一种基于Apriori的搜索建议关键字提取算法

奚 婷 杨 燕
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 随着因特网技术的飞速发展,人们开始频繁地利用网络寻找、获取所需的资源,而传统的搜索引擎返回的结果数量庞大且呈线性排列,用户很难在短时间内找到所需的资源。文本聚类具有较强的灵活性和自动处理能力,成为解决问题的重要手段。以Lingo算法为主要研究对象,针对Lingo聚类算法提取标签时无法提取多个句子中标签的问题,引进Apriori算法来寻找主题,并将其作为搜索建议关键字,来较好地解决这个问题。

关键词 搜索引擎,Lingo,Apriori,建议关键字
中图法分类号 TP391.1 **文献标识码** A

Algorithm to Extract Search Suggested Keyword Based on Apriori

XI Ting YANG Yan
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract With the rapid development of Internet, users make use of network to achieve the resources frequently. But search engines return a big and linear list which is uncomfortable for users to use. Text clustering which is flexible and automatic has become one important measure to solve the problem. This paper used a kind of clustering algorithm named Lingo as major object. And aiming at a problem in Lingo algorithm that the label can't belong to more then one sentence, the paper employed Apriori algorithm to find label as search suggested keyword to make up the problem.

Keywords Search engine, Lingo, Apriori, Suggested keyword

1 引言

信息爆炸增加了查询与获取有用信息的难度。传统的搜索引擎对用户的需求需要比较明确的关键字,但这对用户是不现实的,因为有时用户对自己所需要的东西也只有比较模糊的概念,这就只能在搜索引擎返回的很长的列表中逐一地去寻找,非常浪费时间。这时,就需要对搜索引擎结果进行聚类处理,并且为每个类找到一个标题,使用户可以非常快捷地找到需要的资料。

现在国内外对搜索引擎结果的聚类已经有了一定的研究,其中Scatter/Gather[1]是第一个在信息检索上使用聚类方法的系统;Vivisimo[2]是一个商业化的搜索引擎系统,它提供了可读性较好的类标签。国内对聚类系统的研究比较晚,Bbmao[3]是国内最流行且最有潜质的聚类搜索引擎,以社会搜索见长;ICC[4]是交互式中文检索结果聚类系统,为用户的交互提供相应和个性化的可视化搜索页面。

聚类可以发现有用的数据分布和隐含的数据模式,同时可以不依赖背景知识而直接发现有用的簇。当前,主要使用的聚类方法是:划分聚类以及层次聚类。而本文借鉴了一种概念比较新颖的聚类算法——Lingo算法,它先提取出类标签,再为每个标签分配文档。在研究时发现,它的标签提取有一定的缺陷,只能提取出在同一个句子且是互相相邻的词组作为标签。比如,"盗梦空间是一部好电影。…大家对它的评价非常的高。"Lingo算法只能找到"评价"和"电影"两个标题,而找不到"电影评价"这个标题。所以本文结合数据挖掘中的关联规则Apriori算法,设计出了ALingo算法来弥补这个缺陷。

2 系统框图

本文实现系统的框图如图1所示。

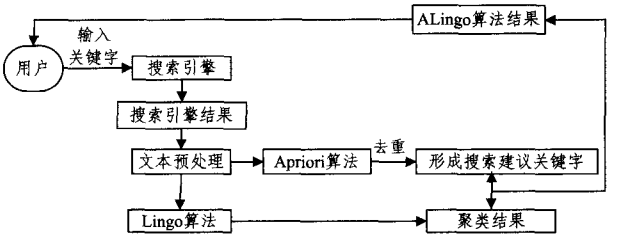

图1 系统框图

搜索引擎结果聚类的主要过程是:用户提出查询条件,系统将查询关键字传送给搜索引擎的服务接口,获取搜索引擎的结果,对其进行数据预处理,并将处理结果分别做两种处理。首先,将处理结果经过Lingo聚类算法进行聚类处理,形

到稿日期:2012-01-05 返修日期:2012-06-07 本文受国家自然科学基金面上项目(61170111),国家自然科学基金委主任项目(61152001),国家自然科学基金重点项目(611734002),中央高校基本科研业务费专项资金(SWJTU11ZT08)资助。

奚 婷(1986—),女,硕士,主要研究方向为数据挖掘,E-mail:xt1985happy@126.com;杨 燕(1964—),女,博士,教授,主要研究方向为数据挖掘、计算智能、集成学习。

