计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200154

基于关系图谱的科技数据分析算法

张寒烁　杨冬菊
北方工业大学大规模流数据集成与分析技术北京市重点实验室　北京 100144
北方工业大学云计算研究中心　北京 100144
(hanshuo_1994@foxmail.com)

摘　要　随着科技数据量的不断增长,各科技部门积累了大量科技项目的科技管理数据。对于大量结构化数据,需要针对分散数据进行整理、分析,最终按需求提供数据查询与抽取服务。由于在关系数据库中关联关系的分析效果不佳,为了提高分析效率,文中引入了关系图谱进行数据处理。首先,提出了一种基于词频的实体搜索与定位算法来提取实体和关系,并构建关系图谱;其次,对关系图谱进行分析,提出了一种基于改进 FP-growth 的图数据频繁项挖掘算法;然后,设计了基于图数据的数据筛选流程,进行数据的筛选、分析,并定义评分矩阵,对待筛选数据情况进行评价,最终给出分析意见,且数据筛选的评价标准可以进行自定义;最后,结合构建的关系图谱,将算法进行了实际应用,并将其封装为服务。实验结果表明,提出的基于改进 FP-growth 的频繁项挖掘算法相比传统 FP-growth 算法在时间上有 10%～12%的提升,数据筛选过程的准确率达到 97%左右。

关键词:关系图谱;数据分析;图谱构建;人员关系图谱;数据挖掘;服务应用
中图法分类号　TP391

Technology Data Analysis Algorithm Based on Relational Graph

ZHANG Han-shuo and YANG Dong-ju
Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data,North China University of Technology,Beijing 100144,China
Research Center for Cloud Computing,North China University of Technology,Beijing 100144,China

Abstract　With the continuous growth of scientific and technological data,various science and technology departments have accumulated a large number of scientific and technological management data of scientific and technological projects. For a large amount of structured data,it is necessary to organize and analyze the distributed data,and finally provide data query and extraction services according to requirements. The analysis of relationships in relational databases is not effective. In order to improve the efficiency of analysis,relational graphs are introduced for data processing. Firstly,an entity search and localization algorithm based on word frequency is proposed,and the entities and relationships are extracted to construct the relationalgraph. Secondly,an improved FP-growth algorithm for frequent item mining of graph data is proposed in order to solve the frequent item screening problem in the graph data. Then,a data filtering process based on graph data is designed. In addition,this paper defines the scoring matrix,evaluate the screening data,and finally give an analysis opinion. The evaluation standard of data screening can be customized. Finally,combined with the constructed relational graph,the algorithm is applied in practice and encapsulated as a service. Experimental results show that the improved FP-growth-based frequent item mining algorithm has 10%～12% improvement over the traditional FP-growth algorithm. The accuracy of the data screening process designed in this paper reaches 97%.

Keywords　Relational graph,Data analysis,Graph construction,Construction of human relation graph,Data mining,Service application

1　引言

随着计算机技术的发展,“互联网＋”时代的网络数据量呈爆发式增长。在科技创新与科技管理改革过程中,各个科技部门积累了大量科技项目申报、评审、过程管理等科技管理数据。随着对数据分析、数据挖掘的需求日益增加,数据的挖掘与深度利用成为了近年来的研究热点,尤其是对数据之间关系的挖掘更利于拓展数据价值。在科技管理过程中,不仅需要考虑评审专家的领域、职称、单位等信息,更重要的是需要回避一段时间内有过项目合作的、同单位的、曾是同事关系的专家,因此,对科技数据进行处理,结合数据中蕴含的关系进行数据的筛选、提取,最终为顶层应用提供相关服务,成为

到稿日期:2019-12-25　返修日期:2020-05-28　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划课题(2019YFB1405103)
This work was supported by the National Key Research and Development Project of China(2019YFB1405103).
通信作者:杨冬菊(yangdongju@ncut.edu.cn)

