文本数据主题挖掘与关联搜索研究

朱卫星[1] 徐伟光[2] 何红悦[2] 李 雯[1]
(解放军理工大学信息管理中心 南京 210007)[1] (解放军理工大学指挥信息系统学院 南京 210007)[2]

摘 要 文本数据是存储和交换信息最自然的方式,文本挖掘技术可以发现海量文本数据中隐藏的潜在知识模式。研究了文本数据主题挖掘与关联搜索技术,首先通过文本解析提取、分词预处理和索引等进行文本信息处理,然后利用基于潜在语义关系的主题发现模型挖掘大量文本数据中隐藏的主题信息,最后利用主题模型计算关键词间的关联程度进行查询扩展,从而实现关联搜索。实现了一个文本数据挖掘与关联搜索的原型系统,对Tancorp数据集进行主题发现和关联搜索,并以视化和网页同步显示关联搜索的过程。

关键词 文本挖掘,主题发现,关联搜索
中图法分类号 TP391.9 **文献标识码** A

Research on Text Data Topic Mining and Association Search

ZHU Wei-xing[1] XU Wei-guang[2] HE Hong-yue[2] LI Wen[1]
(Information Management Center, PLA University of Science and Technology, Nanjing 210007, China)[1]
(Institute of Command Information Systems, PLA University of Science and Technology, Nanjing 210007, China)[2]

Abstract Text data is the most natural way of storing and exchanging information. Text mining technology can discover knowledge patterns hidden in massive text data. The text data mining and related search technology were studied in the paper, Firstly, text information is extracted by text parsing and extraction, word preprocessing and indexing. Then the theme information model based on latent semantic relations is used to mine the hidden topic information in large amount of text data. Finally, the topic model is used to calculate the relevance degree of keywords. In order to achieve the associated search, a prototype system of text data mining and association search is implemented. Subject discovery and association search were performed on Tancorp dataset, and the process of association search was displayed synchronously with visualization and Web page.

Keywords Text mining, Topic discovery, Association search

1 研究背景

在众多数据类型中,文本数据是存储和交换信息最自然的方式,尤其值得关注和研究。文本挖掘技术可以发现海量文本数据中隐藏的潜在知识模式,进而支撑辅助决策,其应用范围相当广泛。数据挖掘[1-2]作为理论研究于20世纪80年代后期出现,在90年代有了突飞猛进的发展。文本数据挖掘是数据挖掘中的一个分支,又称文本知识发现,是指从大量文本数据中抽取隐含的、前所未知的、潜在有用的、可理解的模式和知识的过程[3]。文本数据挖掘的研究热点逐渐由文本的表示模型[4-5]、文本特征提取和降维[6-7]向文本主题深层挖掘研究转移,如互联网上的Web挖掘[8]、自动问答[9]、主题检测[10]等。

本文利用基于潜在语义的主题发现模型在文本语料库上进行训练,学习出语料库中的主题以及主题词间的关联关系;与此同时,对语料库进行倒排索引,并利用主题词间的关系对用户原始查询进行扩展,实现关联搜索,从而提高查询的查全率,改善搜索结果排序。