成聚类结果；其次，将处理结果经过 Apriori 算法处理，经去重后形成搜索建议关键字。结合这两种结果，形成 ALingo 算法结果，同时将其返回给用户。

2.1 Lingo 聚类算法

现在一般的聚类算法都是先对数据进行聚类，然后再提取出标签，这样，很难提取到高质量的标签。Lingo 聚类算法完全颠覆了这个想法，它是先提取出类标签，再将文档一一分配到相对应的标签下，这样可以取得较好的结果。

Lingo 算法主要分为两步，即类标签提取及文档分配。

2.1.1 类标签提取

类标签提取主要包括抽象概念的发现、词组匹配和标签提取。

抽象概念的发现通过奇异值分解获得文档集合的潜在语义关系，Lingo 是将词-文档矩阵 A 经过奇异值进行分解，分解为 3 个矩阵 U,Σ,V，如式(1)所示。

$$A=U\Sigma V^{T} \tag{1}$$

U 的列向量就代表抽象概念。对 A 进行降维处理得到 A_k，k 为降维因子，A_k 是 A 的最佳近似矩阵。A_k 的确定方法如式(2)所示。

$$\| A_k \|_F / \| A \|_F \geqslant q \tag{2}$$

式中，q 是一个控制参数，q 越大，k 值也就越大，最终得到的类别也就越多。

词匹配和标签提取主要解决的问题是类标签的确定。由于抽象概念和关键词组是在同一个向量空间中，因此 Lingo 是使用 Cosine 距离来计算抽象概念和关键词组之间的相似度的。Cosine 距离如式(3)所示。

$$Conine(d_j,q)=\frac{\overline{d_j}\cdot\overline{q}}{|\overline{d_j}|\times|\overline{q}|}=\frac{\sum_{i=1}^{t}w_{i,j}\times w_{i,q}}{\sqrt{\sum_{i=1}^{t}w_{i,j}^2}\times\sqrt{\sum_{j=1}^{t}w_{i,q}^2}} \tag{3}$$

式中，$\overline{d_j}$ 代表一个文档的向量，$\overline{d_j}=(w_{1,j},w_{2,j},\cdots,w_{k,j})$ 表示第 k 个关键词在文档 j 中的权重，$\overline{q}$ 代表查询语句的向量，$\overline{q}=(w_{1,q},w_{2,q},\cdots,w_{k,q})$。

构造矩阵 P 来表示关键词和词组关系，它可以通过将词组和关键词看作伪文档来构造。因此，抽象概念和关键词、词组的关系可以通过式(4)来表示。

$$M=U_k^{T}P \tag{4}$$

选择结果矩阵中每行超过标签相似度阈值且最大者为类标签。

2.1.2 文档分配

定义矩阵 Q，每一个行向量表示一个类标签，C 为文档分配矩阵，具体构造如式(5)所示。

$$C=Q^{T}A \tag{5}$$

结果矩阵 C 的 C_{ij} 表示第 j 个文档属于第 i 个类别的程度。定义一个文档分配阈值 H，当 $C_{ij}>H$ 时，将第 j 个文档分配到第 i 类。

2.2 ALingo 算法

本文提出一种新的算法——ALingo 算法。ALingo 算法结果由两部分组成，即 Lingo 算法的结果和 Apriori 算法的结果。而 Lingo 算法的结果即为聚类结果，且 Apriori 算法的结果作为搜索建议关键字返回给用户，利用 Apriori 的结果弥补 Lingo 算法的缺陷，从而达到缩短用户搜索时间的目的。

2.2.1 Lingo 标签提取的问题

Lingo 选择候选标签算法是一种基于后缀树的词组发现算法，是用一组特殊结构的数组来实现的。主要分为两步，第一步是找到左完全和右完全词组，第二步就将它们联系起来。在这个步骤中有几个特点，如词组必须出现一定的次数；不能超过句子的界限，即类标签中的每个词组只能出现在同一个句子中，不能跨越句子。所以它存在一个缺点，就是类标签的组成词是不能存在于多个句子中的。

2.2.2 搜索建议关键字

针对上述问题，提出了搜索建议关键字的概念，即用户在搜索时，提供给用户搜索结果的高精度的主题，更加明确资源的内容，进行再次搜索，可以达到缩短搜索时间的目的。这与我们使用一些知名搜索工具时会列出相关搜索一样，只是本文是对于搜索出的文档用另一种方法找出主题，将其作为搜索建议关键字，帮助用户更好地了解到文档主题，缩短用户查找的时间。

