ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2000/00(00)-0000-00
doi: 10.3778/j.issn.1673-9418.

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

基于可扩展主题建模技术的多源数据分析框架*

唐爽[1,2], 张灵箫[1,2], 赵俊峰[1,2,3+], 谢冰[1,2,3], 邹艳珍[1,2,3]

1. 北京大学 信息科学技术学院, 北京 100871
2. 高可信软件技术教育部重点实验室, 北京 100871
3. 北京大学(天津滨海)新一代信息技术研究院, 天津 300450

Analytical Framework for Multisource Data based on Extensible Topic Modeling*

TANG Shuang[1,2], ZHANG Lingxiao[1,2], ZHAO Junfeng[1,2,3+], XIE Bing[1,2,3], ZOU Yanzhen[1,2,3]

1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
2. Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing 100871, China
3. Peking University Information Technology Institute (Tianjin Binhai), Tianjin 300450, China
+ Corresponding author: Phn +86-13601389906, E-mail: zhaojf@pku.edu.cn

TANG Shuang, ZHANG Lingxiao, ZHAO Junfeng, et al. Analytical Framework for Multisource Data based on Extensible Topic Modeling. Journal of Frontiers of Computer Science and Technology, 2000, 0(0): 1-000.

Abstract: With the continuous development and application of information technology, many information systems have accumulated a large amount of multi-source heterogeneous data. A large part of these data is structured data which is high-dimensional, low quality and unmarked. It's difficult to extract feature and refine knowledge from this kind of data. Therefore, the analysis and processing of such structured data are one of the important tasks in the big data analysis. Topic modeling is a very important method in text processing and data mining. It is an unsupervised learning algorithm that was originally used to model unstructured natural language text. It can effectively extract topic information from text semantics, extract feature and reduce dimensionality. But topic modeling is still not well applied in the processing of complex multi-source data, especially structured data. This paper presents a framework based on extensible topic modeling technology for structured and unstructured multi-source data analysis. According to the characteristics of multi-source data, this framework analyzes the multi-source data by data importing, data analysis

*The National Natural Science Foundation of China under Grant No. 61472007 (国家自然科学基金); the National Natural Science Foundation of China under Grant No. 91646107 (国家自然科学基金).
Received 2000-00, Accepted 2000-00.

and data visualization three steps. On this basis, a multi-source data analysis tool is implemented. Finally, the experiment of two data sets proves the effectiveness of the multi-source data analysis framework.

Key words: topic modeling; LDA; structured data analysis; visualization

摘 要: 随着信息技术的不断发展和应用，大量信息系统积累了海量多源异构数据，这些数据中有很大一部分都是结构化数据，这些数据大多具有高维度、低质量、无标注等特点，难以进行特征提取与进一步的知识提炼。因此，针对这类结构化数据进行分析处理是大数据分析的重要工作之一。主题建模是文本处理和数据挖掘中的一个非常重要的方法，它是一种无监督学习算法，最初用于对无结构的自然语言文本进行建模，可以有效地从文本语义中提取主题信息，以进行特征提取和降维分析，然而主题建模技术尚不能很好应用在关系复杂的多源数据，尤其是结构化数据的处理中。本文提出了一个基于可扩展主题建模技术的针对结构化与非结构化多源数据分析框架，针对多源数据的特点，通过数据导入、数据分析、数据可视化三个步骤对多源数据进行基于主题建模技术的数据分析，并在此基础上实现了一个多源数据分析工具，最后通过两个数据集的实验证明了本文所提的多源数据分析框架的有效性。

关键词: 主题建模技术;LDA;结构化数据分析;可视化

文献标志码: A **中图分类号:** ***

1 概述

随着大数据相关技术的不断发展和应用，数据的获取和存储变得越来越方便，计算机系统中积累了来自各种行业海量的数据[1-3],这些数据包括日常生产、业务、交易等过程中的记录，从互联网中搜集到的数据集，还有来自自动监测系统的监测指标等。对这些数据进行分析处理，能够获取大量有价值的信息[4-7]。由于现有的大量系统都基于 SQL 数据库，这些数据中有很大一部分都是结构化数据，分析处理这些结构化数据是一个重要工作[8,9]。这些结构化数据具有高维度、低质量、无标注等特点[1,8]，因此从原始数据中采用无监督的方式进行特征抽取，并对原始数据进行信息提炼和降维是提高分析效率和效果的必要手段。

主题建模（Topic Modeling）技术[10]最初是从自然语言文本中抽取主题信息的一种技术，该技术假设主题是一组语义相关的词语，而文章由多个主题混合而成。由于主题建模拥有对数据原始特征进行抽象的能力，它实际上成为了一种通用性的高级特征抽取方法。除此之外主题建模还是一种无监督学习方法，它能对无标注数据进行分析，因此本文选择主题建模方法来处理前面所提到的大量高维度，低质量，无标注的结构化数据。目前已有大量研究工作致力于将主题建模技术应用到结构化数据分析中去，例如基于电子商务交易数据的用户画像[11]，基于诊疗记录的临床路径模式发现[12]等。