2 文本主题挖掘与关联搜索

2.1 基本原理

文本主题挖掘与关联搜索基本原理示意图如图1所示。

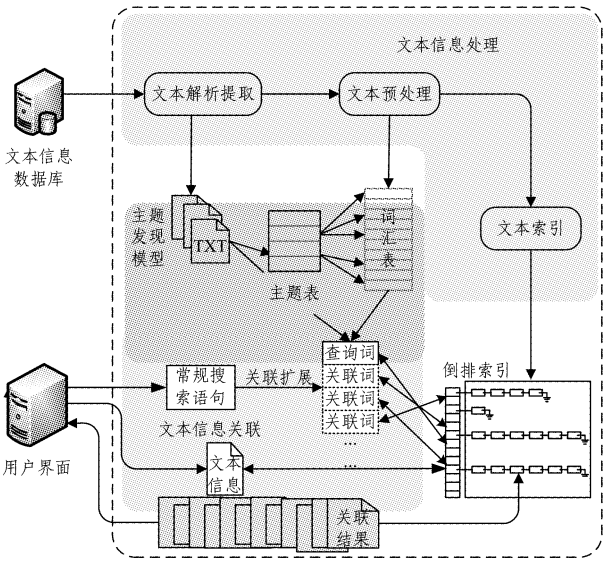

图1 系统基本原理示意图

文本数据存储于服务器端的数据库中。文本信息处理模块从数据库获得文本信息,从中解析提取出文本内容,然后通过预处理将文本内容中的中文字串切分成具有实际意义的词汇,再交由文本索引子模块建立倒排索引,以便后期快速搜索。主题发现模型建立信息文档层、词汇层以及连接二者的

朱卫星(1978—),男,博士,副教授,主要研究方向为数据挖掘。

虚拟主题层，通过迭代式的主题发现算法得出描述三者之间详细关联关系的主题模型，为后期文本信息关联打下基础。在文本信息关联模块中，用户输入常规搜索语句后，系统在主题模型的基础上找出关联词汇，对常规搜索语句进行扩展，最后搜索出常规搜索无法得到的关联搜索结果。

2.2 文本信息处理

文本数据以长字符串的形式存储在多种格式的文档中，因此需要对文档进行预处理，包括文本解析提取、文本预处理和文本索引。

(1)文本解析提取

文本解析提取过程主要是分析文本文档以提取各种格式文档中的字符串，该过程不仅面向互联网的 Web 网页，还面向各种文档类型，例如 XML，PDF，Microsoft Word 或者 Excel 等类型的文档。

(2)文本预处理

提取出文本字符串后，需对文本字符串进行预处理以选择合适的词来建立索引。文本预处理首先将文本中包含的词分析出来，即分词。英文单词天然地被空格隔开，很容易切分；而中文的词是由连续的单个字符组成，因此中文分词相对困难。由于在语义表达方面并不是所有词的表达能力都是同等的，因此除分词之外，文本预处理还包括删除停用词、提取词干、选择索引词和建立词典等操作。

(3)文本索引

文本预处理的结果将被用于构建关联搜索引擎的核心——索引。索引是一种便于用词检索的数据结构。常见的索引数据结构有倒排文件、后缀树和签名文件等。倒排文件对词的搜索非常有效，在大多数应用中比后缀树和签名文件的效果好，在文本检索中应用得最为广泛，本文采用开源工具 Lucene 实现文本数据的倒排索引。

2.3 基于潜在语义关系的主题发现模型

主题发现模型是一种全概率生成模型，其内在结果清晰，可以利用高效的概率推断算法进行计算。主题发现模型参数空间的规模与文档数量无关，更适合大规模文本信息的关联分析。

(1)模型构建

主题发现模型是一种利用概率对文本主题信息进行建模的方法。如图 2 所示，一条文本数据通常由若干个隐含主题组成，而这些主题由特定词汇体现。因此，可将隐含主题看作词汇的概率分布，单个文档表示为这些隐含主题特定比例的随机混合。主题发现模型是由词层、文档层和文档集合层组成有向概率图模型。(α,χ)是文档集合层的参数。文本数据集合中 α 用于描述隐含主题间的相对强弱，隐含主题自身的概率分布用 χ 表示。随机变量 θ 是文档层参数，θ 的分量表示目标文档中每个隐含主题的权重。(z,ω)是词层的参数，z 表示目标文档的隐含主题在每个词上的份额，ω 是目标文档的特征词向量(见图 3)。

