究**

在语义关联发现方法的研究中，本课题组在类型单一化、挖掘策略选择和基于模式增长的语义关联挖掘等环节进行了深入地研究，使得传统数据挖掘中的模式增长挖掘算法能够适应TOG图特殊的图结构。

1）**类型单一化的研究**。

一个对象在RDF图中可能被定义为多个类型，例如一个类型为“人”的对象，其类型同时也可能会被定义为“教师”。对象的多类型属性在语义数据集中普遍存在。但对象的多类型问题将影响TOG图的构建。一个多类型的对象将使得TOG图无法用五元组进行描述，并且给TOG图的挖掘带来极大的困难。我们提出了一组启发式规则用于界定复杂情况下对象的类型问题。当对象的类型在单个或多个RDF图中被定义多次时，我们将区分对象的“局部类型（Local Type）”和全“局类型（Global Type）”，并通过统计分析区分对象的“权威类型（Authoritative Type）”和“非权威类型（non-Authoritative Type）”，最终根据上下文确定对象的单一类型。类型单一化的研究降低了语义关联挖掘的复杂性。

2）**挖掘策略选择**。

语义关联的图模型中涉及链接模式，对语义关联的挖掘实际上是对链接模式的挖掘以及发现链接模式的实例化信息的过程。我们提出了两种语义关联挖掘方法：（1）分阶段挖掘方法。在挖掘的第一阶段，我们通过经典的频繁模式挖掘算法从TOG中挖掘出部分或所有链接模式，在挖掘的第二阶段，在挖掘出了链接模式中选取部分模式，在RDF图中寻找对这些模式的实例化子图，作为最终挖掘出的语义关联。（2）合并式挖掘方法。在该方法中，对链接模式和语义关联的挖掘在同一步骤中并行进行。在利用频繁模式挖掘算法挖掘链接模式时，语义关联在计算链接模式支持度时被挖掘出来。最终挖掘结果是TOG图中的所有关联模式和RDF图中的所有语义关联。两种挖掘方法适用于不同的场景：分段式挖掘方法由于在第二阶段挖掘前引入了模式选取，因此更适用于语义关联的聚焦挖掘，即在给定的条件下（如用户兴趣等）挖掘相关的部分语义关联集；合并式挖掘方法适合于完整语义关联集的挖掘，但面对大规模RDF数据集时，其时间复杂度和空间复杂度均较高。

1. **基于模式增长的语义关联挖掘**

在频繁模式挖掘算法中，我们选择了经典的基于模式增长（Pattern-Growth-Based）的模式挖掘算法。我们选择并修改了gSpan算法。传统的gSpan算法适用于无向的简单图，而TOG图为有向带标签图，且TOG图中可能包括自环、多边等特殊情况。我们对gSpan的修改使之可以用于TOG图中的链接模式与语义关联的挖掘。此外，我们在挖掘过程中提出了使用最小支持度和最大边两个参数控制链接模式挖掘的规模，并通过实验对时间性能和空间性能进行了广泛的分析。

**1）数据预处理**

互联网上提供下载的数据集的格式往往是多种多样的，常见的后缀形式有rdf、ttl、nt等。无论是哪种格式，都可以利用一些开源工具来进行解析，如jena，解析出来的结果就是数据集中包含的三元组的集合。

语义关联主要关心的是对象之间的链接关系，然而并不是所有的三元组都与链接模式相关，例如一个三元组可能描述一个注释，也可能是scheme层一个格式，这些数据与链接模式并不相关，我们必须过滤掉与链接模式不相关的三元组。

下面定义了三个条件，用来决定一个三元组是否应该保留下来用于创建类型对象图。给定一个三元组，如果不满足以下任一条件，那么需要丢弃它。

* 1. 是两个对象间的链接，在中和可以是空白节点，但是不能是字面量（Literal），是用户定义的对象属性。
  2. 是对象的类型定义，是用户自定义的类。
  3. 是对象属性定义域和值域的定义。

条件1）和条件2）中获得的三元组是直接所需的，因为它们是描述对象链接和对象类型的三元组。条件3）中的三元组是用于推理对象类型信息的。经过数据的预处理留下的三元组可以用于构建类型对象图了。

**2）TOG图的生成**

经过数据预处理，保留下来的三元组可用于创建类型对象图。在类型对象图模型中，一个对象必须只有单一类型。然而实际上，一个对象可以定义多种类型。例如在Semantic Web Dog Food数据集中，一篇被接收的论文可以被定义为swc:Paper，也可被定义为swrc:InProceedings。由于多类型的存在，将会导致挖掘复杂度的升高，尤其在子图同构测试阶段将会消耗很多时间用于同构图的测试，为提高挖掘效率，在建立TOG时必须避免这种现象。下面我们制定一些规则用于单一化对象的类型。

