第23卷 第3期 石家庄铁道大学学报(自然科学版) Vol. 23 No. 3
2010年9月 JOURNAL OF SHIJIAZHUANG TIEDAO UNIVERSITY (NATURAL SCIENCE) Sep. 2010

工程软件的小世界效应探究

赵正旭， 龙 瑞， 郭 阳， 刘贾贾

(石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043)

摘要:工程信息具有形式复杂和内容分散的特点,其结构的关联性和数据的兼容性直接影响着数字化工程信息资源的有效管理和充分利用。特别是在工程信息的长期安全保存和可靠调用过程中,存在着信息和其使用系统之间的兼容问题和对信息定期检查、移存、转化的时间和效率问题。目前国内外对这两类问题还没有可行的解决方法和技术。根据"六度分离"现象和小世界网络模型,对工程软件的输入输出格式之间的转换关系进行实验分析,验证了其小世界效应,进一步通过 Protégé 构建的本体查询格式转换的路径,为解决大量级和分散性工程信息的移存提供了可靠的科学依据。

关键词:六度分离;小世界效应;数据移存;信息保护

中图分类号:TP393 **文献标识码**: A **文章编号**: 2095-0373(2010)03-0001-06

0 引言

随着计算机和网络技术的飞速发展,信息量的增长远远超过了人们对信息获取的增长。根据文献[1]、文献[2]介绍,全世界每年约新增加 5 Exabyte(即 5×1018 byte)量级的信息。原始和无序的信息不但不产生价值,反而会加剧信息增长和利用之间的矛盾,甚至会造成信息超载而知识缺乏的社会困局。这不但影响着一个国家经济发展的潜力和速度,而且也关系着一个国家信息资源的安全。尤其在工程领域,如汽车、造船、航空、军事科技等领域的设计和制造,研究如何长期(不短于产品的寿命期,甚至长达50多a)保存产品的原始设计数据和生产过程的工程信息,以满足产品在其寿命期内的维修、故障排除、改型等需求。目前通用的保存方法主要是基于传统图纸形式的电子文档、微缩胶片、穿孔卡片等。这种方法文档容易老化或损坏,不便于在网络上调用和传输,无法记录和保留后续对产品改进的信息。比较先进的方法是基于 STEP ISO 10303 标准,把设计和制造信息,包括相关维修、故障排除、改型等信息都融合到三维 CAD 模型中[3]。可是,因为生成这些信息的系统如 CAD,CAE,CAM,PDM 等的寿命一般不超过10 a,比多数产品的寿命要短的多,即多年以后生成这些信息的系统多数将不复存在,所以今后要利用这些信息将遇到数据和系统之间的兼容问题。第二个趋势是以 ASD-STAN 的 LOTAR 研究计划为代表,主要考虑如何应付企业管理、产品设计及其生产过程汇总的信息量及其分散性的增加,研究对工程信息的生成方法以及对数字化工程信息实行长期保存并保证数据的可靠性,以满足今后不同时期和不同用户的多方面需求。其研究重点是产品的三维几何数据和结构信息的生成方法、保存过程和数据模型的调用等。目前生产企业每年都会积累大量的类似三维 CAD 模型的工程主控模型,这些模型都是通过加写保护后按其本地数据格式存储。要确保其长期的安全保存和数据可靠性,必须对这些模型进行定期检查、移存、转化。但这些定期的处理过程不但有前面提及的兼容问题而且将遇到对信息进行处理的时间和效率问题。现以工程软件格式之间的转换为例,来研究当一个软件不复存在,如何选择用别的软件来代替其打开和保存原有的数据,需要通过哪些格式之间的转换能用现有的软件来管理信息,以及保存为哪种格式最优。

收稿日期:2010-05-17

作者简介:赵正旭 男 1960年出生 教授

基金项目:国家自然科学基金资助项目(60873208)

1 小世界效应

匈牙利作家F·Karinthy在1929年提出了“小世界现象”的论断[4]。他认为,地球上的任何两个人都可以平均通过一条由六位联系人组成的链条而联系起来。在20世纪60年代,美国哈佛大学社会心理学教授斯坦利·米尔格兰姆(Stanley Milgram)通过设计一个连锁信件实验,提出了著名的“六度分隔”(Six Degrees of Separation)假说[5-6],大意为任何两个欲取得联系的陌生人之间最多只隔着6个人,便可完成两人之间的联系。当年,米尔格兰姆给内布拉斯加州奥马哈市随意选择的300多人发信,要求他们把他的这封信寄给波士顿市一个独一无二的“目标”人,分别由每个人独自联系。米尔格兰姆告诉每个发信人有关目标人的信息,包括姓名、所在地、职业,如果发信人不认识这个目标人,他们把这封信寄给他们认为有可能认识目标人的熟人。依此类推形成了发信人的链条,链上的每个成员都力图把这封信寄给他们的朋友、家庭成员、或同事熟人,以便使信件尽快到达目标人。米尔格兰姆发现,有60个链条最终到达目标人,链条中平均步骤大约为6,米尔格兰姆由此得出结论:任意两个人都可通过平均6个熟人联系起来。这就是六度分离理论的产生经过。