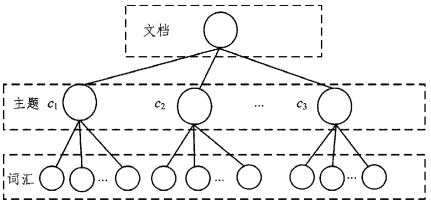

图 2 主题发现模型

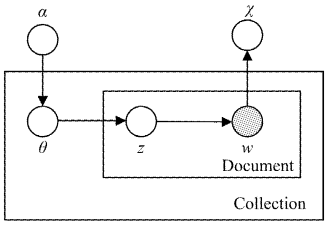

图 3 主题发现模型参数

假设隐含主题有 T 个，则在所给文档中的第 i 个词汇 ω_i 出现的概率为：

$$P(w_i)=\sum_{j=1}^{T}P(w_i|z_i=j)P(z_i=j)$$

其中，ω_i 是文档的第 i 个特征词，j 表示第 j 个隐含主题，$z_i=j$ 表明 ω_i 取第 j 个隐含主题。$P(w_i|z_i=j)$是 ω_i 属于 j 的概率，$P(z_i=j)$是 j 属于所给文档的概率。假设 D 个文档具有 T 个隐含主题，分别以 W 个唯一性的词汇表示。令 $\varphi_w^{(z=j)}=P(w|z=j)$，$\psi_{z=j}^{(d)}=P(z=j)$，则在文档 d 中词汇 w 出现的概率为：

$$P(w|d)=\sum_{j=1}^{T}\varphi_w^{(z=j)}\cdot\psi_{z=j}^{(d)}$$

为使模型易处理训练语料之外的新数据，主题发现模型在 $\psi^{(d)}$ 上作 Dirichlet(α)的先验概率假设；同时，在 $\varphi^{(z)}$ 上亦作对称的 Dirichlet(χ)的先验概率假设，以便于对模型参数进行推理，如下式所示：

$$w_i|z_i,\varphi^{(z)}\sim Discerte(\varphi^{(z_i)}),\varphi^{(z_i)}\sim Dirichlet(\chi)$$
$$z_i|\psi^{(d)}\sim Discerte(\psi^{(d_i)}),\psi^{(d_i)}\sim Dirichlet(\alpha)$$

其中，χ 是根据文档的主题抽取样本获得的词汇出现频率，α 是根据文档的分类获得的主题被抽样的次数。主题及词汇被使用的程度由 α 和 χ 的具体取值决定，但是不同的词汇和主题被使用的方式基本相同。对此可以假定所有的 α 取相同的值，所有的 χ 取相同的值，即对称的 Dirichlet 分布。

(2)模型参数计算

MCMC 提供了从后验分布直接抽样本值的近似迭代方法，而 Gibbs 抽样简化了 MCMC 算法。从词汇对于主题的后验概率 $P(w|z)$出发，使用 Gibbs 抽样计算出 φ 和 ψ 的值，得到词汇的概率分布。使用 Gibbs 抽样的关键是构造目标概率分布函数。主题发现模型只需要对变量 z_i 抽取样本。计算后验概率 $P(z_i=j|z_{-i},w_i)$的公式如下：

$$P(z_i=j|z_{-i},w_i)=\frac{\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W_\chi}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i}^{(d_i)}+T\alpha}}{\sum_{j=1}^{T}\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+W_\chi}\cdot\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i}^{(d_i)}+T\alpha}}$$

其中，$z_i=j$ 表示主题 j 包含词汇 w_i(w_i 是词汇 w 在所给文档中的位置权值)；z_{-i} 表示所有 $w_k(k\neq i)$对该主题的分配；$n_{-i}^{d_i}$是 d_i 中所有主题包含的词汇个数；$n_{-i,j}^{(d_i)}$是文档 d_i 中主题 j 包含的词汇个数；$n_{-i,j}^{(\cdot)}$是主题 j 包含除 $z_i=j$ 的所有词汇个数；$n_{-i,j}^{(w_i)}$是主题 j 包含 w_i 的个数。

Gibbs 抽样算法：

(1)初始化马尔科夫(Markov)链。$z_i\in[1,T]$，$i\in[1,N]$，T 为主题个数，N 是文档特征词个数。z_i 的初值随机选取。

(2)for($i=1, i\leqslant N$,i++){根据 $P(z_i=j|z_{-i},w_i)$将词汇分配给主题，计算 Markov 链的下一个状态}。