给定一个语义数据集，在RDF文档中定义的类型有，以下规则用于决定的类型。

**规则4-1：**如果在语义数据集中除了在其它文档中没有定义的类型，那么将被赋值为，其中并且在中有最多的实例。

**规则4-2：**如果在数据集中，有多个文档中定义了的类型，那么将被赋值为 ，其中，并且与其它类型相比被定义为的类型最频繁。

在规则4-1中，对象的类型被赋值为局部中主要的类型，然而在规则4-2中，对象的类型被赋值为全局中主要的类型。如果在RDF数据中每个对象类型信息频繁度相同，那么就随机决定被赋值的类型。另外，还有一种情况需要考虑，对象的类型信息没有被定义。我们将会利用对象属性的定义域和值域决定对象的类型，具体见规则4-3。

**规则4-3：**在文档中的三元组，如果或的类型没有明确定义，的定义域或者值域有明确的定义，那么可以这样决定对象的类型：，。

这里我们只是简单采用对象属性的定义域或值域来决定对象的类型，以后的工作需要采用更具推理能力的算法来决定对象的类型信息，例如在论文[46]中讨论的推理方法。经过对象类型确定后，所有保留的三元组都可转化成链接五元组，进而构建类型对象图。

1. **大规模知识图谱中关联挖掘的研究**

上述关联挖掘方法仍然在大规模知识图谱数据上存在挖掘的局限性。基于模式增长的挖掘算法由于其复杂度较高，通常适用于小型图集，而现有的大规模语义数据集的规模超过了百万甚至十亿三元组。海量语义数据与语义关联挖掘带来了巨大的挑战。本课题在挖掘过程中的语义数据聚类、语义数据划分与挖掘结果合并，以及挖掘前的图压缩等方面进行了深入的研究。

1）**海量语义数据中挖掘语义关联的流程**

海量语义数据中挖掘语义关联的流程可以分为多个步骤。如图3所示，语义数据聚类通过聚类、划分和合并等步骤最终形成挖掘结果。以下分别对流程中各步骤进行简述。

* 语义数据：语义关联发现的输入数据集为一般意义的RDF数据，并且要求数据中包含对象与对象之间关系的三元组。
* 语义数据聚类单元：在对输入的语义数据进行基本的解析之后，该单元将进一步分析不连通的RDF图之间所共有的单边模式，从而将语义数据进行聚类，形成语义数据簇。
* 语义数据建模单元：针对生成的语义数据簇，该单元将进行基本的数据清洗，并将RDF图转化为TOG图。
* 基本标记单元：由于生成的TOG图的规模可能超过内存限制，为了提高挖掘的效率和可扩展性，挖掘过程开始之前将进行TOG图的分块。我们提出了一种边标记方法用于TOG图的分块与合并。对边的标记有两种规则：基本标记规则采用的是接近随机的标记方法。
* 优化标记单元：与基本标记规则不同，优化标记规则通过一系列启发式规则，提高了边标记TOG图划分方法的成功比例，降低了伪失败（False Failure）比例。
* 多划分单元：通过基本标记规则或优化标记规则，TOG图被首先划分为多个分块。该单元将保证分块具有一定的数学性质，并且每个分块均能够载入内存。
* 二划分单元：为了提高分块的均匀性，我们引入了二划分单元，用于将多划分后仍然较大或结构较为复杂度的分块进一步进行层次化划分。该单元与快速挖掘单元联合工作，快速挖掘单元的反馈为本单元提供划分的依据。
* 快速挖掘单元：本单元用于预测每个分块中潜在的链接模式与语义关联的数量级。该数据作为反馈提供给二划分单元，使得较大或结构较为复杂度的分块能够尽早发现并进一步得到划分，从而提高整体划分的均匀性。
* 完整挖掘单元：本单元采用修改后的gSpan算法完整地挖掘出所有分块内的局部链接模式和语义关联。
* 链接模式合并单元：分块操作可能挖掘出的链接模式和语义关联的缺失，本单元起到的作用是将局部的链接模式和语义关联按照一定的算法进行合并，保证挖掘结果的完整性。
* 语义关联发现单元：语义关联作为链接模式挖掘过程中的副产品，在本单元中进行汇总和统计，并作为输出展示给用户。