需要解决的重点问题。在利用关系数据库进行关联分析时,多次查询数据库的方式效率低下,因此,本文引入了关系图谱,将结构化数据的分析转化为图数据的分析,以提高处理效率。

针对上述问题,本文提出了一套针对科技数据的关系图谱构建及图数据分析的算法。首先,通过规则构建科技数据关系图谱;然后,针对图谱进行关联关系的分析,针对需求进行数据筛选,将结构化数据算法应用于图数据分析。本文的主要贡献如下:

1)针对已有结构化数据,提出了一种基于词频的实体搜索算法,用于完成关系图谱构建。

2)针对科技数据的关系图谱数据的频繁项挖掘问题,对FP-growth算法进行了改进,并进行了应用。

3)基于科技数据的关系图谱数据,提出了一种基于图遍历的数据抽取算法,其使用评分对数据进行评价,且支持抽取变量自定义,最后对其进行了应用。

2 相关工作

关系图谱是在大量数据集合中,按照一定的规则或规律,通过特定的算法对数据进行分析,从而发现不同事物中蕴含的关系,并将事物进行关联,最终形成的数据网络[1]。其概念源自于社交关系图谱(Social Graph,SG)[2],相比知识图谱[3],关系图谱更加注重图谱实体间的关系及整体,而不是知识结构[4]。

特定领域的图谱注重概念之间的体系结构与专业性[5],因此在构建时通常使用自顶向下的方式或自顶向下与自底向上相结合的方式,并以关系数据库中的结构化信息为起点,将结构化数据扩展到非结构化数据[6]。在图谱的构建过程中,基于规则的实体抽取方法是最早被使用的[7],其通过人工手动构造有限规则的方法,利用字符串的模式匹配进行实体抽取,并采用自顶向下的思想。使用基于规则的构造方法需要较为准确的模板,但由于规则有限,因此其识别率较差[8]。基于统计学理论的实体命名方法目前较为普遍,其采用了自底向上的思想,利用统计学方法计算中文人名用字的概率以及中文机构名称,并对实体进行提取[9]。

数据分析及可视化方面的相关研究,集中于基于图的分析及应用和结构化数据的分析。图的分析及应用主要集中在图的匹配算法、算法应用等方面。Yu等[10]对比了精确图匹配的相关算法,得出了图匹配技术更适合对具有高度关联性的数据进行复杂模式搜索的结论。Zhang等[11]提出了一种面向模式图变化的增量图模式匹配算法PGC_Inc GPM,有效缩短了图模式匹配的执行时间。在图数据的分析应用方面,Guan等[12]研究了城市安全知识图谱的检索算法,通过关联类图的构建、剪枝和融合操作来进行查询,并最终返回检索结果。Sun等[13]针对旅游数据构建了知识图谱,并根据旅游数据图谱进行了频繁序列的挖掘,设计了多维度路线搜索和排序机制以推荐个性化的旅游路线。在结构化数据分析方面,Zhao等[14]提出复用结构化数据抽取模板对Web页面的结构化数据进行解析,并对Web页面进行分类。Zhang等[15]针对结构化数据提出了学术地图的概念,重点设计了知识图谱的多种使用环境,并使用三元组的形式进行存储。

3 关系图谱构建方法

关系图谱的构建过程包含3方面:实体识别、数据抽取及关系匹配。

关系图谱的表示使用RDF(Resource Description Framework)框架,RDF三元组可表示为$t\langle e_1,r,e_2\rangle$,其中$e$表示实体,$r$表示关系。本文在RDF框架的基础上,增加了时间标签,用于标识该关系的创建日期。最终形成的关系模型如图1所示,关系模型中箭头方向为实体搜索的方向。

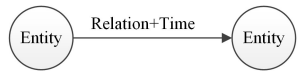

图1 关系模型

Fig. 1 Relational model

构建关系图谱的主要元素为单位、人员和项目这3方面。实体识别的目的是在所有结构化信息数据中对关注的人员、项目、单位进行搜索定位。对于较为普遍的非专有名词,如人员和单位名称,本文使用了目前应用较为广泛且开源的基于HMM模型的实体识别算法[11]。而对于专业性较强的、不易识别的如项目名称一类的专有名词,本文提出了一种基于词频的实体搜索算法。