要将主题建模技术更好地应用于结构化数据，还有很多问题需要解决，主要有以下三个方面：

(1) 结构化数据由二维表来逻辑表达和实现，包含表内字段和表间关联信息，这与由标题和词语集合组成的文档数据有较大差异，无法直接作为主题建模的输入，通常需要进行转化处理。

(2) 朴素的主题建模算法不支持对表内多个字段以及多表关联信息进行建模，因此无法满足对结构化数据的分析需求，需要对其进行扩展。

(3) 主题建模的结果非常抽象，需要良好的可视化方法便于用户理解。

目前常用的主题建模工具[1,2,3]均不能很好地解决以上问题，针对这一情况，本文提出了一个基于可扩展主题建模技术的数据分析框架 DBInsight，它

[1] MALLET: http://mallet.cs.umass.edu

[2] Stanford TMT: https://nlp.stanford.edu/software/tmt/tmt-0.4

[3] Gensim: https://radimrehurek.com/gensim

能够对包括结构化数据在内的多源数据进行主题建模分析，并提供可视化结果展示。本文的主要贡献有：

(1) 提出了针对结构化数据特点进行建模分析的两种扩展主题模型。

(2) 提出了一个基于可扩展主题建模技术的多源数据分析框架。

(3) 根据上述数据分析框架实现了一个数据分析工具，通过对两个现实数据集的分析，证明该框架是可行有效的。

本文后续内容组织如下：第 2 小节介绍扩展的主题模型；第 3 小节详细介绍基于可扩展主题建模技术的多源数据分析框架；第 4 小节展示根据此框架实现的数据分析工具以及实验；第 5 小节是总结和未来工作。

2 扩展的主题模型

2.1 朴素 LDA 主题模型

潜在狄利克雷分布（LDA）[13]是最朴素的主题模型，其基础假设是文章是由多个主题构成的，而每个主题都是词集的一个概率分布。

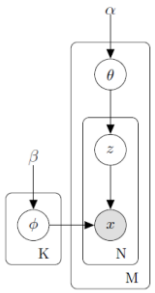


Fig.1 LDA probabilistic graphical model
图 1 LDA 概率图模型

用概率图模型[14]描述 LDA 算法的实例生成过程如图 1，其过程可以描述为：

(1) 从以β为参数的 Dirichlet 分布中抽样K个主题的词语分布ϕ。

(2) 从以α为参数的 Dirichlet 分布中抽样M个文档的主题分布θ。

(3) 对于第i篇文档的第j个词语，首先从第i篇文档的主题分布中抽样一个主题$z_{i,j}$，然后从该主题的词语分布中抽样一个词语$x_{i,j}$。

(4) 重复过程 3，直到生成所有的N个词语。

求解 LDA 模型的普遍方法是吉布斯采样，其流程可以概括为：对z值进行随机初始化，多次迭代进行吉布斯采样（为每一个z重新分配主题），最后对z值进行计数求得分布θ和ϕ。该算法的关键在于为z重新分配主题k'：$\mathrm{p}(z=k'|z_{-i},x)$。在这里我们需要根据其他所有位置上的主题分布计算当前位置上分配到每个主题k的概率$\mathrm{p}(z=k'|z_{-i},x)$，并且将所有$k$个主题的概率合并为一个多项分布并对其抽样，将抽样得到的主题k'赋给当前位置上的z值。下面给出$\mathrm{p}(z=k'|z_{-i},x)$的计算公式：

$$\mathrm{p}\left(z=k^{'}|z_{-i},x\right)\propto\frac{C_{x,k}^{VK}+\beta}{C_{\cdot,k}^{VK}+V\beta}\cdot\frac{C_{m,k}^{MK}+\alpha}{C_{m,\cdot}^{MK}+K\alpha}\qquad(1)$$

公式 1 中k代表当前位置上分配k主题时的概率，x代表当前位置上的特征，m代表当前实例的编号。公式右边由两个因子组成。首先，$C_{x,k}^{VK}$代表所有实例中x分配给主题k的计数，而$C_{\cdot,k}^{VK}$代表所有分配给主题k的任意特征的计数。两式各加上 Dirichlet 先验β做平滑后相除，实际上代表了所有分配了主题k的词中当前的x所占的比例。同理，右侧因子中$C_{m,k}^{MK}$代表文档m中分配给k主题的特征的个数，$C_{m,\cdot}^{MK}$代表实例的所有词个数。两式各加上 Dirichlet 先验α做平滑后相除得到当前实例m中主题k所占的比例。公式 1 语义为：当前位置分配主题k的概率等于当前实例中主题k的占比乘以主题k中当前特征的占比，实际上就代表了实例从主题到特征的生成过程。