图3.关联挖掘的工作过程

2）**知识图谱聚类**

本课题组提出了通过聚类将可能包含共同链接模式的多个TOG图进行合并的方法。一个给定的RDF数据集中可能存在多个不连通的RDF图，进而扩展为多个不连通的TOG图。通过分析TOG图之间共同的单边链接模式（1-edge Link Pattern），我们定义了TOG图的连接关系（Connection），并根据连接关系将可能包含共同链接模式的多个TOG图进行合并。该方法使得语义关联的挖掘可以采用分治法，在每个聚类（Cluster）中独立完成，而无需在整个数据集上进行，且单个聚类中的挖掘结果可以直接合并为最终挖掘结果。

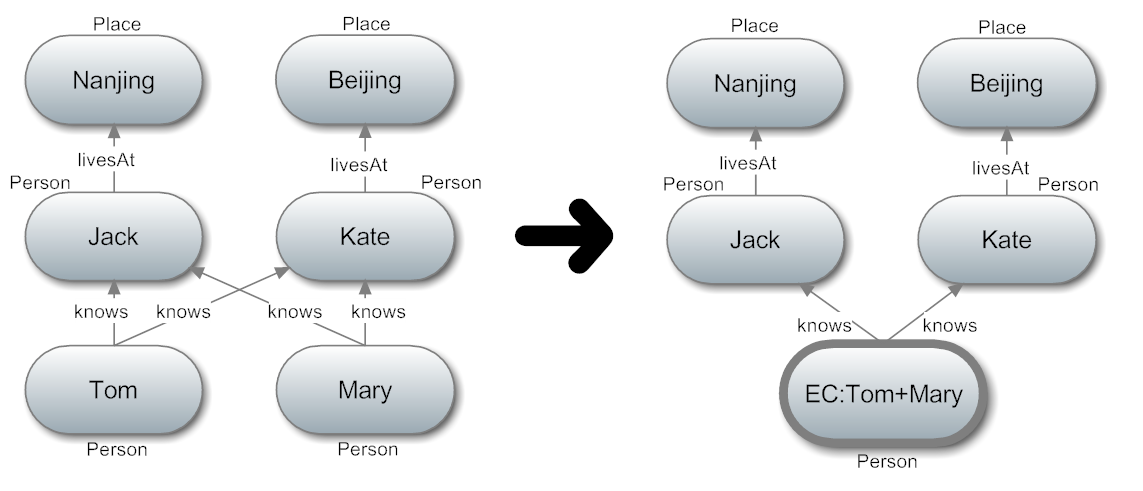
3）**语义数据划分与挖掘结果合并**

本课题组提出了通过分块解决大规模TOG图挖掘的方法。传统频繁模式挖掘算法在处理大规模图数据集时效率不高，且大多算法均假设图数据集可以存放在内存中。虽然TOG图的聚类方法可以降低挖掘的时间和空间复杂度，但单个TOG聚类中的图规模仍然可能超出内存范围。因此我们提出了一种基于边标记（Edge-labeling）的图分块方法。在该方法中，我们使用一种边标记规则为TOG中每条边分配一个标签，该标签体现了其主语对象和宾语对象的类型。在给定内存大小和TOG图规模的情况下，我们将建立TOG图的多个分块（Partition），并依次将TOG图中的每条边按照其标签划分至合适的分块。在划分过程中，我们定义了边与分块之间的连接度（Connectivity），每条边将依次划分至连接度最高的分块中。基于边标记的划分方法并不适用于任意RDF数据集：给定一种边标记规则，当某个RDF数据集中存在大量同种类型的边时，将产生大量具有同样标签的边，这将进一步导致最终存在一个或多个无法放入内存的大型TOG分块，我们将此情况称为该标记规则的失效案例（Failure Case）。我们同时也发现伪失效案例（False Failure），即某种边标记规则无法对某数据集进行有效划分，但存在其他的边标记规则可以进行有效划分。我们提出了一种基于启发式方法的优化标记规则，用于减少伪失效案例的发生。相比于传统模式挖掘中的图分块算法，我们证明了基于边标记的分块方法具有以下优点：1）基于边标记的分块方法保证了最终挖掘结果的完整性，而传统图分块方法存在挖掘结果缺失的问题；2）基于边标记的分块方法保证了局部模式即为全局模式，而传统图分块方法无法保证这一性质。该性质简化了分块后的挖掘结果合并过程。

4）**TOG图压缩**

语义数据的聚类、划分与合并方法可以解决海量语义数据集上的语义关联挖掘问题。但是该方法的复杂性较高，且需要良好的硬件环境支持。本课题组在研究聚类与划分算法的同时，也观察到语义数据中存在大量可压缩的图结构。本课题组提出了一组针对语义数据图结构的压缩策略。实验证明该组压缩策略可以有效地提高海量语义数据上语义关联挖掘的效率。