本文提出的基于词频的实体搜索算法,首先需要通过对结构化数据的分析来构造词频词典;然后使用词典对抽样数据进行分析;最后识别实体并定位实体位置。基于词频的实体搜索算法如算法1所示。

算法1 基于词频的实体搜索算法

```
输入:词组集合 G,待搜索字符串 V
输出:"false"或"true"
1. begin /*按词频制作词典*/
2. cut(G)→temp[] /*将集合进行 G 分词*/
3. count(temp[])→list[]/*统计词频*/
4. WHILE vari<n/*n 为常数*/
   /*取前 n 项作为词典*/
5. DO list [i]→dict[i]
6. END
   /*进行字符串匹配*/
7. compare(V,dict[])→res
8. RETURN res
9. END
```

算法1中,输入G为训练所使用的项目名称集合,V为从关系数据库中获取的一行待搜索的数据,输出res为对字符串V进行搜索后的结果。字符串匹配为包含关系,即词典中的词语出现在字符串V中即视为匹配成功,并停止匹配。

定位信息后需要进行信息抽取工作,具体规则如规则1所示。

规则1 将数据库中所有需要提取的表集合记为T,定义由表A至实体的映射f,对于表A中的一行数据$Data(a)$,a为表A中主键的一个数值,有$Tablename(A)\rightarrow Entitylabel(X)$,$A\in T$,$Tablename(A)$为$A$所在的数据表名称,$Entitylabel(X)$为构建实体的标签,$X$为需要构建的一个实体;有

$Data(a) \rightarrow Entity(X)$, $a \in A$, $Entity(X)$为 a 所对应的实体，将 Data(a)所含数据完整映射至 $Entity(X)$实体中。数据抽取规则如图 2 所示。

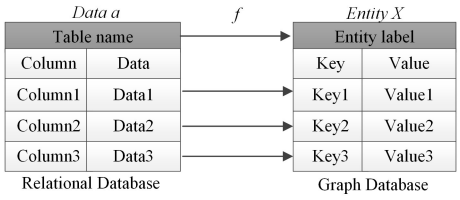

图 2 数据抽取规则

Fig. 2 Data extraction rules

在构建关系实例的过程中，结合已有文献[16-18]，采用基于预定义的关系模板进行数据匹配的方法，对提取实体的数据表中的数据进行数据匹配。关系来源于数据库中基于已有数据抽取制作而成的关系词典。目前定义的语义关系共有 6 种，即参与人员、项目联系人、项目负责人、项目专家组、项目首席以及承担单位，其主要用于描述实体间的主要关系，是关系实例完整表达的重要组成部分。构建关联关系过程中对照关系词典来完整描述关系模型，实现实体间的关联。关系中的时间标签主要用于表述时间变化，构建关联实例时针对每条关系，添加创建时间。

在构建关系图谱的过程中，根据两实体中的所有属性进行数据匹配。使用表 1 所列数据构建的关系实例如图 3 所示(属性未在图中标出)，关系方向为实体搜索的方向。

表 1 实体数据示例

Table 1 Example of entity data

Project	Project field	Year	Start time	End time	Unit type	Project unit	Unit nature
Project1	field	2015	2015/1/1	2019/8/31	Commitmentunit	NCUT	College

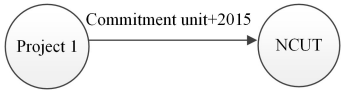

图 3 关系实例示例

Fig. 3 Example of relation instance

4 关系图谱分析方法

关系图谱分析的本质为基于图的遍历或数据抽取，对得到的数据进行挖掘与分析。利用关系图谱数据关联关系时分析效率高的特点，结合实际需求，提出了两类针对关系图谱的数据分析算法，即关联规则挖掘和数据筛选。