“六度分离”在学术上称为“小世界现象”或“小世界效应”。研究表明,在看似庞大的网络中各要素之间的间隔实际上是非常“近”的,大家在世界上通过一步一步的社会相识寻找到目标的这个链子理论普遍存在于各种社会、经济网络中,科学家们把这种现象称为小世界效应(Small-world effect)。小世界效应的精确定义还在讨论中,目前一个较合理的解释是:若网络中两点间的平均距离 随网络大小(网络中结点数)呈对数增长,即 ,且网络的局部结构上仍具有较明显的集团化特征,则称该网络具有小世界效应。这里的平均距离具有广泛的含义,例如在上述信件传递实验中,平均距离就是平均传递次数6。

现主要根据小世界现象,对工程软件的输入输出格式之间的转换关系进行实验分析,验证其小世界效应,为解决工程信息中大量级和分散性数据的移存提供科学依据。

2 网络统计量

网络统计量主要包括度和度分布、平均最短路径、聚类系数等等,这些统计量可以深刻地揭露网络的内部特性,下面先分别对这些统计量做简单的介绍。

2.1 度与度分布

度(degree)是单独节点的属性中简单而又重要的概念。节点的度是指与该节点相关联的边的条数,也就是指与该节点连接的其他节点的数目。度分布(degree distribution)是指网络中各节点具有的度的分布。

2.2 平均最短路径

网络的平均最短路径(average shortest paths)可以对网络的连通性进行较好地描述。网络中两个节点i和j之间的距离d_{ij}定义为连接着两个节点的最短路径上的边数。网络的平均最短路径长度L定义为任意两点之间的最短路径的平均值。

2.3 聚类系数

聚类系数(clustering coefficient)表征的是网络的聚类特征,也就是群落特性。一般假设网络中的节点i与k_i条边关联,即与另外k_i个节点相连。显然,在这k_i个节点之间最多可能有$k_i(k_i-1)/2$条边。而这k_i个节点之间实际存在的边数是E_i。那么这k_i个节点之间实际存在的边数是E_i与总的可能的边数$k_i(k_i-1)/2$之比就定义为节点i的聚类系数C_i,而对网络中所有节点的聚类系数取平均值,就是整个网络的聚类系数C。

3 格式转换的小世界特性

3.1 网络仿真

根据收集的近百种工程软件,统计具有输入和输出格式,并且其输入或输出格式不止用于一种软件,

结果得到了 721 种格式,涉及到 71 种软件。实验中,把格式作为节点,软件作为边,输入格式到输出格式用有向边相连。根据软件之间格式的输入输出转换关系建立网络,图 1 所示是工程软件数据格式网络图;图 2 表示了节点度的分布概率。