总体来说，海量语义数据中可能会产生两种不同类型的可压缩结构：1）对等结构。传统数据挖掘在图挖掘中常会关注图中顶点间存在的拓扑结构或标签等属性上的相似度，因为通过对顶点相似度的计算可以进一步判断出图中是否存在结构或文本上的冗余。在以TOG图为模型的语义数据中，通常会存在大量具有相同拓扑结构的节点。我们通过节点在TOG图中的相邻关系定义了节点的“家族”：即所有与某节点直接相邻的节点集。当TOG图中两个节点具有相同的“家族”，就称这两个节点为“对等点”，表示它们在TOG图中处在完全对等的拓扑结构位置中。一组“对等点”可以被压缩为一个单一的“超级节点”，从而消除TOG图在拓扑结构上的冗余。如图4左所示，Tom和Mary具有完全相同的“家族”，在对等压缩策略下，Tom和Mary将被压缩为同一个“超级节点”，如图4右所示。压缩后的TOG图在节点数和边数上均有所减少。2）依赖结构。此类可压缩结构来源于节点间在语义上的依赖关系。当TOG图中某个节点的“家族”中仅包含一个成员时，表示该节点上在图中的相关信息极为有限，仅为该节点唯一的家族成员。换言之，该成员实际上决定了该节点的语义。我们将此类结构定义为TOG图中节点间的依赖-被依赖关系，并进一步提出了依赖结构压缩策略。TOG中的被依赖节点可以作为一个“超级节点”代表其自身以及所有对其依赖的节点。如图5所示，Tom和Mary分别依赖于Jack和Kate，因此可以被压缩至Jack和Kate这两个“超级节点”的内部结构中。依赖结构压缩策略是一种迭代压缩策略，由于图5中的Jack和Kate节点在第一轮压缩后分别依赖于另两个节点，因此它们在第二轮压缩中将被进一步压缩掉。TOG图在多轮压缩后将成为仅包含两个节点和两条边的最小图。



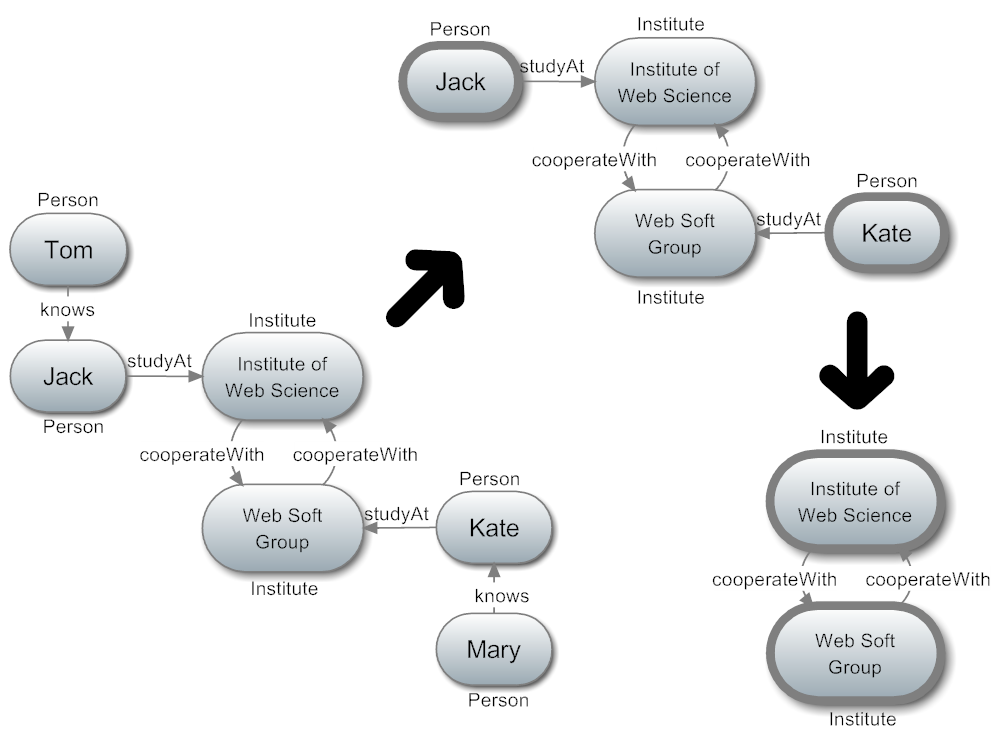
图4. 对等结构压缩策略示例

图5. 依赖结构压缩策略示例

由于最终形成的超级节点在内部结构上存在差异，对等结构和依赖结构压缩策略分别适用于不同的语义挖掘任务。本课题组对两者的适用性做了深入地探讨，并通过在海量语义数据上的实验证明了对等结构压缩策略在提高语义关联挖掘效率上具有明显的作用。

