异构信息空间中支持多模态融合实体搜索的多层时态数据模型

杨 丹[1] 陈 默[2] 孙良旭[1] 王 刚[1]

(辽宁科技大学软件学院 鞍山 114051)[1] (东北大学计算中心 沈阳 110819)[2]

摘 要 面对异构信息空间中具有时间信息的大量相互关联的异构实体数据如作者、论文、产品、电影等,提出一个以实体及关联关系为中心的多层的时态数据模型,即多层的时态实体关联网络MTE-Network,它能有效捕捉异构实体和关联关系的时间信息。基于此时态数据模型,提出了实体搜索的多模态融合的查询模型,其支持用户搜索异构信息空间中的任何类型的实体及相关实体,支持在实体级、实体聚类级和时间轴上的实体搜索,并且满足用户多模态融合实体搜索的信息需求。在真实数据集上的实验结果证明了该时态数据模型和查询模型的可行性和有效性。

关键词 异构信息空间,多模态融合,实体搜索,时态数据模型

中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.029

Multi-layer Temporal Data Model Supporting Multi-modality Fusion Entity Search in Heterogeneous Information Spaces

YANG Dan[1] CHEN Mo[2] SUN Liang-xu[1] WANG Gang[1]

(Software College, University of Science and Technology Liaoning, Anshan 114051, China)[1]

(Computing Center, Northeastern University, Shenyang 110819, China)[2]

Abstract Faced with a large number of associated heterogeneous entities such as projects, authors, papers, products, movies with temporal information in heterogeneous information spaces, an entity and association centered multi-layer temporal data model, named multi-layer temporal entity network MTE-Network, was proposed which captures temporal information of entities and associations effectively. Moreover, multi-modality query model of entity search was proposed which supports users to search entities and related entities of any entity class in heterogeneous information spaces, and supports entity search on entity level, entity aggregation level and time line, and it can satisfy the information needs of users' multi-modality entity search. The experimental results on real data set demonstrate the feasibility and effectiveness of the proposed temporal data model and query model.

Keywords Heterogeneous information spaces, Multi-modality fusion, Entity search, Temporal data model

1 引言

异构信息空间中的实体以及实体关联关系通常具有时间信息,如某个总统的任职时间、某两个作者的合作时间等,并且这些实体不是保持静态的,它们随时间演化,如合并、分裂、被添加或删除。而已存在的以实体为中心的数据模型缺少对实体、实体关联关系的时间信息的描述。若不考虑实体、关联关系的时间维度,则:(1)当实体、关联关系发生变化时,必须创建新的元数据来捕捉这种变化;(2)无法支持各种时间轴上的实体搜索,如对实体过去状态的查询请求。

实体间关联关系是进行语义实体搜索的基础和前提条件,并且在异构信息空间中人们经常使用一种类型的实体来查找另外一种实体类型的实体,即利用不同实体类型实体间的关联关系语义线索进行多模态融合的实体搜索。例如,在社交媒体网络中,某个用户想要找兴趣相同或相似的人来做朋友推荐,他/她可能通过查找照片网络中的照片实体来找到与自己有相同摄影爱好并且喜欢风景摄影的人作为朋友。也就是说,各种实体类型的实体搜索是密切相关的,前提是如何建模这些异构实体及它们间的关联关系来支持多模态融合的实体搜索。已有的典型的异构信息空间数据模型如异构信息网络[1]中虽然将链(link)即实体关联关系的类型区分为异构的和同构的,但是并没有区分关联关系的不同粒度,只关注单个实体间的关联关系,缺少异构实体集合间即实体聚类级的关联描述。

本文针对异构信息空间中实体数据和实体搜索的特征,提出以实体及关联关系为中心的多层时态数据模型,即多层的时态实体关联网络MTE-Network(Multi-layer Temporal Entity Networks)和实体搜索的多模态融合的查询模型。该

到稿日期:2014-06-11 返修日期:2014-09-27 本文受国家自然科学基金项目(61402213,61402093),中央高校基本科研业务费专项资金项目(N120316001),辽宁省教育厅科学研究项目(L2013120)资助。

