用户兴趣实例模型与 K _ means 算法的改进*

何 兴 无

(重庆电子职业技术学院 计算机二系，重庆 400021)

摘 要：Web 检索越来越重要，但检索的效率和准确性始终成为当前面临的主要问题。本文提出了一种用户兴趣模型，以用户感兴趣的实例文档作为用户兴趣的表示方法。接着提出一种以实例文档为聚类中心的 K _ means 聚类算法。实验证明具有较好的准确性和较高的效率。

关键词：用户模型；K _ means 算法；实例文档；个性化服务

中图分类号：TP301.6；TP181　　文献标识码：A　　文章编号：1672-6693(2006)02-0038-04

An Example of User Profile Based on the Interest of User and the Improvement of K _ means Algorithm

HE Xing-wu

(The Second Dept. of Computer, Chongqing Electronic Profession College, Chongqing 400021, China)

Abstract: The Web of retrieve has become more and more important in recent days, but the veracity and efficiency are always the focussed problem. A user profile is firstly presented in this paper, and use the Web example to represent the interest of users. Then, a K _ means algorithm that uses the Web example as the clustering centroid is presented. Finally, the paper shows the results of the experiments to prove that the new algorithm is more veracious than old K _ means algorithm.

Key words: user profile; K _ means algorithm; exampled file; personalized service

随着信息技术的高速发展，因特网的普及，人们每天可以从网上获取大量的信息。但由于 Web 的半结构化特征使得在为特定用户检索特定信息时，变得非常困难。在加之因特网上大量 Web 页面日益增加，导致对网上信息资源获取变得难上加难[1]。目前，人们主要是利用传统的搜索引擎进行网上的信息查询，它需要用户按照要求的格式输入查询串。这类基于关键词检索的引擎在一定程度上满足了网络用户的信息需求，但它不具有智能性[2]，不能学习用户的兴趣。对具有特定专业兴趣，信息需求在相当长一段时间内保持不变或变化不大的用户，只能不断在网上反复查询相同的内容，这造成了许多不必要的时间浪费。正是在这样的需求驱动下，基于个性化服务的 Web 检索技术得到了长足的发展。

在一些信息过滤系统中，人们使用关键词串来构造用户模型，那么包含这些词的信息就认为是满足用户兴趣的信息，即相关信息。使用这种用户兴趣表示方法，常常会造成不正确的匹配。因为这些关键词往往无法清晰地表达兴趣主题，一个词可能有多种含义(多义现象)，相同的概念又能用不同的词来表述(同义现象)。因此本文提出了一种改进的用户模型，以用户感兴趣的实例材料作为用户兴趣的表示方法。接着提出一种以实例文档为聚类中心的 K _ means 聚类算法。

1 基于实例表示的用户模型的创建

个性化的 Web 服务系统中使用用户模型(user profile)来描述用户兴趣主题。但在实验中，人们发现要清晰地、准确地描述用户兴趣是非常困难的。不仅信息的内容，而且信息的新颖性、熟悉程度、紧迫性等，都是构造用户模型时需要考虑的重要因素。

* 收稿日期：2005-11-11

作者简介：何兴无(1970-)，男，四川渠县人，讲师，硕士研究生，研究方向为个性化服务、Web 挖掘。

Groft,Pazzani 等人在总结他们的实验结论时也认为,影响信息处理效率的首要因素在于相关知识的表示,算法复杂性的影响不占主要因素[3,4]。所以信息过滤系统中,用户模型的构造非常重要。

基于 Groft,Pazzani 等人的研究结果,在成熟的 VSM 技术的基础上,本文提出一种实例化的用户模型。该模型主要以大量的实例文档组成用户的兴趣集合。在这里,实例是指用户所感兴趣的信息示例。

1.1 基于实例化的用户模型的基本思路

1)系统获取用户实例文档。用户可以主动提交实例信息或系统自动收集用户的实例文档。实例文档一定是能反应用户当前较高兴趣度的文档。建立初始用户模型。

2)过滤系统依据用户模型进行信息检索,将相似度高的信息发送给用户。

3)当用户对分发信息进行评价后,过滤系统收集用户反馈,将用户评价为相关度最高的那些实例文档加入用户模型中。新加入的实例档与已有实例档做相似度计算。如果相似度大到某种程度,可以认为该实例文档所表达的用户兴趣已经存在,可以不加入该实例文档信息。

4)反复执行 2)和 3),当用户模型中的实例文档数达到某个限定值后,相应地去掉用户评价为相关度最低的实例文档。这些去掉的实例文档是相似度计算结果与用户评价最不一致的信息,即不能较好地描述用户兴趣的实例信息。这样用户模型中的实例文档数保持为某个常数。

1.2 用户模型的表示

传统的用户模型使用由一组词组成的矢量来表示用户的个人兴趣信息。这些信息又往往来自用户反馈的实例信息,在实例信息和表示用户个人兴趣信息的矢量之间很难表达完全一致的用户个人兴趣信息。这就是说一个表示用户个人兴趣信息的矢量很难与用户个人真正的兴趣信息完全一致。这也是导致最终个性化检索系统检索结果偏离用户兴趣的主要原因之一。