1. **关联索引与搜索服务**

本项目拟按照如下方案建立关联的索引与搜索服务：

1） **利用语义关联模式发现语义关联**

本项目拟通过数据挖掘的方式从大规模语义数据中获取典型的语义关联模式。在本项目中，语义关联模式是指在多个数据源的语义数据中反复出现的典型的拓扑结构，而两个资源之间的语义关联是指匹配某种语义关联模式的RDF图，而不是简单的资源间的有向路径。语义关联是语义关联模式的具体化。语义关联模式不仅应当在结构上保持一定的出现频率，同时在语义上也应当具有典型性。这是本项目中的语义关联模式与普遍意义上的典型拓扑结构最大的区别。例如，在分析了多个数据源的语义数据后，系统可能会综合得到如图2所示的拓扑结构。该结构是一种星形结构，并且要求中心节点是Organization概念的实例（Instance），与中心节点相连的节点为Person概念的实例，且节点之间的有向边表示一个人与一个组织之间的“会员(memberOf)”关系。这种拓扑结构内在地反应了会员之间 “共事”的语义关联。当这样的拓扑结构在数据集中出现的频率高于某一阀值时，我们就将其视为一个典型的语义关联模式。

     这些典型的语义关联模式将作为语义关联搜索系统的知识，帮助进一步发现数据集中的语义关联。

**2）利用语义关联模式索引语义关联**

首先，我们将单独索引通过挖掘得到的语义关联模式。对已挖掘出的语义关联模式进行索引有利于提高语义关联模式的挖掘效率，并剔除那些重复的语义关联模式（例如，新发现的语义关联模式从图的角度来说实际上是某已知语义关联模式的子图）；其次，我们将索引语义关联。由于不同的语义关联之间可能存在一定的相似度，即基于相同的语义关联模式，拥有相同的拓扑结构，差异仅仅在于发生关联的资源不同，或者描述关联的数据来源不同。因此，在索引的过程中，我们将索引语义关联之间的差异，而不是关联本身。这种差异性索引（Discriminative Index）将极大地降低索引的空间开销并提高查询的效率。此外，语义关联索引中还包含每个语义关联对应的数据来源。来源信息可以用于评估语义关联的可信度。

**3）语义关联的表现方式**

对于语义关联的表现方式，拟将语义关联的结构信息和字面信息转换为类似自然语言的方式以供用户阅读。当一个语义关联较为复杂时，拟采用摘要的方式对关联的表示进行精简。语义关联摘要应当能够在有限的长度内尽可能多地表达语义关联的含义。

**4）语义关联检索的反馈机制**

本项目拟在语义关联检索界面上实现与用户的人机交互。用户可以对检索结果提供反馈，例如对某个语义关联的典型性进行评价，评价的结果将用于调整语义关联搜索系统对语义关联模式的典型性分析；另外，用户还可以对感兴趣的语义关联进行标注，例如为复杂的语义关联进行命名（Naming）、添加标签（Tagging）、注释（Comments）等等，或者对语义关联的可信度发表评价。

综上所述，本项目将最终在语义搜索引擎Falcons中实现一个语义关联搜索服务，用于验证本项目研究方法的有效性。用户可以输入任意两个资源的URI，假如系统已在数据集中发现该资源间存在某种语义关联，则向用户返回该关联、并接受用户的反馈。图3描述了一个语义关联搜索服务的模块与数据流程。其中：