杨 丹(1978-),女,博士,副教授,CCF会员,主要研究方向为数据集成、数据空间,E-mail:asyangdan@163.com;陈 默(1983-),女,博士,讲师,CCF会员,主要研究方向为空间数据处理;孙良旭(1979-),男,硕士,副教授,主要研究方向为数据集成;王 刚(1978-),男,博士,副教授,主要研究方向为Web数据处理。

时态数据模型不仅能有效地描述实体本身特征，还能捕捉实体间语义关联关系及实体聚类级关联关系。

2 相关工作

时态数据模型(temporal data model)的研究由来已久，早期的相关工作主要集中在将关系数据库中的实体关系(ER)模型扩展成为时态 ER 模型[2,3]，以及对半结构化文档 XML 的时态 XML 模型[4]研究。随着语义 Web 的发展，出现了针对 RDF 图模型，将其扩展成为时态 RDF 图模型[5-7]的研究。此外，近年来随着维基百科等在线知识库的构建和发展，提出了获取知识的同时要考虑时间维度即时间的知识模型，如文献[8]使用维基百科信息框(infoboxes)中的正则表达式来获取时间的事实(facts)。PRAVDA[9]使用一种结合文本的模式和基于图的排序技术来获取时间的事实。文献[10]基于时间特性的约束和弱监督算法，提出一个联合推理框架来对知识库中的事实时间范围化。本文提出的异构信息空间中的时态数据模型与时态数据库(temporal database)不同，允许一个实体(对应于数据库中的一个元组)的不同属性值具有不同的时间戳。时态数据库一般采用增加一个或两个列来存储有效时间或事务时间，因此在时态数据库中，每一行记录可以是历史的或当前的。而针对异构信息空间中实体数据的特点，本文提出的时态数据模型允许一个实体(对应于时态数据库中的一个元组)的不同属性值具有不同的时效性，即一个实体可以包含具有不同时间戳的属性(值)。

与本文研究背景相似的是在个人数据空间、数据空间中的数据模型相关工作，iDM[11]数据模型将异构的个人信息作为一个单独的资源视图(resource view graph)，这个图代表了一个用户的整个数据空间。文献[12]提出用一个称作三元组库(triple base)的三元组集合来建模来自不同的数据资源的数据。每个三元组的形式是(实例，属性，值)或(实例，关联，实例)，因此一个三元组库描述了实例和关联关系的集合。$lgDM$[13]以实体作为基本的数据单位，提出由实体关联数据图 G_D 和模式图 G_S 组成的分层的图模型。但是上述相关工作没有考虑实体、关联关系的时间信息，并且忽略了异构信息空间中实体聚类级的关联关系。

3 多层的时态数据模型

3.1 相关定义

异构信息空间中的时间信息无处不在，如文档的时间元信息(文档的创建时间、修改时间)、文档内容中与实体相关的时间上下文、时间记录(temporal record)的元组时间戳和属性时间戳、类型为时间型的实体属性的值如出生日期、论文发表年代等。因此，异构信息空间可被看作是一个由不同实体类型的具有时间信息的实体组成的实体库(entity repository)。

定义 1(具有时间信息的实体，简称时态实体 Temporal Entity)　异构信息空间中的具有时间信息的实体表示为四元组：$e=\langle id, label, A, E_{type}\rangle$，其中，$id$ 表示实体的唯一标识；$label$ 是实体的名称；E_{type} 表示实体 e 所隶属的实体类型；A 表示实体属性(值)集合，即 $A=\{(type_1, attribute_1, v_1, t_1), \cdots, (type_n, attribute_n, v_n, t_n)\}$，每个实体对应一组类型是 $type_1-type_n$ 的属性模式 $attribute_1-attribute_n$；v_1-v_n 是这些属性在 t_1-t_n 时刻的属性值，其中时间戳 t_1-t_n 可以相同也可以不同。

实体的同一个属性在不同时间戳下的属性值 $v_1, v_2, \cdots, v_n$ 序列体现了实体的演化性。与实体相关联的最小和最大时间戳组成的时间段[Min(t)，Max(t)]表示了信息空间中该实体的存在时间间隔。