4.1 关联规则挖掘

关联规则挖掘计算的数据基础为关系图谱数据，其目的在于解决针对数据源由关系数据变为图数据后带来的频繁项挖掘问题。例如，对于团队关系分析问题，通过合理的项集分析，得到符合需求的团队分析结果。

首先进行图谱搜索，并进行筛选。在图数据库中，由实体 E_1 进行广度搜索，设搜索深度 Δd 为 2，得到图谱集合 G 中全部与 E_1 有关联的实体及关系集合 V，V 集合中每行包含 3 个实体的标签及属性信息和 2 个关系的全部信息，集合 V 的结构如表 2 所列。定义实体标签序列 E，序列为“标签 1，标签 2”。本文设计的实体及关系筛选算法如算法 2 所示，使用定义的实体标签序列对所得到的实体及关系进行剪枝、筛选，其中实体包含自身的属性。频繁项挖掘流程如图 4 所示。

表 2 集合 V 的结构示例

Table 2 Example of collection V structure

Entity	Relation	Entity	Relation	Entity
E_1	R_1	E_2	R_2	E_3
E_1	R_1	E_2	R_3	E_4
…	…	…	…	…

算法 2 实体及关系筛选算法

输入：搜索得到的实体及关系集合 V，实体标签序列 E

输出：筛选后的集合 V

```
1. begin
/ * 比对是否符合定义序列 * /
2. WHILE vari<length(V)/ * 行数 * /
3. DO IF V[i][3]≠E[1] THEN
4. DELETE(V[i])/ * 不符合序列则删除 * /
5. END
6. ELSE IF V[i][5]≠E[2] THEN
7. DELETE(V[i])/ * 不符合序列则删除 * /
8. END
9.  END
10. RETURN V
11. end
```

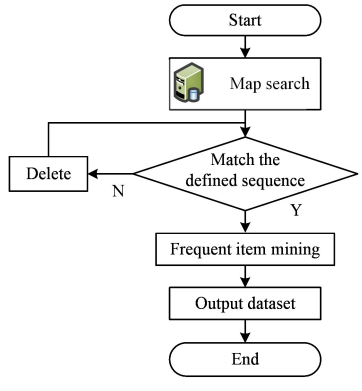

图 4 频繁项分析流程图

Fig. 4 Frequent item analysis flowchart

筛选后的集合中保留的项即为数据挖掘的输入数据。按照科技数据关系图谱的构建规则，将数据筛选后，则得到与实体 E_1 相关且符合定义序列的关系。按照 E_2 实体唯一标识和其他实体唯一标识的分类方法进行数据的汇总，结果如表 3 所列。

表 3 频繁项挖掘源数据的示例

Table 3 Example of frequent item mining source data

Object	Item
P1	I1, I2, I5
P2	I2, I4
…	…

对表 3 所列的数据进行团队关系分析。针对上文所述的科技数据，本文对 FP-growth 算法的数据挖掘部分进行了一定的改进，提出了 OFP-growth 算法。本文中对人员进行频繁项挖掘的数据中人员的重复性较低，因此构造得到的 FP-Tree 中单子树数据较多。由于传统算法递归调用耗时较长，因此在分析算法后对单路径情况进行特殊讨论。OFP-growth 的具体步骤如算法 3 所示。

1）扫描一遍事物数据库 TD，对于每项计数，去掉不满足 MinSup 的项。将项集表 L 中的数据项按照支持度计数递减排序。

2）第二次扫描事物数据库 TD，构建 FP-Tree，建立根节点 root，将 TD 中的事物插入。若节点不存在，则新增节点并设置计数为 1；若节点已存在，则计数加 1。