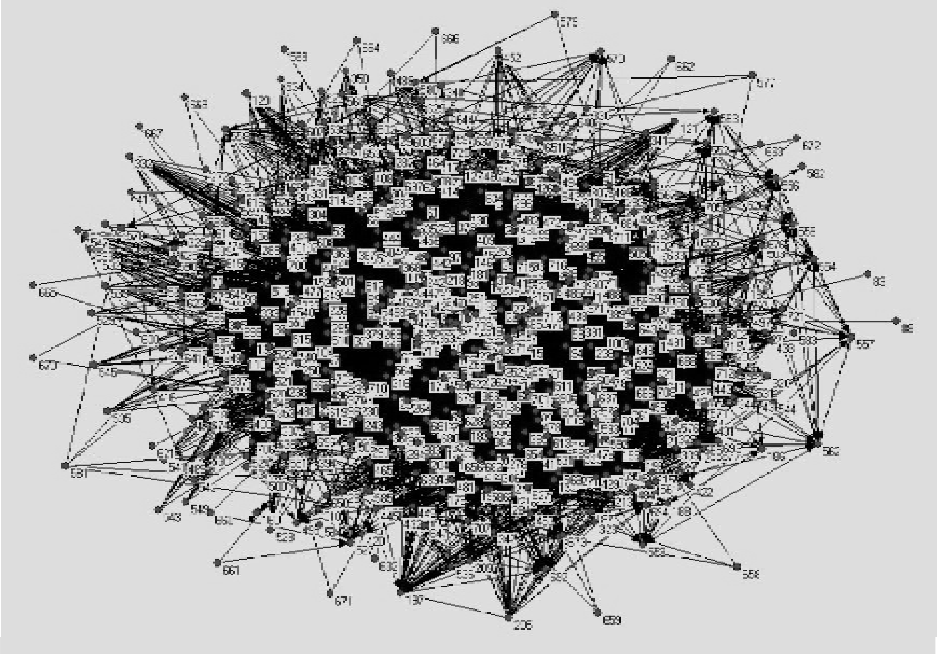

图 1　工程软件数据格式网络图

根据实验进行分析,得出的结果:平均度为 54.241 33,平均路径长度为 2.282 31,聚类系数为 0.674 907 1。

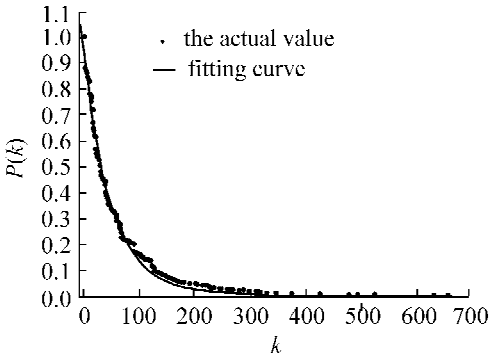

图 2　节点度的概率分布情况

3.2　格式转换具有小世界效应

现有软件的版本升级以及新软件层出不穷,软件格式种类更是名目繁多,但是在应用的时候并没有感觉到需要安装所有的软件,相反,高版本的软件能打开其低版本的格式,新软件的数据通过格式的转换能被常用的软件打开。由上述实验的数据可知格式转换的特征路径长度 L 只有 2.282 31,也就是说,网络的平均距离 L 是随着网络大小 N 呈对数增长的,它明显具有小世界效应。

3.3　格式转换具有集团化、聚类的特征

软件格式种类繁多,但这些格式并不是毫无规律的。每个类型的软件都会有自己的主题,比如绘图类软件都会有 *.jpg、*.jpeg、*.bmp 等格式;视频类软件都会有 *.avi、*.wmv、*.rm 等格式。由实验数据可知格式转换的聚类系数 C 高达 0.674 907 1。

3.4　格式转换中关键节点的效应

格式转换是一个有向网络,从输入格式指向输出格式,对于有些只有输入或输出单一类型的格式,而有些具有多个软件的输入和输出两种类型的格式,造成节点与节点之间的不平等,后者中的某些节点成为网络的关键节点。首先,关键节点都是具有输入和输出两种格式的,所以是双向的;其次,目前软件格式的制定没有统一的规范,格式之间的联系具有不规则特性,对于关键节点指向哪些格式,很多时候是根据需要而编程制定的;再次,关键节点反映出马太效应,关键节点能聚类更多的节点与其相连或通过最短路径转换与其相连,这就是富人的马太效应。

4 格式转换的路径检索

通过上述对格式转换的网络仿真显示其符合小世界效应的特征,是小世界网络,这对工程信息的长期安全保存中的信息移存提供了依据,但是在信息移存过程中还有一个重要的问题,就是格式之间的转换通过哪种软件,这就是所要寻找的路径。为此,用 Protégé 构建了关于工程软件的本体,在此本体的基础上对格式转换进行路径查询。

Protégé 是由斯坦福大学的 Stanford Medical Informatics 开发的一个开放源码的本体编辑器,它是用 Java 编写的,用来帮助本体开发人员和领域专家对领域知识进行建模,使构建本体知识库的过程易于操作和管理,降低了本体构建的高昂成本和维护代价。

本体的构建中,软件的类目如下:

Software(软件)
- 3dsMax
- ACDSee
- ...

Format(格式)
- *.max
- *.drf
- ...

Version(版本)
- 8.0 sp2
- 3.1(SR-1)
- ...

Manufacturer(制造商)
- Autodesk
- ACDSystem
- ...

...