2.2 多视图 LDA 主题模型

朴素 LDA 主题模型只考虑单种词语，不适合处理多表关联的结构化数据。针对关系型数据库中常见的多表关联关系，将朴素主题模型扩展到多视图主题模型，将多个表看作描述同一实例的不同视图，从而在主题中包含属于多个视图的关联特征。这种扩展后的模型能充分利用不同视图之间互补的特性提高建模效果。

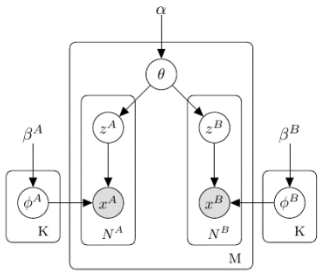


Fig.2 Multi-view LDA probabilistic graphical model
图 2 多视图 LDA 概率图模型

多视图 LDA 主题模型的概率图模型如图 2，它在朴素 LDA 主题模型的基础上增加了多个视图下的主题词语分布。图 2 中，α是主题分布的超参数，θ表示每个实例（文档）的主题分布。每个视图有独立的主题分配变量z，产生的特征x，主题-特征分布ϕ，以及超参数β。同时，所有视图中的特征最终是由同一个主题分布θ抽样产生的，该θ代表了模型不同视图间的共同优化目标。图 2 展示了包含两个不同视图的模型，而多视图的 LDA 可以扩展到包含任意多个不同视图的场景，可以适应包含任意数量的关联表的数据。对于包含n个关联表的数据，每一个实例的实例主题分布是唯一的，它主导了n类不同特征的生成，每类特征都包含各自的ϕ，每一个主题包含n个对应的主题-特征分布。

多视图 LDA 模型的求解过程与朴素 LDA 类似，主要区别在于为z重新分配主题时只计算该视图下的特征占比，以图 2 中视图A为例，其核心概率计算公式如下：

$$\mathrm{p}^{\mathrm{A}}\left(z=k^{'}\,|z_{-i},x\right)\propto\frac{C_{x^{A},k}^{V^{A}K}+\beta^{A}}{C_{\cdot,k}^{V^{A}K}+V^{A}\beta^{A}}\cdot\frac{C_{m,k}^{MK}+\alpha}{C_{m,\cdot}^{MK}+K\alpha}\quad(2)$$

公式 2 中左边的因子用于估算当前视图中的主题-特征分布，右边的因子用所有视图中特征计数估算实例-主题分布，该因子在计算每个视图中特征主题分配概率时形式都相同，实际上起到了在各个视图之间传递信息，达到“共识”的目的。

2.3 多属性 LDA 主题模型

朴素 LDA 主题模型无法考虑数据集中非文本信息，如连续数值等。而结构化数据中每一个表内存在大量的非文本数据，为了充分利用这些非文本数据，将朴素主题模型扩展为多属性主题模型，其核心思想为：将数据表中每一个字段都看作描述主题特征的一种属性，根据多个属性能划分出更准确的主题。它的思路是在朴素 LDA 中增加代表属性的随机变量。

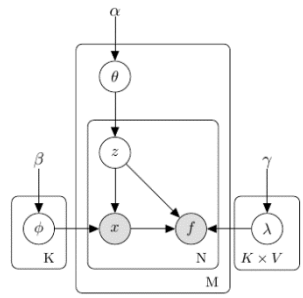


Fig.3 Multi-attribute LDA probabilistic graphical model
图 3 多属性 LDA 概率图模型

多属性 LDA 主题模型的概率图模型如图 3，图中f表示描述属性的随机变量，λ代表f所服从的概率分布。注意由于f所服从的概率分布由x和z共同决定，不同的主题和特征对应这不同的λ。因此包含λ的方框总共重复$K*V$次，代表模型中总共包含$K*V$个不同的λ。这里K代表主题个数，V代表特征的种类数。此外，还为λ引入了概率分布的先验γ。由于不同特征服从不同的概率分布，甚至可能为不同类型的值（离散型或者连续型），因此很多情况下往往需要对不同的特征分别确定其合适的γ先验，其分布是λ分布的共轭先验分布。例如在电子病历的检验检查数据中，不同检查项目的取值范围是非常不同的，如果将检查项目结果作为不同的特征，那么对于V类特征自然应该有不同的V种先验。