3）进行频繁项集挖掘时，判断此 FP-Tree 是否为单子树，若是则不进行递归调用，并通过对每个项进行组合得到结果。

算法 3 OFP-growth 算法

输入：事物数据集 data，最小支持度 minsup

输出：频繁项集合 F

```
1. begin
2. WHILE vari<data
3. DO count(L)/对每项计数 * /
4. DELETE(TD,minsip)/ * 删除不满足 minsup 的项 * /
5. END
6. list_ high2low(L) →L/ * 排序 * /
7. WHILE vari<data
8. DO buile_FP-Tree()→fp-tree/ * 构建 fp-tree * /
9. END
  / * 如果是单子树 * /
10. IF FP-Tree_only(i) THEN
11. combine(fp-tree,minsup,L) →F
12. ELSE select_tree(fp-tree) →F
13. RETURN F
14. end
```

4.2 数据筛选

数据筛选是根据定制需求，通过对图谱中所展示的关系进行分析以及对数据进行对比，对符合条件的实体进行筛选，最终通过算法给出最为合适的抽取意见，如在专家抽取过程中，通过合理的图谱分析，给出最合理的专家抽取意见。

数据筛选算法的基本原理是基于图的广度遍历，将通过遍历得到的关系、属性结果进行整合。在设计数据筛选算法时，本文结合已构建的关系图谱及分析需求两方面因素，针对性地设计了筛选策略，其中包括属性重复筛选策略以及数据关联筛选策略。

在筛选前，本文定义了针对数据的评分矩阵 ***dist***，将所有待评价数据进行汇总，将唯一识别标识添加至 ***dist*** 矩阵中，用于评价计分；同时，引入时间标签，将最后更新时间计算至评分中。

针对属性重复的比较策略进行实体属性的比较。属性中的相似性比对文本均为短文本，采用编辑距离进行文本相似性判别，具体公式如下：

$$Edit(i,j)=\begin{cases} i, & j=0 \\ j, & i=0 \\ \min\begin{cases} Edit(i-1,j)+1 \\ Edit(i,j-1)+1 \\ Edit(i-1,j-1)+w_{sub}(a_i,b_j) \end{cases}, & \text{other} \end{cases}$$

针对属性相似性对比的需求，获取待筛选的集合 e 以及需要进行数据对比的集合 m。定义属性重复标记算法如算法 4 所示。

算法 4 属性重复标记算法

输入：比对集合 m，待筛选集合 e

输出：筛选后的集合 e，评分矩阵 **dist**

```
1. begin / * 标记属性重复的项 * /
2. WHILE vari<length(m)/ * 行数 * /
3. DO WHILE var j<length(e)
4. DO IF Edit(m[i],e[j])==0
5. 0→dist[i][1] / * 可自定义分数 * /
6. END
7. END
8. END
9. RETURN e,dist
10. end
```

针对图谱中相关联情况筛选问题，输入需要比对的数据以及待筛选的集合，定义算法 5。

算法 5 图谱关联标记算法

输入：比对集合 m，待筛选集合 e

输出：筛选后的集合 e，评分矩阵 **dist**

```
1. begin / * 标记图谱中有直接关联的项 * /
2. WHILE vari<length(m)/ * 行数 * /
3. 2→deep / * 深度可自定义 * /
4. / * 使用 m 的唯一标识进行广度遍历 * /
5. DO BFS(deep,m[i]) →temp
6. WHILE var j<length(e)
7. IF temp==j
8. 0→dist[i][2] / * 可自定义分数 * /
9. END
10. END
11. END
12. RETURN e,dist
13. end
```

定义评分标准公式模板 $Score(id)$ 如下：

$$Score(id)=w_1 dist[i][1]+w_2 dist[i][2]+\cdots+w_n dist[i][n], i\leqslant lenght(dist), n<lenght(dist[])$$

其中，i 小于 ***dist*** 矩阵的行数，n 小于 ***dist*** 矩阵的列数。

5 实验与分析

本文的实验环境如下：内存为 DDR3 6 GB，操作系统为 Windows7 64 位，编译环境为 Python3.6，图数据库为 Neo4j，关系数据库为 Oracle。

5.1 科技数据分析系统的处理架构

科技人员关系图谱的整体架构由 3 层逻辑层组成，即数据存储层、数据处理层以及服务层，如图 5 所示。数据存储层包含关系数据库及图数据库。关系数据库存储源数据，包括