软件类目之间通过对象属性(如定义的:has Input Format、has Output Format 等)和量词(some、only、exactly)联系起来。

在 Protégé 构建的软件本体中,目前我们只用到软件和格式两大类目,还未考虑版本、制造商等属性,而且对于数据保存的时间、软件的寿命以及格式的规定年限都设定为理想状态,即在同一时间段。

对于格式转换的路径查询,在 Protégé 中我们可以通过两种方法进行查询,一是基于格式的查询,二是基于软件的查询,下面以 *.anm 到 *.pca 为例简单介绍一下这两种方法的思路。基于格式的查询:

(1)查找有输入格式为 *.anm 的所有软件,查询语句为:has Input Format some *.anm,查询结果为集合 A。

(2)查找集合 A 的所有输出格式,查询语句为:the Output Format is belong of some (A1 or A2 or … or Ai),查询结果为集合 M。

(3)遍历集合 M,看格式 *.pca 是否属于集合 M。如果属于 M,结束查询;如果不属于 M,查找有集合 M 中任意一种输入格式的软件,查询语句为:has Input Format some (M1 or M2 or … or Mi),查询结果为集合 A_M。路径长度 $L = L + 1$(L 初始值为 0)。

(4)同(2)。

(5)同(3)。

最终找到 *.anm 到 *.pca 的路径,结束查询。

基于软件的查询:

(1)查找有输出格式为 * . pca 的所有软件,查询语句为:has Output Format some * . pca,查询结果为集合 A 。

(2)查找有输入格式为 * . anm 的所有软件,查询语句为:has Input Format some * . anm,查询结果为集合 B。

(3) 比较集合 A 和集合 B。如果有交集记为集合 C,则集合 C 即为格式转换所需的软件,结束查询;如果没有交集,则查找集合 B 的所有输出格式,查询语句为:the Output Format is belong of some (B1 or B2 or … or Bi),查询结果为集合 M。路径长度 $L = L + 1$(L 初始值为 0)。

(4)查找有集合 M 中任意一种输入格式的软件,查询语句为:has Input Format some (M1 or M2 or … or Mi),查询结果为集合 B_M 。

(5)同(3)。

(6)同(4)。

最终找到 * . anm 到 * . pca 的路径,结束查询。

从 * . anm 到 * . pca 转换的可视化结果如图 3 所示。

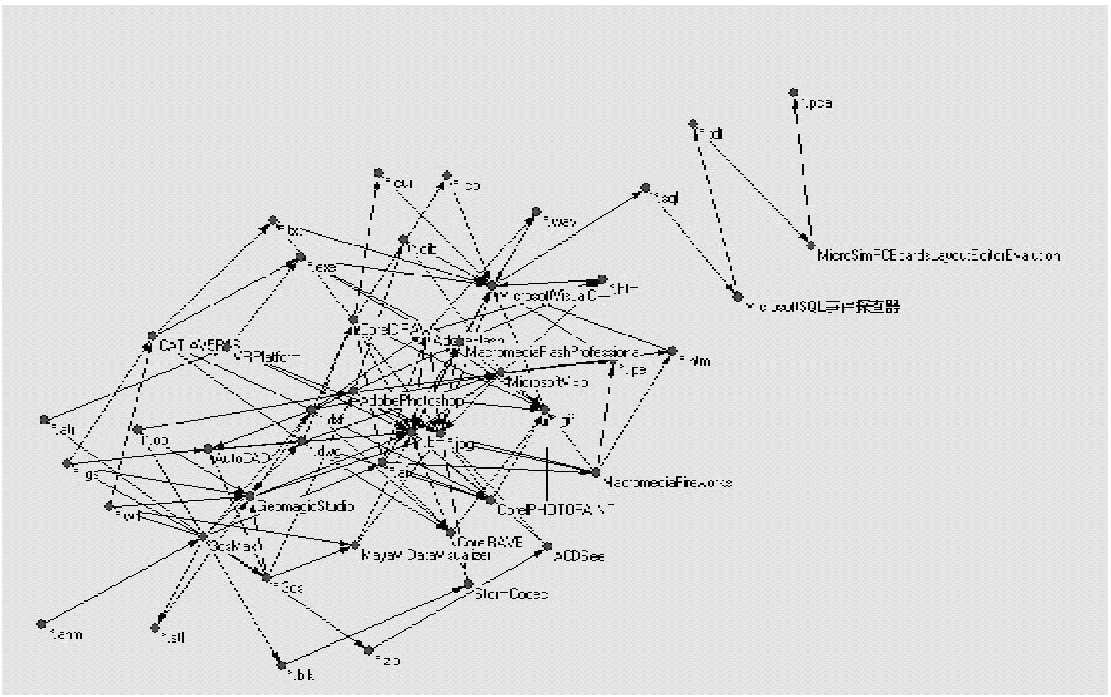

图 3　从 * . anm 到 * . pca 的转换路径

5　运用小世界特性规范工程软件的设计

由以上实验和数据分析可知,该网络具有小世界效应的两个特征:短路径 L 和高聚类系数 C ,称该网络为小世界网络(Small World Net)。结合实际情况,运用小世界网络特性来规范工程软件的设计。