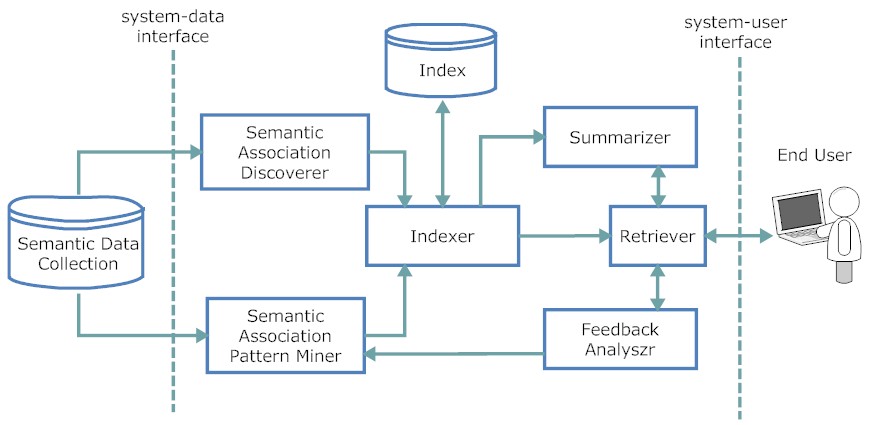
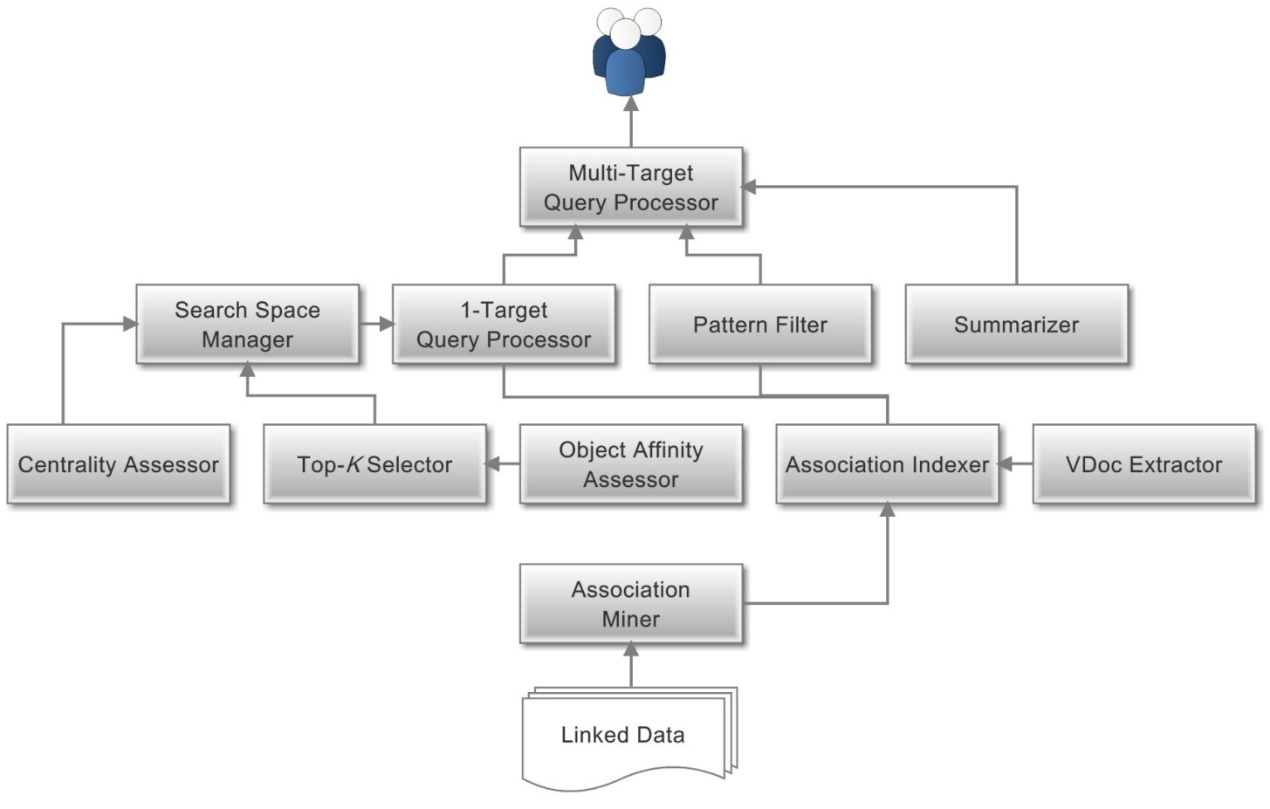
1. Semantic Data Collection: 语义数据集。在本项目中，我们将使用Falcons已收集到的大规模语义数据。
2. Semantic Association Pattern Miner: 语义关联模式挖掘模块，用于发现典型的语义关联模式。
3. Semantic Association Discoverer: 语义关联发现模块，用于通过语义关联模式发现语义关联。
4. Indexer: 索引模块，用于索引语义关联模式及语义关联。
5. Index: 索引文件。
6. Retriever: 检索模块，用于接受用户的查询并生成结果。
7. Summarizer: 摘要模块，用于生成用户可理解的关联摘要。
8. Feedback Analyzer: 反馈分析模块，用于接受用户反馈，并在分析后提交给关联模式挖掘模块。