人员、单位等结构化数据信息。图数据库存储处理后的关系型数据，包括实体、关系信息。

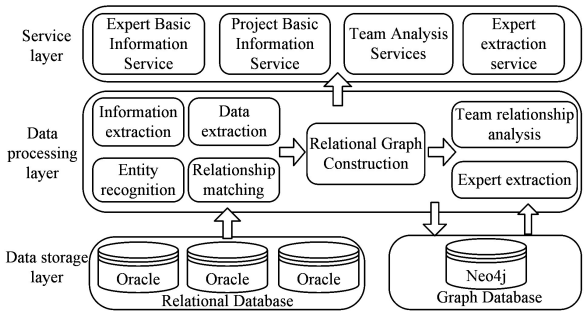

图 5 科技人员关系图谱的整体架构

Fig. 5 Overall structure of relational graph of technological data

数据处理层会对源数据进行分析、处理，并对关系图谱进行分析。对源数据的处理流程分为两大部分：信息抽取和图谱构建与演化。对关系图谱的数据分析包括团队分析和专家抽取。团队分析的结果最终在服务层中进行数据可视化展示；专家抽取则针对需要进行专家抽取服务的项目进行专家评估，综合评价专家的经验、领域以及各项回避规则后，按得分递减排序，再按序推荐所需数量的专家。

此架构的顶层为服务层，为各类应用系统提供服务支撑，包含专家基本信息查询服务、专家评审信息服务、项目基本信息查询服务及专家抽取服务等。服务层中专家抽取服务的数据支撑来源于数据分析层中对图数据库进行的分析。

5.2 关联规则挖掘算法实验与应用

本文首先针对改进的 FP-growth 算法进行实验验证，并与传统的 FP-growth 算法进行对比。实验数据来源于网络中的公开数据库[1)]。本次实验选择了 T10I4D100K 数据集作为测试数据。

在 T10I4D100K 数据集中，选取了 50000 种事物。因为事物中的支持度大多较低，数值均小于 1%，所以本文选取了数值为 0.5%～0.8%的最小支持度区间，最终结果如图 6 所示。

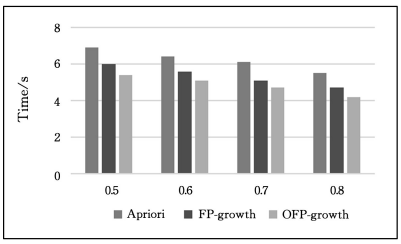

图 6 3 种算法在公开数据集中不同支持度结果的对比

Fig. 6 Comparison of results of three algorithms with different support levels based on public dataset

相比传统的 FP-growth 算法，改进后的算法在支持度较小时，运行效率有一定程度的提升，提升了约 9%～11%；相比 Apriori 算法，提升了 23%。

合作单位数据为结构化数据，按照算法 2 所述，从一个实体开始进行搜索，搜索后将符合筛选条件的数据按照 E_2 实体的唯一识别码进行分类，即可得到关联规则挖掘的初始数据。

针对合作单位的数据，选取了近 2000 条数据进行测试，选取支持度为 0.2%～0.4%的数据，同时进行了 10 次重复实验，以 10 次实验结果的平均值进行算法时间对比，结果如图 7 所示。相比传统的 FP-growth 算法，改进后的算法针对科技数据的运行效率大约提升了 10%～12%；相比 Apriori 算法，改进后的算法针对科技数据的运行效率提升得较为明显，约提升了 15%～22%。

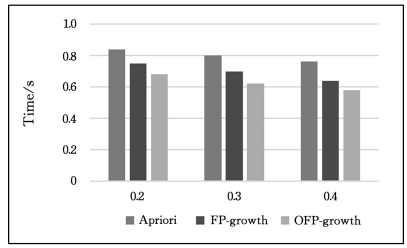

图 7 3 种算法在非公开数据集中不同支持度结果的对比

Fig. 7 Comparison of results of three algorithms with different support levels based on non-public dataset