现有的软件中格式种类繁多,其中有些格式能被多种软件输入输出,而有些格式只能被单一软件输入输出;有些格式有输入输出两种形式,而有些格式只有输入或输出一种形式,对于这种情况,构建的有向网络模型中,其中只有 45.065 703% 的节点对能够找到路径,这对寻找任意两个节点间的路径问题还有一定的困难。虽然不能重新设计现有软件的所有格式,但是根据小世界网络原理,可以在格式转换中给那些输入和输出格式只用于自身的软件添加格式让其与网络中的关键节点相连,从而减少特征路径长度,并且让更多的节点之间能够转换,提高可靠性。对于未来软件版本升级和新软件格式制定,根据小世界网络的特性来规范。首先,基于现有的格式转换关系构建一个网络,已证实此网络符合小世界效应的

两个特性,是小世界网络。其次,根据构建的有向网络,为了让网络中任意两个节点之间都能找到路径,在制定软件格式的时候给每个格式都设定有输入和输出两种类型。再次,实际中有些软件的格式只能被自身软件使用,形成独立的小规模网络,和别的软件格式之间没有联系,这是不符合小世界网络特性的,所以在制定软件格式时,必须要杜绝这种情况,即每个软件必须要有一种格式的输入或输出能供另一种软件使用,同理,这两种软件也不能形成小规模的网络,必须要和别的软件格式相连,以此类推,最终任意的两个软件格式都能通过软件进行转换。

通过运用小世界网络特性制定的软件格式,保证了在信息移存时能够找到任意两个数据格式之间的最短路径,从而通过软件进行格式转换,提高了信息移存的效率,对于信息的长期安全保存提供了保障。

6 结束语

小世界效应的应用提高了解决问题的效率和速度,特别是在大量级信息的保存和调用中。在工程信息的长期保存问题上,通过建立的网络仿真,证明了其具有小世界效应,为信息移存提供了保障。但是现有的软件,其格式种类繁多,并且有些格式只供一种软件使用,有些格式只有输入或输出一种形式,这在网络中会使很多节点之间没有路径,影响小世界网络的特性。如果在设计软件的时候根据小世界网络模型来设计数据格式,那么对信息的保存和移存更便利。本文提出的研究结果为建立一个通用的最佳小世界网络数据模型提供了可靠的科学依据,从而为工程信息资源的安全保存和高效检索、为提高工程信息管理中的兼容性和时效性,创建高效的元数据模型及其设计方法奠定了技术基础。

参 考 文 献

[1]National Institute of Standards and Technology. Long term knowledge retention (LTKR): Archival and representation standards [R/OL]. Gaithersburg: NIST, 2006. [2010-02-26]. http://edge.cs.drexel.edu/LTKR/.

[2]马张华.信息组织[M].北京:清华大学出版社,2003.

[3]National Institute of Standards and Technology. The Role of ISO 10303 (STEP) in long term data retention: Long Term Knowledge Retention Workshop[R]. Gaithersburg: NIST, 2006.

[4]Braun T. Hungarian priority in network theory [J]. Science,2004,304:1745.

[5]Milgram S. The small world problem [J]. Psychol Today, 1967,1(1):60-67.

[6]Jeffrey Travers, Stanley Milgram. An Experimental Study of the Small World Problem [J]. Sociometry,1969, 32(4):425-443.

Research into Small World Effect in Engineering Software

Zhao Zhengxu, Long Rui, Guo Yang, Liu Jiajia

(School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: Engineering information is of sophisticated formats and diverse contents, its domain context and data compatibility directly affect the effective management and full utilization of information resources in digital engineering practice. In particular, for long term retention, preservation and utilization of engineering information, there have been compatibility problems in between data and its host systems and the problems in lead-time and efficiency for regular data check, migration and transformation, for which there has so far been no practical solutions and available satisfactory methods. This paper researches the so-called "six degrees of separation" phenomenon and small-world network model and thereby analyzes the conversion relationship of the engineering software input and output format via experiments, and verifies its small-world effect. By constructing the conversion path of ontology query format, it establishes reliable benchmarking for migrating large-scale and diverse engineering information.

Key words: six degrees separation; small world effect; data migration; information retention