多属性 LDA 的求解在朴素 LDA 的基础上要增加对特征值的考虑，以增加一种属性f为例，其核心概率计算公式如下：

$$\begin{aligned}\mathrm{p}(z_i=k|z_{-i},x,f)\propto&\frac{C_{x,k}^{VK}+\beta}{C_{\cdot,k}^{VK}+V\beta}\cdot\frac{C_{m,k}^{MK}+\alpha}{C_{m,\cdot}^{MK}+K\alpha}\\&\cdot\mathrm{p}\left(f_i|z_i,x_i,f_{-i}^{z_i,x_i}\right)\end{aligned}\quad(3)$$

可以看到，为z分配新的主题时还要考虑特征值f。公式 3 中第三项因子代表当前位置上的属性f在该位置上特征x_i和主题分配z_i确定的情况下，给定其他该主题和该特征下的属性值，该属性取当前值的概率。该项的计算由该特征的先验分布决定，增加多种属性就在公式中再添加与此类似的对应项。

以上提出的多视图 LDA、多属性 LDA 两种算法分别是针对结构化数据的多表关联关系以及表内多字段对应关系进行主题建模分析的扩展算法。在实际应用中两者还能进一步结合成为多视图多属性 LDA 算法，能够较好地满足结构化数据分析需求，同时该图模型还能够进一步扩展，该算法理论上支持对任意多视图，每个视图内任意多特征同时进行分析，具有可扩展性，因此本文称之为可扩展主题模型，运用该模型进行主题建模分析的技术称为“可扩展主题建模技术”。

3 多源数据分析框架

3.1 问题分析

本文将多源数据的分析流程分为三个部分：数据导入、数据分析以及数据可视化。

(1) 数据导入：将外部不同来源的数据转化为主题建模算法的输入格式，并根据需求进行预处理，消除不同来源数据的差异性。为了支持多源数据灵活组合，系统应提供通用的数据接口，以接入不同来源的数据。

(2) 数据分析：为了增加分析框架的适用范围，该框架支持多种不同的主题建模算法，并且能够灵活地修改算法参数。

(3) 数据可视化：将主题建模的结果进行可视化，方便用户快速地了解主题建模的结果。由于主题建模算法有许多种，工具还应该支持为特定算法扩展单独的可视化方法。

3.2 框架总体设计

DBInsight 框架主要分为三个部分，如图 4 所示：

(1) 数据导入：数据导入过程分为三个步骤：一是访问外部数据源，这里需要用户提供访问数据源所需的连接信息；二是将外部数据转化为主题模型算法标准文档格式（文档集，包含多篇由名称和词语集合组成的文档）；三是对文档进行预处理，例如 html、xml 格式解析，长文本分词，去除停用词等。

(2) 数据分析：数据分析的核心是主题建模算法。数据分析时用户能够选择要分析的文档集，要使用的主题建模算法，并且设置算法所需参数，例如 LDA 算法需要设定参数α、β以及主题数目。为了提高框架的可用性，本框架对算法扩展提供良好的支持，方便用户添加新的建模算法。

(3) 数据可视化：根据建模结果的特点，选择适当的可视化方式。除了对基本分析结果可视化以外，框架还支持在分析结果上进一步的深入分析，并将分析结果可视化。对于特定算法可视化的支持，通过提供扩展接口实现。

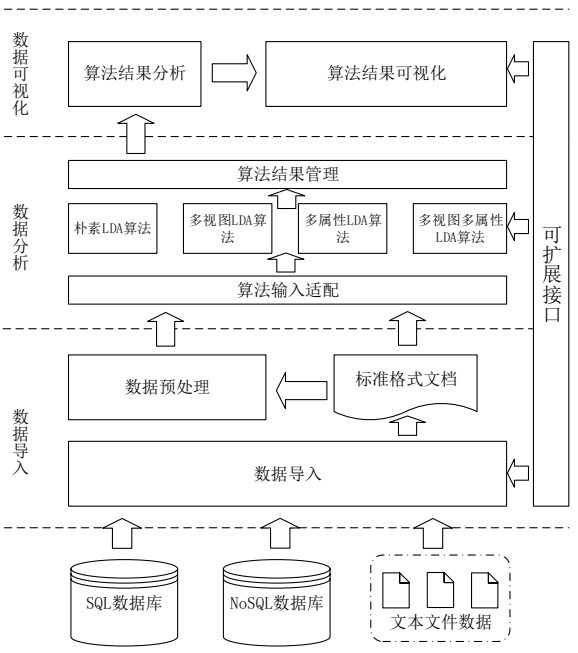


Fig.4 DBInsight Frame Diagram
图 4 DBInsight 框架图

3.3 数据导入

数据导入部分主要分为三个步骤：访问数据源、导入数据、数据预处理。其中前两个步骤将外部多源数据转化为标准文档格式，第三个步骤在标准格式的基础上进行预处理。

3.3.1 访问数据源并导入数据

访问数据源以及导入数据的流程如图 5 所示。DBInsight 框架使用数据源,数据块以及导入器三个概念对这一流程进行建模。其中数据源指用户输入的外部数据来源信息，利用这些信息能建立到外部数据源的连接。数据块即本文定义的标准文档格式的数据。导入器是一段程序，用来连接到数据源，并导入数据转化为数据块。