定义 2(具有时间信息的关联关系，简称时态关联关系 Temporal Association)　时态关联关系表示为四元组 $ass=\langle R_{type}, S_{ource}, T_{arget}, Label\rangle$，其中，$R_{type}$ 表示关联关系所属的关系类型；S_{ource}，T_{arget} 分别是构成二元关联关系的两个端点；$Label$ 是关联关系标签，由关联关系名 $assName$、关联强度值 s、约束集合 $Constraints$ 和时间信息@T 组成，即 $Label::=assName, s, [Constraints=\{c_1, c_2, \cdots, c_n\}], [@T]$，其中约束集合和时间信息是可选项；关联关系的时间信息可以是一个时间点(t)，也可以是一个时间间隔(ts)，即 $T\in\{t, ts\}$。

定义 3(时态约束，Temporal Constraint)　时态实体或时态关联关系必须满足时间语义上的约束，如实体属性值的取值范围限制、关联关系(间)的特定时间限制等。

时态关联关系的时态约束可以作用在单个的关联关系上，如美国总统一届的任期是 4 年；也可以是作用在关联关系之间上，如奥巴马任美国总统与奥巴马是参议员两个关联关系在时间上是互斥的。后一种时态约束表示了时态关联关系间的参照完整性。一个关联关系间的时态约束的例子如图 1 所示，其中实体间带箭头的线表示实体间的时态关联关系；虚线表示时态关联关系间的时态约束。

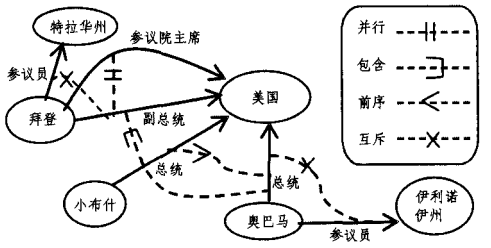


图 1　关联关系的时态约束例子

定义 4(组，group)　由多个同类型实体组成的语义关联的实体集合，表示语义相似或用途相同的多个同构实体组成的集合，即 $group_i=\{e_1, e_2, \cdots, e_n\}$。

组是实体聚类级的组织单位。组的生成可依据聚类算法得到的聚类簇集 $C_1, C_2, \cdots, C_n$，其中每个簇中的实体为一组。

定义 5(多层的时态实体关联网络，MTE-Network)　由 $n(n\geqslant 1)$ 层实体关联网络组成，即 $MTE\text{-}Network=\{N_k: k=1,2,\cdots,n\}$，其中 k 是实体类型的 id。而每层的实体关联网络 N_k 由 $m(m\geqslant 1)$ 个相互关联的组(group)组成，即 $N_k=\{G_k, Edge_k, L_G, L_E\}$，其中 V_k、$Edge_k$ 分别是顶点(组)集合、边(组间关联)集合；L_G、L_E 分别是顶点和边上的标签集合。而每个组中的实体、实体关联关系组成了实体关联图，即 $G_k=\{E_k, A_k, L_V, L_E\}$，其中，$E_k$ 和 A_k 分别是组中的顶点(实体)集合、边(实体间关联)集合；L_E、L_A 分别是顶点和边上的标签集合。

按照实体类型不同将异构实体及其关联关系建模为 n 层的实体关联网络。一种同构实体为一层，在每层中采用基于实体级关联和实体聚类级关联相结合的组织模型。图 2 给出了一个由作者、论文和会议 3 种类型的实体组成的多层的时态数据模型例子。在图 2 中，虚线椭圆表示 1 个组，如作者组、论文组和会议组。

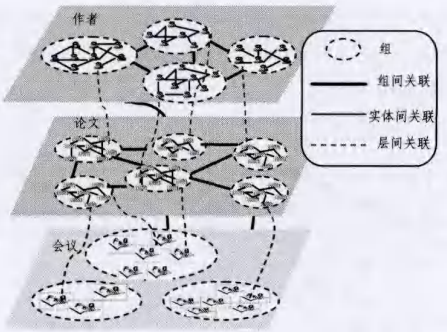


图 2　多层的时态数据模型 MTE-Networks 的例子

3.2　时间信息的捕捉

捕捉实体及关联关系的时间信息可以帮助我们在时间轴上分析这些实体，进而进行时间感知的实体搜索。例如，查询“Halevy's papers between 2010 and 2013”、查询“Halevy 's current job title”等。在 MTE-Neworks 数据模型中实体的属性值、关联关系被赋予一个时间间隔$[t_{start}, t_{end}]$来表示它们在某个开始时间 t_{start} 和结束时间 t_{end} 之内一直延续存在；或用瞬时的时间点 t 来表明它们在时间点上的存在。而未发生的、将来的未知时间值或一直持续的时间值则设置成一个开放式的固定值 Until_Changed。例如，某作者实体从 2010 年至今是副教授的时间间隔表示为[2009，Until_Changed]。