当数据筛选模块最终应用于关系分析后，本文使用 Restful 风格将算法进行封装，并制定了统一的接口标准，提供团队分析的 Web 服务。

5.3 数据筛选算法的实验与应用

结合科技关系图谱的数据，将数据抽取算法应用于专家抽取，用以进行测试。通过对人员合作关系、单位关系等不同信息的综合评价，给出专家对本项目的综合评分，最终进行评审专家的推荐。

对专家进行评分的因素包括：单位关系、合作关系、职称与领域。其中，单位关系与合作关系有绝对否决权。评分矩阵 ***dist*** 用于存放专家单项评分数据，且对应专家列表中的每位专家，前两列的初始值为 1。评分矩阵的定义如表 4 所列。

表 4 评分矩阵的初始定义

Table 4 Initial definition of scoring matrix

ID	Score	Score
ID_1	1	1
ID_2	1	1
…	…	…

依次将算法 4、算法 5 用于同单位专家标记、有合作专家标记，并将被标记专家分数置为 0，通过算法 4 及算法 5 的处理后，将所有必须进行回避的专家进行标记，并在最终的专家评分中剔除被标记项。

针对专家领域（专业）相似性判别，将新申报项目中的“项目领域”项用于专家领域的相似性判别。结合算法 4 所使用的编辑距离，定义专家领域（专业）相似性判别过程中的评分标准如下：

$$Score\ edit=\begin{cases}2, & edit=0\\ \dfrac{1}{x}, & edit>0\end{cases}$$

在计算完上式后，将 *Score edit* 存入评分矩阵。

对高校教师职称、研究职称及工程系列职称对应评分进行划分，并将得分存至评分矩阵 ***dist*** 中。划分标准为：正高级及同级别职称为 1 分；副高级及同级职称为 0.7 分；中级及同级职称为 0.5 分；初级及同级职称为 0.3 分。

1) http://fimi.ua.ac.be/data

最后，将专家数据的最终更新时间计入评分中，若更新年份为同年则记 1 分，若为去年则记 0.9 分，依次类推至 0 分。

针对专家抽取过程，将评分标准公式 $Score(id)$ 修改为：

$$Score=\tanh((dist[i][1]+dist[i][2]+dist[i][3])(dist[i][4]+dist[i][5])),i\leqslant lenght(dist)$$

设定最低阈值，若评分均不合格则不进行抽取，若评分合格则最终按照专家评分由高到低进行抽取。

通过自动进行专家抽取、人工进行验证的方式，对实际数据进行测试，其中验证人员均为本实验室的研究生。此次测试以 20 个项目的数据为测试用例，模拟较高相似度的实体数据进行测试。在人工验证过程中，对照原有结构化数据，针对所抽取专家的正确性进行鉴别，如对领域不相同的专家被抽取、曾有合作的专家被抽取等数据抽取的判定标准进行判断。测试项目的专家抽取准确率如表 5 所列。

表 5 专家抽取结果的正确率

Table 5 Accuracy of expert extraction result

（单位：%）

	Accuracy	Control Group Accuracy
Real data	100	97
Experiment data	95	90

表 5 中，对比组不包含图数据的关联分析，仅通过查询原有结构化数据进行分析。最终，本文方法的准确率相比原有方法提升了 3.5%。

将数据抽取算法组合应用于专家抽取后，对专家抽取模块进行 Restful 封装，供 Web 端进行功能调用。

结束语 在数据关系复杂的环境中，本文将较难进行关系分析的结构化数据转化为关系数据，将数据表分析转化为图分析，有针对性地改进了原有的 FP-growth 算法，并提出了一种对图数据的关联规则的挖掘流程，结合图的遍历、剪枝，最终将该方法应用于团队分析实例。本文提出了基于关系图谱的数据筛选方法，该方法被应用于专家评分，最终实现了专家推荐。本文研究的关系图谱构建及分析方法在实验中得到了良好的效果，后续进行了服务的封装，并且在实际系统中进行了应用。

结合实际反馈，本文方法还有提升与优化的空间，如可以引入语义分析，提高图谱构建的效率与智能程度；继续优化数据挖掘算法，以提升数据挖掘的效率。

参 考 文 献

[1] XU F. Research on spam speech recognition based on user social relationship graph [D]. Wuhan: Huazhong University of Science and Technology, 2017.