图6.关联搜索服务体系结构

1. **多目标关联检索**

****除了采用3.1.6中提及的关联搜索体系以外，本课题还将探索多目标关联检索体系。该体系能够使得用户可以搜索多个对象之间的存在的复杂关联。多目标搜索的体系结构如图7所示：

**图7. 多目标关联搜索体系结构**

1. Linked Data： 语义数据集或知识图谱。
2. Association Miner: 语义关联模式挖掘模块，用于发现知识图谱中的语义关联。
3. Centrality Assessor: 中心度评估模块。通过中心度的计算（PageRank、HITS等链接分析算法）评估知识图谱中的对象、子图、语义关联和关联模式的重要性。
4. Top-K Selector: 基于中心度评估结果，选择具有重要性的前K个搜索结果。
5. Object Affinity Assessor: 对象亲近性评估模块。该模块用于计算对象之间的亲近性，用于后续模块的对象选择策略。
6. Association Indexer: 关联索引模块。用于有效地索引语义关联模式挖掘模块中获得的海量语义关联。
7. VDoc Extractor:虚拟文档抽取模块。用于从知识图谱中为每个对象抽取出能够描述其语义的相关文本信息，并将其自身的文本信息与其邻居的文本信息放入半结构化的虚拟文档中。
8. Search Space Manager：搜索空间管理模块。用于有效的管理搜索空间，对搜索空间中的检索对象、涉及到的语义关联进行裁剪，使得系统的搜索空间始终处于可维护状态。
9. 1-Target Query Processor: 单目标查询处理器。该模块基本采用3.1.6中所述的搜索体系，能够通过用户输入的关键字查询到匹配的单个语义关联。
10. Pattern Filter: 模式过滤器。该模块能够使得用户通过制定模式的方式，对海量搜索结果进行过滤。
11. Summarizer：摘要模块，用于生成用户可理解的关联摘要。
12. Multi-target Query Processer：多目标查询处理器。该模块综合以上所有模块的搜索结果，能够允许用户输入多个查询目标对应的关键字，并将其解析为多个单目标查询，最终将所有单目标查询得到的语义关联进行综合，向用户返回匹配多目标查询的搜索结果。

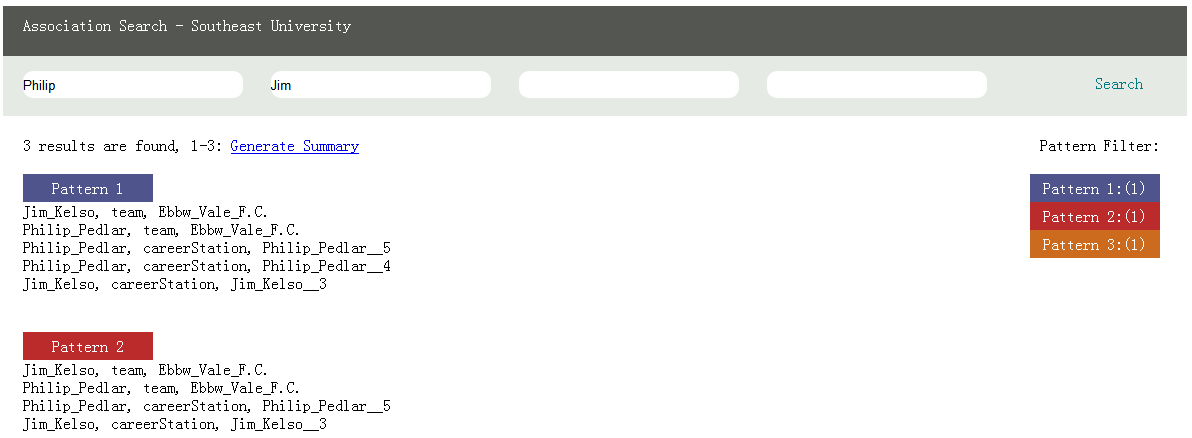
多目标关联搜索的界面如图8所示：

图8. 多目标搜索界面

多目标关联搜索的流程如图9所示。多目标关联搜索过程包含三个搜索步骤：

（1）第一步：用户输入一个多目标关联搜索，其中包含*t* 个子查询： ，该搜索将被分解为*t* 个单目标子查询。每个单目标子查询将会被提交至对象虚拟文档的索引模块，并通过传统的关键字搜索方式获得搜索结果。通常，搜索结果包含海量匹配对象。为了减小搜索空间，将通过第二步top-k策略进一步处理匹配到的对象并找到候选关联。

（2）第二步：上一步的输出为*t*个top-k匹配的对象集合。本步骤中将找出对应的候选关联。候选关联是指在每个对象集合中至少引用到一个对象的关联。该步骤实际上是一个单目标搜索过程。所有候选关联均会被输出到第三步中。