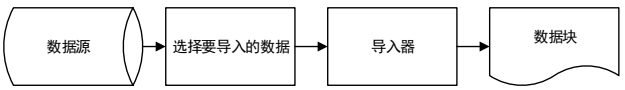


Fig.5 Access data source, import data
图 5 访问数据源、导入数据流程

框架引入数据块的概念来提高可扩展性。如果系统直接从外部数据源读取数据进行算法分析，一个支持m种数据源，n种分析算法的系统需要编写$m \cdot n$份代码来实现。而有了数据块，将数据块作为数据导入的标准输出格式，算法分析的标准输入格式,只用编写$m+n$份代码就能实现工具的功能。要添加一种新的数据源，只需要提供新的导入器即可，避免了为每一种算法进行适配。

导入器的主要工作是将外部数据源中的数据按照用户输入的映射关系映射为文档数据（标题和词语）。框架中提供了对 SQL 数据库、NoSQL 数据库[15]以及文本文件进行导入的方法。对于 SQL 数据库，采用选择表格和字段的方式得到数据库中字段到文档标题、词语的映射关系；对于 NoSQL 数据库，需要用户输入对应的数据库查询语句和查询结果到文档的映射关系；对于文本文件的支持和传统主题建模分析工具类似，将一个文件视为一个文档，文档的标题与文件名相同，文件内容就是文档的内容。

3.3.2 数据预处理

从外部数据源直接转化而来的数据块，根据需求可以进行进一步的预处理,例如带 html 标签的数据可能需要对标签进行解析，获取所需要的数据。而分词，去停用词，大小写统一化这些方法是文本数据预处理的常用方法，框架也提供支持。这些预处理操作主要是对原数据块中文档词语的进一步处理,图 6 展示了对数据块进行分词预处理的效果。

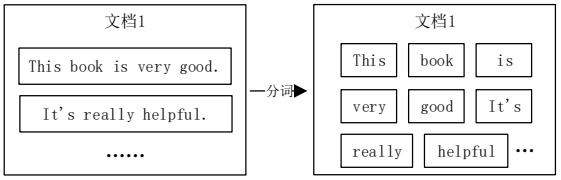


Fig.6 Preprocess: data chunk segmentation
图 6 数据块分词预处理

3.4 数据分析

数据分析是主题建模工具的核心内容。从前面的数据导入部分得到了数据块这一种标准格式的文档数据，数据分析就是将数据块作为输入数据，进行主题建模分析，并将分析结果以概率分布的形式保存下来，其主要流程如图 7 所示。

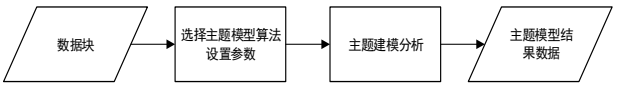


Fig.7 Preprocess: data chunk segmentation
图 7 数据分析主要流程

作为主题建模分析的通用框架，框架支持算法扩展。当用户需要添加一种新的算法时，只要保证算法的输入格式是框架提供的标准文档格式，算法的输出格式是框架提供的标准结果格式，框架就能将新算法添加到算法选项中。

框架还对结构化数据分析的场景进行了针对性的优化。对结构化数据进行分析时常常涉及到对表内多字段和多表中不同字段的分析，这种需求能够从输入数据格式得到，因此框架能够根据输入数据自动为用户选择要使用的分析算法。例如用户选择了同一表内多个字段作为输入,则采用多属性 LDA 进行建模分析，如果选择了多个表下的字段，则采用多视图 LDA 或者多视图多属性 LDA 进行分析，这种自动选择降低了框架的使用门槛，用户无需具有多视图、多属性 LDA 的相关知识就能使用框架对结构化数据进行建模分析。

为了提高可用性，框架将一次分析过程看作一项分析任务，并且支持任务管理操作。用户创建分析任务进行建模分析，并且可以暂停一项正在进行的分析任务或者继续一项暂停中的分析任务，还可以取消一项分析任务。不同分析任务之间可以并行处理，这提高了分析工作的效率。

3.5 数据可视化

经过主题建模分析，得到了主题建模的结果，主题-词语分布和文档-主题分布。数据可视化部分就是设计分析和可视化方法，将主题建模结果更好地呈现给用户。其中两个基本的分布信息作为主题建模最直接的结果，框架分别从主题和文档的角度进行可视化。而在主题建模结果的基础上，框架支持进一步的应用分析，并对应用结果进行可视化。