(3)对第(2)步迭代多次，使 Markov 链接近目标分布，取 z_i 的当前值作为样本记录下来。对于每个单一样本，按下式估算 φ 和 ψ 的值：

$$\hat{\varphi}_w^{z=j}=\frac{n_j^w+\chi}{n_j^{\cdot}+W_\chi},\psi_{z=j}^{(d)}=\frac{n_j^{(d)}+\alpha}{n^{d\cdot}+T\alpha}$$

其中，n_j^w 表示主题 j 包含词汇 w 的个数；$n_j^{\cdot}$ 表示主题 j 包含所有的词数；$n^{d\cdot}$ 表示文档 d 中所有主题包含的词数；$n_j^{(d)}$ 表示文档 d 中分配给主题 j 的词数。

2.4　文本关联搜索

在文本关联搜索中，采用向量空间模型(Vector Space Model，VSM)对文本文档进行建模，VSM 考虑词的频率信息而舍弃了词的顺序信息。一般情况下，查询向量与文档向量类似，区别在于查询向量中的权重表示词对用户的重要程度，权重为 1 表示期望在文本中出现的词，而 0 表示不希望出现的词。例如，查询 q 用向量$(q_1,q_2,\cdots,q_m)$表示，则：

$$q_i=\begin{cases}0, & \text{查询向量中不包含词汇 } t_i\\ 1, & \text{查询向量中包含词汇 } t_i\end{cases}$$

这种查询方法只能检索出用户定义的查询词，不具有查询关联扩展能力。为了进行关联搜索，可利用主题发现模型计算出文本中词语的关联程度，以进行查询扩展。首先定义文档中词的关联关系。若 $c_i(t)$表示主题 i 中包含词汇 t 的个数，$c_i(q)$表示主题 i 中包含查询 q 中所有词汇的总个数，则查询 q 与词汇 t 之间的关联程度可采用下式计算。

$$sim(q,t)=\frac{\sum_{i=1}^{n}c_i(t)c_i(q)}{\max_t(\sum_{i=1}^{n}c_i(t)c_i(q))}$$

将两个词语的相似度作为查询中不包含词语的权值，修改公式可得：

$$q_i=\begin{cases}sim(q,t_i), & \text{查询中不包含词 } t_i\\ 1, & \text{查询中包含词 } t_i\end{cases}$$

得到文档向量和查询向量后，就可以计算文档向量与查询向量之间的相似度，进而评价文本与查询的相关程度，一般用两个向量之间的夹角的余弦值来计算向量间的相似度。

$$sim(q,d_j)=\frac{d_j\cdot q}{|d_j|\times|q|}$$

其中，$d_j\cdot q$ 表示两个向量的点积，a_j 表示向量 d_j 的模。

采用关联向量空间模型的优点是可以对查询向量中的词赋予权重，从而改善检索的精度，并可以利用量化的相似度对检索结果进行排序；同时，在原始向量空间模型上加入了词汇间的关联关系，实现了关联搜索。

3　系统实现

基于上述技术研究，开发实现了一个文本数据挖掘与关联搜索的原型系统(见图 4)。在该系统中，对 TanCorp 数据集进行主题挖掘，并提供关联搜索。为直观展示关联搜索的过程与效果，除传统网页式显示搜索结果外，还采用基于力导向布局算法进行图形化的同步显示。