（3） 第三步：上一步的输入结果为t个单目标关联搜索结果。这些结果将通过计算交集的方式得到最终搜索结果。多目标搜索的伪代码在算法4中列出。

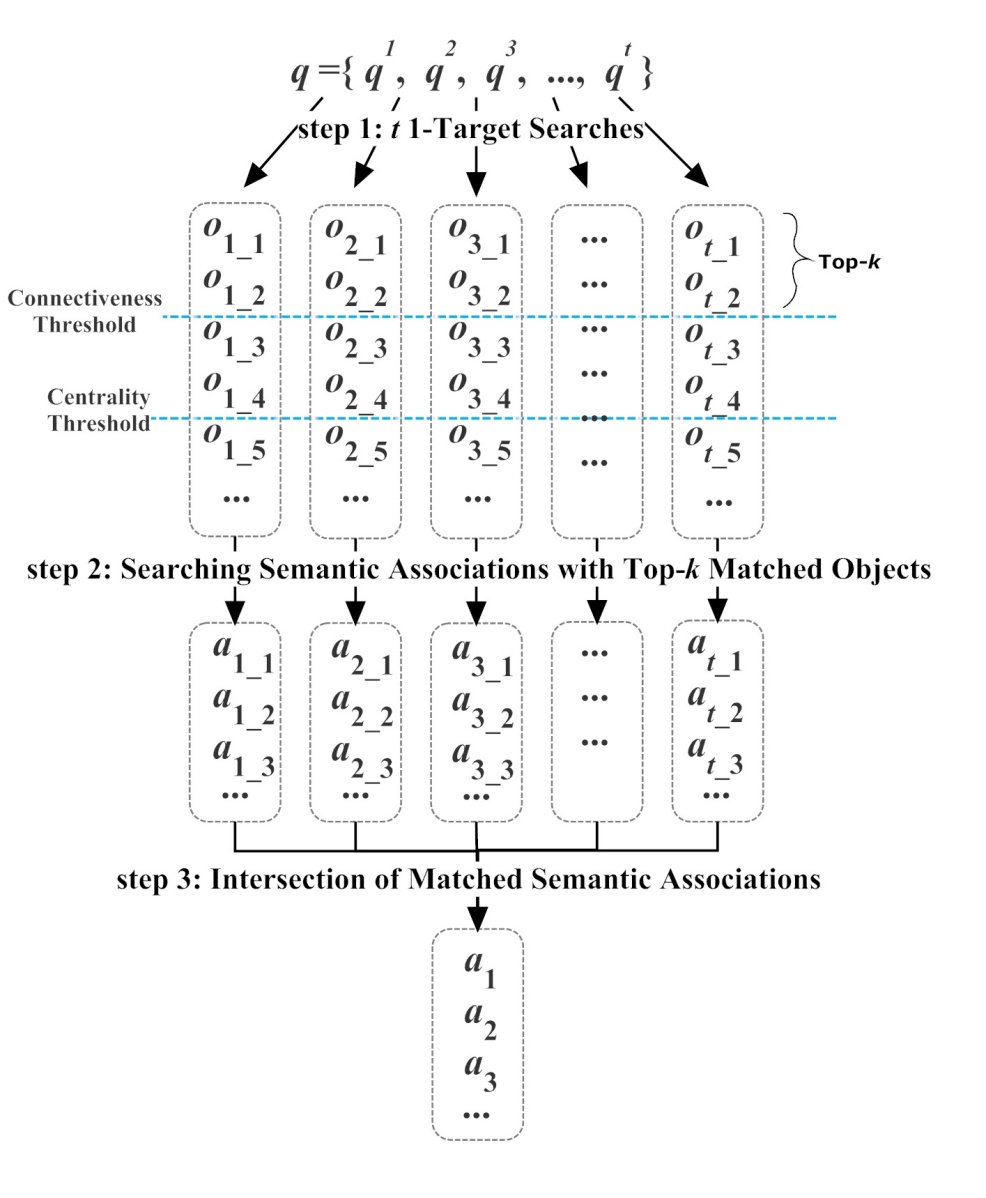


图9. 多目标关联搜索流程

|  |
| --- |
| **算法4 : 多目标关联搜索** |
| **Input**: An MT search: , the object set: , the set of discovered association:   1. For to , find matched objects with keyword search on the Virtual Document of each objects that match keywords in ; 2. Reduce with top-*k* selection strategy; 3. For each , find those associations that refer to at least one object in ; 4. For *j* = 1 to *k*:    1. quick sort all associations in each ;    2. find , which contains least Semantic Associations; 5. For each Semantic Association :    1. for *j* = 1 to *k*: use binary search to check whether ;    2. if every () contains , put into the final result set ;   **Output**: Final Result Set |