4 工具实现与应用展示

4.1 DBInsight 框架的工具实现

根据前面提出的 DBInsight 框架，本文实现了一个基于主题建模的数据分析工具，工具的方法流程见图 8。

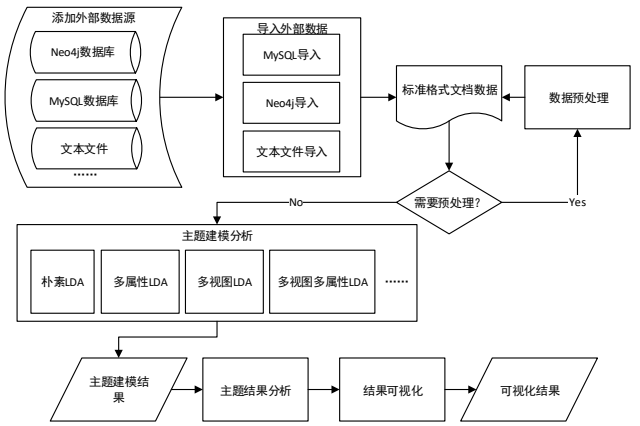


Fig.8 Method process of tool
图 8 工具方法流程

工具支持 MySQL[4]为代表的结构化数据源、Neo4j[5]为代表的 NoSQL 数据源以及文本文件数据源，并提供了朴素 LDA 算法以及多属性 LDA、多视图 LDA 算法、多属性多视图 LDA 算法三种扩展算法，主题建模结果通过 web 页面渲染图表进行可视化。

4.2 工具界面和使用展示

图 9 是工具的主界面，此界面展示了工具中的数据块信息。

用户能够添加新的数据块，用户输入连接信息后工具能够预览数据源信息，方便用户选择要导入的数据。

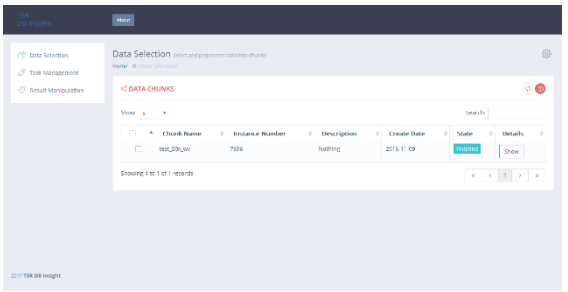


Fig.9 Main interface of tool
图 9 工具主界面

数据导入完成后，用户可以创建新的建模分析任务，如图 10 所示，用户需要选择要分析的数据块以及设置参数。

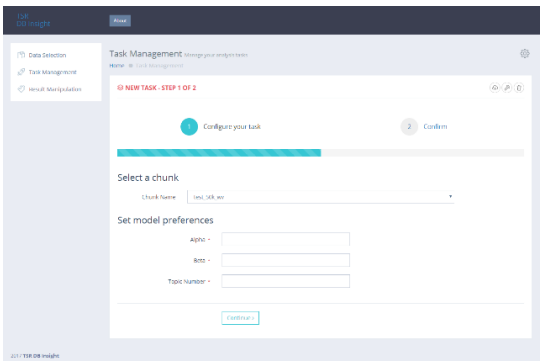


Fig.10 Interface for creating analysis task
图 10 创建分析任务界面

建模分析完成后，用户能够得到建模结果的可视化图表，如图 11 所示。

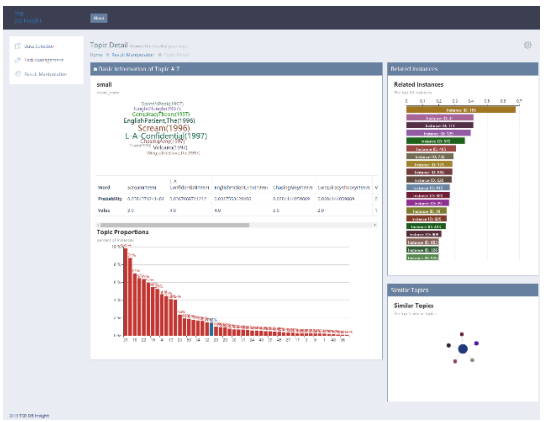


Fig.11 Interface for result visualization
图 11 结果可视化界面

[4] MySQL website. URL: https://www.mysql.com

[5] Neo4j website. URL: https://neo4j.com

用户分析数据时，首先添加数据源，导入数据；然后选择要使用的分析算法，设置算法参数，创建分析任务；分析完成后，通过可视化界面查看分析结果。工具提供了完整的图形化界面支持，因此用户不需要编写具体代码，只需要按照提示选择和输入信息即可完成分析工作。

目前常用的主题建模分析工具，它们大多是一些开源的开发工具包，因此没有图形用户界面支持，同时这些工具只支持对文本数据进行分析，输出结果也只是主题分布，没有可视化图表。使用这些工具对 SQL 等结构化数据进行分析时，用户需要自己编写将 SQL 数据转化为文档文件的代码，然后调用工具进行算法分析，然后再使用或者编写可视化工具进行可视化展示。