因此,用户模型直接采用用户感兴趣的实例文档表示,形式为

$$U=(d_1,d_2,d_3,\cdots,d_n)$$

其中 U 表示用户模型。$d_i(i=1,2,\cdots,n)$ 为实例文档向量。

2 改进基于实例表示的 K_means 算法

作为数据挖掘的一种重要手段,聚类在 Web 文档的信息挖掘中起着非常重要的作用。文档聚类是将文档集合分成若干个簇,要求簇内文档内容的相似性尽可能大,而簇之间文档的相似性尽可能小。文档聚类可以揭示文档集合的内在结构,发现新的信息,因此广泛应用于文本挖掘与信息检索等方面。文档聚类算法一般分为分层和分割二种,普遍采用的是基于分割的 K_means 算法[5,6]。

K_means 算法中文档表示模型采用向量空间模型(VSM),其中的词条权重评价函数一般使用 TF_IDF表示。相似性度量一般使用基于距离的计算方法。K_means 算法具有可伸缩性和效率极高的优点,从而被广泛地应用于大文档集的处理。针对 K_means 算法的缺点,许多文献提出了改进方法,但是这些改进大多以牺牲效率为代价,且只对算法的某一方面进行优化,使执行代价很高。

本文提出相应的解决方法,即基于实例表示 K_means 算法。实验表明改进后的 K_means 算法不仅保留了原算法效率高的优点,而且聚类的准确度有了较大提高。

2.1 K_means 算法简介[7]

K_means 算法以 k 为参数,把 N 个对象分为 k 个簇,以使簇内具有较高的相似度,而簇间的相似度较低。具体算法如下:

1) 在 N 个对象中随机的选取 k 个对象作为初始的聚类中心;

2) 把其余 N-k 个对象归到距离最近的聚类中;

3) 重新计算每一个聚类的中心;

4) 重复 2)和 3),直到每一聚类的中心不再改变。

K_means 算法除生成 k 个聚类外,还生成每个聚类的中心。具有较好的可伸缩性和很高的效率,适合处理大文档集[8]。尽管同样基于划分的 K_medioids聚类算法对存在孤立点的文档集能得到较好的聚类结果,但其效率很低,执行代价很高。为此,本文在下面提出了 K_means 算法的一种改进算法,能较好地按照用户的兴趣模型准确地从文档集中检索出有用的文档。

2.2 对 K_means 算法的一种改进算法

基本思路是将实例文档加入待聚类文档,并以之为聚类中心。采取按相似程度由高到低,用 K_means 算法进行聚类,所有文档完全聚类,亦即被划分,或待聚类文档与各聚类中心的相似程度低到某

种程度时结束计算。其主要依据的原理有：一方面，K_means是一种基于划分的聚类算法，由于K_means算法在选择初始聚类中心时是随机选取k个点，一旦k个点选取不合理，将会误导聚类过程，得到一个不合理的聚类结果。本文在分析聚类结果对初值依赖性的基础上，对初值选取方法进行了分析和研究，采取以能表达用户兴趣的实例文档作为初值进行类中心搜索。从实验结果中可以发现，改进后K_means得到的聚类结果更加准确、稳定；另一方面，用K_means算法进行文档聚类时在稳定性方面存在问题，实验结果偶尔出现较大的偏差。通过分析表明，导致这种偏差产生的原因在于数据的分散性。一旦存在少量数据远离高密度的数据密集区，在进行K_means聚类计算时，是将聚类均值点（簇中所有数据的几何中心点）作为新的聚类种子进行新一轮聚类计算，此时新的聚类种子将偏离真正的数据密集区。本文在结合Web检索的实际情况下，采取以下思路解决文档聚类时在稳定性方面存在问题：当文档与聚类中心的距离太大时，不将该文档聚类；如果所有文档均不能聚类时，则停止聚类操作。

假设有k个实例文档和N-k个待聚类文档。具体算法描述如下。

1）将k个实例文档加入待聚类文档，并将k个实例文档作为算法的初始聚类中心。

K_means算法一般采用向量空间模型[9,10]，因此需要计算出每篇文档的特征向量。思路和其方法为：把文本看作是由一组词组成的矢量空间，每个文档d表示为其中的一个范化特征矢量空间$V(d_i)=\{(t_1,w_1),(t_2,w_2),(t_3,w_3),\cdots,(t_n,w_n)\}$，其中$t_i$为第$i$个关键字，$W_i$为第$i$个特征项的权重，其计算公式有多种，如TF_IDF公式、计算信息增益、计算互信息量、计算文本证据权等方法。目前采用的比较多的是TF_IDF公式：

$$W(t,d)=\frac{tf(t,d)\times\lg\left(\frac{N}{n_t}+0.01\right)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times\lg\left(\frac{N}{n_t}+0.01\right)\right]^2}} \quad (1)$$