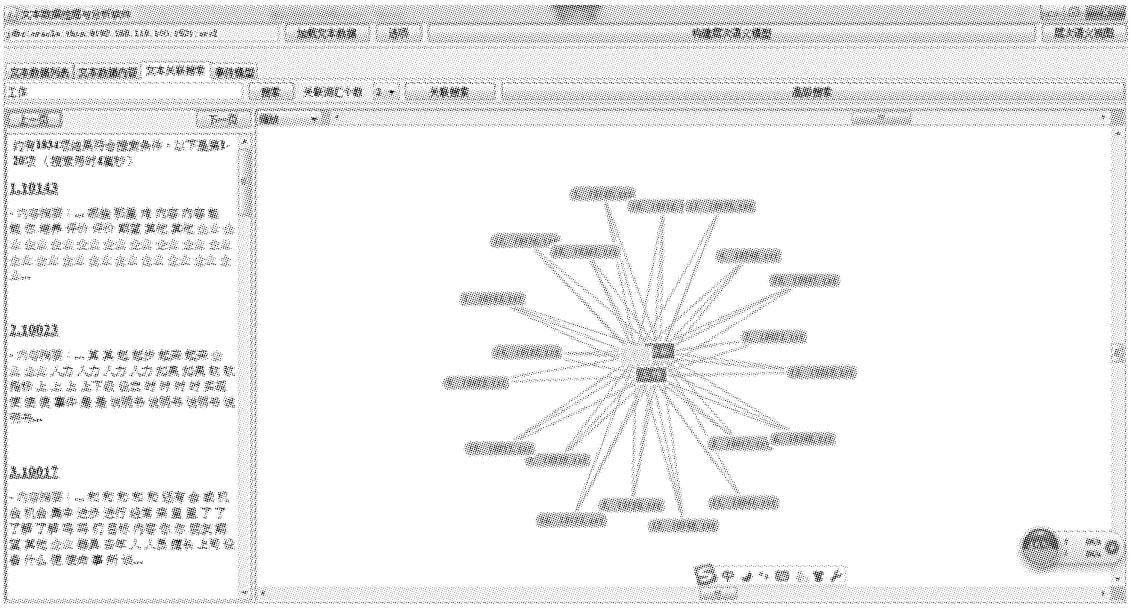

图 4　文本数据挖掘与关联搜索的原型系统

结束语　本文研究了基于语义关系的文本数据主题挖掘与关联搜索技术，利用主题模型计算关键词间的关联程度进行查询扩展，实现了一个文本数据挖掘与关联搜索的原型系统，并对 Tancorp 数据集进行主题发现和关联搜索。

参 考 文 献

[1]　曹波伟，薛青. 面向军事基础数据的数据挖掘研究[C]//2009 年系统仿真技术及其应用学术会议(CCSSTA'2009)论文集. 2009.

[2]　CORMEN T H，LEISERSON C E，RIVEST R L，et al. Introduction to Algorithms(Second Edition)[M]. The MIT Press，2001.

[3]　FELDMAN R，DAGAN I. KDT-Knowledge Discovery in Textual Database[C]//Proceedings of the 1st Annual Conference on Knowledge Discovery and DataMining. 1995：112-117.

[4]　MOTHE J，CHRISMENT C，DKAKI T. Information mining-use of the document dimensions to analyze interactively a document set[C]//European Colloquium on Information Retrieval Research. 2001：6-20.

(下转第 456 页)

4.2 处理数据光谱

实验利用 R 语言动态提取出特征向量矩阵函数 M_STAR$imDat，将处理数据存储成. csv 格式文件，读取对应参数信息数据列，选取图 2 巡天数据库中的 M1 型星光谱数据文件 spec-56647-M31010N33M1_sp09-107. fits，限定最大数据范围值后利用 plot()函数选取 type="s"参数，如图 3 所示，绘制流量光谱图。归一化后频率如图 4 所示。