具体的实现方法为，引入数据挖掘关联规则方面的 Apriori算法，对搜索结果进行处理，因为 Apriori 是频繁出现的词组的组合，并没有句子的界限限制，所以其可以找到跨越句子界限的主题，较好地弥补了 Lingo 算法的缺陷，提高了主题的精度，缩短了用户查询的时间。

Apriori 算法是 R. Agrawal 和 R. Srikant[5] 于 1994 年提出的关联规则挖掘频繁项集的原创性算法。它主要是找到给定数据集的关系，挖掘出频繁项集，对于文本来讲，找到在各个文档中出现次数最多的词组，就有了一定的语义意义，也就是文本的主题，并且它是没有句子边界限制的。按照这种方法，就可以找到存在于不同句子中的主题，且可将其作为搜索建议关键字，作为 Lingo 算法的一种补充，方便用户有更多的选择，更快捷地找到需要的文档。

但是 Apriori 算法处理的初始数据是非常繁乱的，并且充斥着重复的因素，低层次的频繁项集与高层次的频繁项集往往存在包含的关系，这样的包含关系对于文档的主题发现是没有意义的。例如低层次的频繁项集“数据挖掘”和高层次的频繁项集“文本数据挖掘”，显然“文本数据挖掘”精度要高，而且更能够概括文档的内容，所以在处理结果时，应保留高精度的文档主题，这就要做一个去重处理，具体方法为：

假设有频繁项集 A 和 B，它们分别为 a 层频繁项集和 b 层频繁项集，且满足以下两个条件：

$$A\Leftrightarrow B \text{ 且 } a>b \tag{6}$$

去除频繁项集 B，直到没有重复的频繁项集。

举一个简单的例子，得到 1 层频繁项集 $K1$，2 层频繁项集 $K1K2$，3 层频繁项集 $K1K2K3$，可以看出 $K1\subset K1K2\subset K1K2K3$，则可以将 1 层和 2 层频繁项集去除，保留 3 层频繁项集，即保留了高精度的主题，所得的结果即可作为搜索引擎结果的搜索建议关键字。

3 实验结果

本文使用的搜索引擎是 Lucene，它是 apache 基金项目组的一个子项目，是一个开放源代码的全文检索引擎工具包，可

以为应用程序提供索引和搜索功能。Lucene 经过一系列的系统部署及程序实现可对文档建立索引并进行查询。

Lucene 搜索的结果需要进行数据预处理，如果语言是中文，步骤是：分词，去除停用词；如果是英文，就主要包括还原词干，去除停用词。经过预处理的数据，就可以进行聚类操作了。

本文实验使用的数据集是经过 Lucene 查询的关于 data mining 的文档，数据格式如图 2 所示。第一行为文本的 URL，第二行为文本的标题，第三行为文本的内容。

```
"http://en.wikipedia.org/wiki/Data_mining",
"Data mining - Wikipedia,the free encyclopedia",
"Article about knowledge-discovery in databases(KDD)"
```

图 2　实验数据格式

将这 100 个文档经过 Lingo 算法处理，得到若干个类标签，如图 3 所示。

```
Text Mining
Data Mining Group
Data Mining Tools
Daata Management
Data Mining Techology
Introduction to Data Mining
Oracle Data Mining
Predictive Modeling
Center
Data Mining Institute
Data Warehousing
People
Statistical Data Mining
Wikipedia
Approaches
Association
Community
Data Miners
Data Mining Project
Data Mining Resources
Data Mining and Analytic
Technologies
Data-mining Software
Diretory
Downloads
```

图 3　Lingo 算法得到的类标签

为了能够清楚地显示本文的改进，在这 100 个文档中人为地加入了一句话 happy ending，并且一定要加入句号，以显示句子的界限。用 Apriori 算法处理文档得到的结果如图 4 所示。

```
第1项频繁项集：machine
第1项频繁项集：process
第1项频繁项集：through
第1项频繁项集：analysis
第1项频繁项集：analytic
第1项频繁项集：business
第2项频繁项集：data find
第2项频繁项集：data high
第2项频繁项集：data text
第2项频繁项集：get happy
第1项频繁项集：learning
第1项频繁项集：modeling
第1项频繁项集：patterns
第1项频繁项集：products
第1项频繁项集：reserach
第1项频繁项集：services
第1项频繁项集：software
第1项频繁项集：together
第2项频繁项集：areas data
第1项频繁项集：automated
```