3.3　多粒度、多元化的关联关系

针对异构信息空间中的关联关系具有不同种类即异构、不同粒度的特点，将多层的时态实体关联网络 MTE-Network 中的链即关联关系分为如下几种。

(1)组内关联(intra-group link)，描述一个组内同构实体之间的同构或异构关联关系，如图 1 中较细实线表示的多个作者之间的合作关联关系、论文之间的相同主题和会议之间的相同领域关联关系等。

(2)组间关联(inter-group link)，描述同层的或不同层的组之间的聚类级关联关系。一组实体与另一组实体组之间的关联关系表示了聚类级实体间的关联关系。例如，研究机器学习的一组作者和研究数据库的一组作者之间的关联关系；在 VLDB 会议上发表论文的一组作者与在 SIGMOD 会议上发表论文的一组作者之间的关联关系。如图 2 中较粗实线表示的同层的作者组间、论文组间和会议组间的关联关系，不同层间的作者组、论文组间的关联关系，以及论文组、会议组间的关联关系。

(3)层间关联(inter-layer link)，描述不同层的异构实体之间的关联关系。如图 2 中虚线表示的作者、论文和会议等不同实体类的实体间的关联关系。

其中，同一层上的链描述同构实体级和同构实体聚类级的各种关联关系。而层与层之间的链描述异构实体级和异构聚类级实体间的关联关系。

4　多模态融合的查询模型

本文提出的多层的时态实体关联网络 MTE-Network，能够支持如下的多种查询模型进行语义实体搜索。

4.1　实体关联查询

实体关联查询(entity associated query，eaq)是具有如下语义的查询：eaq：$\forall e$：$pq_1\ ass_1\ pq_2 \cdots ass_n\ pq_n \wedge ass_i = k_i$。

其中，k_i 是查询关键字；pq_i 表示一个谓词查询；ass_i 是实体间的关联关系。例如，用户输入关键字查询“social network，SIGMOD2013”想要查询发表在会议 SIGMOD2013 上的关于“social network”的所有论文。

4.2　相关实体查询

相关实体查询(related entity query，req)是具有如下语义的查询：req：$\forall e, k_i \rightarrow e_i \wedge e\ ass_i\ e_i \wedge ass_i \in \{$intra-group link，inter-group link$\}$。

其中，k_i 是查询关键字并且映射到某个实体 e_i；ass_i 是实体 e_i 参与的任意关联关系，即由实体 e_i 作为顶点构成的 intra-group 链或 inter-group 链。相关实体查询的结果是返回同一层实体网络中所有与 e_i 相关的实体集合。例如，用户输入关键字查询“Halevy”来查询作者 Halevy 相关的作者实体，如 Halevy 的论文和作者、研究方向相同的其他作者等。

4.3　多模态融合查询

多模态融合查询(multi-modality fusion query，mmq)是具有如下语义的查询：mmq：$\forall e, k_i \rightarrow e_i \wedge e\ ass_i\ e_i \wedge ass_i \in \{$inter-layer link，inter-group link，intra-group link $\} \wedge Type(e) \neq Type(e_i)$。

其中，k_i 是查询关键字并且映射到某种类型的实体 e_i；ass_i 是实体 e_i 参与的层与层实体间的任意关联关系，即由实体 e_i 作为顶点构成的 inter-layer 链、intra-group 链或 inter-group 链；实体 e 和实体 e_i 所属的实体类型不同。多模态融合实体查询的结果是返回位于不同层实体网络上的所有相关实体集合。例如，研究社会网络的用户想要搜索作者实体进行论文审稿人(reviewer)推荐，他/她可能通过在论文实体中查询“social network”，找到题目具有“social network”关键字或主题是“social network”的论文，发表这些论文的作者可能是用户感兴趣的作者。