其中$W(t,d)$为词t在文档d中的权重，而$tf(t,d)$为词t在文档d中的词频，N为训练文本总数，n_t为训练文本集中出现词t的文档数，分母为归一化因子。

2）计算每个待聚类文档与每个聚类中心的距离。计算方法也有多种，本文使用它们夹角余弦来度量相似度。计算公式为

$$\cos(W,d_i)=\frac{\sum_{i=1}^{n}w_ix_i}{\sqrt{\sum_{i=1}^{n}w_i^2\sum_{i=1}^{n}x_i^2}} \quad (2)$$

其中$W=(w_1,w_2,w_3,\cdots,w_n)$，$d_i=(x_1,x_2,x_3,\cdots,x_n)$，均为文档向量。

3）把待聚类文档与聚类中心的距离较小的文档归到聚类中；

4）重新计算每一个聚类的中心。每一个聚类的聚类中心的计算方法可以使用求平均值的方法。计算公式为

$$C=\frac{1}{N}\sum_{d_i\in D}V(d_i) \quad (3)$$

其中C为聚类的中心向量；N为聚类集合中的文档数；D为聚类集合中的文档向量。

5）重复2）、3）和4），直到每一聚类的中心不再改变或每个待聚类文档与每个聚类中心的距离太大（没有聚类的文档表明不符合用户的兴趣）或聚类文档数超过一定数量（即检索的文档数量达到足够）。

3 实验与结果分析

实验用的文档是从网易的网站上收集了各类电子文献2000余篇，按照用户的兴趣分成5类，分别选出若干示例，构造一个基于实例的初始用户模型。考虑到在实际应用中可以依据网页聚类的相关度对网页排序。因此查准率对用户来讲更为重要。所以本文只对查准率进行了统计分析。查准率=返回相关网页数/返回网页数；试验中对分别使用传统的K_means检索方法和基于实例的K_means方法的准确率进行比较。实验结果表明改进后的K_means算法与原K_means算法相比，准确率比原算法有了普遍提高，提高了近20%，说明此算法具有较高的准确性。随着系统采用实例文本数的增加，对用户兴趣的描述就更为精确，准确率便随之提高。部分文档的错误归类检索通常是由于文本相似度比较造成的。在算法执行过程中部分文档会认为与用户兴趣无关，而不能聚类，这不仅提高了准确性，还提高了系统运行的速度。

4 结论

本文结合个性化Web检索的实际，提出了一个基于实例表示的K_means算法。由于使用用户兴趣实例作为聚类中心，克服了传统方法随机选取聚

类中心的方法所带来的不足。该算法大大提高了检索的有效性。而且克服了传统方法由于孤立点造成的聚类中心偏离,最终导致检索分类的结果偏离用户真实兴趣的错误情况。

另外,本文认为使用实例描述用户兴趣模型能更准确地描述用户的兴趣,因为实例毕竟是用户兴趣表现的最原始数据。该模型对那些能直接利用两文档的具体信息进行相似比较的方法其优势更加明显。而普通的用户兴趣描述模型,如向量空间模型(VSM)使用从实例中提取用户兴趣的方法,本文认为这可能造成用户兴趣描述的失真、偏离。

使用实例描述用户兴趣模型的方法,有利于兴趣模型描述与检索方法的分离,从而有利于系统设计。

今后的工作应放在寻找更准确的文档相似度比较方法,主要是直接运用文档间具体信息的方法。

参考文献:

[1] CHEN L C,LUH C J,JOU C C. Generating page Clippings From Web Search Results Using a Dynamically Terminated Genetic Algorithm[J]. ELSEVIER on Information Systems, 2005,30(4):299-316.

[2] 马燕,邹显春,包骏杰,等. 一种互联网智能元搜索引擎模型的设计[J]. 重庆师范大学学报(自然科学版), 2004,21(3):15-18.

[3] BELKIN N J,CROFT W B. Information Filtering and Information Retrieval:Two Sides of the Same Coin[J]. Communication of M,1992,35(12):29-38.

[4] PAZZANI M,BILLSUS D. Learning and Revising User Profiles:the Identification of Interesting Web Sites[J]. Machine Arning,1997,7(3):313-331.

[5] KAUFMAN L,ROUSSEEUW P J. Finding Groups in Data: An Introductionto Cluster Analysis[M]. New York:John Wiley & Sons,1990.

[6] FASULO D. An Analysis of Recent Work on Clustering Algorithms[R]. Washington:University of Washington,1999.

[7] 万小军,杨建武,陈晓鸥. 文档聚类中 K_means 算法的一种改进算法[J]. 计算机工程,2003,29(2). 102-103, 157.

[8] STEINBACK M,KARYPIS G,KUMAR V. A Comparison of Document Clustering Techniques[R]. Linkoping:Dept. of Computer and lnformation Science,1995.

[9] RIISBERGEN G J V. Information Retrieval[M]. 2nd ed. London:Buttersworth,1989.

[10] KOWALSKI G Information Retrieval Systems-Theory and Lmplcmen-ration [M]. Netherlands:Kluwer Academic Publishers,1997.

(责任编辑 游中胜)