图 4　Apriori 的初始处理结果

可以看到这个结果是非常繁乱的，用上文介绍的去重方法进行处理，可得最终结果。将文档主题和 Lingo 结果及 Apriori 结果进行对比，如表 1 所列。

从表 1 可以清楚地看出，Apriori 和 Lingo 都在一定程度上提取出了文档的主题，覆盖了一定的文本内容，同时又有相辅相成的作用，两种结果都提取出对方所没有提取出的标题，例如 Ariori 提取出了 Web mining 而 Lingo 没有，Lingo 提取出了 Wikipedia 而 Apriori 没有。ALingo 算法将 Lingo 算法结果和 Apriori 算法结果相结合，由 Apriori 的结果作为搜索关键字后，解决了 Lingo 算法漏选标题的问题，提高了算法标签的覆盖面，可以提供给用户更多的选择，使用户更快、更好地了解文档的内容。

表 1　Lingo 算法结果和 Apriori 算法结果对比

文档主题	Lingo 聚类结果	Apriori 结果
Wikipedia	Happy ending	Data mining happy ending
Text Mining	Text Mining	knowledge discovery
Web Mining	Data Mining Group	data mining sql
Knowledge Discovery	Data Mining Tools	web data mining
Software	Data Management	text data mining
Applications	Data Mining Technology	areas data mining
Solutions	Introduction to Data	data mining group
Techniques	Mining	high data mining
Introduction	Oracle Data Mining	data mining tools
Analytic Technologies	Predictive Modeling	data mining databases
Data Warehousing	Center	oracle data mining
Oracle Data Mining	Data Mining Institute	data mining modeling predictive
Data Mining Project	Data Warehousing	
Data Mining Group	People	data mining trends
Data Management	StatisticalData Mining	article data mining
SQL Server	Wikipedia	analysis data mining
Introduction	Approaches	business data mining
Predictive Modeling	Association	data mining process
Concepts	Community	data mining patterns
	Data Miners	data mining products
	Data Mining Project	data mining research
	Data Mining Resources	data mining services
	Data Mining and Analytic	data mining software
	Technologies	automateddata mining
	Data-mining Software	microsoft data mining
	Directory	data mining algorithms
	Downloads	data mining conference
	Knowledge Discovery	databases information
		Datamining management
		datamining techniques
		datamining technology
		Datamining information
		data mining applications
		datalmining international
		data mining learning machine
		customers

同时，Lingo 算法得到的新标签为 happy ending，而 Apriori 算法的结果中有 Data mining happy ending know-ledge discovery 这个标签。由于人为加入 happy ending 这个单独的句子，而 Apriori 算法的标题中同时出现了 data mining 和 knowledge discovery，这些词组跨越了多个句子，是对于这些文档的共同标题的汇总，因此本文 ALingo 算法将 Apriori 算法引入，解决了 Lingo 算法提取标签的句子限制问题，增加了标题的精度。

Lingo 算法的优点是实现了模糊聚类，即一个文档能够同时出现在多个相关的类别中。例如用户想查询的关键字为 data mining，若只有 Lingo 算法，那么只可以看出这些文档中有关于 happy ending 的文档和关于 knowledge discovery 的文档，它们是分开的，这两类中是否有相同的文档却不得而知；而 ALingo 算法的结果加入了 Apriori 算法的结果后，即可以看出文档中有关于 happy ending knowledge discovery 的文档，即它们之间是有相同文档的。如果用 happy ending knowledge discovery 作为搜索关键字，即可快速地寻找到需要的文档，缩短了查询时间。

以上说明可以看出，本文提出的 ALingo 算法将 Lingo 结果和 Apriori 结果整合成新方法，不仅解决了 Lingo 算法遗漏主题的问题，同时也解决了 Lingo 算法提取类标签时对标签句子界限的限制问题，并提出了搜索建议关键字的新概念，提高了算法的标签覆盖面和精度，缩短了用户的搜索时间，较 Lingo 算法具有较多优势。

结束语 由于目前大多数的搜索引擎返回的结果都是线性的，从大量的结果中找到需要的文本是很困难的，因此很多研究者采用聚类的方式对搜索引擎结果进行聚类。Lingo 算法就是其中一种比较新颖的算法，本文主要针对 Lingo 算法在提取类标签时无法提取跨越句子的类标签的问题，设计出了一种名为 ALingo 的算法，提出了搜索建议关键字的概念，从而增加了提取标签的精度，为用户进行二次检索提供了线索，提升了用户找到所需文本的速度。