4.4　组查询

组查询(group query，gq)是具有如下语义的查询：gq：$\forall group$：$k_i \rightarrow g_i \wedge ass_i \in \{$inter-group link$\}$。

其中，k_i 是查询关键字并且映射到某个组 g_i，ass_i 是组 g_i 参与的任意组间的关联关系，即 inter-group 链。组查询的结果是满足查询条件的所有组集合。例如，用户想要搜索研究数据空间方向的所有作者实体。

4.5　关联关系查询

关联关系查询(association query，aq)是具有如下语义的查询：aq：$\forall ass$：$k_i \rightarrow e_i$，$k_j \rightarrow e_j \wedge\ e_i\ as_i\ e_j$。

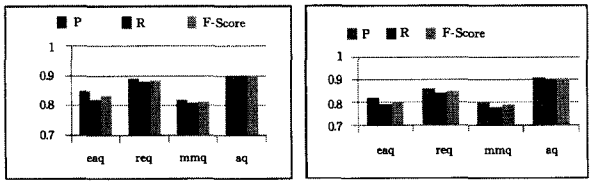
其中，k_i 和 k_j 是查询关键字并且分别映射到两个实体 e_i、e_j，ass_i 是实体 e_i、e_j 间的任意关联关系。关联关系查询的结果是这两个实体间的所有可能关联关系集合。

5　实验评价

实验所使用机器的配置是 3.16 GHz Pentium 4 CPU、4GB内存。实验数据集基于两个真实数据集 DBLP[1] 和 IMDB[2] 进行扩展。DBLP 数据集包括论文(P)、作者(A)、会议(V)和主题(T)4 种实体类；IMDB 数据集包括电影(M)、演员(A)、导演(D)、电影制片厂(S)4 种实体类。两个数据集都包括组内关联、组间关联和层间关联 3 种类型的关联关系。实验查询集包括实体关联查询(eaq)、相关实体查询(req)、多模态融合查询(mmq)和关联关系查询(aq)4 种类型的关键字查询。

查询的有效性：实验评价查询的平均准确率(P)、召回率(R)和 F 值(F-Score)。实验结果如图 3 所示。从图 3 可知，

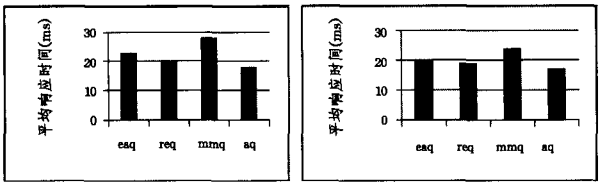
mmq 类型的查询平均准确率最低在 80%左右，其次是 *eaq* 类型的查询，其准确率在 83%左右，而 *req* 和 *aq* 类型的查询准确率在 86%到 91%之间。



(a)DBLP (b)IMDB

图 3 查询的平均准确率、召回率和 F-值

查询的响应时间：各种类型的查询平均响应时间的实验结果如图 4 所示。从图 4 可知，在这些查询种类中 *mmq* 查询的平均响应时间最长，其次是 *eaq*，*req* 和 *aq* 查询的平均响应时间较短。



(a)DBLP (b)IMDB

图 4 平均查询响应时间

使用 DBLP 数据集实验比较不同数量层(即实体网络层数)对多模态融合实体查询准确率的影响。比较以下 5 种实体关联网络下查找作者实体的平均准确率、召回率和 F 值。实验结果如图 5 所示。

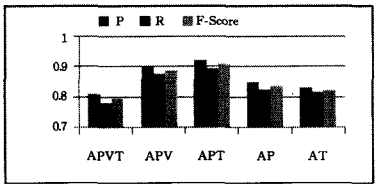
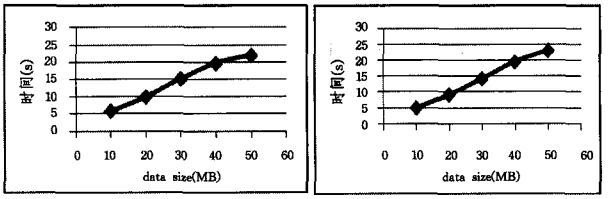