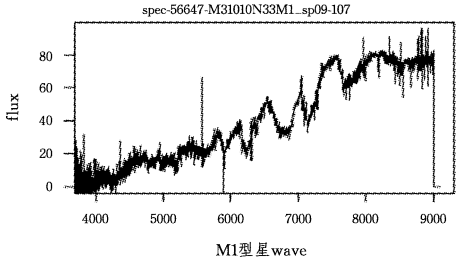

图 3 R 语言读取并绘制的 LAMOST 巡天数据 M1 型星流量光谱图

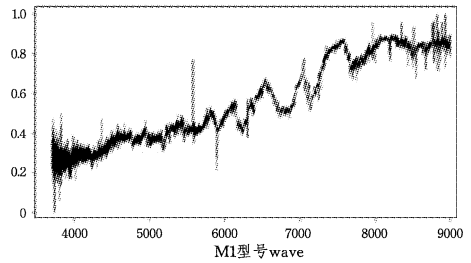

图 4 归一化数据 M1 型星流量光谱图

4.3 网格聚类分析光谱数据

从 LAMOST 巡天大数据中选取 M 型星的 205760 条恒星光谱数据，按 10 种类型进行网格聚类，分析聚类结果将恒星光谱数据聚集到较大的簇中，将拥有相同物理特征的数据聚集在一起，找出不符合恒星光谱数据的分布的光谱，充分保留光谱数据的物理特征，最后进行光谱数据的不同特征聚集离群数据分析。

结束语 天文数据挖掘是从飞速扩容的天文观测大数据信息数据库中提取隐含的、未知的以及具有应用价值的信息模式。中国天文学界研究已从单纯的数据获取提升到引领国际同行共享观测数据的高度[9]，过去的 10 年间我国天文学家增强了观测能力，大型巡天 LAMOST 之后的项目相继建成，如 FAST，HXMT，SVOM 等将投入观测，现已增加获取的天文数据达到 TB 量级，不久将会突破 BP 量级，下一步将融合天文数据科学应用开展相关工作。

(1)在之前已在 FORTRAN 语言、C 语言、IDL 语言、PYTHON 中应用 FITSIO 软件包研究的基础上，再引入 R 语言开发平台，充分发挥强大的统计、分析、数据挖掘的性能优势。

(2)目前天文大数据领域的研究，结合云计算获取海量的数据处理，依靠得力的软件工具读写 FITS 文件完善天文学领域统计分析数据的能力。

(3)R 语言应用于天文大数据挖掘中，利用 RFITIO 软件包对天文数据构造可视化光谱，在低维空间提取样本的主要特征点，高效地获取并挖掘天体信息。

参 考 文 献

[1] CUI X Q, ZHAO Y H, CHU Y Q, et al. The large sky area multi-object fiber spectroscopic telescope (LAMOST)[J]. RAA, 2012, 12(9): 1197.

[2] LUO A L, ZHAO Y H, et al. The first data release(DR1) of the LAMOST regular survey[J]. RAA, 2015, 15(8): 1104.

[3] 柯大荣，赵永恒. 一种图象传输系统及其 FITS 数据基本格式[J]. 现代图书情报技术，1994，10(2)：25-26.

[4] 崔辰州，李文，等. FITS 数据文件的检索和访问[J]. 天文研究与技术，2008，5(2)：117-119.

[5] 郭平，王可，罗阿理，等. 大数据分析中的计算智能研究现状与展望[J]. 软件学报，2015，26(11)：3011.

[6] 孙善武，王楠，欧阳丹彤. 基于聚类分析的业务流程模型抽象[J]. 计算机科学，2016，30(5)：104.

[7] 赵永恒. 大规模天文光谱巡天[J]. 中国科学：物理学力学天文学，2014，44(10)：1041-1045.

(上接第 413 页)

[5] GHANEM M, CHORTARAS A, GUO Y, et al. A grid of infrastructure for mixed bioinformatics data and text mining[J]. Computer Systems and Applications, 2005, 34(1): 116-130.

[6] KARANIKAS H, TJORTJIS C, THEODOULIDIS B. An approach to Text Mining using Information Extraction[C]// Proceeding of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Database. Lyon, France, 2000: 13-16.

[7] HU Q, YU D, DUAN Y, et al. A novel weighting formula and feature selection for text classification based on rough set theory [C]// Proceedings of Natural Language Processing and Knowledge Engineering. 2003: 638-645.

[8] KOSALA R, BLOCKEEL H. Web Mining Research: A Survey [C]// ACM SIGKDD. 2000: 1-15.

[9] LI H, YAMANISHI K. Mining from Open Answers in Questionaire Data [C]// Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2001: 443-449.

[10] PONS-PORRATA A, BERLANGA-LAVORI R, RUI-SHULCLOPER J. Topic discovery based on text mining techniques [J]. Information Processing and Management, 2007, 43(3): 752-768.