相对于这些现有的主题建模分析工具，本文实现的工具自动化地完成了数据导入和结果可视化的工作，降低了主题建模分析的知识门槛，提高了分析工作的效率。

4.3 分析结果展示

本部分主要介绍工具在两个结构化数据集上的分析结果。

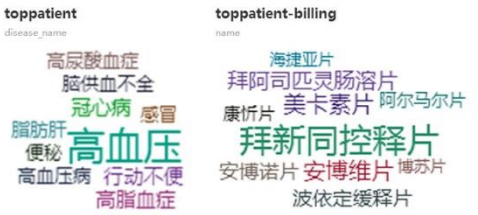
4.3.1 北京某医院门诊记录数据集

该数据集包含了北京某医院 2009 至 2012 年的门诊记录数据，包含诊断信息，用药信息，检验检查信息等。数据集的形式为 SQL 数据库，是标准的结构化数据。本实验用到了病人诊断表以及病人用药表的数据，希望得到诊断和用药之间的关联关系。

本实验中文档的定义：选择病人 id 作为文档 id；分别选择了两种视图的词语，一是病人诊断表中的疾病名称，二是病人用药表中的药品名称。参数设置为：α=0.1，β=1，主题数目$\boldsymbol{k}$=20。

图 12 展示了其中一个主题的信息。其中 toppatient 对应的视图就是疾病名称，toppatient-billing 对应的视图是药品名称。该主题展示了其中包含的两种视图词语之间的关联关系，即疾病和药品的关联关系。

从疾病对应的词语分布信息可以看出，这是关于高血压的主题。而药品的分布信息中占比较高的药物拜新同、安博维（厄贝沙坦）、美卡素（替米沙坦）都是用于高血压治疗的药物，因此可以判断该主题展示的疾病和用药的关联信息是有效的。



Word	高血压	行动不便	冠心病	高脂血症
Probability	0.743024374269	0.0495197968713	0.0423561684802	0.0344321982085

Word	拜新同控释片	安博维片	美卡素片	拜阿司匹灵肠溶片
Probability	0.217412513047	0.106549442671	0.0951137074224	0.0812113572565

Fig.12 Multi-view topic-word distribution
图 12 多视图主题-词语分布信息

4.3.2 MovieLens 数据集[16]

本数据集是 GroupLens Research 采集的一组从 20 世纪 90 年末到 21 世纪初由 MovieLens 用户提供的电影评分数据。其中包含电影评分、电影元数据以及用户的个人信息。数据集存储在 SQL 数据库中，是标准的结构化数据。本实验主要用到了用户电影评分的数据表，本实验中文档的定义为：选择 MovieLens 用户 id 作为文档名称，选择用户评论的电影名称作为文档词语，额外选择用户对电影的评分作为词语的属性。参数设定为：α=0.1，β=1，主题数目$\boldsymbol{k}$=50，评分属性的分布设定为均匀分布。

Word	Scream(1996)	L-A-Confidential(1997)	EnglishPatient,The(1996)	ChasingAmy(1997)	ConspiracyTheory(1997)
Probability	0.0384371241481	0.0367666711212	0.0317553120406	0.0284144059869	0.0284144059869
Value	3.0	4.0	4.0	5.0	2.0

Fig.13 Multi-attribute topic-word distribution
图 13 多属性主题-词语分布信息

图 13 是从结果中选择的一个主题的信息，可以看出这个主题主要是关于惊悚类和爱情类电影的，因为占比较高的电影有《L.A. Confidential（洛城机密）》、《The English Patient（英国病人）》，

《Scream（惊声尖叫）》,《Conspiracy Theory（连锁阴谋）》,《Chasing Amy（猜·情·寻）》它们都具有惊悚或者爱情的元素。同时根据评分这一属性信息，发现该偏好主题表现为对惊悚类电影不喜以及喜爱爱情电影（几部惊悚类电影评分都较低，而爱情片评分高）。

5 总结和未来工作

5.1 文本工作总结

本文从对结构化数据进行主题建模分析的应用场景出发，发现了对结构化数据进行主题建模分析存在的问题。并针对这些问题进行分析并设计了分为数据导入，数据分析，数据可视化三个部分的数据分析框架，支持多种扩展方式。并在框架的基础上实现了一个数据分析工具，该框架和工具降低了主题建模分析的知识门槛，简化了操作流程，提高了主题建模分析工作的效率，同时对结果的可视化使用户更好地了解分析结果，提升了分析的价值。

5.2 未来工作

未来的工作可以从三个方面进行：一是提供编程开发的 API，让框架方便地集成到其他项目中；二是实现更高效的数据处理模式，例如流式处理，在线训练，分布式计算等；三是加入对更多数据源和分析算法的支持。

References:

[1] Wu Xindong, Zhu Xingquan, Wu Gong-Qing, et al. Data mining with big data[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1):97-107.