图 5 不同层数的网络多模态融合查询比较

- APVT：包含作者、论文、会议和主题 4 层实体关联网络；
- APV：包含作者、论文和会议 3 层实体关联网络；
- APT：包含作者、论文和主题 3 层实体关联网络；
- AP：包含作者和论文 2 层实体关联网络；
- AT：包含作者和主题 2 层实体关联网络。

从图 5 可知，3 层实体关联网络的平均准确率要高于 2 层实体关联网络的平均准确率；4 层实体关联网络 APVT 的准确率要低于 3 层实体关联网络的准确率，由于 4 层实体关联网络的情况下在引入了更多关联信息的同时也引入了不必要的噪音，因此对查询带来了负面的影响。此外，注意到 3 层实体关联网络 APT 的平均准确率要高于 APV 的平均准确率，说明在搜索作者实体时主题信息要比会议信息重要。

MTE-Network 的可伸缩性：实验比较不同数据量(10M～50M)下时态数据模型的索引(包括实体类属性索引、关联关系索引和属性值倒排索引)建立时间，如图 6 所示。从图 6 可以看出，在两个数据集上随着数据量的增大，总体执行时间缓步增长，呈现 sub-linear 增长的形势，因此 MTE-Network 具有较好的可伸缩性。



(a)DBLP (b)IMDB

图 6 多层的时态实体关联网络的可伸缩性

结束语 本文提出了异构信息空间中以实体和关联关系为中心的多层的时态数据模型 MTE-Network 和多模态融合的语义实体搜索查询模型。MTE-Network 能够捕捉时态实体间以及聚类级时态实体间的时态关联关系，有效地为异构信息空间中的实体搜索提供语义支持。实验结果表明了所提出的时态数据模型的可行性和有效性。

参 考 文 献

[1] Han Jia-wei. Mining heterogeneous information networks by exploring the power of links[C]// Conferernce of Algorithmic Learning Theory. 2009:3-30

[2] Gregersen H, Jensen C S. Temporal entity-relationship models-a survey[J]. IEEE Transaction on Knowledge and Data Engineering, 1999, 11(3):464-497

[3] Bettini C. Review-temporal entity-relationship models-a survey [OL]. http://dblp.nui_trier.de/db/journals/dr/Bettingiqq.html

[4] Rizzolo F, Vaisman A A. Temporal XML: modeling, indexing, and query processing[J]. VLDB Journal, 2008, 17(5):1179-1212

[5] Gutierrez C, Hurtado C A, Vaisman A A. Introducing time into rdf[J]. IEEE Trans. Knowl. Data Eng., 2007, 19(2):207-218

[6] Pugliese A, Udrea O, Subrahmanian V S. Scaling RDFwith time [C]// Proc. of WWW. 2008:605-614

[7] Guti′errez C, Hurtado C A, Vaisman A A. Temporal RDF[C]// Proc. of ESWC. 2005:93-107

[8] Wang Y, Zhu M, Qu L, et al. Timely Yago: harvesting, querying, and visualizing temporal knowledge from Wikipedia[C]// Proc. of EDBT. 2010:697-700

[9] Wang Y, Yang B, Qu L, et al. Harvesting facts from textual web sources by constrained label propagation[C]// Proc. of CIKM. 2011

[10] Partha Pratim T, Wijaya D, Mitchell T. Coupled temporal scoping of relational facts[C]// Proc. of WSDM. 2012

[11] Dittrich J P, Antonio M, Vaz Salles, et al. iDM: A Unified and Versatile Data Model for Personal Dataspace Management[C]// Proc. of VLDB conference. 2006, 367-378

[12] Sarma A D, Dong Xin-luna, Halevy A Y. Data Modeling in Dataspace Support Platforms[C]// Proc. of Conceptual Modeling: Foundations and Applications. 2009:122-138

[13] 杨丹，申德荣，聂铁铮，等. 数据空间中数据模型及实体关联关系挖掘的研究[J]. 小型微型计算机系统，2012，33(5)：936-939

[1] http://dblp.uni-trier.de/xml/

[2] http://grouplens.org/datasets/movielens/