参 考 文 献

[1] Hearst M A, Pedersen J O. Reexamining the Cluster Hypotheses: Scatter / Gather on Retrieval Results[C]// Proceedings of the Nineteenth Annual International ACM SIGIR Conference. Zurich, June 1996: 76-84

[2] Vivisimo Web search engine[OL]. http://www.vivisimo.com

[3] Bbmao[OL]. http://www.bbmao.com

[4] Wei L, Gui R X, Shen H, et al. Interactive Chinese Search Results Clustering for Personalization[J]. Lecture Notes in Computer Science, 2005, 3739: 678-681

[5] Han J W, Michline K. Data mining concepts and Technniques (Scond Edition)[M]. Beijing: China Machine Press, 2007: 151-154

[6] Stanislaw O. An Algorithm for Clustering of Web Search Results[D]. Master theses, degree of Master of science, Poznan university of technology, Poland, 2003: 36-60

[7] Stanislaw O, Dawid W. Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data[D]. Institute of Computing Science, Poznan University of Technology, 2004: 1-9

[8] Song C F, Shi B. An algorithm to cluster the search results based-on the association rules[J]. Journal of Shandong University, 2006, 41(3): 62-63

[9] Qu S N, Wang Q, Zou Y, et al. Research on text clustering algorithm based on association rule[J]. Application Research of Computers, 2008, 25(4): 986-988

[10] Tao H, Hong Y C. The implement of searching engine for educational resources using text clustering[C]// Granular Computing, 2009, GRC 09. IEEE International Conference. 2009: 260-263

(上接第 144 页)

献[4,5]相比，聚类准确率高，而运算时间短，因此本文设计的相似性度量方法对 The Internet Traffic Archive 的 3 个数据集是适合的，能够取得较好的聚类效果。

结束语 通过对 Web 日志进行预处理，生成包含访问页面及其访问时间信息的用户会话序列，采用基于动态规划的全序列比对方法度量两个用户会话的相似性，并建立以用户会话为顶点、相似度值为边权的加权无向图，然后利用基于 NJW 的自动谱聚类算法进行用户会话的聚类。实验结果表明，本文所提用户会话相似性度量方法及其聚类算法与同类算法相比，在较短时间内可以取得较高的准确性。本文所述的方法稍加完善即可用于一些网站的个性化推荐服务的应用开发。

参 考 文 献

[1] Zhang F, Chang H Y. Research and development in web usage mining system-key issues and proposed soluteions: a survey[C]// Proceedings of the First IEEE International Conference on Machine Learning and Cybernetics. 2002: 986-990

[2] Shahabi C, Zarski A M, Shah J. Knowledge discovery from users web-page navigation[C]// Proc of 7th Int Conf on Research Issues in Data Engineering. Birmingham: IEEE Computer Society Press, 1997: 20-29

[3] Fu Y, Sandhu K, Shih M. A generalization-based approach to clustering of Web usage session[M]. Web Usage Analysis and User Profiling. New York: Springer-Verlag, 2000: 21-38

[4] 宋江春，沈钧毅. 一种新的 Web 用户群体和 URL 聚类算法的研究[J]. 控制与决策，2007，22(3)：284-288

[5] 程舒通. Web 点击流的频繁模式聚类算法[J]. 计算机技术与发展，2007，17(9)：18-20

[6] 杨风雷，阎保平. Web 用户行为模式挖掘研究[J]. 微电子学与计算机，2008，25(11)：146-149

[7] 易明. 基于序列模式的个性化 Web 页面推荐模型[J]. 现代图书情报技术，2008，168(8)：42-47

[8] 许欢庆，王永成. 基于用户访问路径分析的网页预取模型[J]. 软件学报，2003，14(6)：1142-1147

[9] Wang Jun, de Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]// Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006: 501-508

[10] Charter K, Schaeffer J, Szafron D. Sequence Alignment using FastLSA[C]// Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'2000). 2000, 6: 239-245

[11] Gündüz S, Tamerözsu M. A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003

[12] Ng A Y, Jordan M L, Weiss Y. On spectral clustering: Analysis and an algorithm[M]// Advances in Neural Information Processing Systems14. MIT Press, 2002: 849-856

[13] Sanguinetti G, Laidler J, Lawrence N D. Automatic determination of the number of clusters using spectral algorithms[C]// roc of IEEE Machine Learning for Signal Processing. USA, 2005: 28-30

[14] The Internet Traffic Archive[OL]. http://ita.ee.lbl.gov/index.html