[2] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential[J]. Health Information Science and Systems,2,1(2014-02-07), 2014, 2(1):3.

[3] Townsend A M. Smart cities : big data, civic hackers, and the quest for a new utopia[M]. W.W. Norton & Company, Inc, 2013.

[4] Mishra B K, Hazra D, Tarannum K, et al. Business Intelligence using Data Mining techniques and Business Analytics[C]// System Modeling & Advancement in Research Trends. IEEE, 2017:84-89.

[5] Sukumar P, Robert L, Yuvaraj S. Review on modern Data Preprocessing techniques in Web usage mining (WUM)[C]// International Conference on Computation System and Information Technology for Sustainable Solutions. IEEE, 2016:64-69.

[6] Saoudi M, Euler R, Kechadi T. Data Mining Techniques Applied to Wireless Sensor Networks for Early Forest Fire Detection[J]. 2016:1-7.

[7] Niu Pengfei. Application of Web Data Mining In E-commerce[J]. Computer & Network, 2015(7), 30-31.

[8] Deng Xiaodui. Research on the Structured Data Mining Algorithm and the Applications on Machine Learning Field[C]// International Conference on Social Science and Technology Education. 2016:82-84.

[9] Larson D, Chang V. A review and future direction of agile, business intelligence, analytics and data science[J]. International Journal of Information Management, 2016, 36(5):700-710.

[10] Srivastava A, Sahami M. Text Mining: Classification, Clustering, and Applications[M]. 2009.

[11] Ovsjanikov M, Chen Y. Topic Modeling for Personalized Recommendation of Volatile Items[M]// Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010:483-498.

[12] Huang Zhengxing, Wei Dong, Lei Ji, et al. Discovery of clinical pathway patterns from event logs using probabilistic topic models[J]. Journal of Biomedical Informatics, 2014, 47(2):39.

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[14] Jordan M I. Graphical Models[J]. Statistical Science, 2004, 19(1):140-155.

[15] Han Jing, Haihong E, Le Guan, et al. Survey on NoSQL database[C]// International Conference on Pervasive Computing and Applications. IEEE, 2011:363-366.

[16] Harper F M, Konstan J A. The MovieLens Datasets: History and Context[M]. ACM, 2016.

附中文参考文献：

[7] 牛朋飞. Web 数据挖掘技术在电子商务中的应用[J]. 计算机与网络, 2015(7):30-31.

TANG Shuang was born in 1995. He is an M.S. candidate at Peking University. His research interests is software engineering.

唐爽(1995–), 男, 湖北人, 北京大学硕士研究生, 主要研究领域为软件工程。

ZHANG Lingxiao was born in 1990. He received the Ph.D. degree in software and software theory from Peking University in 2017. His research interests include software engineering, data mining, etc.

张灵箫(1990–), 男, 浙江人, 2017 年于北京大学获计算机软件与理论专业博士学位, 主要研究领域为软件工程, 数据挖掘等。

ZHAO Junfeng was born in 1974. She received her Ph.D. degree in software and software theory from Peking University in 2005. Now she is an associate professor at Peking University. Her re-search interests include software engineering, knowledge engineering and service computing, etc.

赵俊峰(1974–), 女, 福建泉州人, 2005 年于北京大学软件与理论专业获得博士学位, 现为北京大学信息科学技术学院副教授, 主要研究领域为软件工程, 知识工程, 服务计算等。发表论文 30 余篇, 主持或承担过国家自然科学基金项目、国家 863 项目、北京重大科技成果转换项目等近 20 项。

XIE Bing was born in 1970. He received his Ph.D. degree from School of Computer, National University of Defense Technology in 1998. Now he is a professor and Ph.D. supervisor at Peking University. His research interests include software engineering and formal methods, etc.

谢冰(1970–), 男, 湖南湘潭人, 1998 年于国防科技大学计算机学院获得博士学位, 现为北京大学信息科学技术学院软件所教授, 博士生导师, 主要研究领域为软件工程, 形式化方法等。发表论文 50 余篇, 主持国家 863 重点项目多项, 获国家科技进步二等奖。

ZOU Yanzhen was born in 1976. She received the Ph.D. degree in software and software theory from Peking university in 2010. Now, she is an associate professor at Peking university. Her re-search interests include software engineering, software reuse, informationa retrival, etc.

邹艳珍(1976–), 女, 辽宁盖州人, 2010 年于北京大学软件与理论专业获得博士学位, 现为北京大学信息科学技术学院副教授, 主要研究领域为软件工程, 软件复用, 信息检索等。