[2] AMIT S. Introducing the Knowledge Graph: Things, Not Strings, Official Blog [OL]. [2019-06-14]. http://googleblog.blogspot.co.uk/.

[3] TANG Y, CHEN G H, HE C B, et al. Knowledge Map and Its Application in the Field of Academic Information Services [J]. Journal of South China Normal University (Natural Science Edition), 2018, 50(5): 110-119.

[4] LING X, WELD D S. Fine-grained entity recognition [C]// Proc of the 26th Conf on Association for the Advancement of Artificial Intelligence. Menlo Park, CA: AAAI, 2012: 94-100.

[5] YIN L, YUAN F, XIE W B, et al. Research Progress and Challenges of Correlation Maps [J]. Computer Science, 2018, 45(S1): 1-10, 35.

[6] JIANG B C, WAN G, XU J, et al. Construction of large-scale geographic knowledge maps of multi-source heterogeneous data [J]. Journal of Surveying and Mapping, 2018, 47(8): 1051-1061.

[7] YAN J H, WANG C Y, CHENG W L, et al. A retrospective of knowledge graphs [J]. Frontiers of Computer Science, 2018, 12(1): 55-74.

[8] NATHAN E, BADER D A. Incrementally updating Katz centrality in dynamic graphs (Article) [J]. Social Network Analysis and Mining, 2018, 8(1): 1-26.

[9] LI X, TUR G, HAKKANI-TUR D, et al. Personal knowledge graph population from user utterances in conversational understanding [C]// Spoken Language Technology Workshop. IEEE, 2015.

[10] YU J, LIU Y B, ZHANG Y, et al. Overview of Large Scale Graph Data Matching Technology [J]. Journal of Computer Research and Development, 2015, 52(2): 391-409.

[11] ZHANG L X, WANG W P, GAO J L, et al. Incremental Graph Pattern Matching for Pattern Graph Changes [J]. Journal of Software, 2015, 26(11): 2964-2980.

[12] GUAN J, WANG W, QI Q H. Multi-Keyword Streaming Parallel Retrieval Algorithm Based on Urban Security Knowledge Map [J]. Computer Science, 2019, 46(2): 35-41.

[13] SUN W P, CHANG L, BIN C Z, et al. Recommendations of Tourism Routes Based on Knowledge Mapping and Frequent Sequence Mining [J]. Computer Science, 2019, 46(2): 56-61.

[14] ZHAO Z B, JIA Y F, YAO L, et al. Research on Web Page Classification Technology with Rich Structured Data [J]. Journal of Computer Research and Development, 2013, 50(S1): 53-60.

[15] ZHANG Y, JIA Y D, FU L Y, et al. AceMap Academic Map and AceKG Academic Knowledge Atlas——Visualization of Academic Data [J]. Journal of Shanghai Jiaotong University, 2018, 52(10): 1357-1362.

[16] ZHENG W G, CHENG H Y, XU J, et al. Interactive natural language question answering over knowledge graphs [J]. Information Sciences, 2019, 481: 141-159.

[17] SHI D X, LI H, YANG R S, et al. Excavation of daily frequent behavior patterns of users [J]. Journal of National University of Defense Technology, 2017, 39(1): 74-80.

[18] FADER A, SODERLAND S, ETZIONI O. Identifying relations for Open information extraction [C]// Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1535-1545.

ZHANG Han-shuo, born in 1994, postgraduate. His main research interests include service computing, cloud computing and big data.

YANG Dong-ju, born in 1975, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include service computing, data integration, cloud computing, cloud storage, and their applications in industry data center.